Frontiers of Large Language Models: Empowering Decision Optimization, Scene Understanding, and Summarization Through Advanced Computational Approaches

November 2023

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Science in Computer Science
by Joaquim de Curtò i Díaz

Supervisor: Dr. Carlos Miguel Tavares de Araujo Cesariny Calafate

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

"The future is already here – it's just not evenly distributed." - W. Gibson.

Acknowledgements

First and foremost, I would like to express my profound gratitude to my advisors Carlos T. Calafate and Gemma Roig. Their insightful feedback, unwavering support, and encouragement were invaluable to the completion of this thesis. Their mentorship has been instrumental in my academic journey, and their relentless pursuit of knowledge and discovery has inspired me to push the boundaries of my own research.

I am immensely grateful to all the members of the Grupo de Redes de Computadores at Universitat Politècnica de València, especially Pietro Manzoni, Juan Carlos Cano and Enrique Hernández. Your collaborations, contributions, and inspiring discussions have significantly enriched this work.

I would also like to extend my gratitude to the team at GOETHE-University Frankfurt am Main, particularly the members of the group Computational Vision and Artificial Intelligence and the Center for Cognition and Computation, as well as the logistics group at TU Darmstadt. Your collaborative spirit and insightful perspectives have been invaluable in shaping this work.

My heartfelt thanks go to the Centre for Intelligent Multidimensional Data Analysis at HK Science Park, the Mathematical Institute at the University of Oxford and the EE Department at City University of Hong Kong. Your stimulating intellectual environment and generous support have greatly facilitated

my research.

I am deeply grateful to my wife, I. de Zarzà, whose unwavering support and constant companionship have been my towers of fortitude. Your love, understanding, and patience have been a great source of comfort throughout this journey.

Finally, my heartfelt gratitude goes to my family for their unconditional love, encouragement, and constant faith in my abilities. Your support has been a guiding light, providing me with the strength and determination to navigate through this academic endeavor.

Abstract

The advent of Large Language Models (LLMs) marks a transformative phase in the field of Artificial Intelligence (AI), signifying the shift towards intelligent and autonomous systems capable of complex understanding and decision-making. This thesis delves deep into the multifaceted capabilities of LLMs, exploring their potential applications in decision optimization, scene understanding, and advanced summarization tasks in diverse contexts.

In the first segment of the thesis, the focus is on Unmanned Aerial Vehicles' (UAVs) semantic scene understanding. The capability of instantaneously providing high-level data and visual cues positions UAVs as ideal platforms for performing complex tasks. The work combines the potential of LLMs, Visual Language Models (VLMs), and state-of-the-art detection pipelines to offer nuanced and contextually accurate scene descriptions. A well-controlled, efficient practical implementation of microdrones in challenging settings is presented, supplementing the study with proposed standardized readability metrics to gauge the quality of LLM-enhanced descriptions. This could significantly impact sectors such as film, advertising, and theme parks, enhancing user experiences manifold.

The second segment brings to light the increasingly crucial problem of decision-making under uncertainty. Using the Multi-Armed Bandit (MAB) problem as a foundation, the study explores the use of LLMs to inform and guide strategies in dynamic environments. It is postulated that the predictive power of LLMs

can aid in choosing the correct balance between exploration and exploitation based on the current state of the system. Through rigorous testing, the proposed LLM-informed strategy showcases its adaptability and its competitive performance against conventional strategies.

Next, the research transitions into studying the goodness-of-fit assessments of Generative Adversarial Networks (GANs) utilizing the Signature Transform. By providing an efficient measure of similarity between image distributions, the study sheds light on the intrinsic structure of the samples generated by GANs. A comprehensive analysis using statistical measures, such as the test Kruskal–Wallis, provides a more extensive understanding of the GAN convergence and goodness of fit.

In the final section, the thesis introduces a novel benchmark for automatic video summarization, emphasizing the harmonious integration of LLMs and Signature Transform. An innovative approach grounded in the harmonic components captured by the Signature Transform is put forth. The measures are extensively evaluated, proving to offer compelling accuracy that correlates well with the concept of a good summary.

This research work establishes LLMs as powerful tools in addressing complex tasks across diverse domains, redefining decision optimization, scene understanding, and summarization tasks. It not only breaks new ground in the applications of LLMs but also sets the direction for future work in this exciting and rapidly evolving field.

Resumen

El advenimiento de los Large Language Models (LLMs) marca una fase transformadora en el campo de la Inteligencia Artificial (IA), significando el cambio hacia sistemas inteligentes y autónomos capaces de una comprensión y toma de decisiones complejas. Esta tesis profundiza en las capacidades multifacéticas de los LLMs, explorando sus posibles aplicaciones en la optimización de decisiones, la comprensión de escenas y tareas avanzadas de resumen de video en diversos contextos.

En el primer segmento de la tesis, el foco está en la comprensión semántica de escenas de Vehículos Aéreos No Tripulados (UAVs). La capacidad de proporcionar instantáneamente datos de alto nivel y señales visuales sitúa a los UAVs como plataformas ideales para realizar tareas complejas. El trabajo combina el potencial de los LLMs, los Visual Language Models (VLMs), y los sistemas de detección objetos de última generación para ofrecer descripciones de escenas matizadas y contextualmente precisas. Se presenta una implementación práctica eficiente y bien controlada usando microdrones en entornos complejos, complementando el estudio con métricas de legibilidad estandarizadas propuestas para medir la calidad de las descripciones mejoradas por los LLMs. Estos avances podrían impactar significativamente en sectores como el cine, la publicidad y los parques temáticos, mejorando las experiencias de los usuarios de manera exponencial.

El segundo segmento arroja luz sobre el problema cada vez más crucial de la toma de decisiones bajo incertidumbre. Utilizando el problema de Multi-Armed Bandits (MAB) como base, el estudio explora el uso de los LLMs para informar y guiar estrategias en entornos dinámicos. Se postula que el poder predictivo de los LLMs puede ayudar a elegir el equilibrio correcto entre exploración y explotación basado en el estado actual del sistema. A través de pruebas rigurosas, la estrategia informada por los LLMs propuesta demuestra su adaptabilidad y su rendimiento competitivo frente a las estrategias convencionales.

A continuación, la investigación se centra en el estudio de las evaluaciones de bondad de ajuste de las Generative Adversarial Networks (GANs) utilizando la Signature Transform. Al proporcionar una medida eficiente de similitud entre las distribuciones de imágenes, el estudio arroja luz sobre la estructura intrínseca de las muestras generadas por los GANs. Un análisis exhaustivo utilizando medidas estadísticas como las pruebas de Kruskal-Wallis proporciona una comprensión más amplia de la convergencia de los GANs y la bondad de ajuste.

En la sección final, la tesis introduce un nuevo benchmark para la síntesis automática de vídeos, enfatizando la integración armoniosa de los LLMs y la Signature Transform. Se propone un enfoque innovador basado en los componentes armónicos capturados por la Signature Transform. Las medidas son evaluadas extensivamente, demostrando ofrecer una precisión convincente que se correlaciona bien con el concepto humano de un buen resumen.

Este trabajo de investigación establece a los LLMs como herramientas poderosas para abordar tareas complejas en diversos dominios, redefiniendo la optimización de decisiones, la comprensión de escenas y las tareas de resumen de video. No solo establece nuevos postulados en las aplicaciones de los LLMs, sino que también establece la dirección para futuros trabajos en este emocionante y rápidamente evolucionante campo.

Resum

L'adveniment dels Large Language Models (LLMs) marca una fase transformadora en el camp de la Intel·ligència Artificial (IA), significat el canvi cap a sistemes intel·ligents i autònoms capaços d'una comprensió i presa de decisions complexes. Aquesta tesi profunditza en les capacitats multifacètiques dels LLMs, explorant les seues possibles aplicacions en l'optimització de decisions, la comprensió d'escenes i tasques avançades de resum de vídeo en diversos contexts.

En el primer segment de la tesi, el focus està en la comprensió semàntica d'escenes de Vehicles Aeris No Tripulats (UAVs). La capacitat de proporcionar instantàniament dades d'alt nivell i senyals visuals situa els UAVs com a plataformes ideals per a realitzar tasques complexes. El treball combina el potencial dels LLMs, els Visual Language Models (VLMs), i els sistemes de detecció d'objectes d'última generació per a oferir descripcions d'escenes matisades i contextualment precises. Es presenta una implementació pràctica eficient i ben controlada usant microdrons en entorns complexos, complementant l'estudi amb mètriques de llegibilitat estandarditzades proposades per a mesurar la qualitat de les descripcions millorades pels LLMs. Aquests avenços podrien impactar significativament en sectors com el cinema, la publicitat i els parcs temàtics, millorant les experiències dels usuaris de manera exponencial.

El segon segment arroja llum sobre el problema cada vegada més crucial de la presa de decisions sota incertesa. Utilitzant el problema dels Multi-Armed

Bandits (MAB) com a base, l'estudi explora l'ús dels LLMs per a informar i guiar estratègies en entorns dinàmics. Es postula que el poder predictiu dels LLMs pot ajudar a triar l'equilibri correcte entre exploració i explotació basat en l'estat actual del sistema. A través de proves rigoroses, l'estratègia informada pels LLMs proposada demostra la seua adaptabilitat i el seu rendiment competitiu front a les estratègies convencionals.

A continuació, la recerca es centra en l'estudi de les avaluacions de bondat d'ajust de les Generative Adversarial Networks (GANs) utilitzant la Signature Transform. En proporcionar una mesura eficient de similitud entre les distribucions d'imatges, l'estudi arroja llum sobre l'estructura intrínseca de les mostres generades pels GANs. Una anàlisi exhaustiva utilitzant mesures estadístiques com les proves de Kruskal-Wallis proporciona una comprensió més àmplia de la convergència dels GANs i la bondat d'ajust.

En la secció final, la tesi introdueix un nou benchmark per a la síntesi automàtica de vídeos, enfatitzant la integració harmònica dels LLMs i la Signature Transform. Es proposa un enfocament innovador basat en els components harmònics capturats per la Signature Transform. Les mesures són avaluades extensivament, demostrant oferir una precisió convincent que es correlaciona bé amb el concepte humà d'un bon resum.

Aquest treball de recerca estableix els LLMs com a eines poderoses per a abordar tasques complexes en diversos dominis, redefinint l'optimització de decisions, la comprensió d'escenes i les tasques de resum de vídeo. No solament estableix nous postulats en les aplicacions dels LLMs, sinó que també estableix la direcció per a futurs treballs en aquest emocionant i ràpidament evolucionant camp.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Artificial Intelligence (AI), propelled by significant advancements in computational resources and modeling paradigms, has transgressed the conventional confines of decision-making, scene understanding, and summarization [1, 2]. The ensuing period encapsulates a transformative phase that steers the direction of AI research towards autonomous systems exhibiting sophisticated comprehension and decision-making capabilities. An instrumental factor in this transformation is the rise of Large Language Models (LLMs) [3]. The research encapsulated within this thesis delves into the extensive capabilities of LLMs and their applicability across varying contexts, particularly in decision optimization, scene understanding, and advanced summarization tasks.

In an era where the volume of visual data is growing exponentially, Unmanned Aerial Vehicles (UAVs) offer a versatile platform to acquire and interpret complex visual cues instantaneously. In this work, we synergize LLMs and Visual Language Models (VLMs) with state-of-the-art detection pipelines for comprehensive scene understanding from UAVs. This integration enables the generation of semantically rich, zero-shot literary text descriptions of scenes [4], evaluated on metrics such as the GUNNING Fog readability index. The potential impact of this work spans various sectors, including advertising, film, and theme parks, underscoring the multifarious applications of LLMs in the

real world [5].

The challenges of decision-making in dynamic, uncertain environments provide a fertile ground for testing the predictive powers of LLMs [6]. Leveraging the theoretical foundation of the Multi-Armed Bandit (MAB) problem, we then propose an LLM-informed strategy that strikes a fine balance between exploration and exploitation based on the system's current state. This research introduces a new non-stationary bandit model with variable reward distributions and demonstrates how LLMs can be used to guide the choice of bandit amidst this volatility [7]. These findings demonstrate the robustness and adaptability of LLMs in dynamic decision-making scenarios, challenging traditional strategies such as epsilon-greedy and upper confidence bound (UCB).

Further delving into the exploration of advanced learning techniques [8], the research shifts towards an in-depth examination of Generative Adversarial Networks (GANs) [9, 11]. Understanding the nuances of GAN convergence and the goodness of fit can shed light on the model's performance. We address this by proposing the Signature Transform [12, 13], a robust and efficient similarity measure between image distributions generated by GANs. This approach not only provides a more comprehensive understanding of GAN behavior but also proves to be more efficient than existing techniques based on deep neural networks, such as Fréchet Inception Distance (FID) and Multi-Scale Structural Similarity Index Measure (MS-SSIM). The paper augments these assessments with statistical measures such as the test Kruskal–Wallis to evaluate the goodness of fit of GAN sample distributions.

Finally, the thesis extends the exploration of VLMs [14] towards video summarization [15, 18, 19]. Traditional video summarization methods suffer from a key limitation - the need for human annotators. To address this, this work proposes a new benchmark for assessing visual summaries using the Signature Transform. Building upon the premise that uniform random sampling can deliver effective summarization, we present a novel technique rooted in the harmonic components captured by the Signature Transform. The results demonstrate a strong correlation with the notion of a good summary, pointing towards a promising new direction in automatic video summarization.

Through this body of work, the thesis firmly establishes LLMs as versatile tools in tackling complex tasks across diverse domains. By integrating ad-

vanced computational approaches, the research not only forges new frontiers in the applications of LLMs but also sets a promising trajectory for future exploration in this rapidly evolving field [20, 21]. The transformative power of LLMs is deeply interwoven into the fabric of AI, gradually redefining how we optimize decisions, understand scenes, and summarize content. The exciting interplay between mathematical intricacies and practical applications signifies the profound impact of LLMs, propelling AI towards a new frontier of innovation and discovery [22].

1.1 Objectives

The main objective of this thesis is to explore and expand the horizons of AI through the application of LLMs in autonomous systems for enhanced decision-making, detailed scene understanding, and comprehensive summarization. To achieve this broad aim, the thesis is structured to accomplish several interlinked objectives:

- To integrate LLMs with UAV technology, thus enabling the autonomous systems to not only capture and process visual data but also to generate real-time, semantically rich textual descriptions that are immediately accessible and actionable.
- To leverage the predictive prowess of LLMs in dynamic and uncertain environments, thereby refining the decision-making processes through a model that effectively balances exploration and exploitation strategies.
- To enhance the understanding and evaluation of GANs by introducing and validating novel assessment methodologies that are computationally efficient and robust.
- To develop and propose a new benchmark for video summarization using the Signature Transform, aimed at overcoming the limitations of traditional methods that rely heavily on human annotation.

These objectives are addressed through a combination of theoretical research and practical experimentation, with each chapter dedicated to exploring these aims within the context of the broader goal of advancing AI through the use of LLMs.

1.2 Methodology

The methodological approach of this thesis is built on a rigorous, multi-phase process designed to address the complex challenges inherent in AI and LLM applications. The research is grounded in both qualitative and quantitative analysis, encompassing a series of structured objectives to develop a comprehensive understanding of the field. The methodology is divided into the following key phases:

1. **Literature Review:** A thorough review of existing research to identify gaps in current knowledge and to provide a solid theoretical framework for the thesis.
2. **Technological Integration:** Development and implementation of advanced computational models, particularly LLMs, within autonomous systems such as UAVs, exploring their capabilities in real-world scenarios.
3. **Algorithmic Development:** Crafting and refining algorithms for decision-making, scene understanding, and summarization, with a focus on leveraging the unique strengths of LLMs.
4. **Experimental Evaluation:** Conducting experiments to test the developed models and algorithms, and to evaluate their performance using a range of metrics and real-world data sets.
5. **Statistical Analysis:** Applying statistical methods to analyze the results of experiments, ensuring that findings are robust, reliable, and generalizable.
6. **Critical Analysis:** Interpreting the outcomes of the experimental phase, drawing insights and understanding their implications in the broader context of AI.
7. **Synthesis and Reporting:** Integrating the insights gained from the analysis to articulate comprehensive conclusions, and to suggest avenues for future research.

The following research questions guide the investigative journey of this thesis:

- How can LLMs be integrated with UAV technology to enhance real-time scene understanding and narrative generation?
- In what ways can LLMs inform and optimize decision-making strategies in dynamic and uncertain environments?

- What novel methodologies can be developed to assess the convergence and performance of GANs more effectively and efficiently?
- How can the Signature Transform be applied to create benchmarks for video summarization that surpass the limitations of human annotation dependency?

By answering these questions, the thesis aims to contribute substantively to the field of AI, pushing the boundaries of current technologies, and providing a blueprint for future research endeavors.

Chapter 2

Background

The domain of AI has experienced a revolution in recent years, fueled by exponential growth in computational power and data availability, culminating in the advent of LLMs. These models, such as GPT-3, have showcased a profound ability to understand and generate human-like text, making them an essential tool in a myriad of applications, ranging from natural language understanding and generation to more complex tasks like scene understanding, decision optimization, and summarization.

In the realm of LLMs, the thesis explores their applicability in multiple facets. Initially, it scrutinizes their role in enhancing UAVs' scene understanding. UAVs, equipped with cutting-edge image and sensor technology, capture vast amounts of visual data. This data, when processed and understood with the aid of LLMs, opens a wide array of applications such as monitoring, reconnaissance, and inspection tasks, among others. Yet, semantic understanding of scenes remains a complex task due to the dynamic and multifaceted nature of real-world environments. Bridging this gap requires an integration of VLMs with LLMs, thereby allowing us to tap into a more comprehensive understanding of visual scenes.

In the vein of decision-making under uncertainty, we delve into a well-established problem in Reinforcement Learning (RL) and stochastic optimization - the MAB problem. In a dynamic environment, where the underlying probabil-

ity distributions change over time, traditional strategies fall short. This necessitates a deeper exploration of how LLMs can be used to devise effective strategies, shaping decisions based on the evolving context.

Another crucial area explored in this thesis involves Generative Adversarial Networks (GANs), which have been pivotal in generating synthetic data indistinguishable from real data. Despite their success, evaluating GANs' performance remains challenging due to the high dimensionality and complexity of the data. Therefore, the study involves using the Signature Transform, a mathematical tool used to encode sequential data, for assessing the goodness-of-fit of empirical distributions generated by GANs.

Lastly, this work takes on the challenge of video summarization, a task of paramount importance considering the ever-increasing volume of video data. The utility of video summarization extends across domains, from aiding content retrieval in large video databases to enhancing the efficiency of video surveillance systems. Here, we harness the capabilities of LLMs, coupled with the Signature Transform, to automate the process of video summarization.

The core of this research lies in investigating the effectiveness of LLMs in these applications and developing computational methodologies that can lead to significant advancements in each of these domains. The overarching aim is to leverage the power of AI to improve the efficiency and effectiveness of systems, be it in interpreting environmental scenes, optimizing decisions in a stochastic setting, understanding the complex structure of generated samples, or succinctly summarizing large volumes of video data. By doing so, the thesis contributes significantly to expanding the frontiers of LLMs.

2.1 Large Language Models (LLMs)

In the burgeoning field of AI, LLMs [4, 2] have emerged as a powerful tool for understanding and generating human-like text, providing a breakthrough in diverse applications such as scene understanding, decision optimization, and summarization.

LLMs, including the likes of OpenAI's GPT-3, are typically instantiated as transformer models, a class of models that relies on self-attention mechanisms. The architecture of transformer models is designed to handle sequential data, making it apt for processing and understanding textual information.

Formally, let's consider a sequence of tokenized text, $\{x_1, x_2, ..., x_T\}$, where each x_z corresponds to a word or sub-word in the text. The goal of a transformer model is to learn the distribution $p(x_1, x_2, ..., x_T)$, which can be decomposed using the chain rule of probability as the product of conditional probabilities:

$$p(x_1, x_2, ..., x_T) = \prod_{z=1}^{T} p(x_z | x_1, ..., x_{z-1}). \tag{2.1}$$

To learn these conditional probabilities, transformer models use self-attention mechanisms. At a high level, the self-attention mechanism computes a weighted sum of all tokens in the input, where the weights determine the contribution of each token to the representation of the current token.

More specifically, the attention weights are calculated as:

$$\alpha_{zo} = \text{softmax}\left(\frac{Q_z K_o}{\sqrt{d_k}}\right), \tag{2.2}$$

where Q_z, K_o are the query and key vectors for tokens z and o, and d_k is the dimensionality of the key vectors. The softmax function ensures that the weights sum to 1, and the scaling factor of $\sqrt{d_k}$ is used to control the magnitudes of the gradients.

Once the attention weights are computed, the output for a given token z is given by:

$$h_z = \sum_o \alpha_{zo} V_o, \tag{2.3}$$

where V_o is the value vector for token o.

The advantage of the self-attention mechanism is that it can model dependencies between tokens regardless of their positions in the sequence, making transformer models very powerful at modeling long-range dependencies in text.

Training these models typically involves maximizing the log-likelihood of the observed data, which corresponds to the objective function:

$$L = \sum_{z} \log p(x_z | x_1, ..., x_{z-1}). \tag{2.4}$$

Despite the power of LLMs, their application in different domains requires careful consideration. For instance, when used for scene understanding in UAVs, LLMs need to be fused with VLMs due to the constraints of the multi-modal problem. When addressing the MAB problem, LLMs should be leveraged to incorporate the evolving context into decision-making. In the case of GANs, the Signature Transform can be used along with LLMs to understand the complex structure of generated samples. For video summarization, the combination of LLMs and Signature Transform can be utilized to automate the process effectively. This research aims to expand the frontiers of LLMs in these domains.

2.2 Visual Language Models (VLMs)

As AI continues to evolve, VLMs [21] have carved out an indispensable niche for themselves within the broader AI research landscape. These models have taken the power of LLMs and extended it into the realm of visual inputs, creating systems capable of understanding and interpreting not just textual data, but also images and other visual inputs.

A prominent exemplar of VLMs is OpenAI's CLIP (Contrastive Language–Image Pretraining) [3]. This model simultaneously learns to understand and translate between images and text by capitalizing on the vast expanse of internet text–image pairs, which act as its training data.

The fundamental idea behind CLIP is to bring images and text into a shared semantic space. More specifically, CLIP learns to map an image and a text snippet into the same high-dimensional vector space such that the cosine similarity between the image and the text snippet vectors corresponds to the semantic similarity of the image–text pair.

Formally, let $f_{\text{image}}(\cdot)$ and $f_{\text{text}}(\cdot)$ denote the image and text encoders of CLIP, respectively, where each encoder maps its input into a d-dimensional vector. Given a mini-batch of N image–text pairs $(x_z, y_z)_{z=1}^{N}$, where x_z and y_z denote the image and the text snippet of the z-th pair, respectively, CLIP is trained by minimizing the following contrastive loss:

$$L = -\sum_{z=1}^{N} \log \frac{\exp(f_{\text{image}}(x_z)^\top f_{\text{text}}(y_z)/\tau)}{\sum_{o=1}^{N} \exp(f_{\text{image}}(x_z)^\top f_{\text{text}}(y_o)/\tau)}, \tag{2.5}$$

where τ is a temperature parameter that controls the sharpness of the distribution.

An interesting aspect of CLIP is that, unlike traditional supervised learning models which learn to map from input to output, it learns to map both images and text into a shared latent space. This, combined with its use of a transformer architecture (similar to LLMs), makes CLIP a powerful tool for a wide range of tasks, including zero-shot classification and generation of textual descriptions of images.

The fundamental premise and structure of CLIP, as well as the impressive results that it delivers, serve as a key motivating factor for the present research. By fusing LLMs and VLMs like CLIP, there are significant opportunities to create hybrid models that can handle complex tasks involving both text and images, such as scene understanding in UAVs. Through careful integration and experimentation, this research aims to explore and harness the potential of such models.

2.3 Semantic Scene Understanding

Semantic Scene Understanding (SSU) is an important aspect of computer vision and AI systems, underpinning various tasks including autonomous navigation, image segmentation, video analysis, and even social robot interactions. With the advent of LLMs and their associated technologies, we can further extend the capabilities of SSU by connecting visual perception with the human language semantics.

At its core, SSU is about understanding the context, entities, and relationships in a given scene. It involves not just object detection or recognition, but a comprehensive understanding of the entities' roles, their properties, and the activities taking place in the scene. Thus, it encapsulates tasks like object detection, semantic segmentation, instance segmentation, and scene graph generation.

2.3.1 Role of Large Language Models in SSU

LLMs and VLMs have emerged as a powerful tool for semantic understanding, capable of handling complex sentence structures, idioms, and even demonstrating an understanding of factual knowledge encoded in text. Transferring these capabilities to SSU allows for better, more nuanced understanding of visual scenes.

Consider a scene with several entities and interactions. The goal of SSU, in this case, is to accurately identify and understand these entities and their relationships. By integrating LLMs, we can extend this understanding to a linguistic level. For instance, if the scene depicts a dog chasing a ball, the SSU model can recognize not only the entities (the dog and the ball) but also their relationship (the chasing action), and the LLM can help in generating a human-understandable description of the scene.

2.3.2 Mathematical Approach

Mathematically, the integration of LLMs with SSU can be viewed as a joint probabilistic model. Consider a scene S and its textual description D. Our goal is to maximize the joint probability $P(S, D)$, which can be decomposed into the product of conditional probabilities $P(S|D)$ and $P(D|S)$. Here, $P(S|D)$ can be seen as the task of generating a scene given a description (e.g., in tasks like image generation), and $P(D|S)$ as the task of generating a description given a scene (e.g., in tasks like image captioning).

$$P(S, D) = P(S|D) \cdot P(D|S). \tag{2.6}$$

Training a model to maximize this joint probability can, in theory, lead to a model that can both understand and generate descriptions of scenes, paving the way for more human-like understanding of visual data.

In practice, one common approach is to use neural networks (e.g., Convolutional Neural Networks (CNNs) for scene understanding and Transformers for language modeling) and train these networks using backpropagation and optimization techniques like stochastic gradient descent.

2.3.3 Applications in Autonomous Systems

In the context of autonomous systems, SSU plays a crucial role in decision-making. For instance, in autonomous vehicles, understanding the scene is vital for safe navigation. Similarly, in surveillance systems, understanding activities in a scene can help detect anomalous behavior. By integrating LLMs into these systems, we can enhance their ability to comprehend scenes, thereby leading to safer, more effective systems.

The ability to translate visual information into natural language could also make these systems more accessible and user-friendly. For example, an autonomous system could provide a natural language description of what it "sees", making its operations more transparent and understandable to human users.

Hence, our research focuses on leveraging the power of LLMs and VLMs to augment SSU in autonomous systems, to improve their performance and adaptability in various environments.

2.4 Multi-Armed Bandits

The MAB problem [26] represents a central challenge in the field of RL, focusing on the perennial trade-off between exploration and exploitation. This problem gets its name from the metaphor of playing a slot machine with multiple levers (or "arms"), where each lever provides a different unknown reward.

Formally, consider a learner interacting over T time steps with an environment that presents K options (or "arms") at each time step. At each time step $t = 1, 2, ..., T$, the learner selects an arm $a_t \in \{1, 2, ..., K\}$, and receives a reward $r^t_{a_t}$ that is sampled from a stationary distribution specific to the selected arm a_t. The goal of the learner is to select arms over the course of T time steps so as to maximize the total reward.

Different strategies, or algorithms, exist to tackle this challenge, with Epsilon-Greedy and Upper Confidence Bound (UCB) being popular approaches.

The algorithm **Epsilon-Greedy** introduces a simple mechanism to balance exploration and exploitation. At each time step, it selects a random arm with probability ϵ (exploration), and selects the arm with the highest empirical mean reward with probability $1 - \epsilon$ (exploitation). Formally, the selected arm a_t is given by:

$$a_t = \begin{cases} \arg\max_{a \in \{1,2,...,K\}} \bar{r}_a^t, & \text{with probability } 1 - \epsilon, \\ \text{Uniform}(\{1, 2, ..., K\}), & \text{with probability } \epsilon, \end{cases} \tag{2.7}$$

where $\bar{r}_a^t$ denotes the empirical mean reward of arm a up to time t.

The **UCB** algorithm, on the other hand, attempts to tackle the exploration-exploitation dilemma by assigning an "optimism" value to each arm, leading to a preference for arms that have either been profitable in the past or have been little explored. Specifically, it selects the arm with the highest upper confidence bound at each time step, where the upper confidence bound of an arm is defined as the sum of its empirical mean reward and a confidence term. Formally, the selected arm a_t is given by:

$$a_t = \underset{a \in \{1,2,...,K\}}{\arg\max} \left(\bar{r}_a^t + \sqrt{\frac{2 \log t}{n_a^t}} \right), \tag{2.8}$$

where n_a^t denotes the number of times arm a has been selected up to time t.

While this problem has been studied extensively, it still forms the basis for much of the experimentation and advancement in RL [28]. In particular, our research explores the use of these techniques in the context of autonomous systems, studying the possibility of implementing multi-armed bandits with LLMs to tackle the problem of exploration and exploitation in non-stationary environments.

2.4.1 Multi-Armed Bandits and LLMs

A problem frequently encountered in the field of RL is the dilemma of exploration and exploitation, which is most notably exemplified by the MAB problem. In this scenario, the task is to choose, repeatedly and over a sequence of time steps, from a set of strategies with uncertain outcomes, in order to maximize the total reward. This context encapsulates a trade-off between "exploration", where new or less-known strategies are chosen to gather more information, and "exploitation", where strategies with the highest expected reward based on current knowledge are chosen.

The MAB problem can be modeled as follows: we have a set of K slot machines (or "arms"), each providing a random reward when pulled, according to an

unknown probability distribution. At each time step $t = 1, 2, ..., T$, we choose one arm $a_t \in \{1, ..., K\}$ to pull, and receive a reward $r_{a_t}(t)$.

The goal is to find a strategy, or "policy", which specifies the probability of pulling each arm at each time step, in order to maximize the expected total reward over T time steps.

Formally, let $R_a(t)$ denote the random reward received when pulling arm a at time t, and let $\pi(t, a)$ denote the probability of pulling arm a under policy π at time t. The expected total reward is given by:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}\pi(t,a)R_a(t)\right]. \tag{2.9}$$

In practice, the MAB problem and its variants are found in many aspects of AI and Machine Learning (ML), such as online advertising, web page design, clinical trials, and more. Recently, with the rise of LLMs, we have started to explore their potential applications in tackling the exploration-exploitation dilemma.

LLMs, such as GPT-3, can generate human-like text and understand the context, making them excellent candidates for building decision-making systems. These models can be used to learn the value function of each action (arm) based on past observations, predict the reward, and choose the next action accordingly. This adds a level of sophistication to traditional bandit algorithms, as the model can also take into consideration the context in which the decision is being made.

Moreover, the ability of LLMs to generate diverse yet contextually relevant responses can be used to address the exploration aspect of the problem. They can generate novel actions or strategies that haven not been tried before but seem promising given the context, thereby aiding in the discovery of potentially better strategies.

In conclusion, the use of LLMs in the MAB framework introduces a promising approach to tackle the exploration-exploitation dilemma, facilitating a more nuanced and efficient strategy exploration and enabling a more robust and adaptable decision-making process.

2.5 Generative Adversarial Networks

GANs are a class of AI algorithms used in unsupervised ML, conceived by [9]. The objective is to generate new data that mimics the distribution of the training set. These innovative models comprise two parts: a generator network and a discriminator network, each serving its specific function in the generative process.

2.5.1 Architecture of GANs

The architecture of a GAN [10] consists of two primary components: the generator (G) and the discriminator (D). The generator's role is to sample from an arbitrary noise distribution (z), and its objective is to generate data that resembles the training data as closely as possible. The discriminator, on the other hand, aims to distinguish between the real data drawn from the training distribution and the fake data produced by the generator.

Formally, the generator maps the latent space to the data space, $G : z \to x$, where z is drawn from a prior noise distribution $z \sim p_z$. The discriminator is a binary classifier, $D : x \to [0, 1]$, that outputs the probability that x comes from the real data rather than G.

2.5.2 GAN Training

Training GANs involves a two-player min-max game where the generator is trying to fool the discriminator by generating real-looking data, and the discriminator is trying to correctly classify real vs. generated data. Given a generator G and a discriminator D, the value function $V(G, D)$ for this min-max game is:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \quad (2.10)$$

Here, the first term $\mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)]$ corresponds to the expectation of the logarithm of the discriminator outputs on the real data. This term is maximized when the discriminator correctly classifies real data as real. The second term $\mathbb{E}_{z \sim p_z(z)}[\log(1-D(G(z)))]$ corresponds to the expectation of the logarithm of (1 - discriminator outputs on the generated data). This term is maximized when the discriminator correctly classifies fake data as fake.

The generator's goal is to minimize this function to generate better fake data that the discriminator classifies as real. The two networks are thus engaged in a continuous adversarial game, where the generator is continuously learning to create better fakes, and the discriminator is learning to get better at distinguishing them from the real data. The training typically continues until a state of equilibrium is achieved, where the generator produces perfect fakes, and the discriminator is left guessing at random whether a given sample is real or fake.

Actual research employs GANs for a variety of applications, most notably for generating synthetic data that can be used for instance to improve the training of models in data-scarce environments, a common scenario in autonomous systems. The flexibility and power of GANs make them a valuable tool for addressing such challenges.

2.6 Video Summarization

Video summarization [15] is an important area of research within computer vision that seeks to create a shorter version of a video while maintaining its essential content and semantic meaning. The goal is to provide a compressed representation that can facilitate quick understanding of the video's content without needing to watch the entire video. Applications range from surveillance systems to content recommendation, video archives, and beyond.

There are two primary forms of video summarization: keyframe-based and segment-based. In keyframe-based summarization, representative frames from the video are selected to create a summary, while in segment-based summarization, short clips from the video are concatenated to produce a summary.

2.6.1 Problem Formulation

Given a video V comprised of a sequence of n frames $V = \{f_1, f_2, ..., f_n\}$, the goal of video summarization is to select a subset of keyframes or segments S such that the summary adequately represents the original content. The key challenge lies in determining which frames or segments are 'important' or 'interesting'.

2.6.2 Keyframe-Based Summarization

Formally, the task of keyframe-based summarization [16] can be described as a binary labeling problem. Each frame f_o is associated with a binary variable $y_o \in \{0, 1\}$. If $y_o = 1$, the frame f_o is included in the summary, otherwise it is not. The objective function for this task can be defined as:

$$S^* = \arg\max_{S \subseteq V} U(S, V). \tag{2.11}$$

Here, $U(S, V)$ is a utility function measuring how well the summary S represents the original video V. Various methods have been proposed to define and optimize this utility function, often involving notions of representativeness and diversity. The problem is usually subject to a length constraint, which restricts the summary to a specified fraction of the video length.

2.6.3 Segment-Based Summarization

Segment-based summarization [17], on the other hand, involves selecting a set of segments or clips from the video. In this case, the binary variables y_o are associated with segments (comprising one or more frames) rather than individual frames. A similar objective function can be defined as above, but the optimization process is typically more complex due to the need to determine both the position and length of each segment.

2.6.4 Approaches for Video Summarization

Approaches for video summarization can be categorized into supervised, unsupervised, and RL methods. In supervised methods, a model is trained on a dataset of videos with human-annotated summaries. In unsupervised methods, the summary is created based on some predefined criteria, such as visual feature clustering or attention mechanisms. RL methods model video summarization as a sequential decision-making problem and learn a policy to select frames or segments based on a reward function.

Our research in video summarization employs a combination of these methods, with a particular focus on the Signature Transform for their adaptability to diverse video content. Video summarization is of paramount importance for efficient information extraction in various autonomous systems, especially when dealing with vast amounts of video data. This research contributes to

the state of the art, enhancing the capacity to quickly understand and make decisions based on video content.

2.7 Signature Transform

The Signature Transform [12, 13] is a powerful mathematical tool rooted in rough path theory, an area of stochastic analysis. It has gained popularity in the field of ML for its capability to succinctly capture the key features of sequential or path-like data, including time series and video data.

2.7.1 Conceptual Overview

Conceptually, the Signature Transform provides a systematic way to encode the effect of a path. The signature of a path, in its infinite-dimensional form, can capture the entire path up to arbitrary precision, encapsulating not only the position and velocity of a path but also higher-order interactions.

2.7.2 Mathematical Definition

Let Ω denote the space of paths of length T on $\mathbb{R}^d$, where d is the dimension of the path. A path $\omega \in \Omega$ is a continuous function $\omega : [0, T] \to \mathbb{R}^d$. The signature of a path ω up to level N is defined as:

$$S^N(\omega) = (1, S^1(\omega), S^2(\omega), ..., S^N(\omega)), \tag{2.12}$$

where $S^k(\omega)$ is the k-th iterated integral of the path ω, defined as:

$$S^k(\omega) = \int_{0 \le t_1 \le ... \le t_k \le T} d\omega_{t_1} \otimes ... \otimes d\omega_{t_k}. \tag{2.13}$$

Here, $\otimes$ denotes the tensor product, and the integral is interpreted in the sense of Young's integration for rough paths.

2.7.3 Computational Aspects

In practice, paths are often discretized and the signature is truncated at a certain level N. This results in a finite-dimensional feature vector that can be computed efficiently.

2.7.4 Applications to GANs and Video Summarization

Our research leverages the Signature Transform in the assessment of Generative Adversarial Networks (GANs) and video summarization. For GANs, the Signature Transform can provide a robust measure of the divergence between the distribution of generated samples and the distribution of real data. This can inform the training process and lead to the generation of more realistic samples.

In the context of video summarization, the Signature Transform can be used to extract features from the video frames, providing a compact and informative representation of the video. This signature representation can then be used as input to a video summarization model, enhancing the model's ability to understand the content of the video and select representative frames or segments.

The use of the Signature Transform in these contexts aligns with our overarching theme of advancing autonomous systems through innovative AI methodologies. By providing a versatile tool for capturing intricate path features, the Signature Transform holds great potential for advancing state of the art in these and other application domains.

2.8 Technological Foundations and Application Domains

The methodologies employed in this thesis rest on a dual foundation of advanced AI techniques and their targeted application domains. To elucidate this relationship, it is imperative to distinguish between the technologies used — namely, LLMs, VLMs, and GANs — and the specific fields of application, such as UAV-based surveillance, decision-making in stochastic environments, and multimedia content analysis.

2.8.1 Technologies Used

- **LLMs:** These models form the core of our text-based analysis, prediction, and generation. Their capacity to process and generate human-like text underpins their use in scene understanding and decision-making applications.
- **VLMs:** They extend the capabilities of LLMs to the visual domain, allowing for an integrated understanding of images and text, crucial for applications in UAV-based monitoring and video summarization.
- **GANs:** These architectures are leveraged for their synthetic data generation capabilities, providing a basis for modeling and evaluating the goodness-of-fit in data distributions, particularly useful in image and video-related applications.

2.8.2 Application Domains

Each application domain benefits from a tailored combination of the aforementioned technologies:

- **UAV-based Surveillance:** LLMs and VLMs are synergistically integrated to enhance the real-time processing and interpretation capabilities of UAVs. The combination is justified by the need for a comprehensive and instantaneous understanding of visual data, which LLMs facilitate through descriptive analytics, while VLMs enable the contextualization of images within the surrounding environment.
- **Stochastic Decision-Making:** The dynamic nature of environments in which decisions must be optimized necessitates the predictive power of LLMs. They are utilized to model and forecast outcomes, aiding in the formulation of strategies that account for the uncertainty inherent in such settings.
- **Multimedia Content Analysis:** GANs, paired with the Signature Transform, address the challenge of evaluating and summarizing vast amounts of video data. The use of GANs allows for the generation of new content and the assessment of distributional fidelity, while the Signature Transform offers a compact and computationally efficient representation of video sequences for summarization tasks.

This delineation not only clarifies the utilization of specific AI technologies in distinct application areas but also provides a rationale for their integration,

ensuring that the methodology strategy is both explicit and robust. The subsequent chapters will delve deeper into each application, showcasing the efficacy of these technological combinations in real-world scenarios.

The justification for the combination of AI approaches in each application field is rooted in the complementary strengths of each technology:

- In UAV-based surveillance, the real-time semantic interpretation by LLMs is enriched by the VLMs' ability to contextualize visual data, forming a comprehensive scene understanding framework.
- For stochastic decision-making, LLMs' predictive analytics are critical for adapting to changing conditions and uncertainties, enhancing the strategic decision-making processes.
- In multimedia content analysis, the creative power of GANs to generate lifelike images and the Signature Transform's efficiency in summarization form a powerful duo for analyzing and synthesizing video content.

The thesis presents a coherent methodological strategy that aligns with the fundamental goals of advancing autonomous systems' capabilities through AI.

Chapter 3

Semantic Scene Understanding with LLMs on Unmanned Aerial Vehicles

J. de Curtò, I. de Zarzà and Carlos T. Calafate. (2023). "Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles." Drones, vol(7), 114. DOI: 10.3390/drones7020114

Unmanned Aerial Vehicles (UAVs) are able to provide instantaneous visual cues and a high-level data throughput that could be further leveraged to address complex tasks, such as semantically rich scene understanding. In this chapter, we built on the use of Large Language Models (LLMs) and Visual Language Models (VLMs), together with a state-of-the-art detection pipeline, to provide thorough zero-shot UAV scene literary text descriptions. The generated texts achieve a GUNNING Fog median grade level in the range of 7–12. Applications of this framework could be found in the filming industry and could enhance user experience in theme parks or in the advertisement sector. We demonstrate a low-cost highly efficient state-of-the-art practical implementation of microdrones in a well-controlled and challenging setting, in addition to proposing the use of standardized readability metrics to assess LLM-enhanced descriptions.

3.1 Introduction and Motivation

Unmanned Aerial Vehicles (UAVs) have proven to be an essential asset for practically addressing many challenges in vision and robotics. From surveillance and disaster response to the monitoring of satellite communications, UAVs perform well in situations where seamless mobility and high-definition visual capture are necessary. In this work, we focused on tasks that require a semantic understanding of visual cues and that could guide initial estimates in proposing an adequate characterization of a certain environment. Problems that are of interest include semi-adaptive filming [1] and automatic literary text description. In this setting, we propose a complete pipeline that provides real-time original text descriptions of incoming frames or a general scene description given some pre-recorded videos. The descriptions are well-suited to creating an automatic storytelling framework that can be used in theme parks or family trips alike.

Foundation models are techniques based on neural networks that are trained on large amounts of data and that present good generalization capabilities across tasks. In particular, Natural Language Processing (NLP) has seen a dramatic improvement with the appearance of GPT-2 [2] and its subsequent improvements (GPT-3 [3]). Indeed, Large Language Models (LLMs) and Visual Language Models (VLMs) have recently arisen as a resource for determining widespread problems in disciplines from robotics manipulation and navigation to literary text description, completion, and question answering. We attempt to introduce these techniques in the field of UAVs by providing the vehicle with enhanced semantic understanding. Our approach uses a captioning technique based on CLIP [5, 4], along with the YOLOv7 detector [6], which enhances the captioning output with the object annotations detected and then wires the text into GPT-3.

The descriptions provided are accurate and show a detailed understanding of the scene, and they introduce hallucinated elements that yield sound and consistent seed captions. The literary style allows for the system to be used in a wide variety of situations; for example, a human companion can use the generated text for assistance in writing a script.

The system can be used without fine-tuning in a wide variety of environments, as the base models are trained on large amounts of data. However, to further improve the consistency of the descriptive text, a proper fine-tuning of the detector could be useful when the objects that the system would normally encounter are not present in the COCO object classes [8, 7], or when one wants to emphasize certain aspects of the visual cues; for instance, in an amusement

park, a fine-tuning of the data could add specificity to the descriptions, e.g., providing captions that include trademark imaginary characters or specific attractions, rides, or games.

This thesis proposes, from the point of view of system integration, a novel zero-shot literary text description system using state-of-the-art large-language modules through the use of microdrones (RYZE Tello) and UAVs; additionally, a proposed set of measures is newly introduced in this context to assess the adequacy of the output text for the target audience.

One of the main technical issues of applying LLMs to UAVs is that the data have to be relayed to the computer, where either computation has to take place or a query has to be formulated to use an API. On-board processing is possible, but it is limited due to the amount of GPU memory that state-of-the-art models need. A high-definition camera, well-calibrated and possibly stabilized, is crucial for the optimal behavior of the overall system, as it mainly relies on visual cues for processing the entire pipeline. Another limitation is due to the object detector (YOLOv7) that is used to improve the query formulation prior to using GPT-3; in this particular setting, we used a pretrained model trained on the COCO dataset, but specific training data may be needed for a target application. Furthermore, the object detector could be integrated into the on-board processing using a CORAL board.

The main goal of this chapter is to propose a system that could be used in many real-life applications. The majority of the techniques used have been thoroughly tested in standard datasets before, but there has been little experimentation in real settings with varying conditions and equipment. For testing the system, we used standardized measures originally used to assess texts written by human instructors in the context of the military, education, and so on.

3.2 Contribution and Organization of the Chapter

A low-cost, highly efficient practical implementation of the system was performed through the use of microdrones (e.g., RYZE Tello), which perform real-time video streaming on a ground computer that controls the vehicle. The level of autonomy of the system could be further enhanced by performing part of the computation on-device; for example, by the attachment of a CORAL Dev Board Mini (Google), which only adds 26 g of payload, to the body of the microdrone. This endows the UAV with a TPU (2GB) that can process

on-device real-time detections, for instance, through the use of state-of-the-art models such as SSD MobileNet V2 [9] and EfficientDet-Lite3x [10].

The RYZE Tello drone is a compact and lightweight quadrotor drone designed for use in educational and recreational applications. It is equipped with an Intel processor and a variety of sensors, including a camera, an IMU, and ultrasonic range finders. The drone is capable of autonomous flight using a pre-programmed set of commands and can be controlled remotely using a compatible device, such as a smartphone. It is also equipped with a number of interactive features, such as gesture control and throw-to-fly, which allow users to easily interact with the drone in a variety of ways; that is, the RYZE Tello drone is a versatile and user-friendly platform that is well-suited for a wide range of applications, including education, entertainment, and research.

A more professionally driven, inexpensive prototype appropriate for outdoor use was attempted by the use of an NXP Hover Games Drone Kit with a CORAL Dev Board Mini (Google) and a high-definition camera (see Figure 3.1). It also includes GPS and a Flight Management Unit (FMU) that supports the PX4 autopilot flight stack. Autonomy could be enhanced by the use of a LiDAR lite-v3 for navigation purposes, a lightweight 23 g light-ranging device with a high accuracy and range (40 m). In a well-controlled situation, such as a film studio, a tethered UAV could be used to eliminate the limitation of the battery capacity of the vehicle.

Figure 3.1: RYZE Tello Microdrones and the NXP Hover Games Drone Kit.

The NXP Hover Games Drone Kit is a hardware and software platform designed for the development and evaluation of autonomous drone systems. It includes a quadrotor drone equipped with an NXP S32 processor, a variety of

sensors including an IMU, ultrasonic range finders, and stereo cameras, and a range of peripherals such as LED lights and a buzzer. The kit also includes a software library and sample code for implementing various autonomous flight behaviors such as hovering, takeoff, and landing. It is intended for use by researchers and developers working in the field of autonomous drone systems, and can be used for a wide range of applications, including drone racing, search and rescue, and aerial photography. Overall, the NXP Hover Games Drone Kit is a comprehensive and versatile tool for exploring the capabilities and limitations of autonomous drone systems.

Experimental results based on a UAV testbed show that the proposed pipeline is able to generate accurate state-of-the-art zero-shot UAV literary text descriptions.

The remainder of the chapter is structured as follows: an overview of state-of-the-art approaches that entail the use of foundation models is provided. Next, Section 3.4 addresses the proposed methodology, as well as the background for the prior knowledge needed for the experimental assumptions, while experiments are presented in Section 6.4. Section 3.6 proposes standardized readability metrics to evaluate LLM-generated descriptions. Finally, Section 6.5 provides the conclusions and describes further work.

3.3 Overview and State of the Art

Large Language Models (LLMs) [12, 11, 13] and Visual Language Models (VLMs) [5] have emerged as an indispensable resource to characterize complex tasks and bestow intelligent systems with the capacity to interact with humans in an unprecedented way. These models, also called foundation models, are able to perform well in a wide variety of tasks, e.g., in robotics manipulation [14, 16, 15], and can be wired to other modules to act robustly in highly complex situations, such as in navigation and guidance [18, 17].

LLMs are ML models that are trained on very large datasets of text and are capable of generating human-like text. These models are typically based on neural networks, which are composed of interconnected processing units that are able to learn and adapt through training. The goal of large language models is to learn the statistical patterns and relationships present in the training data and use this knowledge to generate coherent and plausible text.

One of the key features of large language models is their ability to generate text that is difficult to distinguish from text written by humans. These models

are trained on vast amounts of text and, as a result, are able to capture a wide range of linguistic patterns and structures, including syntax, grammar, and vocabulary. This enables them to generate text that is highly coherent and grammatically correct, and these models can thus be used for a variety of tasks, such as translation, summarization, and text generation.

In addition to their language generation capabilities, large language models have also been shown to be effective at a variety of natural language processing tasks, including language translation, question answering, and text classification. In essence, LLMs are a powerful and versatile tool for understanding and working with natural language data.

Visual Language Models (VLMs) are ML models that are trained on large datasets of text and images and are capable of generating natural language text that is coherent and grammatically correct. The goal of VLMs is to learn the statistical patterns and relationships present in the training data and use this knowledge to generate text that is descriptive and informative about the visual content of an image or a set of images.

One of the key features of visual language models is their ability to generate text that is grounded in the visual content of an image or a set of images. This means that the text generated by these models is specifically related to the objects, people, and events depicted in the image and provides descriptive and informative details about these elements. For example, a VLM could be used to generate a caption for an image depicting the occurrence of a particular action.

In addition to generating descriptive text, visual language models can also be used for a variety of other tasks, such as image classification, object detection, and image captioning. These models can be trained to recognize and classify different types of objects and events in an image and can also be used to generate coherent and grammatically correct captions that describe the content of an image.

VLMs are a powerful and versatile tool for understanding and working with both text and image data. By enabling the generation of descriptive and informative text that is grounded in the visual content of an image, these models have the potential to facilitate a wide range of applications, including image and video analysis, content generation, and natural language processing.

Drones, also known as unmanned aerial vehicles (UAVs), have the potential to be used for a wide range of applications involving semantic scene understanding, which refers to the ability of a system to analyze and interpret the meaning

or significance of the objects, people, and events present in a scene. This capability is important for many applications, including robotics, surveillance, and autonomous driving.

One way in which drones can be used for this particular purpose is through the use of on-board sensors and cameras to capture visual data and other types of data about the environment. These data can then be processed and analyzed using ML algorithms to identify and classify the objects and events present in the scene. For example, a drone equipped with a camera and an object recognition algorithm could be used to identify and classify different types of objects in a scene, such as vehicles, pedestrians, and buildings.

In addition to object recognition, drones can also be used for other types of tasks, such as event detection and tracking. For example, a drone equipped with a camera and an event detection algorithm could be used to identify and track the movements of people or vehicles in a scene. This could be useful for applications such as surveillance or traffic monitoring. By enabling the analysis and interpretation of the meaning or significance of objects and events in a scene, drones can provide valuable insights and information for a variety of tasks and scenarios. In this work, we built on the improvements in object detection [19, 20] and model reparameterization [21, 22] to apply LLMs and VLMs in the field of Unmanned Aerial Vehicles (UAVs) [1]. State-of-the-art techniques of captioning [23, 24, 25] have allowed computers to semantically understand visual data, while advances in automated storytelling can now generate realistic storylines from visual cues [26, 27].

3.4 Methodology

UAV real-time literary storytelling refers to the use of Unmanned Aerial Vehicles (UAVs), also known as drones, to generate narrative stories in real-time based on data they collect. This could involve using the UAVs to capture visual data and other types of data about the environment and then processing and analyzing these data using ML algorithms to identify the objects and events present in the scene. The resulting data could then be used to generate a narrative story that describes and explains the objects and events in the scene coherently and grammatically.

One potential application of UAV real-time literary storytelling is in the field of journalism, where UAVs could be used to capture newsworthy events and generate narratives about these events in real time. For example, a UAV could be used to capture images and video of a natural disaster and then generate

a narrative story about the disaster that is based on the data collected by the UAV. This could provide a more immersive and interactive way of reporting on events and could enable journalists to generate stories more quickly and efficiently.

Another potential application is in the field of entertainment, where UAVs could be used to capture data about live events and generate interactive narratives about these events in real time. For example, a UAV could be used to capture data about a sports game and then generate a narrative story about the game that is based on the data collected by the UAV. This could provide a more engaging and interactive way of experiencing live events and could enable users to experience events in a more immersive and interactive way.

UAV real-time literary storytelling offers potential for a wide range of applications, including journalism, entertainment, and education. By enabling the generation of narrative stories in real time based on data collected by UAVs, this technology has the potential to facilitate a more immersive and interactive way of experiencing and understanding events and situations.

CLIP (Contrastive Language-Image Pre-training) is a neural network architecture developed by researchers at OpenAI that can be used for image captioning and other natural language processing tasks. It is based on the idea of pre-training a model on a large dataset of images and text and then fine-tuning it for a specific task, such as image captioning.

CLIP uses a transformer architecture, which is a type of neural network that is particularly well-suited for tasks involving sequential data, such as natural language processing. The model is trained to predict the next word in a sentence given the previous words, using the images as additional context. One key feature of CLIP is that it is able to learn a continuous space of image and text representations, which allows it to generate high-quality captions for a wide range of images. It is also able to learn from a large amount of data, which helps it to generalize to new images and improve the performance in the image captioning task.

The problem of captioning can be formulated as follows: given a dataset of paired images and captions $\{x^z, c^z\}_{z=1}^N$, the aim is to be able to synthesize adequate captions given an unseen sample image. In our approach, we built on recent work that uses the embedding of CLIP as a prefix to the caption and that is based on the next objective, where the captions can be understood as a sequence of tokens $c^z = c_1^z, \ldots, c_\ell^z$, padded to a maximum length ℓ:

$$\max_{\theta} \sum_{z=1}^{N} \sum_{w=1}^{\ell} \log p_{\theta}(c_w^z |\, x^z, c_1^z, \ldots, c_{w-1}^z). \tag{3.1}$$

We consider, as in [4], an autoregressive language model that predicts the consequent token without considering future tokens.

The CLIP embedding is then projected by a mapping network, denoted as F:

$$p_1^z, \ldots, p_k^z = F(\text{CLIP}(x^z)), \tag{3.2}$$

where p_w^z is a vector with the same dimension as a word embedding and then concatenated with the caption embedding. A cross-entropy loss is used to train the mapping F.

YOLO (You Only Look Once) [19, 20] is a real-time object detection algorithm. It is an end-to-end neural network model that is able to detect and classify objects in images and videos. YOLO works by dividing the input image into a grid of cells and predicting the class and location of objects within each cell. The model uses anchor boxes to make predictions at multiple scales, so it can detect objects of different sizes. The model also predicts the confidence of each detection, which helps to filter out false positives.

One of the main advantages of YOLO is its speed. It is able to process images and videos in real time, making it suitable for use in applications such as video surveillance and autonomous vehicles. YOLO has undergone several versions, with each version improving the accuracy and efficiency of the model. YOLOv7 is the latest version of YOLO and includes several enhancements over previous versions.

We propose a general pipeline for UAV real-time literary storytelling (see Figure 3.2) that is based on the previously described captioning technique that utilizes CLIP prefix captioning [5, 28, 4] and that combines the obtained sentence trained with Conceptual Captions [29] with detections given by YOLOv7 [6]. The output of the object detector is processed by a module of sentence formation such that it can be fed into a GPT-3 module, which provides an enhanced literary description. A query formulating the task to be determined by GPT-3 is needed. The system can work in real time on the streaming frames of the vehicle or as a post-processing module once the UAV has landed.

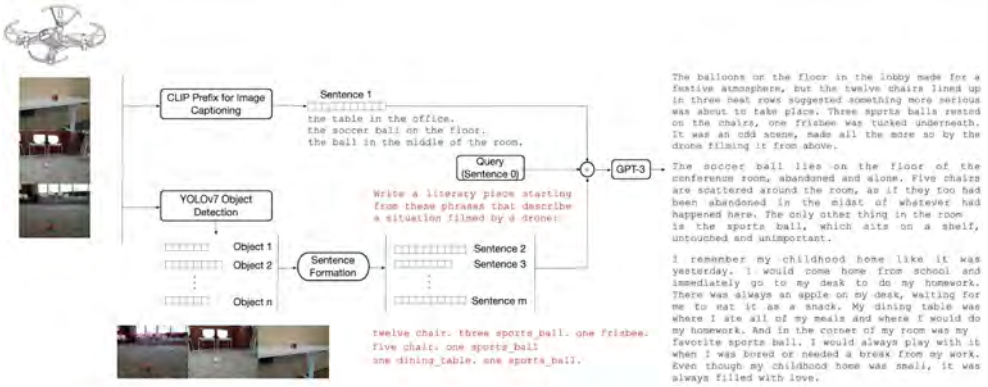


Figure 3.2: UAV real-time literary storytelling.

The pipeline does not require fine-tuning to specific tasks, although it would benefit from such tuning if used in a particular environment where some specific objects need to be identified, e.g., when there is a need to be specific in terms of trademark names.

The main blocks of the architecture are CLIP Prefix for Image Captioning, YOLOv7, and GPT-3.

CLIP Prefix for Image Captioning is a transformer-based architecture that enables the generation of captions while the CLIP and GPT-2 model are frozen. It consists of the training of a lightweight mapping network based on a transformer [31, 30] that translates from the CLIP embedding space to GPT-2.

YOLOv7 is the state-of-the-art object detector in terms of speed and accuracy. It is a generalization of previous YOLO-based architectures with the use of Extended Efficient Layer Aggregation Networks (E-ELANs) [6]. E-ELANs address the problem of controlling the shortest longest gradient path so that the network converges effectively. It uses expand, shuffle, and merge cardinality to continue learning without losing the original gradient path.

GPT-3 is used to enhance the captions by the natural language instruction and prompt engineering. All of our experiments used the API of OpenAI, and the model is surprisingly effective with zero-shot prompts.

Having said that, the chapter has the goal of deploying state-of-the-art LLMs to accomplish the task of zero-shot semantic scene understanding through the use of a low-cost UAV (RYZE Tello or a NXP Hover Games Drone Kit) that

incorporates a high-definition camera. Further integration by the use of a Raspberry Pi Zero W or a CORAL board can move some of the computation on-device with the proper module adaptation, both for object detection and also for the LLM API. In the latter case, a call to OpenAI API is necessary at this stage but advances on the field will soon make it possible to test the trained models directly on-board (e.g., pruning the LLM model to make it fit on memory) without the need to relay the video frames to the computer for further processing. In either way, model pruning can be used to reduce the model size and thus reduce the computational requirements. Another technique would be to use model quantization to reduce the precision of the model and make it more efficient. Additionally, another viable approach is knowledge distillation, where the knowledge of a large teacher model is transferred to a smaller student model for the purpose of using it on a resource-constrained environment.

3.5 Results and Experiment Set-Up

Experiments were conducted on a well-controlled challenging environment with the use of RYZE Tello, streaming the data in real time to a ground computer that processes the frames one by one. Figures 3.3–3.6 illustrate all of the stages of the used methodology for a number of UAV captured stream frames, with contrasting levels of descriptive goodness. The drone captures a particular visual scene that is consequently sent to the ground computer, where a first caption is generated using CLIP Prefix for Image Captioning with beam search. The caption is improved by the output of a YOLOv7 object detector after sentence formation. Finally, a query is formulated together with the resultant caption to generate an enhanced text description by the GPT-3 module. The results are consistent and robust and exhibit original and genuine descriptions of the imagery, and the modules of captioning and object detection are effective and efficient.

CLIP Prefix for Image Captioning presents an average runtime of 206 ms and the YOLOv7 Object Detector presents one of 556 ms, using a GPU Tesla V100-SXM2 (16GB) and a two-core Intel(R) Xeon(R) CPU @ 2.00 GHz. Prompts to the OpenAI API take an average of 3.9 s using the completion model text-davinci-002, where max_tokens = 401 and temperature = 0.9.

Figure 3.3 shows the methodology when the CLIP captioning module and the YOLOv7 object detection produce accurate outputs, and the GPT-3 module produces a very good enhanced literary description.

Figure 3.4 shows the methodology when the CLIP captioning module and the YOLOv7 object detection produce relatively good outputs (detection of the majority of objects—not all or not completely accurate), and the GPT-3 generates a realistic literary description but with the presence of hallucinated elements that provide realism but are not actually in the scene.

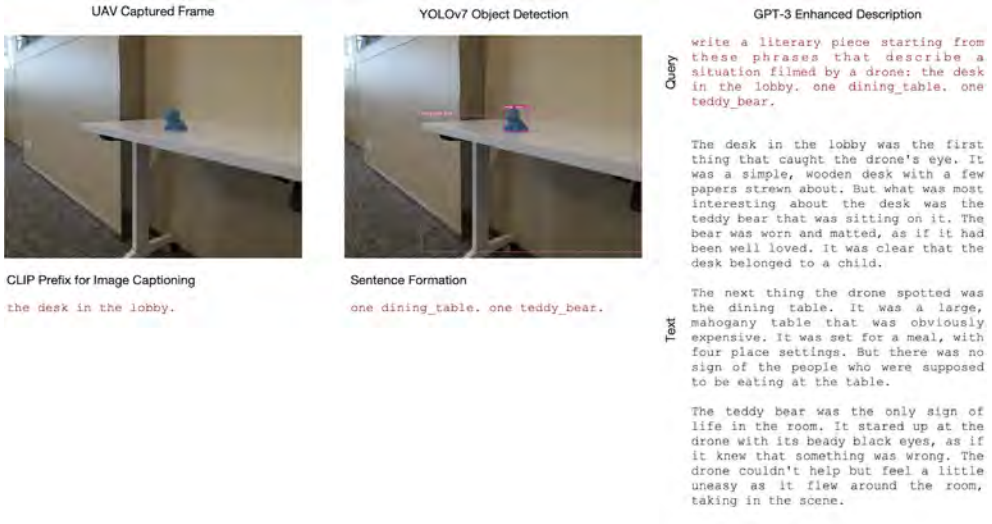
Figure 3.5 shows the methodology when the CLIP captioning module and the YOLOv7 produce somewhat adequate outputs, but they are not particularly accurate, e.g., detecting objects but misclassifying some of them, or generating overly general caption descriptions. The GPT-3 then produces an enhanced description, but not a very accurate one.

Finally, Figure 3.6 shows the methodology when the CLIP captioning module or YOLOv7 object detection fail to describe the scene accurately, and the GPT-3 module generates an erroneous text description.

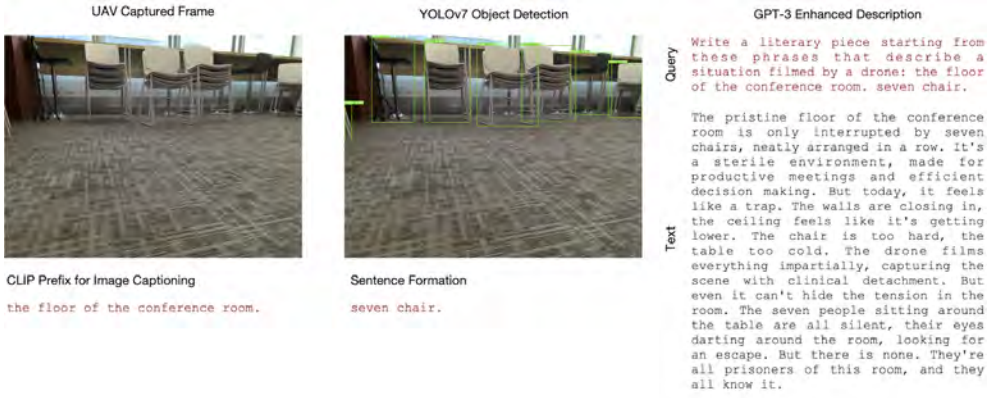
Figure 3.3: UAV captured frame processing and GPT-3. Very good GPT-3 descriptions of the scene.

(a)

Figure 3.4: *Cont.*



(b)

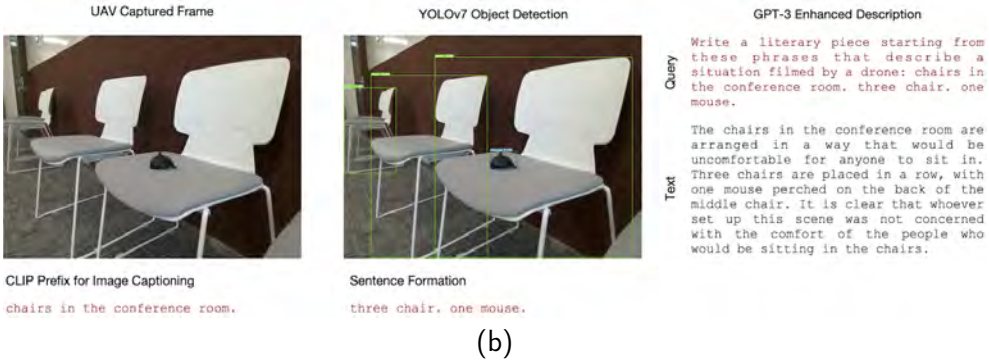


(c)

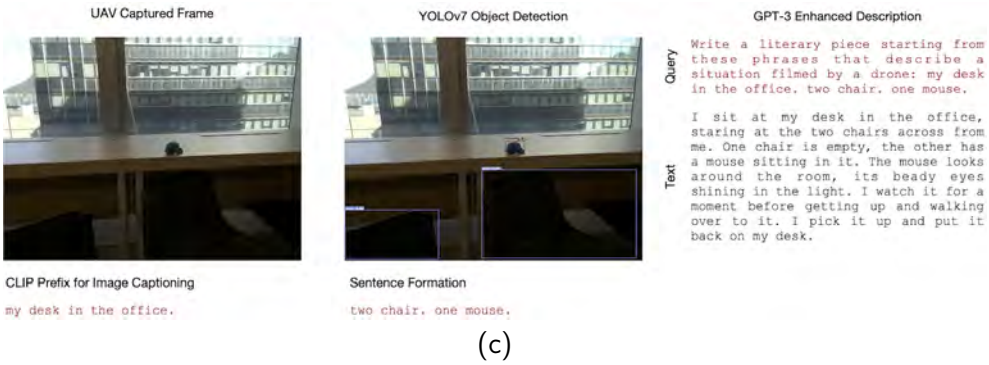
Figure 3.4: UAV captured frame processing and GPT-3. Adequate literary GPT-3 descriptions.

(a)

Figure 3.5: *Cont.*



(b)



(c)

Figure 3.5: UAV captured frame processing and GPT-3. Somewhat good descriptions, but the CLIP captioning module and the YOLOv7 produce inaccurate outputs.

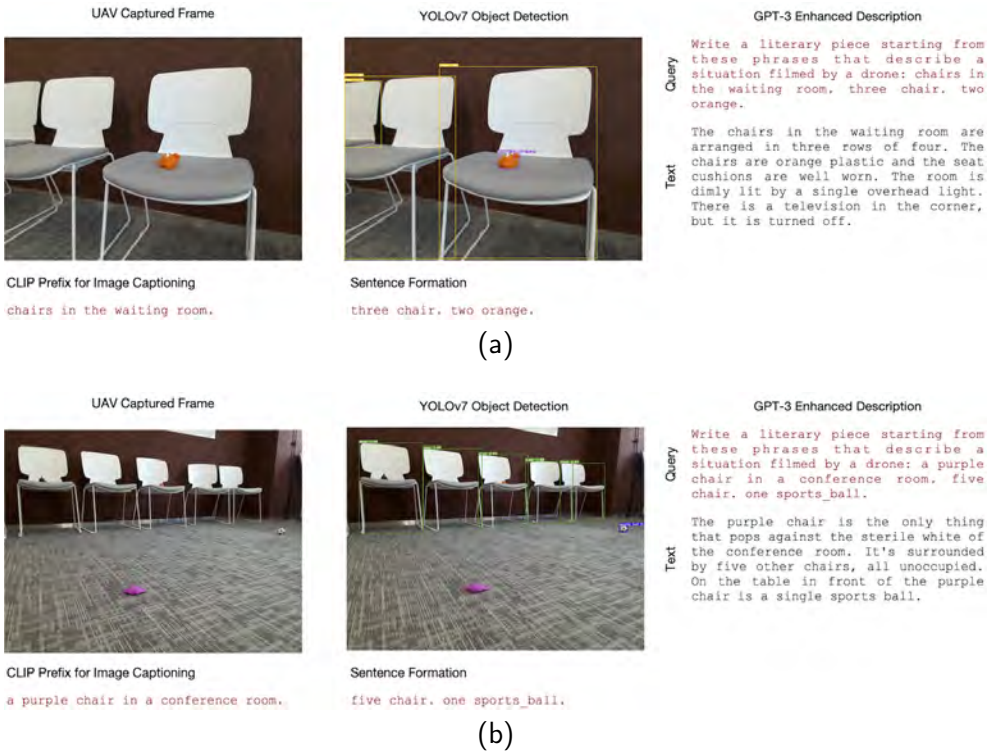


Figure 3.6: UAV captured frame processing and GPT-3. Failure cases.

3.6 Readability Analysis

GPT-3 (short for "Generative Pre-training Transformer 3") is a large language model developed by OpenAI that is trained on a very large dataset of text and is capable of generating human-like text. It is based on a type of neural network called a transformer, which is composed of interconnected processing units that are able to learn and adapt through training. The goal of GPT-3 is to learn the statistical patterns and relationships present in the training data and use this knowledge to generate coherent and plausible text.

One of the key features of GPT-3 is its ability to generate text that is difficult to distinguish from text written by humans. It is trained on a dataset of billions of words and, as a result, is able to capture a wide range of linguistic patterns and structures, including syntax, grammar, and vocabulary. This enables it

to generate text that is highly coherent and grammatically correct, and it can thus be used for a variety of tasks, such as translation, summarization, and text generation.

Readability measures are tools that are used to evaluate the complexity of written text and determine how easy or difficult it is for readers to understand. One common readability measure, for instance, is the GUNNING Fog index, which is a formula that estimates the number of years of education a reader would need to understand a piece of text. The GUNNING Fog index is based on the average number of words per sentence and the percentage of complex words (those with three or more syllables) in the text.

To calculate the GUNNING Fog index, the following steps are followed:

- Count the number of words in a sample of the text;
- Count the number of sentences in the sample;
- Divide the total number of words by the total number of sentences to calculate the average number of words per sentence;
- Count the number of complex words (those with three or more syllables) in the sample;
- Divide the number of complex words by the total number of words, and multiply the result by 100 to calculate the percentage of complex words in the sample;
- Add the average number of words per sentence and the percentage of complex words. The result is the GUNNING Fog index.

The GUNNING Fog index is typically used to evaluate the readability of written materials, such as reports, documents, and articles. It is a useful tool for determining the level of difficulty of a piece of text and ensuring that it is appropriate for a particular audience. For example, a text with a GUNNING Fog index of 8 would be considered suitable for readers with an eighth-grade education or higher.

Such readability measures are useful tools for evaluating the complexity of written text and ensuring that it is appropriate for a particular audience. This can help writers and editors to produce written materials that are clear, concise, and easy to understand and can help readers to more easily comprehend and retain information presented in a text.

A readability analysis of the GPT-3-enhanced text is provided by the use of standardized measures, the one introduced earlier being the most effective. In this chapter, we propose analyzing LLM texts by the following metrics: FLESCH reading ease, DALE CHALL readability, the Automated Readability Index (ARI), the COLEMAN LIAU index, GUNNING Fog, SPACHE, and Linsear Write. The scores obtained by the use of these formulas were designed by linguists to assess the readability of texts to approximate their usability and have been extensively used by, for example, the Office of Education of the United States of America to calibrate the readability of textbooks for the public school system, daily newspapers and monthly magazines to target the appropriate audience, the Department of Defense to help assess the adequacy of technical manuals, and, in general, many US Government Agencies to evaluate the difficulty of a reading passage written in English.

FLESCH reading ease [32] is a simple approach used to assess the grade level of the reader. It is based on the average sentence length and the average number of syllables per word. It is a score in the set $[0, 100]$; the higher the number, the easier the text is to read. According to the scale, $[0, 30]$ means a text is easily understood by a college graduate, $[60, 70]$ means it is easily understood by eighth and ninth graders, and $[90, 100]$ means it is easily understood by a fifth grader.

DALE CHALL readability [33] calculates the grade level of a text sample based on the average sentence length in words and the number of difficult words according to a designated list of common words familiar to most fourth-grade students. Adjusted scores are as follows: <5: Grade 4 and below; $[5, 6)$: Grades 5–6; $[6, 7)$: Grades 7–8; $[7, 8)$: Grades 9–10; $[8, 9)$: Grades 11–12; $[9, 10)$: College; ≥ 10: College Graduate.

The Automated Readability Index (ARI) consists of a weighted sum of two ratio factors: the number of characters per word, and the average number of words per sentence. It assesses the understandability of a text and outputs a value that approximates the grade level needed to grasp the text. For example, the tenth grade corresponds to 15–16 years old, the eleventh grade corresponds to 16–17 years old, the twelfth grade corresponds to 17–18 years old, and greater than twelve corresponds to the level of college.

The COLEMAN-LIAU index [34] is similarly based on the average number of letters per 100 words and the average number of sentences per 100 words. It is like the ARI, but unlike most of the other metrics that predict the grade level, it relies on characters instead of syllables per word.

GUNNING Fog [35] is based on the scaled sum of the average sentence length and the percentage of hard words. It measures the readability of a text passage, and the ideal value is 7 or 8. Texts with a score above 12 are too hard for most people to understand. The measure scores highly with short sentences written in simple language but penalizes long sentences with complicated words.

The SPACHE readability formula [36] is based on the average sentence length and the number of difficult words according to a third grader. It is similar to Dale Chall, but for primary texts until the third grade. To assess the readability of a text, SPACHE is first used, and if the result is higher than third grade, Dale Chall is used.

Linsear Write is a readability formula based on sentence length and the number of words with three or more syllables. Analogous to the previous formulations, it scores a text passage according to the grade level.

Table 3.1 shows the proposed metrics on several example frames. The metrics are computed on unique frames in Row 1–3 and on multi-frame configurations in Row 4–8. We can observe that the storylines generated exhibit a relatively consistent behavior among the statistical indices, where unique frames tend to be ranked at a lower grade level and multi-frame configurations are closer to college level. All SPACHE readability indices are higher than third grade, so Dale Chall has to be considered, where the frames are consistently ranked with a median grade level of $[7, 8]$. Among the measures, GUNNING Fog presents an ideal behavior, as all values are in the range of [7–12], which means that the level of generated texts is comparable to that of established publications in magazines and books, and therefore can be understood by the general public while presenting a rich vocabulary.

3.7 Conclusions

An RIZE Tello drone is a small, lightweight, and low-cost quadrotor drone that is equipped with a camera and is capable of autonomous flight. In this system, the drone is used to capture video footage of a scene and transmit it to a ground computer in real time.

On the ground computer, the video stream is processed using state-of-the-art LLMs together with a module of object detection to produce accurate text descriptions of a scene in the form of captions. These captions can be used to provide a verbal description of the scene for individuals who are deaf or hard

Table 3.1: Readability analysis of a random stream of data captured by RYZE Tello. Score (upper row) and Grade Level (lower row) for each metric.

Frame(s)	FLESCH Reading Ease	Dale Chall	ARI	Coleman Liau	GUNNING Fog	SPACHE	Linsear Write
00	68.36 [8, 9]	6.56 [7, 8]	6.29 [7]	9.10 [9]	9.47 [9]	4.56 [5]	6.1 [6]
01	84.22 [6]	6.06 [7, 8]	3.63 [9, 10]	4.36 [4]	8.18 [8]	3.91 [4]	7.14 [7]
02	84.57 [6]	5.67 [5, 6]	3.80 [9, 10]	5.04 [5]	7.40 [7]	4.18 [4]	6.46 [6]
03-05	71.11 [7]	6.82 [7, 8]	10.27 [16, 17]	7.89 [8]	11.81 [12]	5.82 [6]	13.14 [13]
06-07	82.08 [6]	6.82 [7, 8]	4.36 [5]	7.89 [8]	7.56 [8]	3.64 [4]	6.94 [7]
08-10	74.30 [7]	6.35 [7, 8]	7.05 [13, 14]	7.53 [8]	10.58 [11]	4.87 [5]	9.0 [9]
11-13	75.94 [7]	6.33 [7, 8]	8.54 [9]	7.47 [7]	10.76 [11]	5.33 [5]	11.5 [12]

of hearing, or to provide additional context for individuals who are able to see the video footage.

A pipeline for semantic scene understanding given a stream of UAV data frames was proposed. The methodology does not require fine-tuning; rather, it provides zero-shot text descriptions. The modules consist of state-of-the-art architectures. A captioning module based on CLIP Prefix for Image Captioning is wired through sentence formation to a YOLOv7 object detector, and the generated text is enhanced by prompting GPT-3 natural language instructions. We are the first to provide zero-shot UAV literary storytelling that can stream to a ground computer in real time or after landing (in this latter case, the video would be stored on an SD card, and the RYZE Tello drone needs to be equipped with a board computer, e.g., a Raspberry Pi Zero W or a CORAL board) and that provides state-of-the-art accurate literary text descriptions. Metrics used to assess the readability of LLM texts are proposed, leveraging standardized measures from linguistics.

The system combines the capabilities of an RIZE Tello drone (or an NXP Hover Games Drone) with advanced techniques of computer vision to provide a rich and detailed description of a scene in real time. The system has potential applications in a wide range of fields, including surveillance, search and rescue, and environmental monitoring.

As further work, the trajectory of the drone could be optimized for a certain filming style to help the text description module to obtain better shots for

particularly interesting events that need to be addressed in the storyline. That being said, in the current work, we did not take planning and trajectory issues into consideration and assumed that the UAV is being remotely controlled or is flying using an adequate autopilot policy. In addition, GPS coordinates and positioning information from other sensors such as IMU or LiDAR could be used to further improve the resultant text descriptions by prompting the GPT-3 module with the corresponding trajectories.

There are a number of other ways that the previously described system could be extended or improved upon. Some potential areas of further work include the following.

The accuracy and reliability of the algorithms that handle captioning and object detection can be improved: while current LLMs and object detection algorithms are highly accurate, there is always room for improvement. Further research could focus on developing new techniques or fine-tuning existing algorithms to increase their accuracy and reliability. Other sensors can be added: the RIZE Tello drone is equipped with a camera, but additional sensors, such as LiDAR or RADAR, could allow the system to gather more detailed and comprehensive data about the scene. The drone's autonomy could be enhanced: the RIZE Tello drone is capable of autonomous flight, but further work could focus on developing more advanced autonomy algorithms to enable the drone to navigate more complex environments and perform more sophisticated tasks. Real-time analysis could be implemented: at the moment, the system processes the video stream and generates captions and object detections after the fact. However, implementing real-time analysis could allow the system to provide updates and alerts in near-real time, making it more useful for applications such as surveillance or search and rescue. Finally, applications could be developed for specific domains: the system could be tailored to specific domains by training the captioning and object detection algorithms on domain-specific data and developing domain-specific applications. For example, the system could be used for agricultural monitoring by training the algorithms on data specific to crops and farm machinery.

The ultimate goal is to be able to confer autonomous systems (e.g., UAVs and self-driving cars) with literary capabilities comparable to those provided by human counterparts. Specifically, the use of LLMs and VLMs push the boundaries of system perception and the understandability of events, situations, and contextual information.

Bibliography

[1] Bonatti, R.; Bucker, A.; Scherer, S.; Mukadam, M.; Hodgins, J. Batteries, camera, action! learning a semantic control space for expressive robot cinematography. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021.

[2] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. *Language Models Are Unsupervised Multitask Learners*; Technical Report; 2019. Available online: `https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe` (accessed on 15 December 2022).

[3] Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.

[4] Mokady, R.; Hertz, A.; Bermano, A.H. ClipCap: CLIP prefix for image captioning. *arXiv* **2021**, arXiv:2111.09734.

[5] Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021.

[6] Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.

[7] Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv* **2015**, arXiv:1504.00325.

[8] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.

[9] Sandler, M.; Howard, A.G.; Zhu, M.; Zhmoginov, A.; Chen, L. Inverted residuals and linear bottlenecks: Mobile networks for classification, detection and segmentation. *arXiv* **2018**, arXiv:1801.04381.

[10] Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. *arXiv* **2020**, arXiv:1911.09070.

[11] Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.

[12] Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2022**, arXiv:2104.13921.

[13] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[14] Cui, Y.; Niekum, S.; Gupta, A.; Kumar, V.; Rajeswaran, A. Can foundation models perform zero-shot task specification for robot manipulation? In Proceedings of the 4th Annual Learning for Dynamics and Control Conference, Stanford, CA, USA, 23–24 June 2022.

[15] Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A universal visual representation for robot manipulation. *arXiv* **2022**, arXiv:2203.12601.

[16] Zeng, A.; Florence, P.; Tompson, J.; Welker, S.; Chien, J.; Attarian, M.; Armstrong, T.; Krasin, I.; Duong, D.; Wahid, A.; et al. Transporter networks: Rearranging the visual world for robotic manipulation. *arXiv* **2022**, arXiv:2010.14406.

[17] Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MA, USA, 17–23 July 2022.

[18] Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv* **2022**, arXiv:2204.00598.

[19] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 779–788.

[20] Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

[21] Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.

[22] Tan, M.; Le, Q.V. Efficientnetv2: Smaller models and faster training. *arXiv* **2021**, arXiv:2104.00298.

[23] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6077–6086.

[24] Fang, H.; Gupta, S.; Iandola, F.; Srivastava, R.K.; Deng, L.; Dollár, P.; Gao, J.; He, X.; Mitchell, M.; Platt, J.C.; et al. From captions to visual concepts and back. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015 ; pp. 1473–1482.

[25] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.

[26] Fan, A.; Lewis, M.; Dauphin, Y. Hierarchical neural story generation. *arXiv* **2018**, arXiv:1805.04833.

[27] See, A.; Pappu, A.; Saxena, R.; Yerukola, A.; Manning, C.D. Do massively pretrained language models make better storytellers? *arXiv* **2019**, arXiv:1909.10705.

[28] Li, X.L.; Liang, P. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv* **2021**, arXiv:2101.00190.

[29] Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2556–2565.

[30] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.

[31] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762.

[32] Flesch, R. A new readability yardstick. *J. Appl. Psychol.* **1948**, *32*, 221–233. [CrossRef]

[33] Dale, E.; Chall, J.S. A formula for predicting readability. *Educ. Res. Bull.* **1948**, *27*, 11–28.

[34] Coleman, M.; Liau, T.L. A computer readability formula designed for machine scoring. *J. Appl. Psychol.* **1975**, *60*, 283–284. [CrossRef]

[35] Gunning, R. *The Technique of Clear Writing*; McGraw-Hill: New York, NY, USA, 1952.

[36] Spache, G. A new readability formula for primary-grade reading materials. *Elem. Sch. J.* **1953**, *53*, 410–413. [CrossRef]

Chapter 4

LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments

J. de Curtò, I. de Zarzà, Gemma Roig, Pietro Manzoni and Carlos T. Calafate.(2023). "LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments." Electronics, vol(12), 2814. DOI: 10.3390/electronics12132814

In this chapter, we introduce an innovative approach to handling the Multi-Armed Bandit (MAB) problem in non-stationary environments, harnessing the predictive power of Large Language Models (LLMs). With the realization that traditional bandit strategies, including epsilon-greedy and Upper Confidence Bound (UCB), may struggle in the face of dynamic changes, we propose a strategy informed by LLMs that offers dynamic guidance on exploration versus exploitation, contingent on the current state of the bandits. We bring forward a new non-stationary bandit model with fluctuating reward distributions and illustrate how LLMs can be employed to guide the choice of bandit amid this variability. Experimental outcomes illustrate the potential of our LLM-informed strategy, demonstrating its adaptability to the fluctuating nature of the bandit problem, while maintaining competitive performance against conventional strategies.

4.1 Introduction

In the realm of artificial intelligence (AI) and reinforcement learning (RL), the multi-armed bandit (MAB) problem [1, 2] is a classic decision-making knot that captures the exploration–exploitation trade-off. Traditionally, the MAB problem assumes a stationary setting, where the underlying distribution of each bandit's reward remains constant. However, real-world scenarios often present non-stationary environments, where these reward distributions change over time.

The MAB problem, a classic dilemma of decision theory, exemplifies the balance of exploration and exploitation in RL. It is formulated as a game with a fixed number of slot machines, or 'bandits', each with an unknown probability distribution of rewards. The goal is to develop a strategy for selecting which bandit to play so as to maximize the total reward over a series of plays. While the MAB problem has been extensively studied, the extension to non-stationary environments, where the reward probabilities change over time, poses significant challenges [3, 4]. Traditional strategies often falter in such scenarios, as they are unable to adapt to the evolving reward distributions.

In this study, we delve into the non-stationary multi-armed bandit (NSMAB) problem, where we adapt well-known strategies to handle fluctuating reward distributions. NSMAB poses unique challenges, primarily due to the dynamic nature of the problem, and the need to continuously adapt the decision-making strategy.

While conventional algorithms, such as epsilon-greedy and upper confidence bound (UCB), are adapted to handle non-stationary bandit problems, they often fall short in optimally adjusting to rapid changes in the environment. In the quest for a better approach, we turn our attention to large language models (LLMs), such as GPT-3.5 Turbo from OpenAI. These models have shown remarkable language understanding and problem-solving abilities, and we harness this power to guide our decision-making in the NSMAB setting.

The rise of LLMs [5, 6, 7] has revolutionized many fields of AI, providing solutions that can understand, generate, and learn from human-like text [9, 8]. Leveraging the predictive prowess of LLMs, this work aims to inform and enhance MAB strategies for non-stationary environments. An LLM can provide valuable insights into whether to exploit the current best-performing bandit or explore others that are potentially better suited to the current environment state. By integrating this LLM-informed advice into traditional MAB strategies, we aim to increase the overall effectiveness in non-stationary settings.

The LLM-informed strategy that we propose provides advice on whether to explore or exploit, given the current state of the bandits. This approach effectively leverages the ability of LLM to understand complex scenarios and make informed decisions [10, 11, 12].

In the complex domain of decision making, we often encounter situations that require strategic selection among several alternatives with uncertain outcomes—known as the multi-armed bandit problem. Particularly challenging is the non-stationary variant of this problem, where the probabilities associated with rewards can dynamically change over time. Widely accepted strategies for addressing this problem, such as epsilon-greedy, and UCB (upper confidence bound), seek to balance exploration (seeking out new, potentially superior options), and exploitation (leveraging the currently best-known option). However, performance may vary significantly in non-stationary environments due to the unpredictable nature of the rewards associated with the bandits.

As an alternative approach, we propose a novel strategy that harnesses the predictive capabilities of an LLM, specifically GPT-3.5-turbo-0301 and quantized low-rank adapters (QLoRAs) [13, 14], to guide the decision-making process. Our LLM-informed strategy solicits advice from the LLM, deciding whether to explore or exploit based on the current state of the bandits. Remarkably, we observed that our novel LLM-informed strategy often performs on par with, if not better than, traditional approaches, indicating the potential of integrating advanced AI technologies such as LLMs in real-time decision-making tasks. This contribution advances the current understanding of non-stationary multi-armed bandit problems and opens new avenues for applying LLMs to enhance traditional decision-making strategies in dynamic environments.

The rest of the chapter is organized as follows: In Section 4.2, we delve into the related work, providing a comprehensive overview of both the stationary and non-stationary multi-armed bandit (MAB) problems, the various strategies developed for these settings, and the promising capabilities of LLMs. Section 4.3 introduces the fundamentals of multi-armed bandits, offering a mathematical representation of the problem and discussing its practical applications. Following this, in Section 5.6, we elaborate on our methodology, detailing the adaptation of existing MAB strategies to non-stationary environments, and the innovative incorporation of LLM advice. In Section 4.5, we detail our experimental setup and results, describing the diverse scenarios under exploration, presenting the results along with illustrative figures, and performing a thorough analysis. The subsequent section, Section 7, opens up a broader discussion on the implications of our findings and potential applications of our LLM-informed

framework. Finally, in Section 4.7, we look ahead to future research directions and conclude our study.

4.2 Related Works

The multi-armed bandit (MAB) problem, initially formalized by [1], has been the subject of extensive research due to its inherent need for balancing exploration and exploitation. Several algorithms have been proposed to tackle this problem, each exhibiting specific attributes that render them favorable under different scenarios. For instance, the epsilon-greedy strategy [2] offers simplicity and practicality, guaranteeing eventual convergence to optimal solutions given sufficient time, and a suitable choice of epsilon. The UCB approach [15] is renowned for its optimality in stationary problems, demonstrating logarithmic regret growth over time, effectively minimizing regret in the long run. Finally, Thompson sampling [16] stands out for its probabilistic nature, wherein it favors actions with high uncertainty, or high expected rewards, making it particularly suitable for scenarios with non-stationary rewards [17].

Extensions of the MAB problem to non-stationary environments [18], where the reward probabilities change over time, are less well studied, and yet increasingly relevant in dynamic real-world scenarios [3, 4, 19]. Strategies that adapt to changing reward distributions have been proposed [20, 21], but they often require assumptions about the rate of change or the total number of changes.

The advent of LLMs [5, 22, 6, 13, 7], such as GPT-3 [9], Flan [23] or QLoRA [14], has opened new avenues for AI applications. Their ability to generate human-like text and predict next word probabilities has been exploited in tasks ranging from text completion to more complex decision-making problems [24, 25]. In this work, we explore the potential of LLMs to advise and enhance traditional MAB strategies in non-stationary environments.

4.3 Multi-Armed Bandit

The problem of multi-armed bandit is a classic dilemma from probability theory that describes an agent trying to maximize rewards when faced with multiple options, each with an unknown and potentially different reward distribution. This problem is characterized by the inherent trade-off between exploration (trying out all options to learn more about their rewards) and exploitation (sticking with the option that currently seems the best).

Consider an agent faced with K slot machines, or "one-armed bandits". Each pull of a machine's lever, or "arm", gives a random reward drawn from a stationary and unknown probability distribution specific to that machine. The agent's objective is to develop a strategy to decide which arm to pull at each time step in order to maximize the total reward over a sequence of T time steps.

Let $X_{o,t}$ be the reward from the o-th arm at time t, and let $x_{o,t}$ be the observed reward. We assume $X_{o,t}$ are independent and identically distributed random variables for each o, but the distributions can differ between arms.

The value of an action a is the expected reward:

$$q(a) = \mathbb{E}[X_{a,t}]; \quad \forall t. \tag{4.1}$$

However, the agent does not have access or is agnostic to $q(a)$. Instead, it must estimate the values based on the observed rewards. A natural estimate is the sample average:

$$\hat{q}_t(a) = \frac{1}{N_t(a)} \sum_{\tau=1}^{t} I(A_\tau = a) x_{\tau,a}, \tag{4.2}$$

where $N_t(a)$ is the number of times action a has been selected up to time t, A_τ is the action selected at time τ, and $I(A_\tau = a)$ is an indicator function that is 1 if $A_\tau = a$, and 0 otherwise.

The challenge in the multi-armed bandit problem involves devising a strategy for selecting A_t based on $\hat{q}_{t-1}(1), \ldots, \hat{q}_{t-1}(K)$ that successfully balances exploration and exploitation. A good strategy should allow for pulling all arms sufficiently to obtain an accurate estimate of all $q(a)$, but also aim to minimize the number of pulls on arms that have consistently provided lower rewards. By "inferior arms" we refer to those bandits that, based on past interactions, appear to offer less reward (on average) than other options. The goal is to avoid excessively engaging with these seemingly less lucrative options while ensuring that all bandits have been sampled enough to make an informed judgment about their reward distributions. This, in essence, captures the core challenge of the multi-armed bandit problem.

In the following sections, we will delve into some well-established strategies for this problem, setting the groundwork for our innovative approach. Our unique contribution lies in the development of a new strategy that leverages LLMs in a way that has not been done before. This breakthrough approach aims to

significantly improve upon the current methodologies, providing more effective and efficient solutions, particularly for non-stationary environments.

4.4 Methodology

In this section, we first provide a comprehensive look at the non-stationary multi-armed bandit problem, offering a detailed examination of the inherent complexities and unpredictable elements found in such environments. Then, we delve into the strategies we utilized to tackle the problem. This includes well-known approaches, such as epsilon-greedy and UCB, as well as a novel method that leverages the capabilities of LLMs. We then illustrate the reasons for focusing on these specific strategies, along with a discussion on our innovative LLM-informed strategy.

4.4.1 Non-Stationary Multi-Armed Bandit

In the non-stationary multi-armed bandit problem, there are K bandits or slot machines, each with an unknown reward distribution that may change over time. At each time step t, the agent chooses to play a bandit o and it receives a reward $X_{o,t}$, which is sampled from the bandit's current reward distribution.

The objective of the agent is to maximize the sum of rewards over a sequence of T trials, which is a challenging task due to the exploration–exploitation dilemma, and the non-stationary nature of the bandits' reward distributions.

Moreover, acknowledging that real-world scenarios often involves non-stationary processes, where the reward distributions evolve over time, we extend our methodology to accommodate non-stationary bandits. This extension is facilitated by dynamically modifying the reward functions at specific time intervals, which can involve varying the mean or variance. Given the temporal nature of these reward distributions, it is plausible that the optimal action may not remain constant over time. As such, it is crucial for the agent to maintain an exploratory approach over time, as the most rewarding action may change as the experiment proceeds. This is especially true in our experiment settings, where reward distributions of the bandits may undergo a significant shift halfway through the experiment, thereby also changing the optimal bandit at that point. Thus, our methodology embraces these non-stationary aspects to ensure a more holistic and realistic evaluation of the different strategies.

The key point of the derivation that follows is that using the LLM to inform the strategy in a non-stationary multi-armed bandit problem is analogous to utilizing a sophisticated, data-driven decision rule in a coevolutionary game. It demonstrates how tools from AI can be effectively leveraged to adapt traditional game theoretic models to complex, dynamic settings.

4.4.2 Strategies

In this study, we consider three distinct strategies for tackling the multi-armed bandit problem: the epsilon-greedy strategy, the UCB strategy, and a novel approach we propose and name as the LLM-informed strategy. These three strategies were selected due to their different methods for addressing the exploration–exploitation dilemma, a key challenge in the multi-armed bandit problem. The epsilon-greedy and the UCB strategies are well-known approaches in this field, providing useful benchmarks for comparison, while the LLM-informed strategy introduces an innovative use of AI, specifically LLMs, to this problem space.

Thompson uses a Bayesian approach, updating a probability distribution over each arm's reward distribution, and then choosing an arm to play based on sampling from these distributions. This strategy provides a natural and probabilistic trade-off between exploration and exploitation.

In this study, we chose to focus on epsilon-greedy, UCB, and the novel LLM-informed strategy when dealing with non-stationary environments for the following reasons:

- Comparative simplicity: Both epsilon-greedy and UCB strategies are simpler in their implementation compared to Thompson sampling. These strategies provide clear baselines for comparison, allowing us to measure the impact of the LLM-informed strategy against well-understood and straightforward mechanisms [2].
- Demonstrated effectiveness: While Thompson sampling has its advantages, epsilon-greedy [26] and UCB strategies [15] have been extensively studied and proven effective in a wide variety of scenarios. They provide solid and reliable benchmarks, against which the novel LLM-informed strategy can be compared.
- Novelty of LLM-informed strategy: The main goal of our study was to explore and demonstrate the potential of leveraging LLMs [9] in the multi-armed bandit problem. By focusing on comparing this novel strategy

with simpler, well-known strategies, we aimed to isolate and highlight the impact of LLM advice on problem solving.

- Computation resources: Thompson sampling [27] often requires more computational resources than epsilon-greedy and UCB strategies due to the need to sample from probability distributions during each decision-making step. As our study included large-scale experiments, we decided to exclude Thompson sampling to minimize computational resource consumption.

Another point in favor of omitting Thompson sampling is that applying it to non-binary rewards can be more complex. If the reward distributions are not Bernoulli, then we need to choose and update appropriate prior distributions for the rewards. Depending on the actual reward distributions and the chosen priors, this could involve complex calculations or approximations, which may not be feasible or efficient for large-scale experiments or real-time applications.

Strategy Epsilon-Greedy

The strategy epsilon-greedy is a simple yet effective approach to address the exploration–exploitation dilemma. The strategy can be described as follows:

$$\pi_\epsilon(a|s) = \begin{cases} 1 - \epsilon + \epsilon/K & \text{if } a = \arg\max_{a'} Q(s, a'), \\ \epsilon/K & \text{otherwise.} \end{cases}$$

where $Q(s, a)$ is the estimated reward of action a at state s, K is the number of bandits, and ϵ is a parameter that controls the trade-off between exploration and exploitation.

UCB Strategy

The UCB strategy offers a more sophisticated way to balance exploration and exploitation by taking into account both the estimated reward and the uncertainty of each bandit. The UCB strategy selects the bandit with the highest upper confidence bound on the expected reward:

$$a_t = \arg\max_a \left[Q(s, a) + c\sqrt{\frac{\ln t}{N(s, a)}} \right],$$

where $N(s, a)$ is the number of times that action a has been selected at state s, t is the current time step, and c is a constant that controls the degree of exploration.

LLM-Informed Strategy

We introduce a unique and novel strategy, the LLM-informed strategy, specifically designed to harness the predictive capabilities of LLMs for tackling the multi-armed bandit problem. The core innovation of our approach is to recast the bandit problem as a question, which is then presented to the LLM. Based on the LLM advice regarding exploration or exploitation, we determine the subsequent action. This represents a significant contribution, as it unveils a new path to employ advanced AI technologies in the decision-making process of complex stochastic problems, such as multi-armed bandits.

There are multiple advantages that underpin our approach. Firstly, the strengths of LLMs lie in their ability to understand context, learn from past data, and generate predictions based on complex patterns. This allows the LLM-informed strategy to incorporate more nuanced decision making that is responsive to the trends and changes in the non-stationary environment. Rather than relying on rigid mathematical formulae, the LLM-informed strategy is capable of adapting its decision-making process based on the evolving patterns in the rewards and their distributions, leading to more robust performance in non-stationary scenarios. Secondly, the LLM can process and consider a much larger history of past rewards and decisions than traditional algorithms, potentially leading to more informed decisions. Lastly, the use of LLMs offers an intriguing avenue of investigation into how advanced AI models can be integrated with classic problems in RL, expanding our understanding of how these models can be harnessed in new and innovative ways.

The LLM response is parsed and used to determine the next action. Specifically, if the LLM suggests to "explore", we select a bandit uniformly at random; if the LLM suggests to "exploit", we select the bandit with the highest estimated reward.

In the context of a coevolutionary game [28, 29, 30], the "explore" and "exploit" strategies can be seen as analogous to the decision for an agent (or node) to cooperate or defect. Let us denote the strategy space for the agent as $S = \{\{\text{“explore”}\}, \{\text{“exploit”}\}\}$.

Given this, we can introduce a simplified fitness landscape, denoted as $F : S \times S \to \mathbb{R}$, which encodes the rewards for each combination of strategies. This concept is analogous to the payoff matrix in a standard game theoretic setup. Under our model, which is also applied in our experimental setup, the reward for exploration is considered a random variable $R_{\{\text{explore}\}}$, following a certain distribution that may change over time, signifying non-stationarity. On the other hand, the reward for exploitation is the current estimated mean reward $R_{\{\text{exploit}\}}$ of the best arm. This framework allows us to effectively study and evaluate the performance of the LLM-informed strategy, but it is important to note that real-world scenarios can be more complex:

$$F(s_1, s_2) = \begin{cases} R_{\{\text{explore}\}} & \text{if } s_1 = \{\text{“explore”}\} \text{ or } s_2 = \{\text{“explore”}\}, \\ R_{\{\text{exploit}\}} & \text{if } s_1 = \{\text{“exploit”}\} \text{ and } s_2 = \{\text{“exploit”}\}. \end{cases}$$

In a coevolutionary game, agents update their strategies based on their fitness and the fitness of their neighbors. In the multi-armed bandit context, the strategy recommendation of the LLM can be seen as the agent “observing” the fitness of its neighbors (i.e., the performance of different strategies in the past), and deciding to update its strategy accordingly.

The decision rule for the agent (or the bandit strategy algorithm) can be modeled as a function $D : S \times S \to S$, which takes as input the current state of the game and the LLM recommendation, and outputs the next action:

$$D(s_{\{\text{actual}\}}, s_{\{\text{LLM}\}}) = \begin{cases} \{\text{“exploit”}\} & \text{if } F(s_{\{\text{actual}\}}, s_{\{\text{LLM}\}}) = R_{\{\text{exploit}\}}, \\ \{\text{“explore”}\} & \text{otherwise.} \end{cases}$$

Note that this is a simplistic model and in reality, the decision rule could take into account other factors, such as the degree of uncertainty in the estimated rewards. Moreover, the fitness landscape could be more complex, depending on the specifics of the non-stationary environment. For instance, the reward distribution for exploration might not be the same for all arms, or it might be correlated with the past rewards of the arms.

In sum, leveraging the LLM as a strategy informant for the non-stationary multi-armed bandit problem can be compared to the application of an advanced, data-driven decision protocol in a coevolutionary game. This clearly exemplifies how AI resources can be powerfully harnessed to adapt traditional game-theoretic frameworks to intricate, dynamic environments.

Quantized Low-Rank Adapters

Building upon the foundation laid by low-rank adapters (LoRA) [13], Quantized low-rank adapters (QLoRA) introduces a strategy that efficiently fine tunes large-scale language models while minimizing memory requirements [31, 14]. Much like its predecessor, QLoRA utilizes the concept of adapters, a small set of trainable parameters, while keeping the bulk of the model parameters constant. The process of optimizing the loss function is achieved by passing gradients via the fixed pre-trained model weights to the adapter. However, QLoRA takes a step further by incorporating quantization techniques, which enables a reduction in the numerical precision of the model weights, thus drastically decreasing the memory footprint and computational requirements.

For a given projection $\mathbf{XW} = \mathbf{Y}$ with $\mathbf{X} \in \mathbb{R}^{b \times h}$, $\mathbf{W} \in \mathbb{R}^{h \times o}$, QLoRA follows a similar computation pattern as LoRA:

$$\mathbf{Y} = \mathbf{XW} + s\mathbf{XL}_1\mathbf{L}_2, \tag{4.3}$$

where $\mathbf{L}_1 \in \mathbb{R}^{h \times r}$, $\mathbf{L}_2 \in \mathbb{R}^{r \times o}$, and s is a scalar. The key differentiating factor lies in the handling of these computations; they are executed at significantly lower precision, in line with the QLoRA principle of quantization. This makes QLoRA a highly effective solution for fine tuning larger models on hardware, such as the A100 GPU, without compromising performance levels.

One of the standout innovations in QLoRA is the introduction of the NormalFloat (NF) data type, which is a fundamental component of its 4-bit quantization mechanism. This data type builds upon the concept of quantile quantization [31], an approach that is designed to be information-theoretically optimal. The distinguishing feature of quantile quantization is that it assigns an equal number of values from the input tensor to each quantization bin, effectively working through the estimation of the input tensor's quantile using the empirical cumulative distribution function.

However, quantile estimation is computationally intensive, which represents a significant limitation for quantile quantization. To mitigate this, QLoRA incorporates fast quantile approximation algorithms, such as SRAM quantiles [31], for the estimation process. It is important to note, though, that the inherent approximation errors in these algorithms can result in substantial quantization errors for outlier values, which are often critically important.

This is where the NF4 data type comes in. By leveraging the fact that pretrained neural network weights typically follow a zero-centered normal distri-

bution, the NF4 data type allows for the transformation of all weights to one fixed distribution by scaling the standard deviation σ to fit precisely within the data type's range. This means that both the data type and the neural network weights' quantiles need to be normalized into this range.

Through this normalization process, the NF4 data type facilitates the optimal quantization for zero-mean normal distributions with arbitrary standard deviations σ within a predefined range. This approach effectively sidesteps the issue of expensive quantile estimates and approximation errors, making it a crucial contributor to the efficiency of QLoRA.

Double quantization (DQ) introduces an additional layer of quantization to the quantization constants, achieving further memory optimization. The process uses 8-bit floats for the second layer of quantization. DQ significantly reduces the memory requirements from an average of 0.5 bits per parameter to just 0.127 bits per parameter. It manages to do this while preserving model performance, which demonstrates the power of the quantization approach taken by QLoRA. In order to tackle the problem of out-of-memory errors during GPU processing, QLoRA utilizes page optimizers. These optimizers rely on NVIDIA's unified memory feature, which transfers data between the CPU and GPU on a page-by-page basis, similar to traditional CPU RAM-disk memory paging. By allocating paged memory for the optimizer states, the system can automatically relocate memory from the GPU to CPU RAM when the GPU is out of memory and vice versa when the memory is needed for optimizer updates.

QLoRA integrates these key procedures to process a linear layer in the quantized base model complemented with a LoRA adapter. This methodology primarily hinges on the process of 'double dequantization'. This operation transforms weights, which have undergone two stages of quantization, back into their original computational format, while preserving the memory-saving advantages of quantization.

Defined as doubleDequant, this function dequantizes the input weights that are quantized quantization constants, and subsequently performs a second dequantization on the resulting weights:

$$\text{doubleDequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}, \mathbf{W}^{\text{k-bit}}) = \text{dequant}(\text{dequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}), \mathbf{W}^{\text{4bit}}) = \mathbf{W}^{\text{BF16}}, \tag{4.4}$$

This function allows weights, originally stored in the 4-bit NormalFloat (NF4) format, to be converted back into the 16-bit BrainFloat (BF16) format for computation.

A crucial component of this approach is that, for parameter updates, only the gradients concerning the LoRA adapter weights are necessary, rather than those for the 4-bit weights. This is achieved by calculating the derivative of $\mathbf{X}$ with respect to $\mathbf{W}$ in BF16 precision after dequantization from the storage format. The forward pass of the model can then be expressed as follows, which is analogous to the general formulation introduced in Equation (4.3):

$$\mathbf{Y}^{\text{BF16}} = \mathbf{X}^{\text{BF16}}\text{doubleDequant}(c_1^{\text{FP32}}, c_2^{\text{k-bit}}, \mathbf{W}^{\text{NF4}}) + \mathbf{X}^{\text{BF16}}\mathbf{L}_1^{\text{BF16}}\mathbf{L}_2^{\text{BF16}}. \quad (4.5)$$

To summarize, the QLoRA approach employs two distinct data types: a storage data type (usually 4-bit NormalFloat), and a computation data type (16-bit BrainFloat). This arrangement optimizes memory efficiency while maintaining computational accuracy. The methodology achieves a balance between resource utilization and performance by conducting computations in the higher precision format, while saving memory in the lower precision format during storage.

4.5 Experiments and Results

This section studies the empirical analysis of various multi-armed bandit strategies and introduces a new approach informed by LLMs. We investigate the epsilon-greedy strategy as a base case and further compare it with other traditional strategies, such as UCB and Thompson sampling. As the environment becomes more complex, such as in non-stationary and parametrized bandit distributions, these traditional strategies are put to the test. The results help identify the strengths and weaknesses of each strategy and how quickly they converge to the best action under different circumstances. Moreover, we take a significant leap by introducing the LLM-informed strategy. It harnesses the potential of LLMs, such as GPT-3.5-turbo, Flan-t5-xl or QLoRA, to aid the decision-making process in multi-armed bandit problems. This novel approach seeks to exploit the superior predictive abilities of LLMs, providing insightful recommendations on the best bandit selection strategy based on the current state of the game.

4.5.1 Epsilon-Greedy

We begin the experimentation with the epsilon-greedy strategy, one of the most common ways of balancing the exploration–exploitation trade-off. In this context, a multi-armed bandit is a problem in which you have to choose the most profitable action from a set of choices, based on a series of trials. The "bandit" part of the name comes from a metaphor with slot machines, which are also known as one-armed bandits.

In the simulation in Figure 4.1, we have three bandits, each with a different "true mean" of the reward. The epsilon value determines the proportion of the time that the simulation will explore (choose a random bandit) instead of exploiting (choosing the bandit that currently has the highest estimated mean). After running the simulation with different values of epsilon, then we produce a plot that shows the cumulative average of the rewards over time, on a logarithmic scale. This plot shows how quickly the distinct values of epsilon allow the simulation to converge on the best bandit.

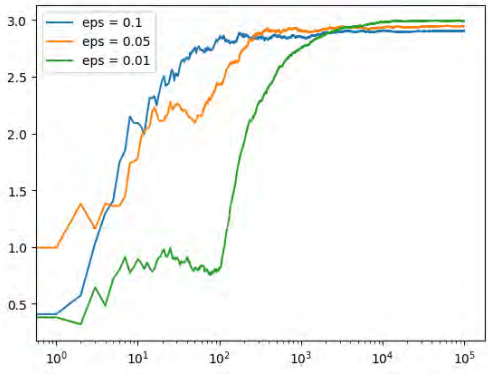


Figure 4.1: Cumulative average of the rewards over time, on a logarithmic scale, for the epsilon-greedy strategy.

4.5.2 Alternative Strategies: UCB and Thompson Sampling

Next we expand the bandit class to include the UCB and Thompson sampling strategies. Note that these strategies require a little more information than epsilon-greedy. For UCB, we need to keep track of the total number of actions taken to compute the confidence bounds. For Thompson sampling, we need to keep track of both the number of successes and failures (modeled here as rewards of 1 and 0) to shape the beta distribution from which we sample.

We will now generate plots for the average rewards over time using the strategies epsilon-greedy, UCB, and Thompson sampling. First, we assume the bandits have binary rewards (either 0 or 1) for simplicity and to align ourselves with the typical use cases of UCB and Thompson sampling. Then, we run experiments with each strategy and plot the cumulative average rewards over time in Figure 4.2, where we show how the average reward evolves over time for each strategy. With this, we can compare how quickly each strategy converges to the optimal bandit, and how they perform relative to each other over the course of many trials.

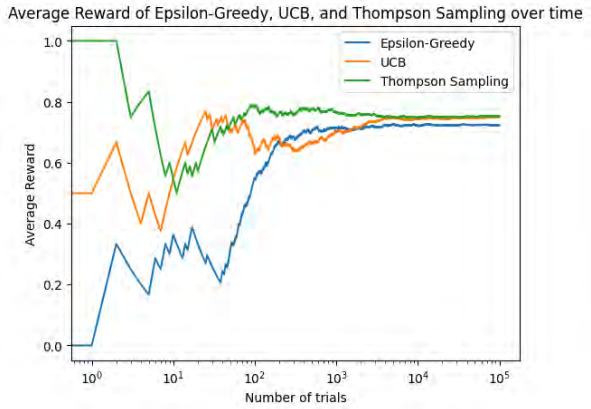


Figure 4.2: Cumulative average of the rewards over time for strategies epsilon-greedy, UCB, and Thompson sampling.

4.5.3 Parametrized Distributions of Bandits

For the next experiments, we focus on parametrized bandit distributions. At first, the reward for each bandit is modeled as a Gaussian distribution with a certain mean. It would be more flexible to allow for arbitrary reward distributions, parametrized by more than just the mean. In Figure 4.3, the rewards for each bandit are generated by drawing from a normal distribution with a distinct mean. We then create three functions that generate rewards according to different normal distributions, and run the experiment using the epsilon-greedy and the UCB strategies. The results show the average reward over time for each strategy, which helps us understand the performance of the distinct strategies. Thompson sampling is typically used for binary rewards and is not included in this plot because our reward functions generate normally distributed rewards.

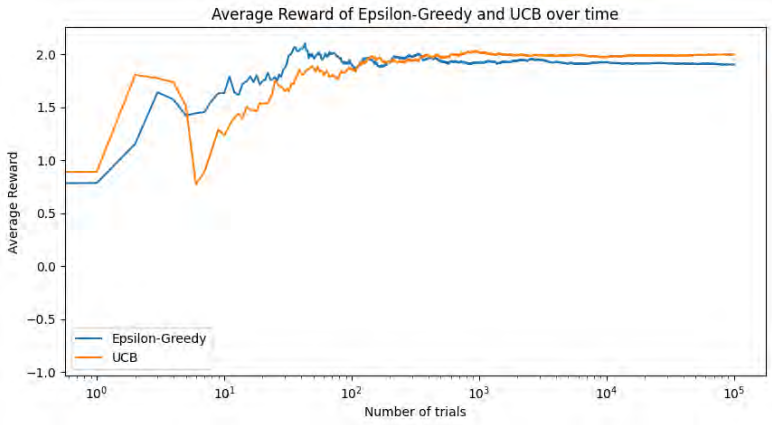


Figure 4.3: Average reward over time for epsilon-greedy and UCB strategies.

4.5.4 Non-Stationary Bandits

From now on, we will focus on non-stationary bandits. Real-world scenarios often involve non-stationary processes, where the reward distributions change over time; an extension can handle such non-stationary bandits. We therefore incorporate non-stationary bandits by modifying the reward functions over time. For instance, we can adjust the mean or variance at certain time steps. However, in a non-stationary environment, it is generally beneficial for the agent to continue exploring, as the optimal action may change over time. Therefore, using strategies that balance exploration and exploitation, such as epsilon-greedy or UCB, becomes more effective in these cases. That is, the reward distributions of the bandits change halfway through the experiment. This means that the optimal bandit may also change at this point.

Figure 4.4 plots the average reward over time for the epsilon-greedy and UCB strategies when facing non-stationary bandits. The vertical dashed line represents the change point where the reward distributions of the bandits shift.

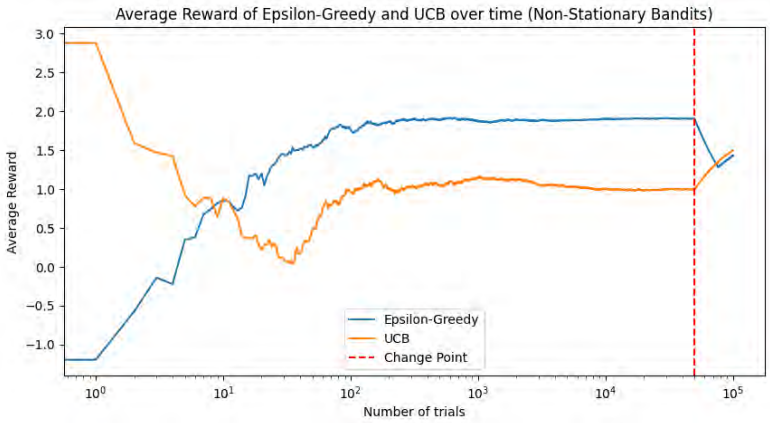


Figure 4.4: Average reward over time for the epsilon-greedy and UCB strategies with non-stationary bandits.

Graphical Display

To visualize the estimated value of each bandit over time, we plot the estimated values in Figures 4.5 and 4.6; they illustrate how the estimated value of each bandit evolves over the course of the experiment.

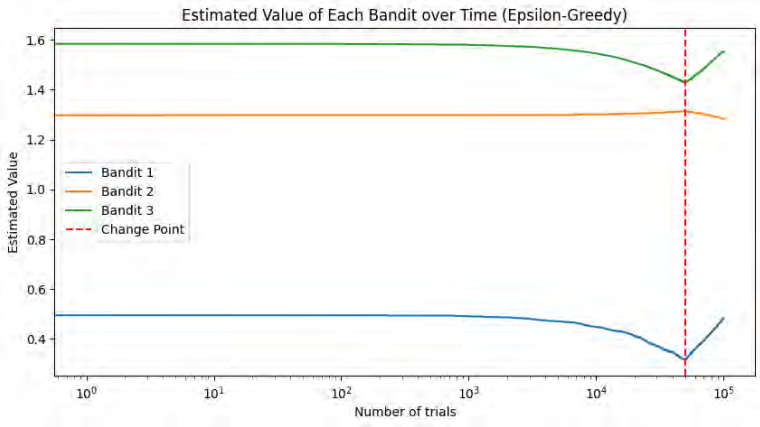


Figure 4.5: Estimated value of each bandit over time for the epsilon-greedy strategy.

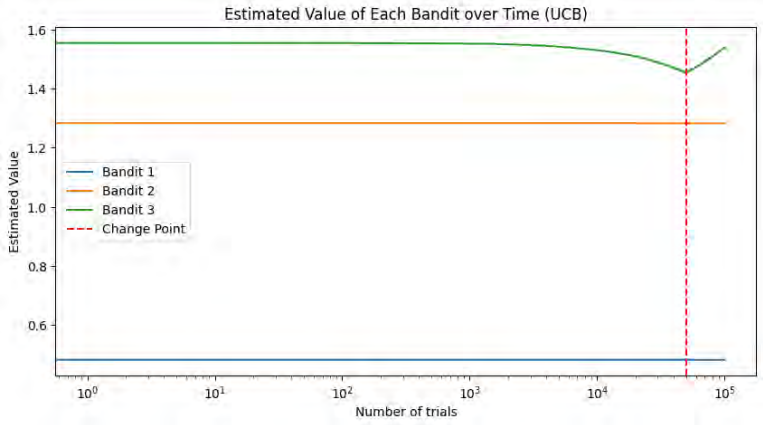


Figure 4.6: Estimated value of each bandit over time for the UCB strategies.

Performance Metrics beyond Average Rewards

Furthermore, we can use other performance metrics. For instance, in addition to plotting the average rewards, we could also compute and display other performance metrics, such as regret, which measures the difference between the rewards we received and the rewards we could have received if we always chose the optimal action, so a smaller regret indicates a better strategy. In this sense, we need to know the optimal bandit at any given time point. In a stationary setting, it is the bandit with the highest expected reward. However, in a non-stationary setting, it could change over time. In the setup, the optimal bandit may change when the reward functions change.

Figure 4.7 shows the regret for the epsilon-greedy and UCB strategies and plots it over time. Please note that, in a non-stationary environment, it could be tricky to define an optimal bandit, especially if the reward distribution changes unpredictably or frequently. Here, we assumed that the change point is known, and we re-evaluated the optimal bandit at the change point, but in a real-world scenario, we might not know when or how the reward distributions change.

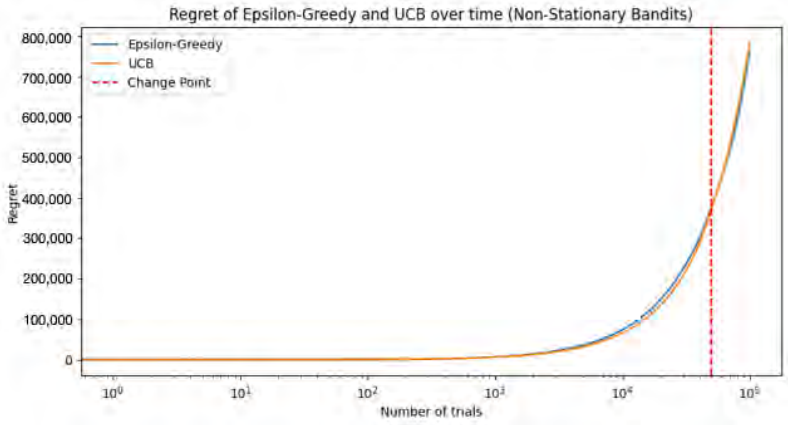


Figure 4.7: Regret for the epsilon-greedy and UCB strategies.

Convergence Analysis

In our convergence analysis, we incorporated functionality to examine the algorithm's convergence across various scenarios. This includes tracking the number of trials required for the algorithm to accurately identify the best bandit as illustrated in Figures 4.8 and 4.9. This process entails recording the selected bandit at each step and checking when it aligns with the bandit possessing the highest mean reward. It is important to remember, however, that the concept of convergence in a multi-armed bandit problem is somewhat more complex, especially when applying an epsilon-greedy approach. As there is always a probability epsilon of selecting a random action, we do not strictly converge to always selecting the optimal action. Rather, it may be more insightful to monitor the evolution of the proportion of instances in which we opt for the optimal action over time.

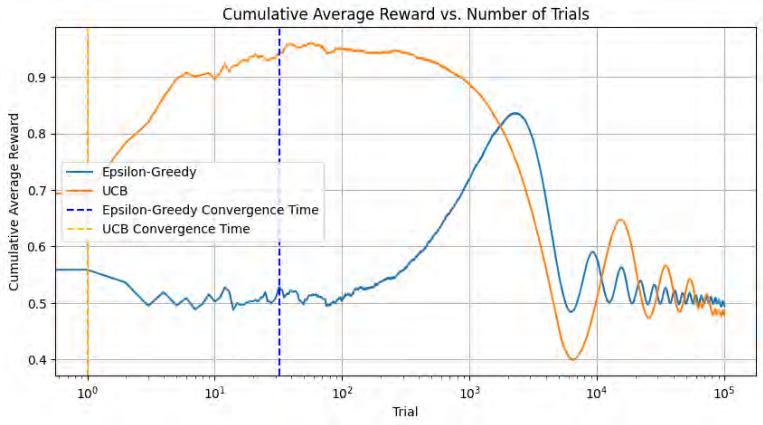


Figure 4.8: Cumulative average reward versus number of trials for epsilon-greedy and UCB strategies.

In Figure 4.8, the dashed lines represent the time at which the respective algorithms first identified the optimal bandit. This is a simplistic measure of convergence and might not fully reflect the learning process, especially in non-stationary settings. Nonetheless, it gives us a sense of when each algorithm begins to catch on to the best choice. For this, we redefine the reward functions to take the time step as an argument and to return values that vary over time. We make a simple change such that the mean of each bandit's reward changes slowly over time. In these new reward functions, the mean reward of each bandit slowly oscillates over time.

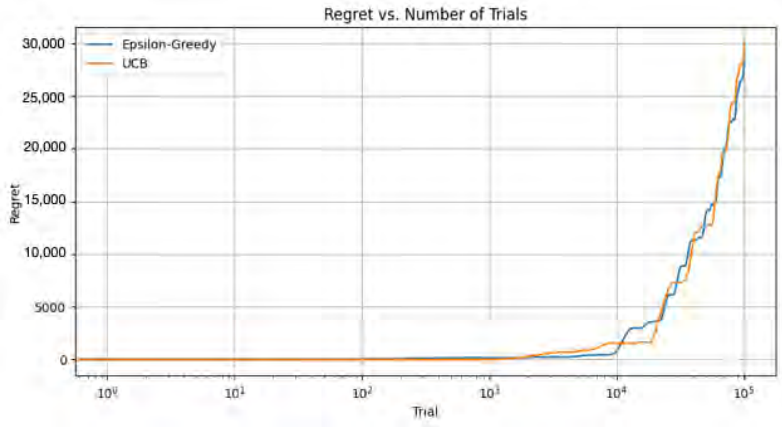


Figure 4.9: Regret versus number of trials for epsilon-greedy and UCB strategies.

In a non-stationary setting, the optimal action can change over time. The above convergence time still refers to the time it first reaches the optimal action, not how well it adapts to changing circumstances.

4.5.5 LLM-Informed Strategy

The utilization of an LLM, such as GPT-3.5-turbo, or Flan-t5-xl [23], can facilitate insightful recommendations for the multi-armed bandit problem. The proposed approach relies on the model's capacity to suggest an optimal strategy (e.g., "epsilon-greedy" or "upper confidence bound") given the present state of the game, including prior results. This process can be formalized as follows, and the flow diagram is shown in Figure 4.10:

1. Game state definition: The game state could encapsulate an array of information, including the total rewards accrued from each bandit, the frequency with which each bandit is selected, and the average reward obtained from each bandit. These data must be translated into a format that can be readily comprehended by the LLM.

2. Strategy recommendation request: This game state information can be utilized to request a strategy recommendation from the LLM. It is crucial to structure the prompt in a manner that clearly articulates the game state and seeks a specific output (e.g., the designation of a strategy).

3. Output interpretation: The LLM output must then be translated back into a form that can be interpreted by the bandit selection algorithm. This could be as straightforward as mapping strategy names to corresponding functions within the code.

4. Recommended strategy implementation: The final step entails utilizing the strategy recommended by the model to decide the next bandit to be selected.

In our research, we specifically focus on rate-limiting requests to the OpenAI API, as well as employing regular expressions to distill strategy recommendations from the LLM output. In this context, we pose a query to the model regarding whether to "exploit" (i.e., select the bandit with the highest estimated mean reward), or "explore" (i.e., select a bandit randomly) in the forthcoming round, given the current state of the bandits.

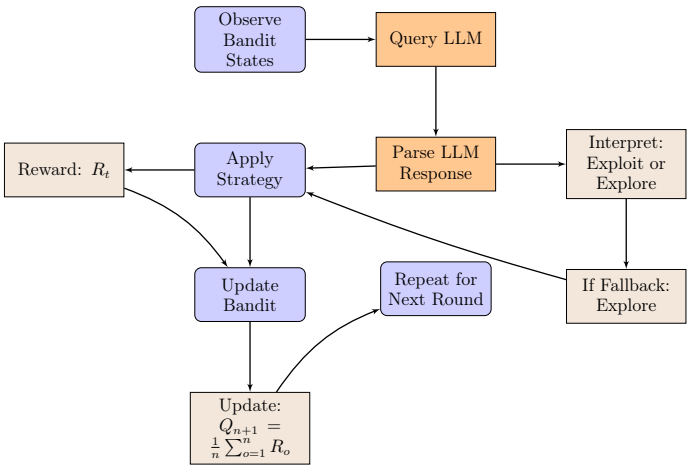


Figure 4.10: Flow diagram of the LLM-informed strategy for the problem of multi-armed bandit.

That is, the usual flow described in Figure 4.10 in a MAB problem encompasses the following:

1. Observe the state of the bandits.

2. Decide on a strategy, either to explore (choose a bandit randomly) or exploit (choose the bandit currently known to give the highest reward).

3. Apply the chosen strategy, meaning pull the arm of a bandit based on the decision in step 2.
4. Receive a reward from the bandit that was chosen.
5. Update the knowledge about the bandit that was chosen, based on the reward received.

In our implementation, we employ several strategic measures to optimize the interaction with the LLM and the execution of the recommendation process.

- Firstly, we incorporate a caching mechanism to store the previous LLM recommendations. By doing so, we eliminate the need for redundant API calls when the state of the game has not changed significantly, thereby conserving resources and increasing efficiency. The state of the game is represented as a string summarizing the pull count and estimated average reward for each bandit, which is then used as the key in the recommendation cache.
- Secondly, our implementation is designed to handle potential exceptions that may occur during interaction with the OpenAI API. Specifically, we implement an exponential backoff strategy, which essentially means that if an API call fails, the system waits for a certain amount of time before retrying, with the wait time increasing exponentially after each consecutive failure. This mechanism provides robustness against temporary network issues or API rate-limiting, enhancing the overall reliability of the system.
- Lastly, we introduce a threshold (ö) for determining significant changes in the bandit state. This is particularly important, as it governs when a new strategy recommendation is required from the LLM. If the change in the bandit state falls below this threshold, the system reuses the previous recommendation, once again avoiding unnecessary API calls. This threshold is a flexible parameter that can be fine tuned to balance the trade-off between responsiveness to changes and minimizing API requests.

In the following analysis, we explore the performance of three strategies in tackling the MAB problem: epsilon-greedy, UCB, and the proposed LLM-informed strategy. We conducted a series of trials, running each strategy through the same sequence of bandits, and then recording their cumulative average rewards over time.

Figure 4.11 represents the evolution of the cumulative average reward for each of these strategies over the course of the trials. Each point on a line represents the average reward that a particular strategy had received up to that iteration, giving us an insight into how quickly and effectively each strategy accrues rewards.

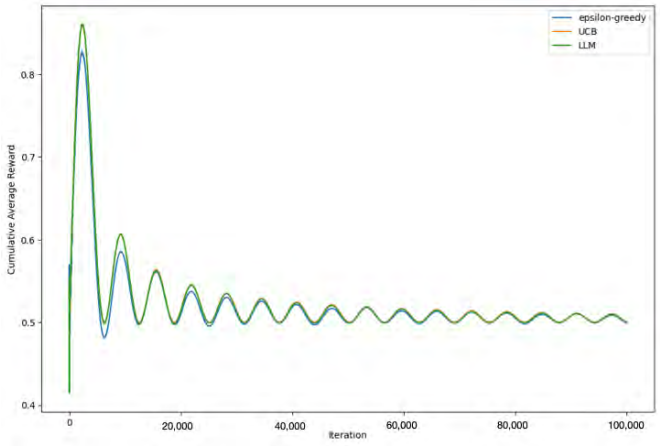


Figure 4.11: Cumulative average rewards over time for epsilon-greedy, UCB, and LLM-informed strategies with ö = 0.1.

Observing the trends in the graph, we can analyze the behavior and effectiveness of the different strategies. The epsilon-greedy and UCB strategies follow conventional approaches with known strengths and weaknesses. The epsilon-greedy strategy provides a balance between exploration and exploitation, while UCB optimizes its choices based on uncertainty and potential for reward. On the other hand, the LLM-informed strategy leverages the predictive power of large language models, in this case, GPT-3.5-turbo-0301. The model offers strategy recommendations based on the current state of the game, which includes the number of times each bandit has been pulled and their average rewards.

4.5.6 Utilizing QLoRA with A100 GPU

The methodology can be implemented in a real system by replacing the calls to OpenAI API model GPT-3.5-turbo with a very recently released LLM model: QLoRA [14], an efficient fine-tuning approach designed for large-scale models. QLoRA facilitates the fine tuning of models as large as 65 billion parameters and inference on a GPU, such as A100, while preserving the performance level of 16-bit fine tuning. Its low memory usage and efficient performance were achieved through a number of innovative strategies [32], such as 4-bit NormalFloat (NF4), double quantization (DQ), and paged optimizers.

We follow a similar methodology as the one adopted with GPT-3.5-turbo-0301 but this time implementing the recommendations through QLoRA.

- In the first step, we defined the state of the game, converting the relevant data into a format comprehensible to QLoRA.
- Then, we made a strategy recommendation request, using the game state information to prompt QLoRA for a strategy.
- After receiving the QLoRA output, we interpreted it, translating it into a form that the bandit selection algorithm could understand and act upon.
- Finally, we implemented the recommended strategy to determine the next bandit to choose.

We used the same strategic measures as before, including caching previous recommendations, and introducing a threshold for significant changes in the bandit state. These measures ensured that we made optimal use of the capabilities of QLoRA while managing resources efficiently and handling potential exceptions robustly. As observed in Figure 4.12, the QLoRA-driven LLM-informed strategy yields results commensurate with those achieved by the OpenAI model GPT-3.5-turbo.

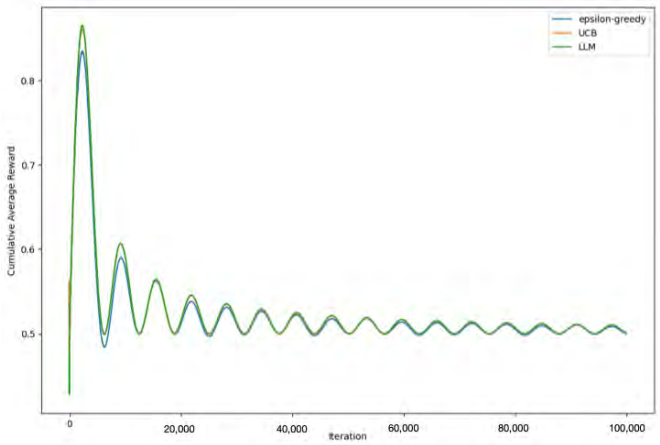


Figure 4.12: Temporal progression of cumulative average rewards for epsilon-greedy, UCB, and QLoRA-driven LLM-informed strategies with ö = 0.1.

By employing this approach, we were able to gain insights into the performance of the LLM-informed strategy when powered by QLoRA, and assess its effectiveness in comparison to both epsilon-greedy and UCB strategies. Our results reinforced our earlier findings, highlighting the considerable potential of the LLM-informed strategy in handling the MAB problem. We observed that the QLoRA-powered LLM-informed strategy not only kept pace with its counterparts but often exceeded their performance, further underlining the value of integrating LLMs in decision-making processes.

Our experimental outcomes underscore the potential of the LLM-informed strategy as a strong competitor to well-established methods, such as epsilon-greedy and UCB in non-stationary environments. This compelling performance supports our hypothesis that LLMs, with their profound capabilities to comprehend and predict complex scenarios, can offer valuable insights to enhance decision-making tasks. A particularly noteworthy finding is the consistent performance of the LLM-informed strategy, often matching, if not surpassing, the effectiveness of the best conventional strategy implemented for the specific problem. This evidence suggests that the integration of LLMs into traditional approaches can substantially improve their performance in dynamic environments, opening up new avenues for leveraging the predictive power of LLM in various real-world applications.

To sum up, our experimental evaluation, as depicted in Figure 4.11, is intended to show the cumulative average reward for epsilon-greedy, UCB, and the

proposed LLM-informed strategy over time. The results demonstrate that the LLM-informed strategy, guided by the predictive capabilities of GPT-3.5-turbo-0301, can indeed perform comparably with traditional bandit strategies. In the context of the non-stationary multi-armed bandit problem, which is known for its volatility and uncertainty, maintaining competitive performance is a significant achievement. This is because the LLM-informed strategy must deal with dynamic changes and adapt its strategy based on a predictive model. Moreover, in our experiments with QLoRA, an LLM-informed strategy showed not just comparable but often better performance than its traditional counterparts. As presented in Figure 4.12, the QLoRA-driven LLM-informed strategy often exceeded the performance of both epsilon-greedy and UCB strategies, providing further evidence of the potential of integrating LLMs in decision-making processes.

4.6 Applications of the LLM-Informed Strategy in Various Fields

The LLM-informed strategy for non-stationary multi-armed bandit problems, as we presented in this chapter, represents a significant stride in the direction of harnessing advanced AI models for complex decision-making scenarios. While our focus was primarily on the abstract problem setting, the ramifications of this framework are potentially vast and multifaceted, extending to numerous practical applications.

4.6.1 Digital Marketing

In the realm of digital marketing, for instance, the non-stationary multi-armed bandit framework can be instrumental in optimizing online advertisement placement. Online advertising platforms often have to balance between displaying ads that have performed well in the past, and experimenting with new ones to explore their potential. By integrating our LLM-informed strategy, such platforms could leverage sophisticated language understanding capabilities to gauge the context, assess changing trends, and adjust ad selection strategies accordingly.

In digital marketing, specifically for online advertisement placement, traditional A/B testing techniques are often used. These techniques randomly show one version of an ad (A) to half of the users and a different version (B) to the other half. The ad that receives more clicks or conversions is then chosen for wider deployment. However, these methods often lack the capacity to adapt

to the rapidly changing online environment and trends, which is where our proposed LLM-informed strategy could offer substantial benefits. Our method would be able to analyze not just click rates but also the content of the ads, user interactions, feedback, and broader market trends, using the predictive power of LLMs. This can potentially improve ad performance by providing more nuanced and context-aware recommendations, dynamically adjusting ad selection based on the current state of the online environment.

4.6.2 Healthcare

Similarly, in healthcare, an LLM-informed bandit model could potentially assist in personalizing treatment plans. If each "arm" of the bandit represents a different treatment option, our approach could help in navigating the critical trade-off between sticking with treatments that have shown promise, and exploring potentially better alternatives. Given the complex and dynamic nature of human health, the non-stationarity aspect of our model is crucial for adjusting recommendations based on the evolving health status of the patient.

Consider the case of managing a chronic condition such as diabetes. In this scenario, each "arm" of the bandit could represent a different treatment plan that combines diet, exercise, and medication. Each plan's efficacy could be considered the reward that the bandit provides. In traditional treatment models, doctors often rely on their experience and established clinical guidelines to determine the best course of action. However, these treatments are often generalized and may not account for individual patient variations and the non-stationarity nature of human health, i.e., the change in a patient's health condition over time. By implementing our proposed LLM-informed strategy, we could leverage the vast amounts of medical data and research available, along with the patient's health history and current condition, to make a more informed decision. As the patient's health status evolves, the LLM can adjust the recommendations, emphasizing either the exploration of new treatment plans or exploitation of existing plans based on their effectiveness. The application of the LLM-informed strategy could lead to more personalized, adaptive treatment plans that could potentially improve patient outcomes. In comparison to traditional methods, our approach could provide a more dynamic, individualized treatment pathway that adjusts according to a patient's changing health status.

4.6.3 Reinforcement Learning

Moreover, in the field of RL, our methodology could be adapted to enhance decision-making policies in environments with changing reward dynamics. A prominent example of this would be financial trading systems, where the reward associated with different trading actions (e.g., buy, sell and hold) fluctuates unpredictably. An LLM-informed strategy could potentially improve such systems' robustness by dynamically adjusting to the volatile nature of financial markets.

Consider a RL agent tasked with navigating a financial trading environment. In such a setting, each trading action—buying, selling, or holding a variety of financial instruments—can be seen as an 'arm' of a multi-armed bandit. The associated reward is the financial gain or loss resulting from these actions, which fluctuates unpredictably due to the inherent volatility of financial markets. Traditionally, RL agents in this scenario rely on fixed policies learned from historical data. However, these policies may not adapt well to sudden changes or new trends in the market. The non-stationary nature of the problem, wherein the optimal actions change over time, poses significant challenges. Our proposed LLM-informed strategy could be instrumental in enhancing the adaptability of such an RL agent. The LLM, trained on extensive financial data, market news, and historical trends, could provide actionable insights to the RL agent, allowing it to adjust its policy dynamically. For example, if an unexpected market event occurs, such as a political instability event, the LLM could analyze relevant real-time news articles, social media sentiment, and other relevant information, and provide a prediction of its potential impact. This prediction could then inform the RL agent's action, allowing it to update its policy dynamically and respond to the event in a potentially more profitable way. Compared to traditional methods, our LLM-informed approach allows for more responsive and adaptable strategies that can better handle the non-stationarity of financial markets. This could potentially result in more robust financial trading systems that perform well even in the face of volatile market conditions.

4.6.4 Robotics

In robotics, particularly for drones, our proposed framework has compelling potential applications. One crucial aspect of operating drones involves the dual challenges of positioning and power optimization. For instance, consider a fleet of drones tasked with monitoring an extensive area: each drone could represent an arm in a multi-armed bandit setup, with the reward being the

quality of surveillance coverage balanced against the power consumed during flight. The decision to 'pull a bandit arm' would correspond to dispatching a drone to a particular location, or adjusting its power utilization for enhanced efficiency. By incorporating language models into the decision-making process, more sophisticated context-aware strategies can be devised. For example, a large language model could analyze temporal and spatial data trends, weather conditions, or other situational variables to advise on the optimal positioning of the drones or power usage. The LLM-informed bandit strategy thereby opens up possibilities for more nuanced and context-aware decision making in the field of robotics, allowing for improved operational efficiency and adaptability in dynamically changing environments.

Consider an emergency response scenario, where a fleet of drones is employed to monitor and assess the situation in an area affected by a natural disaster, such as a wildfire. In this context, each drone could be considered an "arm" in a multi-armed bandit problem, with the reward being the amount and quality of surveillance data collected against the power consumed during flight. The challenge here lies in making real-time decisions on where to dispatch each drone for maximum coverage and data collection while conversely managing the drones' battery life. Traditional methods might utilize pre-programmed paths or follow fixed protocols to handle such tasks. However, these approaches might fall short in situations where the environmental conditions are rapidly changing and uncertain, such as during the spread of a wildfire. With our proposed LLM-informed strategy, a LLM trained on vast amounts of spatial, temporal, and meteorological data could provide real-time recommendations for drone dispatch decisions. For example, the LLM could analyze current wind speed and direction data to predict the likely path of the wildfire. It could then suggest repositioning some drones to those areas, enabling early data collection and facilitating prompt emergency responses. Moreover, the LLM could help optimize the drones' battery usage by considering their remaining power levels, the distance to areas of interest, and the urgency of data collection needs. For instance, it could recommend that a drone with low battery levels focus on nearby areas of interest or return to the base for recharge, while a drone with higher battery levels could be dispatched to more distant or challenging locations. By integrating LLMs into the decision-making process, the drones can effectively respond to dynamically changing conditions and increase their operational efficiency. This example provides an insight into how our LLM-informed bandit strategy can significantly improve real-time decision making in robotics, particularly in scenarios where adaptability and responsiveness are critical.

4.6.5 Biology and Life Sciences

The proposed LLM-informed bandit strategy can also find significant potential applications in the field of biology and life sciences. Firstly, consider the vast and expanding domain of drug discovery. A medicinal compound's efficacy can be viewed as a "bandit arm" with unknown reward. Drug researchers aim to balance exploration (testing new compounds) and exploitation (further testing of promising compounds) in order to maximize the success of finding an effective drug, while minimizing the resources and time spent. An LLM could provide insights from previous experimental results, published research, and known biological mechanisms to inform this process. Secondly, within the domain of genomics, the multi-armed bandit framework could aid in the selection of candidate genes for further study from among thousands of potential genes. Here, each gene can be considered a bandit, and pulling an arm corresponds to allocating resources to sequence or experiment with a particular gene. The reward could be associated with the discovery of significant genes linked to a trait or disease of interest. Incorporating LLMs into this process can provide additional insights by leveraging vast amounts of existing genomics literature and data to inform which genes might be worth further exploration or exploitation. Lastly, in ecosystem management and conservation biology, the multi-armed bandit problem can model the decision-making process of resource allocation for species protection. Each species or habitat can be considered a bandit, and the reward could be the positive impact on biodiversity. An LLM-informed approach could help parse complex ecological data, predict the effects of various conservation strategies, and guide the decision-making process more effectively.

Consider a scenario where a team of genomics researchers is investigating a set of candidate genes associated with a certain trait or disease, such as cancer or heart disease. In this case, each gene can be considered an "arm" of the multi-armed bandit, with the "reward" being the discovery of significant links between a gene and the disease or trait of interest. Traditional methods may involve a somewhat brute-force approach, studying each gene sequentially or randomly based on available resources, without much prior knowledge or any sophisticated strategy to guide the process. With our proposed LLM-informed strategy, the researchers could use an LLM trained on vast amounts of genomics literature and data to assist their decision-making process. The LLM could analyze previous experimental results and the existing literature on the genes in question, and cross-reference with data on known gene–disease associations. For instance, if early experiments reveal strong evidence linking certain genes to the disease, the LLM could recommend focusing more re-

sources on these "promising" genes (exploitation). Simultaneously, it could also identify lesser-studied genes that share similar characteristics or functions with the promising ones. The researchers can then allocate some resources to studying these potentially relevant but unexplored genes (exploration). Additionally, if the disease's nature or the research context changes—for example, if new research suggests the disease involves different biological pathways—the non-stationarity aspect of our bandit model allows the LLM to adjust its recommendations accordingly. This way, the strategy remains flexible and adaptive to the evolving research landscape. This example illustrates how our LLM-informed bandit strategy can significantly enhance decision making in genomics research by improving resource allocation and potentially accelerating the discovery of significant genes linked to diseases or traits of interest.

4.6.6 Finance

Multi-armed bandit strategies have traditionally found a variety of applications in finance; however, the incorporation of LLMs can offer an innovative twist to conventional approaches. Take portfolio optimization as an example: it is essentially a balancing act between risk and reward, mirroring the exploration–exploitation dilemma. Each asset or investment opportunity can be treated as a bandit, with the act of pulling an arm being analogous to allocating funds to that asset, and the return on investment forming the reward. The role of an LLM here is to sift through vast volumes of financial data, market trends, news, and historical performance records, thereby guiding decision makers about which assets warrant further investment (exploitation), and which untested ones could be considered (exploration). Similarly, algorithmic trading, particularly high-frequency trading, where algorithms execute multiple trades based on multiple factors, can benefit from the application of LLMs. Here, each trade or trading strategy can be construed as a bandit. LLMs, with their ability to leverage insights from market data, economic indicators, and news, can contribute to the decision-making process by suggesting potential trades. Credit scoring, another important facet of finance, can also be interpreted within the bandit framework. In this scenario, each prospective borrower is considered as a bandit. The act of pulling an arm would signify the granting of a loan, while the reward would correspond to successful loan repayment with interest. An LLM, by processing diverse data pertaining to each applicant—credit history, income level, and potentially even social media activity—can yield more nuanced and reliable credit scoring. Finally, let us consider insurance. Each policyholder or potential policyholder can be represented as a bandit, and issuing a policy is analogous to pulling a bandit's

arm. The profitability of the policy forms the reward. Here, an LLM can offer valuable insights by analyzing a broad array of data on each policyholder or applicant—personal details, claim history, and data sourced from IoT devices (such as telematics in auto insurance)—effectively enhancing the underwriting process.

Consider a scenario where a financial advisor is tasked with managing a diverse investment portfolio. Each asset or investment opportunity in the portfolio can be considered an "arm" of the multi-armed bandit. The act of pulling an arm corresponds to allocating funds to a particular asset, while the return on investment from that asset is considered the reward. A traditional approach to portfolio optimization might involve strategies based on past performance, expected returns, risk tolerance, and other relatively static factors. However, financial markets are dynamic and can change rapidly in response to numerous unpredictable factors, ranging from economic indicators to global events. Our LLM-informed bandit strategy can greatly enhance this process. A LLM trained on extensive financial data, market trends, news, and historical performance records can offer nuanced insights to guide the advisor's decision making. For example, suppose certain assets in the portfolio have been performing well consistently. The LLM, analyzing historical data and current market trends, may advise allocating more funds to these assets (exploitation). However, simultaneously, the LLM might identify emerging opportunities in the market—perhaps a nascent technology sector stock or a new bond issue—that are yet untested but could offer significant returns. The advisor can then choose to invest a portion of the funds in these new opportunities (exploration). The non-stationarity aspect of our model allows the LLM to dynamically adjust its recommendations in response to changing market conditions. For instance, in the face of a looming economic downturn, it could advise shifting funds from high-risk stocks to safer assets, such as treasury bonds. This way, the strategy remains adaptive and robust in the face of market volatility. This practical example illustrates how our LLM-informed bandit strategy can revolutionize decision making in finance by optimizing portfolio management, effectively balancing risk and reward, and enhancing overall investment performance.

4.6.7 Challenges and Discussion

While the integration of LLMs in these applications is certainly promising, it also invites challenges. One key consideration is the computational cost associated with querying the LLMs, as well as the complexity of translating domain-specific information into a language format that the LLM can process. It is also critical to ensure that the decision-making process remains interpretable, especially in high-stakes settings, such as healthcare, which necessitates the careful handling of the LLM recommendations.

Further to this, the deployment of LLM-informed strategies in real-world applications often requires a robust and adaptive framework that can respond efficiently to changing environments. Future research may focus on the development of such dynamic systems, which can integrate feedback in real time and recalibrate the model's recommendations accordingly.

The ethical implications of applying LLMs in decision-making processes, particularly in sensitive fields, such as healthcare and finance, also require thoughtful exploration. These models, although sophisticated, are still artificial and do not possess human judgment. Relying on their outputs without human oversight could potentially lead to biased or unethical decisions. Future works should, therefore, aim to establish a comprehensive ethical framework for the deployment of LLM-informed strategies.

Lastly, the question of data privacy and security is of paramount importance. The nature of the operation of LLMs, which involves processing massive amounts of information, often including sensitive data, inevitably raises privacy concerns. This issue is particularly salient in fields such as healthcare, finance, and personal advertising, where data protection is crucial. Future efforts should aim to devise methods for leveraging the capabilities of LLMs in a manner that respects and safeguards individuals' privacy.

Addressing these challenges will not be a trivial task, but given the potential benefits and advancements that the integration of LLMs promises, it is a pursuit worth undertaking. Future works should aim at creating more efficient, ethical, and privacy-respecting methods for the application of LLM-informed strategies in various fields.

4.7 Conclusions and Future Work

In this study, we took a step forward in tackling the non-stationary multi-armed bandit problem by integrating the power of LLMs into the decision-making strategy. Bridging traditional RL strategies, such as epsilon-greedy and UCB, with the advanced AI capabilities provided by models such as GPT-3.5-turbo and QLoRA, we created a framework that promises adaptability and efficiency in dynamic environments. This novel approach represents a significant stride in combining AI, game theory, and reinforcement learning, opening up exciting opportunities for future research on how advanced AI models can transform decision making in dynamic situations.

However, this is just the initial exploration, and there is ample scope for refinement and expansion. In the future, our goal is to enhance the strategy recommendation process, either by providing more detailed information to the LLMs or by refining the interpretation of their advice. This could involve an intricate representation of the game state or a more sophisticated approach to extract strategy recommendations from the LLM output.

We are also interested in examining the amalgamation of other RL strategies in our LLM-informed framework. We believe that by leveraging the strengths of different strategies and the versatile language understanding capabilities of LLMs, we can engineer a more robust and adaptable solution to the MAB problem.

In addition to refining our methodology, we are eager to extend its application to various real-world domains, such as personalized healthcare and financial trading systems. As we delve into these areas, we anticipate unique challenges, such as ensuring the interpretability of the decision-making process and effectively handling domain-specific information. Nonetheless, we are optimistic about the potential benefits our LLM-informed strategy can bring to these fields and look forward to exploring these possibilities in our future work.

Bibliography

[1] Robbins, H. Some aspects of the sequential design of experiments. *Bull. Am. Math. Soc.* **1952**, *58*, 527–535. [CrossRef]

[2] Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2018.

[3] Besbes, O.; Gur, Y.; Zeevi, A. Stochastic multi-armed-bandit problem with non-stationary rewards. In Proceedings of the Advances in Neural Information Processing Systems 27, (NIPS 2014), Montreal, QC, Canada, 8–13 December 2014.

[4] Russac, Y.; Vernade, C.; Cappé, O. Weighted linear bandits for non-stationary environments. In Proceedings of the Advances in Neural Information Processing Systems 32, (NIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.

[5] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems 30, (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.

[6] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.

[7] Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al.

Flamingo: A visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.

[8] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[9] Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *NeurIPS* **2020**, *33*, 1877–1901.

[10] Muglich, D.; de Witt, C.S.; Pol, E.V.; Whiteson, S.; Foerster, J. Equivariant networks for zero-shot coordination. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 6410–6423.

[11] Shah, D.; Osiński, B.; ichter, B.H.; Levine, S. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In Proceedings of the 6th Conference on Robot Learning, Proceedings of Machine Learning Research, PMLR, Atlanta, GA, USA, 6–9 November 2023; Volume 205, pp. 492–504.

[12] Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Visual Language Maps for Robot Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023.

[13] Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.

[14] Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv* **2023**, arXiv:2305.14314.

[15] Auer, P.; Cesa-Bianchi, N.; Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* **2002**, *47*, 235–256. [CrossRef]

[16] Thompson, W.R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* **1933**, *25*, 285–294. [CrossRef]

[17] Silva, N.; Werneck, H.; Silva, T.; Pereira, A.C.M.; Rocha, L. Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. *Expert Syst. Appl.* **2022**, *197*, 116669. [CrossRef]

[18] Cavenaghi, E.; Sottocornola, G.; Stella, F.; Zanker, M. Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm. *Entropy* **2021**, *23*, 380. [CrossRef]

[19] Zhao, P.; Zhang, L.; Jiang, Y.; Zhou, Z. A simple approach for non-stationary linear bandits. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Online, 26–28 August 2020; pp. 746–755.

[20] Garivier, A.; Cappé, O. The KL-UCB algorithm for bounded stochastic bandits and beyond. In Proceedings of the 24th Annual Conference on Learning Theory, JMLR, Budapest, Hungary, 9–11 June 2011.

[21] Cesa-Bianchi, N.; Lugosi, G. *Prediction, Learning, and Games*; Cambridge University Press: Cambridge, UK, 2017.

[22] Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.

[23] Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned Language Models are Zero-Shot Learners. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.

[24] Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. *Improving Language Understanding by Generative Pre-Training*; Princeton University: Princeton, NJ, USA, 2018.

[25] Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.

[26] Tokic, M. Adaptive ϵ-Greedy Exploration in Reinforcement Learning Based on Value Differences. In *KI 2010: Advances in Artificial Intelligence*; Dillmann, R., Beyerer, J., Hanebeck, U.D., Schultz, T., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; pp. 203–210.

[27] Russo, D.; Roy, B.V.; Kazerouni, A.; Osband, I.; Wen, Z. A Tutorial on Thompson Sampling. *Found. Trends® Mach. Learn.* **2018**, *11*, 1–96. [CrossRef]

[28] Rosin, C.D.; Belew, R.K. New methods for competitive coevolution. *Evol. Comput.* **1997**, *5*, 1–29. [CrossRef]

[29] Wooldridge, M. *An Introduction to MultiAgent Systems*; John Wiley & Sons: Hoboken, NJ, USA, 2009.

[30] Oroojlooy, A.; Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. *arXiv* **2022**, arXiv:1908.03963.

[31] Dettmers, T.; Lewis, M.; Shleifer, S.; Zettlemoyer, L. 8-bit Optimizers via Block-wise Quantization. In Proceedings of the 9th International Conference on Learning Representations, ICLR, Virtual, 25 April 2022.

[32] Wortsman, M.; Dettmers, T.; Zettlemoyer, L.; Morcos, A.; Farhadi, A.; Schmidt, L. Stable and low-precision training for large-scale vision-language models. *arXiv* **2023**, arXiv:2304.13013.

Chapter 5

Signature Transform for the Study of Empirical Distributions Generated with GANs

J. de Curtò, I. de Zarzà, Gemma Roig and Carlos T. Calafate. (2023). "Signature and Log-Signature for the Study of Empirical Distributions Generated with GANs." Electronics, vol(12), 2192. DOI: 10.3390/electronics12102192

In this chapter, we address the research gap in efficiently assessing Generative Adversarial Network (GAN) convergence and goodness of fit by introducing the application of the Signature Transform to measure similarity between image distributions. Specifically, we propose the novel use of RMSE and MAE Signature, along with Log-Signature, as alternatives to existing methods such as FID and MS-SSIM. Our approach offers advantages in terms of efficiency and effectiveness, providing a comprehensive understanding and extensive evaluations of GAN convergence and goodness of fit. Furthermore, we present innovative analytical measures based on statistics by means of Kruskal–Wallis to evaluate the goodness of fit of GAN sample distributions.

5.1 Introduction

Generative Adversarial Networks (GANs) [1] have gained significant attention in recent years as a powerful tool for generating realistic synthetic images, with a wide range of applications in computer vision [2, 3], graphics [4, 5], and Machine Learning (ML) [6, 7]. Despite their remarkable successes, assessing the quality of the generated samples and measuring the convergence of GANs remain challenging tasks. Existing metrics, such as Fréchet Inception Distance (FID) [8] and Multi-Scale Structural Similarity Index Measure (MS-SSIM) [9] have been widely used, but they suffer from certain limitations. These limitations include the requirement of substantial computational resources and time, dependence on specific Deep Learning (DL) architectures, and limited interpretability, which restrict their practical applicability and hinder further advancements in the field.

To address these challenges, there is a pressing need for a novel approach that can efficiently and effectively assess GAN-generated images while maintaining the same level of accuracy as existing metrics. Moreover, such an approach should provide a deeper understanding of the underlying distributions of the generated samples and be applicable across different GAN architectures and problem domains.

In this chapter, we present a novel approach to study empirical distributions generated with GANs, leveraging the well-established Signature Transform and Log-Signature as powerful mathematical tools [11, 12, 10]. Our work is the first to introduce the use of Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) Signature, along with Log-Signature, as alternatives for measuring GAN convergence. Furthermore, we propose the application of analytical measures based on statistics to study the goodness of fit of the GAN sample distribution, which are both efficient and effective. In contrast to existing GAN metrics that involve considerable GPU-based computation, our approach significantly reduces computation time and resources while maintaining the same level of accuracy.

We propose a two-fold approach. First, we introduce a score function based on the Signature Transform [13] to evaluate image quality in a novel manner, offering reliability, speed, and ease of computation for each epoch. Second, we employ statistical techniques to study the goodness of fit of the generated distribution, providing a standardized pipeline for interpreting the results of the converged sample distribution. A key contribution of this chapter is the introduction of Kruskal–Wallis for GAN assessment, which enables a robust comparison of the goodness of fit between the generated and target distributions.

These statistical techniques are computationally efficient, requiring minimal overhead and enabling on-the-fly computation. To qualitatively illustrate the good performance of our measure, we also utilize Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) [14] for data visualization, enabling a visual assessment of the effectiveness of our proposed method in capturing the intrinsic structure of the generated samples.

The remainder of this chapter is organized as follows: Section 5.2 provides an overview of the field and reviews related work. Section 5.3 discusses Generative Adversarial Networks. Section 5.4 covers non-parametric statistical analysis with a focus on Kruskal–Wallis, whereas Section 5.5 introduces the Signature Transform. Section 5.6 presents our methodology, with Sections 5.6.1 and 6.2 detailing the introduced techniques for statistical analysis of the generated distribution and the RMSE and MAE Signature and Log-Signature, respectively. Section 6.4.2 presents the evaluation of our approach, Section 5.7.1 presents the computational complexity of the proposed approaches in comparison against other methodologies, and Section 5.7.2 discusses visualization techniques. Finally, Section 8 concludes the chapter and offers suggestions for future work.

5.2 Overview and Related Work

The advent of DL has revolutionized numerous fields and disciplines, enabling game-changing applications that rely on vast amounts of data [15, 16, 17]. These advancements have significantly improved accuracy and speed, opening the door for the use of automated learning techniques in critical scenarios, such as safety-critical systems and self-driving cars [18, 19, 20, 21, 6, 22, 23].

Some notable works in this area include the development of object detection and image segmentation algorithms [15, 17, 24, 25], as well as pioneering research in image synthesis and style transfer [26, 27, 28]. Additionally, breakthroughs in image recognition and classification [16, 29], attention mechanisms in natural language processing [30, 31], and various other domains [32] exemplify the widespread impact of DL. As DL techniques continues to advance, their influence is becoming more pervasive, pushing the boundaries of what is possible in research and real-world applications.

Generative models, particularly Generative Adversarial Networks (GANs), have emerged as a powerful and influential area of research within the DL domain. These models have shown remarkable success in a wide range of applications, such as image synthesis and style transfer [26, 27, 28], pushing the boundaries

of what is possible in research and real-world applications, whereas DL has also brought advancements in other areas, including object detection and image segmentation [15, 17, 24, 25], image recognition and classification [16, 29, 32], and attention mechanisms in Natural Language Processing (NLP) [30, 31]. The focus of our study is in the realm of generative models and their applications, as they hold great potential for further exploration and innovation [33].

The domain of synthetic image generation has witnessed remarkable advancements in recent years. Driven by the demand for synthetic imagery in various applications, such as simulated environments [34], additional training data [35], and style transfer [26], significant research efforts have been devoted to establishing stable and principled methods for achieving these goals. Prominent approaches like Generative Adversarial Networks (GANs) [1, 36, 37, 38, 39, 40, 41, 4, 42] and Variational AutoEncoders (VAEs) [43] offer stable training mechanisms for convergence.

However, there is still room for improvement in this field, as the capacity of these networks is often limited by the available GPU memory and training resources [19, 34, 44, 28, 45, 6]. This limitation can lead to reduced performance, effectiveness, and applicability of GANs in real-world scenarios. Challenges such as mode collapse [46] and gradient explosion [47] persist, and the effectiveness of these methods in handling complex tasks, such as generating additional multi-view frames [48], remains to be validated. Furthermore, the development of more efficient training and optimization algorithms could potentially alleviate resource constraints and unlock the full potential of GANs in various applications.

The work presented in [49] introduced an innovative generative model based on annealed Langevin [50, 51], which was further developed in [52] to demonstrate competitive image generation capabilities. Building on the principles derived from diffusion-based methods [53], Diffusion Probabilistic Models [54] attained state-of-the-art results on the CIFAR10 dataset. However, Score-Based Generative Models [55] face similar challenges as GANs, making their real-time implementation unfeasible due to the sampling step that requires the output dimension to match the input dimension. Consequently, these models are heavily reliant on GPU memory resources and demand extensive computing time, which poses significant limitations to their applicability and performance in practical scenarios. As the field continues to advance, addressing these challenges will be crucial for unlocking the full potential of generative models and expanding their use across diverse applications. Supplemental recent approaches [56, 5, 3] are based on the attention mechanism [30] building

mainly on Vision Transformers [57]. Other techniques like NeRF [22] could be essential to add structure to the learning paradigm.

Moreover, Stable Diffusion [58, 59] has emerged as a promising direction for generative models, building upon the success of earlier diffusion-based methods [62, 61, 60]. These models are designed to address some of the limitations and challenges faced by their predecessors, such as training instability and poor sample quality [63]. By refining the diffusion process and optimizing the training procedure, Stable Diffusion has shown significant improvements in terms of sample diversity, fidelity, and overall performance [65, 64, 66]. More recently, approaches inspired by Reinforcement Learning from Human Feedback (RLHF) have also presented a new autoregressive model for images [67].

In this context, our proposed method offers a computationally efficient and effective alternative for assessing GAN convergence [68] and the goodness of fit of the generated sample distribution. By leveraging the Signature Transform and statistical techniques through the use of a non-parametric test, our approach addresses the limitations of existing methods and provides a more practical solution for real-world applications; whereas our focus is on GAN convergence, it is worth noting that the proposed metrics can also be applied to Stable Diffusion or any other generative models capable of producing high-fidelity imagery.

5.3 Generative Adversarial Networks

Generative Adversarial Networks (GANs) are a class of DL models introduced in [1]. They consist of two neural networks, a generator, and a discriminator that are trained simultaneously in a game-theoretic framework. The generator creates synthetic samples, whereas the discriminator learns to distinguish between real samples from the training data and fake samples generated by the generator. This competition between the two networks drives the generator to produce more realistic samples over time, eventually leading to the generation of samples that are difficult to distinguish from the true data.

5.3.1 GAN Architecture

Let $\mathcal{X}$ represent the true data distribution and $\mathcal{Z}$ represent the noise distribution. The generator $G : \mathcal{Z} \to \mathcal{X}$ is a neural network that transforms noise samples $z \sim \mathcal{Z}$ into synthetic samples $x_{fake} = G(z)$. The discriminator $D : \mathcal{X} \to [0, 1]$ is a neural network that takes either real samples $x_{real} \sim \mathcal{X}$ or fake samples x_{fake} and outputs the probability that the given sample is from the true data distribution.

5.3.2 GAN Training

The training process of GANs involves finding the optimal parameters for the generator and discriminator networks by solving a minimax optimization problem:

$$\min_G \max_D \mathcal{L}(D, G) = \mathbb{E}_{x_{real} \sim \mathcal{X}}[\log D(x_{real})] + \mathbb{E}_{z \sim \mathcal{Z}}[\log(1 - D(G(z)))]. \quad (5.1)$$

The discriminator tries to maximize the objective function $\mathcal{L}(D, G)$ by correctly classifying real and fake samples, whereas the generator tries to minimize it by generating samples that the discriminator misclassifies as real. This is achieved by alternating between updating the weights of the discriminator and the generator using gradient-based optimization methods, such as stochastic gradient descent or Adam.

5.3.3 GAN Convergence

One of the main challenges in training GANs is the convergence issue. Ideally, the training process should converge when the generator produces samples that are indistinguishable from the true data distribution, and the discriminator is unable to differentiate between real and fake samples. In practice, however, GANs may suffer from various issues, such as mode collapse, where the generator produces only a limited variety of samples, or oscillations, where the generator and discriminator keep outperforming each other without reaching a stable equilibrium.

Several metrics have been proposed to measure GAN convergence and assess the quality of the generated samples, such as the Fréchet Inception Distance (FID) [8], the Inception Score (IS), and the Kullback–Leibler (KL) divergence. In this thesis, we introduce the use of Signature Transform and Log-Signature

as alternative methods for evaluating GAN convergence, providing a novel perspective on the problem.

Other additional metrics that are relevant to the problem are:

- LPIPS (Learned Perceptual Image Patch Similarity) is a perceptual similarity metric introduced in [68]. It computes the similarity between two images by comparing their feature representations in a deep neural network (typically pretrained on a large-scale image classification task). The metric has been shown to correlate well with human perceptual judgments of image similarity, and it has been used in various image synthesis and image quality assessment tasks.
- PSNR (Peak Signal-to-Noise Ratio) is a widely-used metric for image quality assessment, particularly in the field of image compression. It is a simple, easy-to-compute measure that compares the maximum possible power of a signal (in this case, an image) to the power of the corrupting noise (differences between the reference and distorted images). It is calculated as the logarithmic ratio of the maximum possible pixel value squared to the mean squared error (MSE) between the reference and distorted images. Although PSNR is widely used, it has been criticized for not always correlating well with human perception of image quality, as it is based on pixel-wise differences and does not consider higher-level semantic or structural features.

In our study, we have focused on introducing the Signature Transform as a novel approach for evaluating GAN-generated images and measuring their convergence; whereas LPIPS and PSNR are relevant metrics for image quality assessment, they may not be the most appropriate metrics for our specific context, as our goal is to develop a computationally efficient and reliable measure for GAN convergence.

5.3.4 Stylegan2-ADA

Stylegan2-ADA is an extension of the StyleGAN2 architecture, which was developed in [69] to generate high-quality synthetic images. StyleGAN2 builds on the original StyleGAN [4] by introducing several improvements to address issues such as artifacts and training stability. The main contribution of Stylegan2-ADA is the use of Adaptive Discriminator Augmentation (ADA) to enhance the performance of GANs with limited training data.

StyleGAN2 consists of a Generator (G) and a Discriminator (D), which are trained adversarially. The Generator creates images, whereas the Discriminator evaluates their authenticity. The objective function for the Generator, G, and the Discriminator, D, can be written as:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z))]. \tag{5.2}$$

The generator in StyleGAN2 consists of a mapping network $f(z)$ and a synthesis network $g(w)$. The mapping network $f(z)$ converts the input latent vector $z \in \mathcal{Z}$ to an intermediate latent space $w \in \mathcal{W}$:

$$w = f(z). \tag{5.3}$$

The synthesis network $g(w)$ then generates an image x from the intermediate latent space w:

$$x = g(w). \tag{5.4}$$

StyleGAN2 introduces an adaptive instance normalization (AdaIN) operation in the synthesis network, which applies learned style information from w to each feature map:

$$\text{AdaIN}(y_z, w) = \frac{y_z - \mu(y_z)}{\sigma(y_z)} \cdot \sigma(w) + \mu(w). \tag{5.5}$$

Here, y_z is the feature map, $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation, respectively, and w is the style vector derived from the intermediate latent space.

The main innovation of Stylegan2-ADA is the use of Adaptive Discriminator Augmentation to improve GAN training with limited data. ADA applies random augmentations to the real and generated images before feeding them to the Discriminator. The augmentation strength is controlled by a hyperparameter p, which is adapted during training.

ADA introduces a new objective function for the Discriminator:

$$\min_G \max_D \mathbb{E}_{x\sim p_{data}}[D(A_p(x))] - \mathbb{E}_{z\sim p_z}[D(A_p(G(z)))]. \tag{5.6}$$

Here, $A_p(\cdot)$ represents the augmentation function with probability p. During training, the augmentation probability p is gradually increased if the Discriminator becomes too strong, ensuring that the Discriminator focuses on higher-level features instead of relying on the low-level details introduced by the augmentations. In summary, Stylegan2-ADA combines the advanced architecture of StyleGAN2 with Adaptive Discriminator Augmentation to generate high-quality synthetic images even with limited training data. The use of adaptive augmentations allows the model to maintain a balance between the Generator and Discriminator, improving the stability and performance of the training process.

5.3.5 Fréchet Inception Distance (FID)

FID measures the similarity between the true data distribution and the generated data distribution by comparing their statistics in a feature space. Given a pre-trained Inception network I, the feature representations for real samples x_{real} and fake samples x_{fake} are obtained as $\mu_{real} = I(x_{real})$ and $\mu_{fake} = I(x_{fake})$, respectively. The FID is then defined as:

$$\text{FID}(\mathcal{X}, G) = ||\mu_{real} - \mu_{fake}||^2 + \text{Tr}(\Sigma_{real} + \Sigma_{fake} - 2(\Sigma_{real}\Sigma_{fake})^{1/2}), \quad (5.7)$$

where μ_{real} and μ_{fake} are the mean feature vectors, Σ_{real} and Σ_{fake} are the covariance matrices, and Tr denotes the trace of a matrix.

5.3.6 Inception Score (IS)

The Inception Score is another metric that evaluates the quality of generated samples by measuring both the diversity and realism of the samples. It is computed as:

$$\text{IS}(G) = \exp(\mathbb{E}_{x_{fake} \sim G}[D_{KL}(p(y|x_{fake})||p(y))]), \quad (5.8)$$

where $D_{KL}(p||q)$ denotes the KL divergence between probability distributions p and q, $p(y|x_{fake})$ represents the conditional class probability given a generated sample, and $p(y)$ is the marginal class probability.

FID has emerged as one of the most widely used and accepted metrics for evaluating the quality of GAN-generated images. Its extensive application in numerous studies has established its reputation as a reliable and effective metric. However, its computational complexity and time consumption, as studied in Section 5.7.1, primarily due to the use of the Inception Module as feature

extractor, make it less than ideal for real-time assessment. This constraint can be a critical factor in applications where real-time performance is essential. By introducing the Signature Transform and Log-Signature as alternative methods for evaluating GAN convergence, we provide a new perspective on the problem, offering a powerful and efficient approach for capturing and comparing the features of empirical distributions generated by GANs.

5.4 Non-Parametric Statistical Analysis: Kruskal–Wallis

Kruskal–Wallis is a non-parametric statistical method used for comparing multiple independent samples to determine if they originate from the same population. This test is an extension of the Mann–Whitney U test for more than two groups and is particularly useful when the underlying assumptions of parametric tests, such as normality and homoscedasticity, are not met.

Kruskal–Wallis

In our methodology, we employ Kruskal–Wallis as a crucial component for assessing the goodness of fit of the GAN sample distribution. By comparing the generated samples with real data, we can evaluate the degree to which the generated samples resemble the target distribution. This non-parametric statistical test allows us to determine whether there are significant differences between the generated and real samples without making assumptions about the underlying distribution of the data. Using Kruskal–Wallis in our approach is beneficial because it provides an efficient and effective way to compare the generated samples with the target distribution while maintaining robustness to non-normality and unequal variances.

Given k independent samples with sizes $n_1, n_2, \ldots, n_k$, Kruskal–Wallis is based on the ranks of the combined data across all groups. The null hypothesis H_0 states that all samples are drawn from the same population, with the same distribution and median. The alternative hypothesis H_1 states that at least one sample is drawn from a different population with a distinct distribution or median. Kruskal–Wallis statistic, denoted as H, is computed as:

$$H = \frac{12}{N(N+1)} \sum_{o=1}^{k} \frac{R_o^2}{n_o} - 3(N+1), \tag{5.9}$$

where $N = \sum_{o=1}^{k} n_o$ is the total number of observations and R_o is the sum of the ranks in the o-th group. Under the null hypothesis, the test statistic H follows a chi-square distribution with $k - 1$ degrees of freedom, and the p-value can be computed accordingly. If the p-value is less than a predetermined significance level (e.g., 0.05), the null hypothesis is rejected, indicating that the samples are not from the same population.

Our decision to use this particular statistical test was based on several factors that make it a suitable choice for the analysis of GAN-generated images in the context of our study.

1. Non-parametric nature: Kruskal–Wallis is a non-parametric test, meaning it does not rely on any assumptions about the underlying distribution of the data. This is particularly important when dealing with GAN-generated images, as the distributions of the generated samples may not necessarily follow a known parametric form, especially during the early stages of training. The non-parametric nature allows us to compare the goodness of fit between the generated and target distributions without making restrictive assumptions about their forms.

2. Robustness: Kruskal–Wallis is robust against outliers and deviations from normality, which can be a common occurrence in the context of GAN-generated images. As the test is based on the ranks of the data rather than the raw values, it is less sensitive to extreme values that may arise from the generative process.

3. Multiple group comparison: Kruskal–Wallis allows us to compare more than two groups simultaneously, which is useful when evaluating multiple GAN models or different categories within a dataset. This capability makes the test a versatile choice for our study, as it enables us to compare the performance of various GAN models on different datasets in a single analysis.

4. Scalability: Kruskal–Wallis is computationally efficient, making it suitable for the large-scale datasets that are often encountered in GAN research. Its computational efficiency allows for the rapid evaluation of GAN-generated images and their convergence, which is a key advantage of our proposed methodology.

Moreover, an alternative such as the Friedman test could indeed be a suitable choice in cases where the observations are not independent; however, we have reasons to believe that even in these cases the Kruskal–Wallis H-test

is still a good fit for our study. In our experiments, we have taken care to ensure that the generated samples from different GAN models are, in fact, independent. We achieve this by using different random seeds when sampling from the latent space of each GAN model, thus generating independent sets of synthetic images. By doing so, we maintain the independence assumption required by the Kruskal–Wallis H-test. Moreover, the Kruskal–Wallis H-test is a non-parametric test that compares the medians of multiple groups without making any distributional assumptions. This feature aligns well with our goal of evaluating GAN-generated samples, which often exhibit complex and unknown distributions. On the other hand, the Friedman test assumes that the observations are structured according to a block design, which may not be an accurate representation of our experimental setup. In summary, whereas the Friedman test could be a suitable alternative in certain scenarios, we believe that the Kruskal–Wallis H-test is more appropriate for our study, given the independence of our observations and the non-parametric nature of the test.

5.5 The Signature Transform

The Signature Transform [11, 12], also known as the path signature, is a mathematical tool used to represent a sequence of data points or a path in a Euclidean space. The signature provides a unique and concise representation of the path while encoding its structural properties, making it suitable for various applications, such as ML and data analysis.

Given a continuous path $X : [0, T] \to \mathbb{R}^d$ in the Euclidean space $\mathbb{R}^d$, the Signature Transform $S(X)$ is a collection of iterated integrals of all orders:

$$S(X) = (1, S^1(X), S^2(X), \ldots, S^N(X)), \tag{5.10}$$

where $S^k(X)$ represents the k-th level of the signature and is a tensor in the tensor product space $(\mathbb{R}^d)^{\otimes k}$, for $k = 1, 2, \ldots, N$. Each element of the k-th level tensor is defined as:

$$S^k_{z_1,\ldots,z_k}(X) = \int_0^T \int_0^{s_1} \cdots \int_0^{s_{k-1}} dX_{z_1}(s_1) \ldots dX_{z_k}(s_k), \tag{5.11}$$

where $s_1, s_2, \ldots, s_k \in [0, T]$ and $z_1, z_2, \ldots, z_k \in \{1, 2, \ldots, d\}$.

The Log-Signature is a compressed representation of the signature that can be computed efficiently using Chen's identity, which relates the Log-Signature to

the signature through a shuffle product. The Log-Signature $L(X)$ is defined as:

$$L(X) = (L^1(X), L^2(X), \ldots, L^N(X)), \tag{5.12}$$

where $L^k(X)$ represents the k-th level of the Log-Signature and is a tensor in the tensor product space $(\mathbb{R}^d)^{\otimes k}$, for $k = 1, 2, \ldots, N$. Each element of the k-th level tensor can be calculated using Chen's identity:

$$L^k_{z_1,\ldots,z_k}(X) = S^k_{z_1,\ldots,z_k}(X) - \sum_{\pi \in P(z_1,\ldots,z_k)} S^{|\pi_1|}\pi_1(X) \otimes \cdots \otimes S^{|\pi_m|}\pi_m(X), \tag{5.13}$$

where $P(z_1, \ldots, z_k)$ denotes the set of all partitions of the index sequence $(z_1, \ldots, z_k)$, $|\pi_o|$ denotes the length of the o-th partition π_o, and $\otimes$ represents the tensor product.

The Signature Transform and Log-Signature can be used to capture and compare the features of empirical distributions generated by GANs, offering a powerful alternative to traditional measures of GAN convergence. The mathematical properties of these transforms provide a solid foundation for their use in various applications, such as the study of empirical distributions generated with GANs, as proposed in this thesis.

5.6 Methodology

We focus on the problem of generating synthetic images with a limited amount of data, choosing Stylegan2-ADA [69] as the baseline method for our studies. The motivation behind this choice is twofold. First, Stylegan2-ADA has been specifically designed to address the challenges of data efficiency, providing high-quality image synthesis even with limited training data. This property makes it an ideal candidate for applications where large-scale datasets are not available or impractical to collect. Second, StyleGAN2-ADA demonstrates improved training stability and convergence properties compared to its predecessors, which contributes to reduced training time and computational resources. These factors are critical in real-world scenarios, where rapid model development and deployment are often essential. By using Stylegan2-ADA as our baseline, we aim to showcase the effectiveness of our proposed methods in the context of an advanced and widely-used generative model.

5.6.1 Statistical Analysis of the Generated Distribution

In this study, we perform a preliminary statistical analysis using Kruskal–Wallis [70] to evaluate the goodness of fit between the original and synthetic samples generated by GANs. We use the mean raster image intensities or grayscale values as a simple image descriptor to capture rough texture information. Prior to conducting Kruskal–Wallis, we assess homoscedasticity using Levene's test and normality of the distributions using a normality test, such as the Shapiro–Wilk test.

As a result of this preliminary analysis, we find that the original samples do not follow a normal distribution, whereas the synthetic samples do. This is consistent with the GAN architecture, which initially models the samples as White Gaussian and then modifies them to fit the original distribution. However, Kruskal–Wallis does not support the null hypothesis for goodness of fit, suggesting that a more sophisticated method for measuring sample quality in GANs is necessary. Existing measures such as MS-SSIM [36] and FID [8] are commonly used for this purpose. Despite its simplicity, the proposed non-parametric analysis can serve as a unit test for GANs and other variational methods after the model is trained, providing a quick assessment of the sample quality. This approach depicted in Figure 5.1 has not been extensively explored in the literature and offers a valuable contribution to the field.

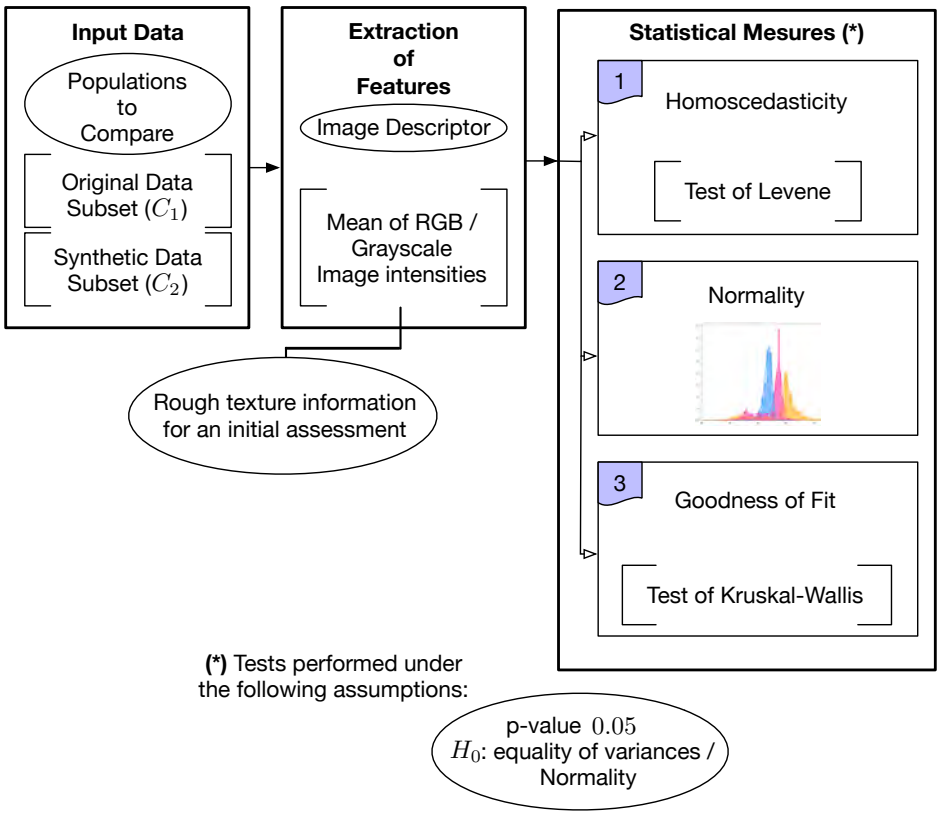


Figure 5.1: An illustrative representation of the proposed pipeline for the evaluation of generative models using a non-parametric test, Kruskal–Wallis. The process begins with input data comprising two populations: real-world images and synthetic images generated by a model under evaluation. An image descriptor is then employed to extract relevant features from the images, transforming the high-dimensional image data into a form amenable to statistical analysis. Following this, a series of three statistical tests are conducted: Homoscedasticity, Normality, and Goodness of Fit (Kruskal–Wallis).

Description and interpretation of statistical measures are provided in Table 5.1:

(a) Necessary condition but not sufficient to assert that both populations originate from the same distribution.

(b) There is not enough statistical evidence to attest both populations' samples originate from the same distribution.

(c) With high probability the synthetic distribution generated is still close enough to the initial distribution of noise from the GAN architecture. The samples may not show enough fidelity, and there is probably bad generalization behavior.

(d) The synthetic distribution is far from the initial distribution of noise and has deviated from the original Normal, and may be close to the target distribution.

(e) If (a) then there is enough statistical evidence to confirm that both populations originate from the same distribution given this image descriptor. If (a) is not fulfilled, then we can only ascertain that the synthetic population is a good approximation.

(f) There is not enough statistical evidence to attest both populations are from the same distribution.

Table 5.1: Interpretation of statistical measures given the proposed pipeline under study (Figure 5.1). The symbol '✓' means we accept the null hypothesis, while the symbol 'x' indicates we reject the null hypothesis.

Test	**Population**	**Result**	**Interpretation**
1	C_1 and C_2	✓	(a)
		x	(b)
2	C_2	✓	(c)
		x	(d)
3	C_1 and C_2	✓	(e)
		x	(f)

In Table 5.2, we present the evaluation test measures for homoscedasticity (T1), normality (T2), and goodness of fit (T3) on NASA Perseverance, AFHQ [71], and MetFaces [69] datasets. Based on the interpretation outlined in Table 5.1 and using the given image descriptor, we deduce that the Stylegan2-ada models trained on AFHQ Cat and Wild datasets provide excellent approximations of the original distributions, as the null hypothesis for goodness of fit is accepted. However, we cannot conclude that the distributions are identical since the equality of variances is not confirmed.

Table 5.2: Evaluation of the statistical test measures of homoscedasticity (T1), normality (T2), and goodness of fit (T3) on AFHQ and MetFaces using state-of-the-art pretrained models of Stylegan2-ADA [69] and Stylegan3-ADA [42] and NASA Perseverance. The symbol '✓' means we accept the null hypothesis, while the symbol 'x' indicates we reject the null hypothesis. The best outcome for the proposed pipeline would be for Test 1 and Test 3 to yield positive results (accepting the null hypothesis), and for Test 2 to yield a negative result (rejecting the null hypothesis). However, an alternate good approximation would be when Test 1 and Test 2 yield negative results (rejecting the null hypothesis) and Test 3 yields a positive result (accepting the null hypothesis).

Model	Dataset		T1	T2	T3
Stylegan2-ADA	NASA Perseverance		x	✓	x
	AFHQ	Cat	x	x	✓
		Dog	x	✓	x
		Wild	x	x	✓
	MetFaces		x	x	x
r-Stylegan3-ADA			x	x	x
t-Stylegan3-ADA			x	x	x

For the AFHQ Dog dataset, additional training is required as the null hypothesis for T2 (normality of the synthetic distribution) is accepted, indicating that the learned distribution is close to the original white noise. A similar conclusion applies to the model trained on the NASA Perseverance dataset, which also needs further training. In the case of MetFaces, the learned distribution is considerably different from the original white noise, but the null hypothesis for goodness of fit is not accepted. This finding suggests several possible interpretations: the model may be overfitting, it might require increased capacity to represent all features of the original distribution, or additional training might be needed.

We have introduced statistical measures and a visualization pipeline to examine and comprehend the data at hand. Nevertheless, the high-dimensional nature

of images, coupled with the sequential aspect of video streams, brings forth a sense of time and space that our current analysis does not accommodate. In fact, the data comprise a series of images captured over a linear time span, following a specific trajectory. To address this aspect, we will employ tools from harmonic analysis in the subsequent section to offer a more comprehensive interpretation.

5.6.2 RMSE and MAE Signature and Log-Signature

The Signature Transform [10, 72, 73, 74, 75] is roughly equivalent to Fourier; instead of extracting information about frequency, it extracts information about order and area.

However, the Signature Transform differs from Fourier by the fact that it utilizes a basis of the space of functions of paths, a more general case to the basis of space of paths found in the preceding.

Following [10], the truncated signature of order N of the path $\mathbf{x}$ is defined as a collection of coordinate iterated integrals

$$\mathrm{S}^N(\mathbf{x}) = \left(\left(\int\limits_{0<t_1<\cdots<t_a<1} \prod_{c=1}^{a} \frac{\mathrm{d}f_{z_c}}{\mathrm{d}t}(t_c)\mathrm{d}t_1 \cdots \mathrm{d}t_a \right)_{1\le z_1,\ldots,z_a\le d} \right)_{1\le a\le N}. \tag{5.14}$$

The Signature is a homomorphism from the monoid of paths into the grouplike elements of a closed tensor algebra; see Equation (5.16). It provides a graduated summary of the path $\mathbf{x}$. These extracted features of a path are at the center of the definition of a rough path [13]; they remove the necessity to take into account the inner detailed structure of the path.

$$\mathrm{S} : \left\{ f \in F \mid f : [x, y] \to E = \mathbb{R}^d \right\} \longrightarrow T(E), \tag{5.15}$$

$$\text{where } T(E) = T(\mathbb{R}^d) = \prod_{c=0}^{\infty} \left(\mathbb{R}^d\right)^{\otimes c}. \tag{5.16}$$

It has many advantages over other tools of harmonic analysis for ML. It is a universal non-linearity, which means that every continuous function of the input stream may be approximated arbitrarily by a linear function of its signa-

ture. Furthermore, among other properties, it presents outstanding robustness behavior to missing or irregularly sampled data, along with optional invariance in terms of translation and sampling. It has recently been introduced in the context of DL to add some structure to the learning process, and it seems a promising tool in Generative Models and Reinforcement Learning, as well as a good theoretical framework. It mainly works on streams of data which could describe from video sequences to our entire life experiences. That is to say, under the correct assumptions and the right application, it could potentially compress all human experiences into a representation that could be stored and processed efficiently. Here, we propose to conduct a preliminary study in terms of harmonic analysis and understand its properties to compare the original and synthetic samples.

The Signature [13, 76, 77, 78, 79] of an input data stream encodes the order in which data arrive without being concerned with the precise timing of its arrival. This property, known as invariance to time reparameterizations [80], makes it an ideal candidate for measuring GAN-generated distributions against an original data stream. Notably, when sampling the GAN model, instances of the latent space are retrieved in no specific order, even though the original data are inherently time-dependent, as recorded video streams or images captured by sensors are constrained by the temporal nature of the physical world. However, GANs are not yet capable of generating data linearly in time and space, making comparisons using other methods potentially biased or unable to capture all relevant cues.

Furthermore, it is essential to note that the number of components in the truncated signature does not depend on the number of data samples under consideration. Specifically, it maps the infinite-dimensional space of data streams, $\mathcal{S}(\mathbb{R}^d)$, into a finite-dimensional space of dimension $(d^{N+1}-1)/(d-1)$, where N corresponds to the order of the truncated signature. This characteristic makes the Signature Transform highly suitable for processing long sequential data with varying lengths or unevenly sampled data.

At the same time, we can introduce the concept of Log-Signature [74, 75], which is a more compact representation than the Signature.

Definition 1 *If $\gamma_t \in E$ is a path segment and S is its Signature, then*

$$\begin{aligned} S &= 1 + S^1 + S^2 + \dots \ \forall c,\ S^c \in E^{\otimes c}, \\ \log(1+x) &= x - x^2/2 + \dots, \\ \log S &= \left(S^1 + S^2 + \dots\right) - \left(S^1 + S^2 + \dots\right)^2/2 + \dots \end{aligned}$$

The series $\log S = (S^1 + S^2 + \ldots) - (S^1 + S^2 + \ldots)^2/2 + \ldots$ *which is well-defined, is referred to as the Log-Signature of* γ.

In practice, the Log-Signature calculation involves a series expansion that is typically truncated at a certain level to obtain a finite-dimensional representation. The choice of the truncation level depends on the specific application and the desired trade-off between computational complexity and the level of detail captured by the Log-Signature. In our experiments, we have chosen a truncation level that balances these considerations and yields satisfactory performance for our GAN evaluation task.

Unlike the Signature, the Log-Signature does not guarantee universality [13], and as a result, it needs to be combined with non-linear models for learning. However, it is empirically more robust to sparsely sampled data. There is a one-to-one correspondence between the Signature and the Log-Signature, as the logarithm map is bijective [12, 74]. This statement also holds true for the truncated case up to the same degree.

In this study, we perform a comparison of the mean signature and Log-Signature for original and synthetic samples at a size of 64×64. We observe that synthetic samples encompass the most relevant information from the original harmonic distribution. We compare against sets of 1000 and 5000 synthetic samples, with each instance considered a path $\mathbf{x}$ of dimension 64 to which we apply the Signature and Log-Signature transforms.

We propose the use of the element-wise mean of the truncated signatures $\tilde{S}^N$, depicted in Figure 5.2, to analyze the convergence of GAN-learned models by employing RMSE (Root Mean Squared Error) and MAE (Mean Absolute Error). We refer to these measures as RMSE and MAE Signature, and RMSE and MAE Log-Signature. For instance, in Figure 5.3, we can observe that the model is achieving good convergence, though it is not capturing all the information present in the original distribution.

RMSE and MAE, when understood through the element-wise mean, can be considered as score functions built upon the Signature Transform, capable of measuring the quality of the generated distribution. This perspective on these measures is important for future applications, as it allows for the possibility of generalizing them to other tasks or even applying them to other transforms. RMSE and MAE Signature and Log-Signature can serve multiple purposes, such as comparing models, monitoring performance during training across several epochs, and analytically detecting overfitting, as demonstrated in Table 5.3, whereas all these measures capture information about the visual

cues present in the distributions, RMSE and MAE Signature, as well as MAE Log-Signature, prove to be more accurate in tracking the convergence of the GAN training procedure. In contrast, the RMSE Log-Signature exhibits less precision.

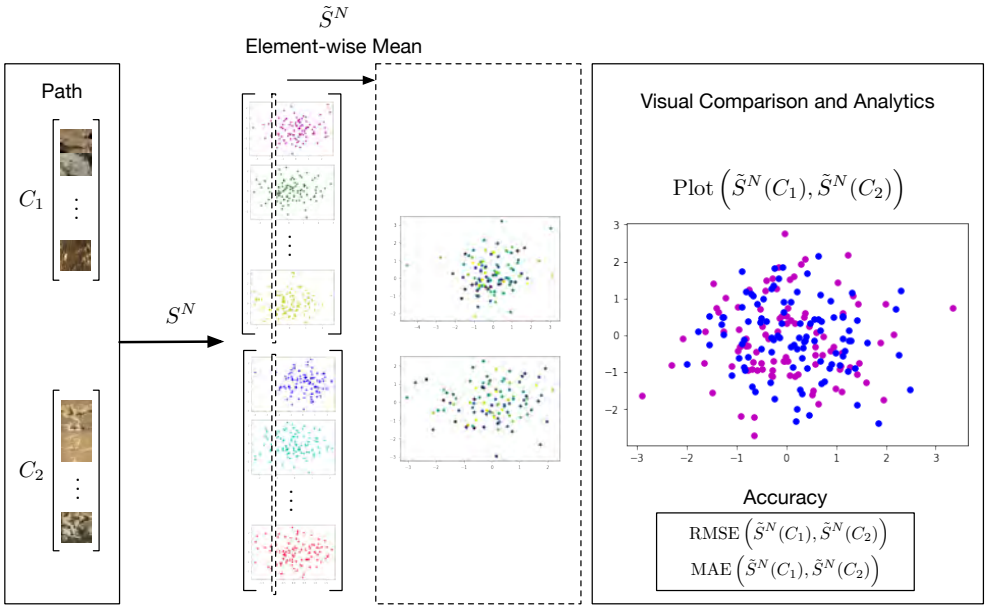


Figure 5.2: Visual explanation of the use of $\tilde{S}^N$ to analyze GAN convergence. Samples are resized at 64×64 and transformed to grayscale previous to the computation of the signatures. The procedure used for Log-Signature $\log \tilde{S}^N$ is analogous. In the rightmost side plot, each color represents a pair of functions: the violet curve illustrates one element-wise mean spectrum, while the blue curve represents the other element-wise mean spectrum. The difference between these two functions is quantified using RMSE or MAE.

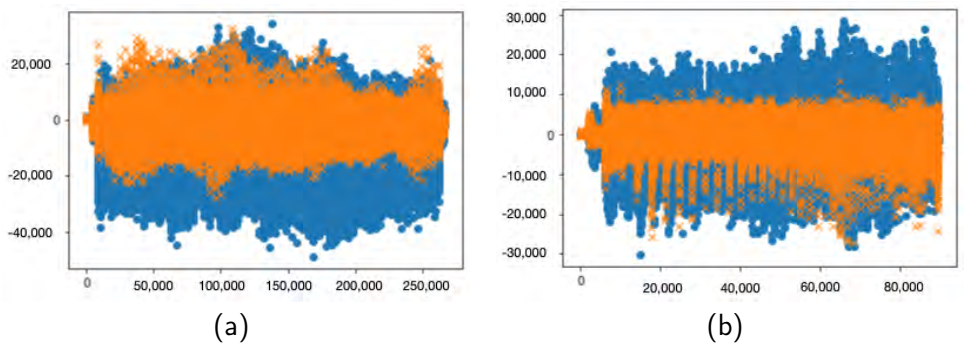


Figure 5.3: Spectrum of the element-wise mean of the Signatures (**a**) and Log-Signatures (**b**) of order 3 and size 64×64 of original ('o' in blue) against synthetic ('x' in orange) samples.

In Table 5.3, we present the RMSE and MAE Signature and Log-Signature values for different iterations of Stylegan2-ADA training. These values are calculated to evaluate the performance of the GAN at various stages of its training process. A closer examination of the table reveals that the 798th iteration of Stylegan2-ADA achieves the lowest RMSE and MAE Signature and Log-Signature values, which indicates the highest accuracy among the listed iterations. This table demonstrates the utility of RMSE and MAE Signature and Log-Signature metrics in tracking the progress of GAN training and identifying the optimal model iteration. By comparing the values across different iterations, we can observe the improvements in GAN performance as it learns to generate more realistic images. Furthermore, the table showcases the effectiveness of our proposed metrics in detecting potential overfitting, as evidenced by the increased RMSE and MAE values in the 983rd iteration. This increase in values suggests a decline in the GAN's performance, likely due to overfitting the training data. In summary, Table 5.3 highlights the value of our proposed RMSE and MAE Signature and Log-Signature metrics in evaluating GAN performance, enabling us to monitor progress, compare different models, and detect overfitting during the training process.

Table 5.3: RMSE and MAE Signature and Log-Signature across several iterations of training of Stylegan2-ADA (lower is better, being the best results highlighted in bold). Our synthetic samples are generated using the model 798 which achieves the highest accuracy on RMSE and MAE Signature and Log-Signature.

Iteration Stylegan2-ADA	**193**	**371**	**596**	**798**	**983**
RMSE Signature	15,617	13,336	12,353	**11,601**	25,699
MAE Signature	11,072	10,686	9801	**9086**	19,481
RMSE Log-Signature	9882	7563	**7354**	7397	15,621
MAE Log-Signature	6467	5955	5724	**5717**	12,063

To provide a more comprehensive understanding of the concepts presented in this section, we will analytically describe the abstraction of a set of images as an unevenly sampled stream of data, for example, a path, and present the definitions used to measure the similarity between image distributions.

A stream of data, $\mathbf{x} \in \mathcal{S}(\mathbb{R}^d)$, can be understood as a discrete representation of a path.

Definition 2 *Let* $\mathbf{x} = (x_1, \ldots, x_n) \in \mathcal{S}(\mathbb{R}^d)$ *be a stream of data. Let* X *be a linear interpolation of* $\mathbf{x}$*. Then the signature of* $\mathbf{x}$ *is defined as*

$$\mathrm{S}(\mathbf{x}) = \mathrm{S}(X), \tag{5.17}$$

and the truncated signature of order N *of* $\mathbf{x}$ *is defined as*

$$\mathrm{S}^N(\mathbf{x}) = \mathrm{S}^N(X). \tag{5.18}$$

This definition of the signature of a stream of data is independent of the choice of linear interpolation of X by the invariance to time reparameterizations [10].

Definition 3 *Given a set of truncated signatures of order N, $\{\mathrm{S}_c^N(\mathbf{x}_c)\}_{c=1}^m$, the element-wise mean is defined by*

$$\tilde{\mathrm{S}}^N(x^{(z)}) = \frac{1}{m}\sum_{c=1}^{m}\mathrm{S}_c^N(x_c^{(z)}), \tag{5.19}$$

where $z \in \{1,\dots,n\}$ is the specific component index of the given signature.

Then, RMSE and MAE Signature, whose results are presented in Tables 5.3 and 5.4, can be defined as follows.

Table 5.4: RMSE and MAE Signature and Log-Signature evaluation and comparison on AFHQ and MetFaces using state-of-the-art pretrained models of Stylegan2-ADA [69] and Stylegan3-ADA [42]. Lower is better, being the best results highlighted in bold.

Model	Dataset		RMSE $\tilde{S}^3$	MAE $\tilde{S}^3$	RMSE log $\tilde{S}^3$	MAE log $\tilde{S}^3$
Stylegan2-ADA	AFHQ	Cat	61,450	45,968	29,201	22,297
		Dog	38,861	30,441	31,686	24,612
		Wild	33,306	25,578	26,622	20,359
	MetFaces		33,247	23,428	25,685	18,071
r-Stylegan3-ADA			34,977	22,799	24,707	16,539
t-Stylegan3-ADA			**30,894**	**19,872**	**21,560**	**13,761**

Definition 4 *Given n components of the element-wise mean of the signatures $\{\tilde{y}^{(c)}\}_{c=1}^n \subseteq T(\mathbb{R}^d)$ from the model chosen as a source of synthetic samples and the same number of components of the element-wise mean of the signatures $\{\tilde{x}^{(c)}\}_{c=1}^n \subseteq T(\mathbb{R}^d)$ from the original distribution, we define the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) by*

$$\mathrm{RMSE}\left(\left\{\tilde{x}^{(c)}\right\}_{c=1}^n, \left\{\tilde{y}^{(c)}\right\}_{c=1}^n\right) = \sqrt{\frac{1}{n}\sum_{c=1}^{n}(\tilde{y}^{(c)} - \tilde{x}^{(c)})^2}, \tag{5.20}$$

and

$$\mathrm{MAE}\left(\left\{\tilde{x}^{(c)}\right\}_{c=1}^n, \left\{\tilde{y}^{(c)}\right\}_{c=1}^n\right) = \frac{1}{n}\sum_{c=1}^{n}|\tilde{y}^{(c)} - \tilde{x}^{(c)}|. \tag{5.21}$$

The case for Log-Signature is analogous.

5.7 Evaluation

We present the results of our proposed measures using several state-of-the-art pretrained models in Table 5.4. For evaluation and testing, we use the standard AFHQ dataset [71] classes 'cat', 'dog', and 'wild', as well as MetFaces [69], in conjunction with their corresponding pretrained models. To compute the RMSE and MAE for $\tilde{S}^N$ and $\log \tilde{S}^N$, we generate 1000 synthetic samples from each model and compare them against the full original dataset. Prior to the Signature Transform, the samples are converted to grayscale and resized to 64×64. Figures 5.4 and 5.5 provide a visual comparison of the spectrum, demonstrating that the trained models effectively learn the empirical distribution of the original data.

The AFHQ dataset comprises high-quality images of animal faces, which are divided into three distinct classes: cats, dogs, and wildlife. This dataset provides a challenging evaluation scenario due to the inherent differences between the classes and the detailed textures present in the animal faces. MetFaces, on the other hand, is a collection of face images derived from various art pieces, including paintings, photographs, and sculptures. It showcases a diverse range of artistic styles, time periods, and image content, making it an ideal dataset to assess the performance of our proposed metrics on more complex and varied data distributions. By evaluating our method on both AFHQ and MetFaces, we aim to demonstrate the adaptability and robustness of our approach across different scenarios and data complexities.

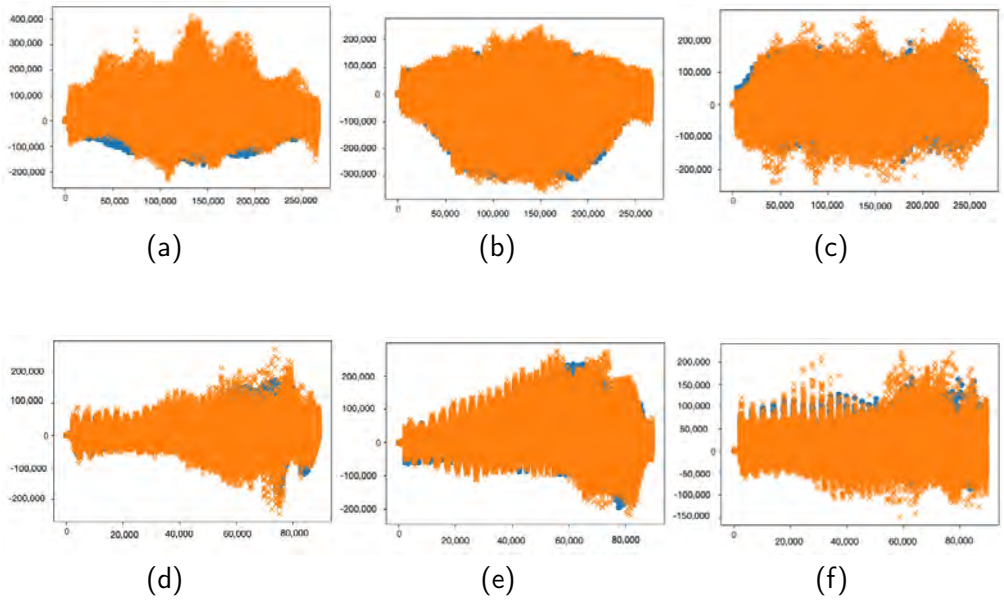


Figure 5.4: Spectrum comparison of the element-wise mean of the Signatures $\tilde{S}^3$ (top) and Log-Signatures $\log \tilde{S}^3$ (bottom) of order 3 and size 64×64 of original ('o' in blue) against synthetic ('x' in orange) samples. (**a**,**d**): AFHQcat, (**b**,**e**): AFHQdog, (**c**,**f**): AFHQwild.

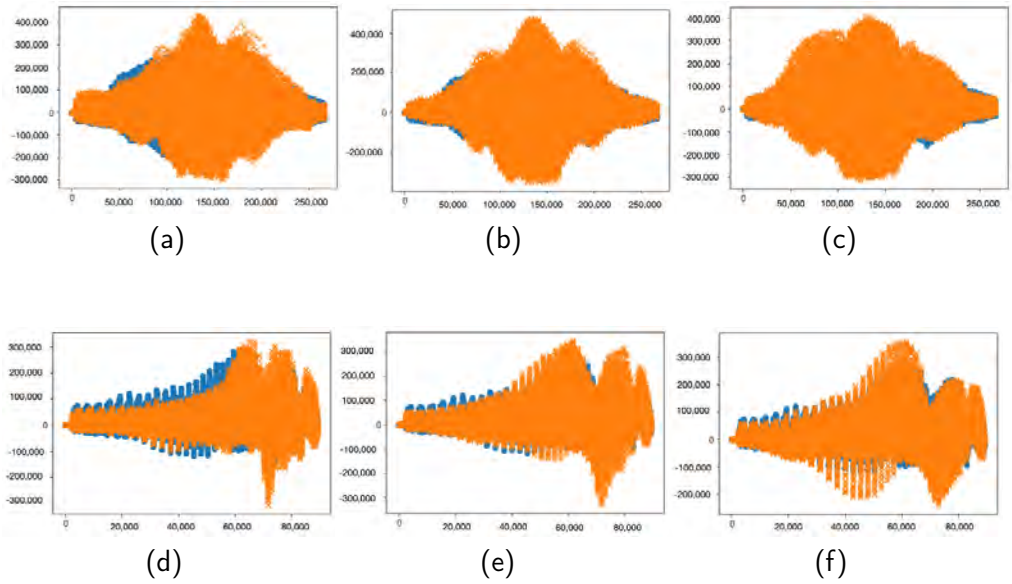


Figure 5.5: Spectrum comparison of the element-wise mean of the Signatures $\tilde{S}^3$ (top) and Log-Signatures $\log \tilde{S}^3$ (bottom) of order 3 and size 64×64 of original ('o' in blue) against synthetic ('x' in orange) samples from MetFaces. (**a**,**d**): Stylegan2-ADA, (**b**,**e**): r-Stylegan3-ADA, (**c**,**f**): t-Stylegan3-ADA.

In Table 5.4, we compare the recently developed models r, t-Stylegan3-ADA [42] against Stylegan2-ADA using MetFaces. We observe that t-Stylegan3-ADA significantly outperforms Stylegan2-ADA and r-Stylegan3-ADA, which is consistent with the FID results reported in [42], as shown in Table 5.5. Here, we can see that FID closely resembles the behavior of RMSE $\tilde{S}^3$. Nonetheless, our metrics are both effective and efficient. A visual comparison of the spectrum of the Signatures for the given dataset can be seen in Figure 5.5. Computation is performed on the CPU in seconds, which is orders of magnitude faster and requires fewer resources than FID or MS-SSIM.

Table 5.5: Evaluation and comparison of FID (as reported in [42]) and RMSE $\tilde{S}^3$ on MetFaces. Lower is better, being the best results highlighted in bold.

Model	**FID**	**RMSE $\tilde{S}^3$**
Stylegan2-ADA	15.22	33,247
r-Stylegan3-ADA	15.33	34,977
t-Stylegan3-ADA	**15.11**	**30,894**

Table 5.4 presents the RMSE and MAE Signature and Log-Signature evaluation results for different GAN models and datasets, including AFHQ and MetFaces. The table showcases a comparison of state-of-the-art pretrained models: Stylegan2-ADA [69], r-Stylegan3-ADA, and t-Stylegan3-ADA [42]. The goal of this comparison is to highlight the performance differences between these models using the proposed metrics. A close inspection of the table reveals that the t-Stylegan3-ADA model consistently achieves the lowest RMSE and MAE Signature and Log-Signature values across all datasets, indicating superior performance in generating synthetic samples that closely resemble the original distributions. This result demonstrates the effectiveness of the t-Stylegan3-ADA model in learning the intricacies of the underlying data distributions and generating high-quality synthetic samples. Additionally, the table illustrates the performance variations between different categories within the AFHQ dataset, with AFHQ Cat and Wild categories having a closer resemblance to the original distributions than AFHQ Dog. This observation aligns with the qualitative assessment of the generated samples visualized in Figures 5.6 and 5.7, providing further evidence of the accuracy of our proposed metrics in capturing the characteristics of the generated samples. That is, Table 5.4 highlights the utility of the RMSE and MAE Signature and Log-Signature metrics in evaluating and comparing the performance of different GAN models across various datasets. By analyzing these metrics, we can gain insights into the quality of the generated samples and their similarity to the original distributions, as well as assess the effectiveness of the GAN models in capturing the essential features of the data.

Table 5.5 presents a comparison of the FID and RMSE $\tilde{S}^3$ metrics on MetFaces for three GAN models: Stylegan2-ADA, r-Stylegan3-ADA, and t-Stylegan3-ADA. The aim of this comparison is to highlight the relationship between the two evaluation metrics and demonstrate the efficacy of RMSE $\tilde{S}^3$ in capturing the performance differences among these models. As observed in the table, the

FID scores and RMSE $\tilde{S}^3$ values show a similar trend, with t-Stylegan3-ADA achieving the best performance in both metrics. This consistency between the two evaluation metrics suggests that RMSE $\tilde{S}^3$ can serve as a reliable alternative to FID in assessing GAN performance. Moreover, the lower RMSE $\tilde{S}^3$ values for t-Stylegan3-ADA indicate that the model generates synthetic samples that are closer to the original distribution compared to the other two models. Notably, the proposed RMSE $\tilde{S}^3$ metric offers significant advantages over FID in terms of computational efficiency and resource requirements. As mentioned in the text, the RMSE $\tilde{S}^3$ computations are performed on the CPU in seconds, making it substantially faster and less resource-intensive than FID or MS-SSIM. This efficiency makes the proposed metric more suitable for practical applications, where rapid evaluation and limited resources may be critical factors. In summary, Table 5.5 demonstrates the effectiveness of the RMSE $\tilde{S}^3$ metric as an alternative to FID for evaluating GAN performance. The strong correlation between the two metrics, coupled with the computational advantages of RMSE $\tilde{S}^3$, showcases its potential as a valuable tool for assessing the quality of synthetic samples generated by various GAN models.

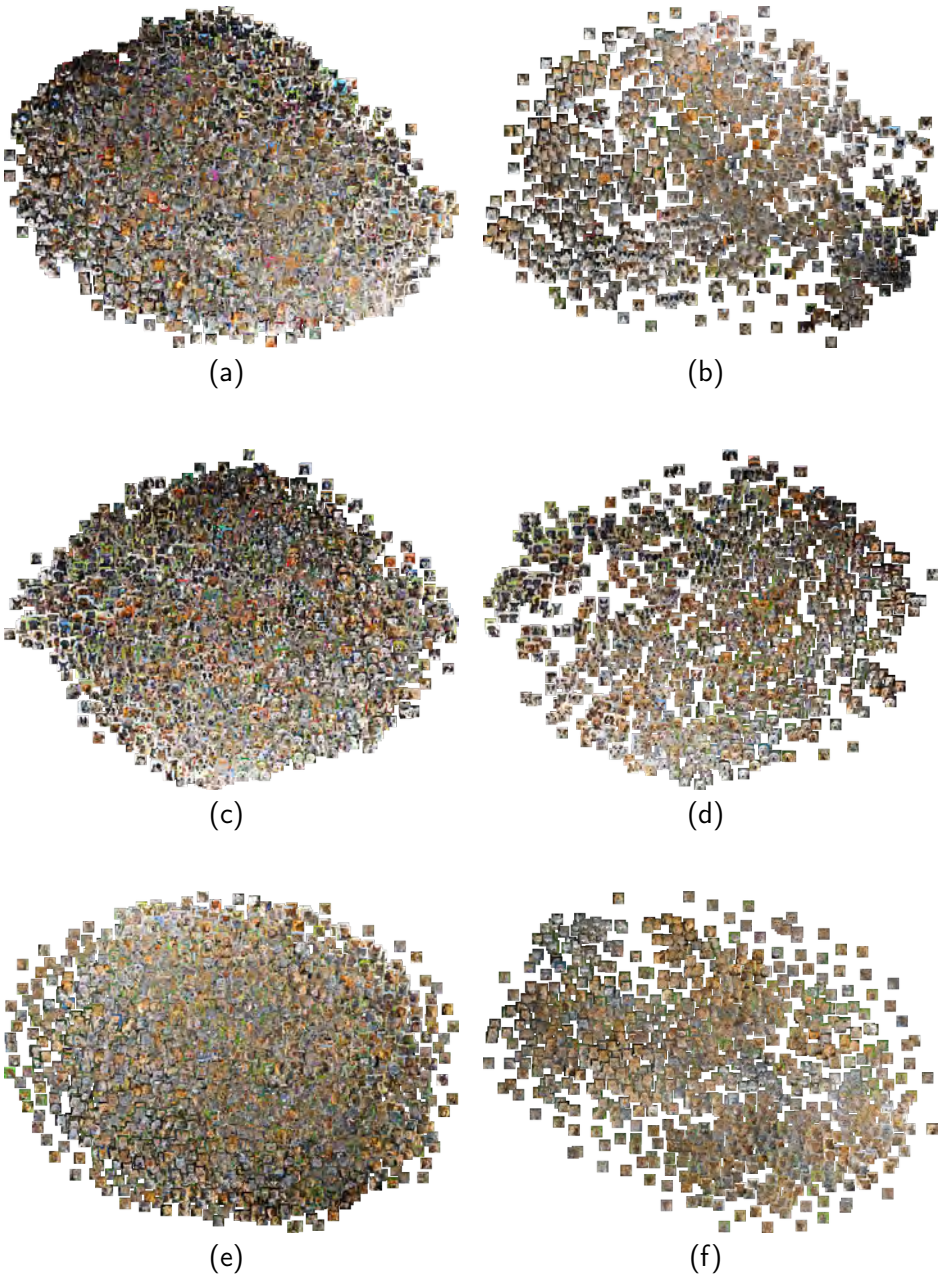


Figure 5.6: Visualization of PCA Adaptive t-SNE on original (**left**) versus synthetic (**right**) samples of AFHQ Cat (**a**,**b**), Dog (**c**,**d**), and Wild (**e**,**f**) using Stylegan2-ada.

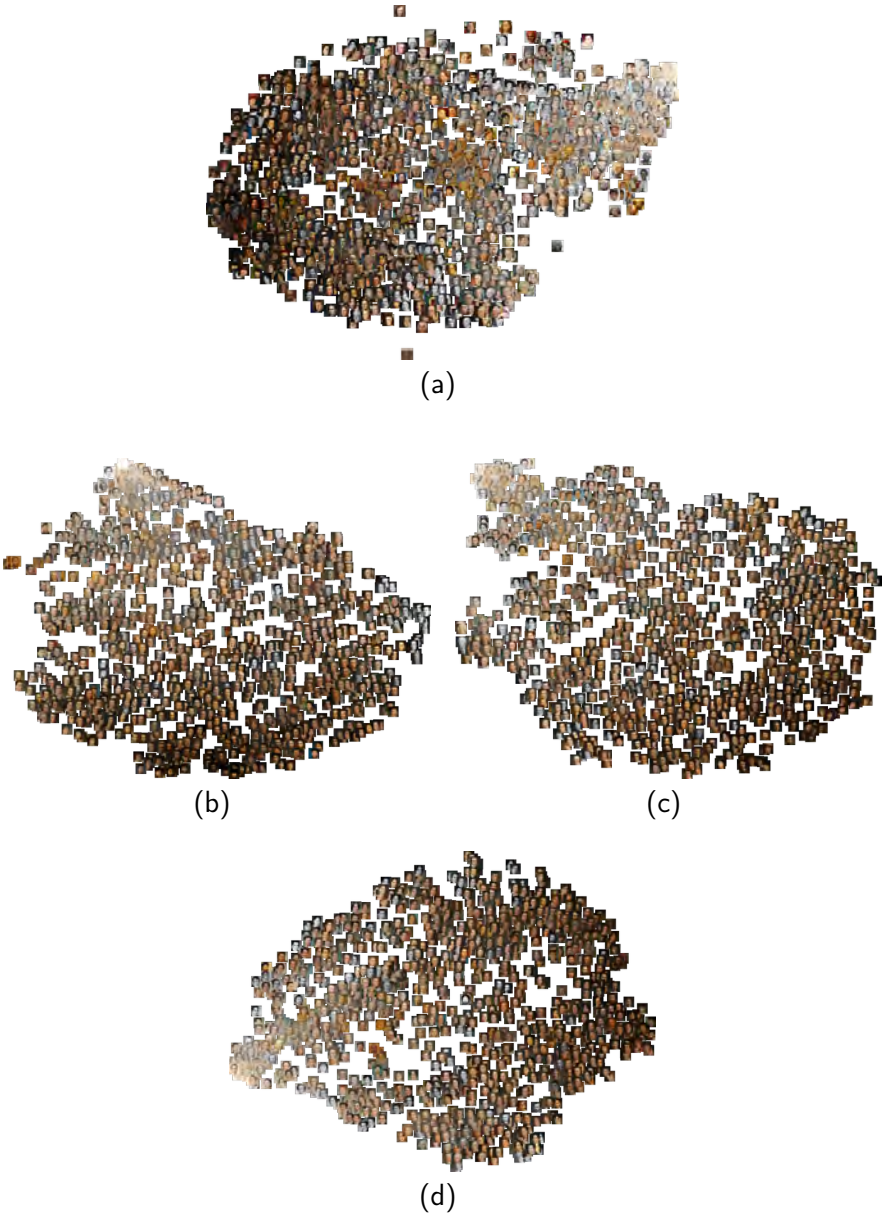


Figure 5.7: Visualization of PCA Adaptive t-SNE on original (**a**) versus synthetic (bottom) samples of MetFaces using Stylegan2-ADA (**b**), r-Stylegan3-ADA (**c**), and t-Stylegan3-ADA (**d**).

5.7.1 Computational Complexity

In this subsection, we elaborate on the computational complexity and time estimates for the element-wise mean of the Signatures and Kruskal–Wallis in comparison to FID, MS-SSIM, LPIPS, and PSNR.

1. Element-wise mean of the Signatures: The computation of the Signature Transform has a time complexity of $\mathcal{O}(LM^2)$, where L is the length of the path and M is the order of the signature. However, since we are computing the element-wise mean of the Signatures, the complexity becomes $\mathcal{O}(NLM^2)$, where N is the number of samples. In practice, the Signature Transform can be efficiently computed using recursive algorithms, which keeps the computational cost low.

2. Kruskal–Wallis has a time complexity of $\mathcal{O}(N \log N)$ for sorting the samples, followed by $\mathcal{O}(N)$ for computing the test statistic, resulting in an overall complexity of $\mathcal{O}(N \log N)$. This complexity is relatively low, especially when compared to more computationally demanding metrics such as FID and MS-SSIM.

In comparison:

1. The FID calculation involves computing the Inception features for each sample, which requires a forward pass through a deep neural network, followed by computing the mean and covariance of these features. The complexity of the forward pass depends on the architecture of the Inception network, but it is generally much higher than the complexity of the Signature Transform and the Kruskal–Wallis. Additionally, FID requires GPU resources to perform these calculations efficiently, further increasing its computational cost.

2. MS-SSIM involves computing the structural similarity index at multiple scales, which requires computing the mean, variance, and covariance for each scale. The complexity of MS-SSIM is $\mathcal{O}(NWH)$, where W and H are the width and height of the images, respectively, whereas this complexity is not as high as FID, it is still higher than the complexities of the proposed methods.

3. LPIPS metric computes the distance between image features extracted from a pretrained deep neural network (e.g., AlexNet or VGG). The complexity of LPIPS is primarily determined by the forward pass through the chosen deep neural network. The complexity of the forward pass depends

on the architecture of the network, but in general, it is higher than the complexity of the Signature Transform and the Kruskal–Wallis. Similar to FID, LPIPS also typically requires GPU resources for efficient computation.

4. PSNR is a simple and widely used metric for image quality assessment. It is computed as the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. The complexity of PSNR is $\mathcal{O}(NWH)$, where W and H are the width and height of the images, respectively. Although the complexity of PSNR is similar to that of MS-SSIM, it is still higher than the complexities of the proposed methods (element-wise mean of the Signatures and Kruskal–Wallis).

To summarize, our proposed methods (element-wise mean of the Signatures and Kruskal–Wallis) have significantly lower computational complexity than FID, MS-SSIM, LPIPS, and PSNR, allowing for faster computation and reduced resource usage. Based on the complexity analysis, we can estimate that our methods can be computed on the CPU in seconds, whereas FID, MS-SSIM, LPIPS, and PSNR require more time and resources, particularly when GPUs are not available.

5.7.2 Visualization

In our study, we first apply PCA to reduce the dimensionality of the data, which helps us to retain the global structure of the dataset. Then, we use t-SNE to visualize the data in a lower-dimensional space, which emphasizes the local differences between samples. This two-step approach allows us to capture both the global and local structures within the data, providing a richer visualization of the generated GAN images compared to using PCA alone.

In Figures 5.6 and 5.7, we visualize the sets of images of AFHQ and MetFaces, both original and synthetic, used in the evaluations in Tables 5.2 and 5.4 using PCA Adaptive t-SNE. The importance here is to observe the overall distribution of the samples, which is well captured by our proposed method. For instance, we can observe that the synthetic samples of AFHQ Cat and Wild closely resemble the original distribution in terms of variability and quality. In contrast, AFHQ Dog demonstrates less variability but still achieves high-quality samples, which aligns with the analytical interpretation of the proposed statistical measures shown in Table 5.2.

In Figure 5.7, we can observe that the synthetic samples generated with t-Stylegan3-ADA exhibit better quality than those produced by Stylegan2-ADA and r-Stylegan3-ADA, and the model is evidently learning the original distribution. Nonetheless, there is potential for improvement in terms of variability and scope. These observations are consistent with the RMSE and MAE Signature and Log-Signature results, as shown in Table 5.4.

That being said, our proposed method relies on the Signature Transform and Log-Signature to evaluate GAN-generated samples, which are independent of PCA and t-SNE. The use of PCA and t-SNE in our study is only to provide a visual representation of the original and synthetic distributions, allowing us to better understand and interpret the quality of the generated images. The sample size and analysis time of the generated GAN images are not affected by the application of PCA and t-SNE for visualization. Our methodology remains efficient and effective in assessing the quality of GAN-generated samples without the need to reduce the dimensionality of the images for the actual evaluation process.

5.8 Conclusions

GAN evaluation has been one of the central research efforts of the community of computer vision during these last years. The ability of the networks to generate high-fidelity samples has inspired researchers all over the world to work on the topic. However, although many variants of the original successful DCGAN architecture are able to generate very realistic samples, neither the advance in proposing metrics to assess the imagery has been effectual, nor the ability of the metrics to guarantee some level of robustness, and overall description of the resultant distribution. The best effort of them being FID suffers from high-computation time and use of GPU resources; it depends mainly on an inception module that extracts features from lots of samples rather than from analytical measures that quantify properly their characteristics.

We are the first to propose the use of the Signature Transform to assess GAN convergence by introducing RMSE and MAE Signature and Log-Signature. The measures are reliable, consistent, efficient, and easy to compute. Additionally, an effective methodology to test the goodness-of-fit according to the original distribution by the use of simple statistical methods is also proposed, being the first to be able to reduce the amount of computation for accurate GAN Synthetic image quality assessment to the order of seconds. Worth mentioning is the proposal of a taxonomical pipeline to systematically assess the

resultant distributions using a non-parametric test. Lastly, we also introduce an adaptive technique based on t-SNE and PCA that, without the need for hyperparameter tuning, puts forward exceptional visualization capabilities.

Future work that could be pursued under these assumptions, among others, is to increase the complexity of the descriptor, extend the proposed score functions on top of the Signature Transform to be used in other tasks or use the metrics inside the training loop to assess convergence and help the networks train faster.

In this study, we presented a novel approach to assess GAN convergence and goodness of fit using the Signature Transform, whereas our methodology provides significant advantages over existing methods, we acknowledge the following limitations:

1. The proposed RMSE and MAE Signature and Log-Signature metrics are based on the Signature Transform, which inherently captures information about the underlying distribution. However, these metrics may not be sensitive to certain aspects of the generated images, such as fine-grained details or specific structures, which could be essential for certain applications.

2. Although our proposed method significantly reduces computation time and resource usage compared to existing GAN evaluation methods, it might still be computationally expensive for extremely large datasets or high-resolution images. Further optimization of the computation process may be necessary to address these challenges.

3. The evaluation of GAN performance based on our proposed metrics assumes that the original and synthetic image distributions are stationary. In cases where the data exhibit non-stationary behavior, the effectiveness of our approach might be compromised, and additional methods or adaptations may be required.

4. The goodness-of-fit methodology proposed in this study relies on statistical methods, which might not always provide definitive conclusions on the quality of the generated samples. In some cases, additional qualitative assessments or domain-specific evaluations may be necessary to obtain a comprehensive understanding of the GAN's performance.

In conclusion, despite these limitations, our study introduces a promising and efficient approach to assess GAN convergence and goodness of fit using the Sig-

nature Transform. Future work may involve addressing these limitations and further exploring the potential of our proposed metrics in other applications and tasks.

Bibliography

[1] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.

[2] Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.

[3] Kang, M.; Zhu, J.-Y.; Zhang, R.; Park, J.; Shechtman, E.; Paris, S.; Park, T. Scaling up GANs for Text-to-Image Synthesis. *arXiv* **2023**, arXiv:2303.05511.

[4] Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 4217–4228. [CrossRef]

[5] Chan, E.R.; Lin, C.Z.; Chan, M.A.; Nagano, K.; Pan, B.; Mello, S.D.; Gallo, O.; Guibas, L.J.; Tremblay, J.; Khamis, S.; et al. Efficient geometry-aware 3D generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16123–16133.

[6] Brock, A.; Donahue, J.; Simonyan, K. Large scale gan training for high fidelity natural image synthesis. In Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2019.

[7] Zhao, L.; Zhang, Z.; Chen, T.; Metaxas, D.N.; Zhang, H. Improved transformer for high-resolution gans. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Virtual, 6–14 December 2021.

[8] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; Volume 30.

[9] Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Multiscale structural similarity for image quality assessment. In Proceedings of the Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Pacific Grove, CA, USA, 9–12 November 2003; Volume 2, pp. 1398–1402.

[10] Bonnier, P.; Kidger, P.; Arribas, I.P.; Salvi, C.; Lyons, T. Deep signature transforms. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Vancouver, BC, Canada, 8–14 December 2019.

[11] Chen, K.-T. Iterated path integrals. *Bull. Am. Math. Soc.* **1977**, *83*, 831–879. [CrossRef]

[12] Lyons, T.; Caruana, M.; Levin, T. *Differential Equations Driven by Rough Paths, Proceedings of the 34th Summer School on Probability Theory, Saint-Flour, France, 6–24 July 2004*; Springer: Berlin/Heidelberg, Germany, 2007.

[13] Lyons, T. Rough paths, signatures and the modelling of functions on streams. In Proceedings of the International Congress of Mathematicians, Madrid, Spain, 22–30 August 2014.

[14] van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *JMLR* **2008**, *9*, 2579–2605.

[15] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.

[16] He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings

of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.

[17] Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.

[18] Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]

[19] Chen, Q.; Koltun, V. Photographic image synthesis with cascaded refinement networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.

[20] Yang, B.; Luo, W.; Urtasun, R. PIXOR: Real-time 3D object detection from point clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.

[21] Parmar, N.; Vaswani, A.; Uszkoreit, J.; Kaiser, L.; Shazeer, N.; Ku, A.; Tran, D. Image transformer. In Proceedings of the 35th International Conference on Machine Learning, ICML, Stockholm, Sweden, 10–15 July 2018.

[22] Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. NeRF: Representing scenes as neural radiance fields for view synthesis. In Proceedings of the ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020.

[23] Park, T.; Efros, A.A.; Zhang, R.; Zhu, J. Contrastive learning for unpaired image-to-image translation. In Proceedings of the ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020.

[24] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.

[25] Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.

[26] Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Las Vegas, NV, USA, 27–30 June 2016.

[27] Gatys, L.A.; Bethge, M.; Hertzmann, A.; Shechtman, E. Preserving color in neural artistic style transfer. *arXiv* **2016**, arXiv:1606.05897.

[28] Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive growing of GANs for improved quality, stability, and variation. In Proceedings of the 6th ICLR International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.

[29] He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.

[30] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.

[31] Wang, T.; Liu, M.; Zhu, J.; Liu, G.; Tao, A.; Kautz, J.; Catanzaro, B. Video-to-video synthesis. In Proceedings of the Annual Conference on Neural Information Processing Systems, NeurIPS, Montréal, QC, Canada, 3–8 December 2018.

[32] de Zarzà, I.; de Curtò, J.; Calafate, C.T. Detection of glaucoma using three-stage training with EfficientNet. *Intell. Syst. Appl.* **2022**, *16*, 200140. [CrossRef]

[33] de Curtò, J.; de Zarzà, I.; Calafate, C.T. Semantic scene understanding with large language models on unmanned aerial vehicles. *Drones* **2023**, *7*, 114. [CrossRef]

[34] Dosovitskiy, A.; Brox, T. Generating Images with Perceptual Similarity Metrics Based on Deep Networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona Spain, 5–10 December 2017; pp. 658–666.

[35] Ratner, A.; Sa, C.D.; Wu, S.; Selsam, D.; Ré, C. Data Programming: Creating Large Training Sets, Quickly. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona Spain, 5–10 December 2017; pp. 3567–3575.

[36] Odena, A.; Olah, C.; Shlens, J. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the CML'17: 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August 2017.

[37] Antoniou, A.; Storkey, A.; Edwards, H. Data augmentation generative adversarial networks. In Proceedings of the 6th International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.

[38] Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; Chen, X. Improved techniques for training GANs. In Proceedings of the Annual Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016.

[39] Mescheder, L.; Nowozin, S.; Geiger, A. The numerics of GANs. In Proceedings of the Annual Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.

[40] Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? In Proceedings of the International Conference on Machine Learning PMLR, Beijing, China, 14–16 November 2018.

[41] Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. In Proceedings of the 7th International Conference on Learning Representations, ICLR, New Orleans, LA, USA, 6–9 May 2019.

[42] Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Virtual, 6–14 December 2021.

[43] Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2014**, arXiv:1312.6114.

[44] Zhao, J.; Mathieu, M.; LeCun, Y. Energy-based generative adversarial networks. In Proceedings of the 5th International Conference on Learning Representations ICLR, Toulon, France, 24–26 April 2017.

[45] Wei, X.; Gong, B.; Liu, Z.; Lu, W.; Wang, L. Improving the improved training of wasserstein gans: a consistency term and its dual effect. In Proceedings of the 6th International Conference on Learning Representations ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.

[46] Arora, S.; Ge, R.; Liang, Y.; Ma, T.; Zhang, Y. Generalization and Equilibrium in Generative Adversarial Nets (GANs). In Proceedings of the 34th International Conference on Machine Learning, PMLR, Seoul, Republic of Korea, 15–17 November 2017; pp. 224–232.

[47] Pascanu, R.; Mikolov, T.; Bengio, Y. On the Difficulty of Training Recurrent Neural Networks. In Proceedings of the 30th International Conference on Machine Learning, Atlanta GA, USA, 16–21 June 2013; pp. 1310–1318.

[48] Flynn, J.; Neulander, I.; Philbin, J.; Snavely, N. DeepStereo: Learning to Predict New Views from the World's Imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5515–5524.

[49] Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Vancouver, BC, Canada, 8–14 December 2019.

[50] Roberts, G.O.; Tweedie, R.L. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* **1996**, *2*, 341–363. [CrossRef]

[51] Welling, M.; Teh, Y.W. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning ICML, Bellevue, WA, USA, 28 June–2 July 2011.

[52] Song, Y.; Ermon, S. Improved techniques for training score-based generative models. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Virtual, 6–12 December 2020.

[53] Goyal, A.; Ke, N.R.; Ganguli, S.; Bengio, Y. Variational walkback: Learning a transition operator as a stochastic recurrent net. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Long Beach, CA, USA, 4–9 December 2017.

[54] Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Virtual, 6–12 December 2020.

[55] Jolicoeur-Martineau, A.; Piché-Taillefer, R.; Combes, R.T.; Mitliagkas, I. Adversarial score matching and improved sampling for image generation. In Proceedings of the International Conference on Learning Representations ICLR, Vienna, Austria, 4 May 2021.

[56] Zhao, Z.; Kunar, A.; Birke, R.; Chen, L.Y. Ctab-gan: Effective table data synthesizing. In Proceedings of the Machine Learning in Computational Biology Meeting, PMLR, Online, 22-23 November 2021; pp. 97–112.

[57] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations ICLR, Addis Ababa, Ethiopia, 26–30 April 2020.

[58] Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.

[59] Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.

[60] Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep image prior. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9446–9454.

[61] Sohl-Dickstein, J.; Weiss, E.A.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the 7th Asian Conference on Machine Learning, Hong Kong, 20-22 November 2015; pp. 2256–2265.

[62] Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371–3408.

[63] Ho, J.; Saharia, C.; Chan, W.; Fleet, D.J.; Norouzi, M.; Salimans, T. Cascaded Diffusion Models for High Fidelity Image Generation. *J. Mach. Learn. Res.* **2022**, *23*, 1–33.

[64] Luo, Z.; Chen, D.; Zhang, Y.; Huang, Y.; Wang, L.; Shen, Y.; Zhao, D.; Zhou, J.; Tan, T. VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 20–22 June 2023.

[65] Wu, J.Z.; Ge, Y.; Wang, X.; Lei, S.W.; Gu, Y.; Hsu, W.; Shan, Y.; Qie, X.; Shou, M.Z. Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation. *arXiv* **2022**, arXiv:2212.11565.

[66] Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv* **2022**, arXiv:2208.12242.

[67] Hua, T.; Tian, Y.; Ren, S.; Zhao, H.; Sigal, L. Self-supervision through random segments with autoregressive coding (randsac). *arXiv* **2022**, arXiv:2203.12054.

[68] Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595.

[69] Karras, T.; Aittala, M.; Hellsten, J.; Laine, S.; Lehtinen, J.; Aila, T. Training generative adversarial networks with limited data. In Proceedings of the Annual Conference on Neural Information Processing Systems NIPS, Virtual, 6–12 December 2020.

[70] Kruskal, W.H.; Wallis, W.A. Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **1952**, *47*, 583–621. [CrossRef]

[71] Choi, Y.; Uh, Y.; Yoo, J.; Ha, J. Stargan v2: Diverse image synthesis for multiple domains. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.

[72] Kidger, P.; Lyons, T. Signatory: Differentiable computations of the signature and logsignature transforms, on both CPU and GPU. In Proceedings of the International Conference on Learning Representations ICLR, Vienna, Austria, 4 May 2021.

[73] Chevyrev, I.; Kormilitzin, A. A primer on the signature method in machine learning. *arXiv* **2016**, arXiv:1603.03788.

[74] Liao, S.; Lyons, T.J.; Yang, W.; Ni, H. Learning stochastic differential equations using RNN with log signature features. *arXiv* **2019**, arXiv:1908.08286.

[75] Morrill, J.; Kidger, P.; Salvi, C.; Foster, J.; Lyons, T.J. Neural CDEs for long time series via the log-ode method. In Proceedings of the 38th International Conference on Machine Learning, ICML, Virtual, 18–24 July 2021.

[76] Kiraly, F.J.; Oberhauser, H. Kernels for sequentially ordered data. *J. Mach. Learn. Res.* **2019**, *20*, 1–45.

[77] Graham, B. Sparse arrays of signatures for online character recognition. *arXiv* **2013**, arXiv:1308.0371.

[78] Chang, J.; Lyons, T. Insertion algorithm for inverting the signature of a path. *arXiv* **2019**, arXiv:1907.08423.

[79] Fermanian, A. Learning Time-Dependent Data with the Signature Transform. Ph.D. Thesis, Sorbonne Université, Paris, France, 2021. Available online: `https://tel.archives-ouvertes.fr/tel-03507274` (accessed on 1 November 2022).

[80] Lyons, T. Differential equations driven by rough signals. *Rev. Mat. Iberoam.* **1998**, *14*, 215–310. [CrossRef]

Chapter 6

Summarization of Videos with the Signature Transform

J. de Curtò, I. de Zarzà, Gemma Roig and Carlos T. Calafate. (2023). "Summarization of Videos with the Signature Transform." Electronics, vol(12), 1735. DOI: 10.3390/electronics12071735

This chapter presents a new benchmark for assessing the quality of visual summaries without the need for human annotators. It is based on the Signature Transform, specifically focusing on the RMSE and the MAE Signature and Log-Signature metrics, and builds upon the assumption that uniform random sampling can offer accurate summarization capabilities. We provide a new dataset comprising videos from Youtube and their corresponding automatic audio transcriptions. Firstly, we introduce a preliminary baseline for automatic video summarization, which has at its core a Vision Transformer, an image–text model pre-trained with Contrastive Language–Image Pre-training (CLIP), as well as a module of object detection. Following that, we propose an accurate technique grounded in the harmonic components captured by the Signature Transform, which delivers compelling accuracy. The analytical measures are extensively evaluated, and we conclude that they strongly correlate with the notion of a good summary.

6.1 Introduction and Problem Statement

Video data have become ubiquitous, from content creation to the animation industry. The ability to summarize the information present in large quantities of data is a central problem in many applications, particularly when there is a need to reduce the amount of information transmitted and to swiftly assimilate visual contents. Video summarization [1, 3, 6, 2, 4, 5, 7] has been extensively studied in Computer Vision, using both handcrafted methods [8] and learning techniques [10, 9]. These approaches traditionally use feature extraction on keyframes to formulate an adequate summary.

Recent advances in Deep Neural Networks (DNN) [12, 13, 11] have spurred progress across various scientific fields [17, 18, 16, 19, 14, 15]. In the realm of video summarization, two prominent approaches have emerged: LSTM- and RNN-based models [21, 20, 22]. These models have demonstrated considerable success in developing effective systems for video summarization. Additionally, numerous other learning techniques have been employed to address this challenge [26, 25, 24, 23].

In this study, we introduce a novel concatenation of models for video summarization, capitalizing on advancements in Visual Language Models (VLM) [27, 28]. Our approach combines zero-shot text-conditioned object detection with automatic text video annotations, resulting in an initial summarization method that captures the most critical information within the visual sequence.

Metrics to assess the performance of such techniques have usually relied on a human in the loop, using services such as Amazon Mechanical Turk (AMT) to provide annotated summaries for comparison. There have been attempts to introduce quantitative measures to address this problem, the most common being the F1-score, but these measures need human annotators and have shown that many state-of-the-art methodologies perform worse than mere uniform random sampling [29].

However, in this work, we go beyond the current state of the art and introduce a set of metrics based on the Signature Transform [31], a rough equivalent to the Fourier Transform that takes order and area into account and that contrasts the spectrum of the original video with the spectrum of the generated summary to provide a measurable score. We then propose an accurate state-of-the-art baseline based on the Signature Transform to accomplish the task. Thorough evaluations are provided, where we can see that the methodologies provide accurate video summaries, and that the technique based on the Signature Transform achieves summarization capabilities superior to the state of the art.

Indeed, the temporal content present in a video timeline makes the Signature Transform an ideal candidate to assess the quality of generated summaries where a video stream is treated as a path.

Section 6.2 gives a primer on the Signature Transform to bring forth in Section 6.2.1 a set of metrics to assess the quality of visual summaries by considering the harmonic components of the signal. The metrics are then used to put forward an accurate baseline for video summarization in Section 6.2.2. In the following section, we introduce the concept of Foundation Models, which serves to propose a preliminary technique for the summarization of videos. Thorough experiments are conducted in Section 6.4, with emphasis on the newly introduced dataset and the set of measures. Section 6.4.1 gives an assessment of the metrics in comparison to human annotators, whereas Section 6.4.2 evaluates the performance of the baselines based on the Signature Transform against another technique. Finally, Section 6.5 delivers conclusions, addresses the limitations of the methodology, and discusses further work.

6.2 Signature Transform

The Signature Transform [32, 34, 33, 35, 36] is roughly equivalent to the Fourier Transform; instead of extracting information concerning frequency, it extracts information about the order and area. However, the Signature Transform differs from the Fourier Transform in that it utilizes the space of functions of paths, a more general case than the basis of the space of paths found in the Fourier Transform.

Following the work in [32], the truncated signature of order N of the path $\mathbf{x}$ is defined as a collection of coordinate iterated integrals

$$\mathrm{S}^N(\mathbf{x}) = \left(\left(\int \cdots \int_{0<t_1<\cdots<t_a<1} \prod_{c=1}^{a} \frac{\mathrm{d}f_{z_c}}{\mathrm{d}t}(t_c)\mathrm{d}t_1 \cdots \mathrm{d}t_a \right)_{1\le z_1,\ldots,z_a\le d} \right)_{1\le a\le N} . \tag{6.1}$$

Here, $\mathbf{x} = (x_1, \ldots, x_n)$, where $x_z \in \mathbb{R}^d$. Let $f = (f_1, \ldots, f_d)\colon [0,1] \to \mathbb{R}^d$ be continuous, such that $f(\frac{z-1}{n-1}) = x_z$, and linear in the intervals in between.

6.2.1 RMSE and MAE Signature and Log-Signature

The F1-score between a summary and the ground truth of annotated data has been the widely accepted measure of choice for the task of video summarization. However, recent approaches highlighted the need to come up with metrics that can capture the underlying nature of the information present in the video [29].

In this work, we leverage tools from harmonic analysis by the use of the Signature Transform to introduce a set of measures, namely, Signature and Log-Signature Root Mean Squared Error (denoted from now on as RMSE Signature and Log-Signature), that can shed light on what a good summary is and serve as powerful tools to analytically quantize the information present in the selected frames.

As introduced in [30] in the context of GAN convergence assessment, the RMSE and MAE Signature and Log-Signature can be defined as follows, particularized for the application under study:

Definition 5 *Given n components of the element-wise mean of the signatures $\{\tilde{y}^{(c)}\}_{c=1}^{n} \subseteq T(\mathbb{R}^d)$ from the target summary to the score, and the same number of components of the element-wise mean of the signatures $\{\tilde{x}^{(c)}\}_{c=1}^{n} \subseteq T(\mathbb{R}^d)$ from the original video subsampled at a given frame rate and uniformly chosen, we define the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) as*

$$\mathrm{RMSE}\left(\left\{\tilde{x}^{(c)}\right\}_{c=1}^{n}, \left\{\tilde{y}^{(c)}\right\}_{c=1}^{n}\right) = \sqrt{\frac{1}{n}\sum_{c=1}^{n}\left(\tilde{y}^{(c)} - \tilde{x}^{(c)}\right)^2}, \tag{6.2}$$

and

$$\mathrm{MAE}\left(\left\{\tilde{x}^{(c)}\right\}_{c=1}^{n}, \left\{\tilde{y}^{(c)}\right\}_{c=1}^{n}\right) = \frac{1}{n}\sum_{c=1}^{n}\left|\tilde{y}^{(c)} - \tilde{x}^{(c)}\right|, \tag{6.3}$$

respectively, where $T(\mathbb{R}^d) = \prod_{c=0}^{\infty}\left(\mathbb{R}^d\right)^{\otimes c}$.

The case for Log-Signature is analogous.

For the task of video summarization, two approaches are given. In the case where the user has annotated summaries available, RMSE $(\bar{S}, \bar{S}_{target})$ is computed between an element-wise mean of the annotated summaries and the target summary to the score. If annotations are not available, a comparison against mean random uniform samples is performed, $\bar{S}$, and mean score and standard deviation are provided. Given the properties of the Signature Transform, the measure takes into consideration the harmonic components that are

intrinsic to the video under study and that should be preserved once the video is shortened to produce a summary. As a matter of fact, both approaches should lead to the same conclusions, as the harmonic components present in the annotated summaries and the ones present in average in the random uniform samples should also agree. A confidence interval of the scores can be provided for a given measure by analyzing the distances in the RMSEs of annotated summaries or random uniform samples, RMSE $(\bar{S}_a, \bar{S}_c)$.

When comparing against random uniform samples, the underlying assumption is as follows: we assume that good visual summaries capturing all or most of the harmonic components present in the visual cues will achieve a lower standard deviation. In contrast, summaries that lack support for the most important components will yield higher values. For a qualitative example, see Figure 6.1. With these ideas in mind, we can discern techniques that likely generate consistent summaries from those that fail to convey the most critical information. Moreover, the study of random sample intervals provides a set of tolerances for considering a given summary adequate for the task, meaning it is comparable to or better than uniform sampling of the interval at capturing harmonic components. Consequently, the proposed measures allow for a percentage score representing the number of times a given methodology outperforms random sampling by containing the same or more harmonic components present in the spectrum.

6.2.2 Summarization of Videos with RMSE Signature

Proposing a methodology based on the Signature Transform to select proper frames for a visual summary can be effectuated as follows: Given a uniform random sample of the video to summarize, we can compare it against subsequent random summaries using RMSE $(\bar{S}, \bar{S}_*)$. We can repeat this procedure n times and choose, as a good candidate, the minimum according to the standard deviation. Using this methodology, we can also repeat the procedure for a range of selected summary lengths, which will give us a set of good candidates, among which we will choose the candidate with the minimum standard deviation. This will provide us with an estimate of the most suitable length. It is important to note that this baseline is completely unsupervised in the sense that no annotations are used, only the metrics based on the Signature Transform. We rely on the fact that, in general, uniform random samples provide relatively accurate summaries, and among those, we choose the ones that are best according to std(RMSE $(\bar{S}, \bar{S}_*)$), which we denote as RMSE $(\bar{S}, \bar{S}_{u_{min}})|_n$. This will grant us competitive uniform random summaries according to the given measures to use as a baseline for comparison against other methodolo-

gies, and with which we can estimate an appropriate summary length to use in those cases.

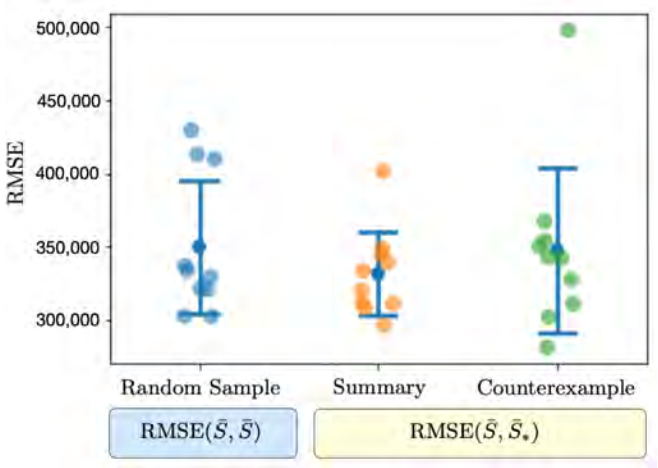


Figure 6.1: Conceptual plot with RMSE $(\bar{S}, \bar{S})$ and RMSE $(\bar{S}, \bar{S}_*)$ standard deviation and mean for two given summaries (our method and a counterexample) of 12 frames using a randomly picked video from Youtube to illustrate how to select a proper summary according to the proposed metric.

Below, we provide a description of the entities involved in the computation of the metrics and the proposed baselines based on the Signature Transform:

- $\bar{S}_*$: Element-wise mean Signature Transform of the target summary to the score of the corresponding video;
- $\bar{S}$: Element-wise mean Signature Transform of a uniform random sample of the corresponding video;
- RMSE $(\bar{S}, \bar{S}_*)$: Root mean squared error between the spectra of $\bar{S}$ and $\bar{S}_*$ with the same summary length. For the computation of standard deviation and mean, this value is calculated ten times, changing $\bar{S}$;
- RMSE $(\bar{S}, \bar{S})$: Root mean squared error between the spectra of $\bar{S}$ and $\bar{S}$ with the same summary length. For computation of standard deviation and mean, this value is calculated ten times, changing both $\bar{S}$ each time;
- RMSE $(\bar{S}, \bar{S}_{u_{min}})|_n$: Baseline based on the Signature Transform. It corresponds to RMSE $(\bar{S}, \bar{S}_*)$, where $\bar{S}_*$ is, in this case, a fixed uniform random

sample denoted as $\bar{S}_u$. We repeat this procedure n times and choose the minimum candidate according to standard deviation, $\bar{S}_{u_{min}}$, to propose as a summary;

- $std()$: Standard deviation.

6.3 Summarization of Videos via Text-Conditioned Object Detection

Large Language Models (LLM) [38, 37, 39, 40] and VLMs [41] have emerged as indispensable resources for characterizing complex tasks and bestowing intelligent systems with the capacity to interact with humans in unprecedented ways. These models, also called Foundation Models [42, 44, 43], excel in a wide variety of tasks, such as robotics manipulation [45, 47, 46], and can be integrated with other modules to perform robustly in highly complex situations such as navigation and guidance [49, 48]. One fundamental module is the Vision Transformer [50].

We introduce a simple yet effective technique aimed at generating video summaries that accurately describe the information contained within video streams, while also proposing new measures for the task of the summarization of videos. These measures will prove useful not only when text transcriptions are available, but also in more general cases in which we seek to describe the quality of a video summary.

Building on the text-conditioned object detection using Vision Transformers, as recently proposed in [51], we enhance the summarization task by leveraging the automated text transcriptions found in video platforms. We utilize a module of noun extraction employing NLP techniques [52], which is subsequently processed to account for the most frequent nouns. These nouns serve as input queries for text-conditioned object searches in frames. Frames containing the queries are selected for the video summary; see Figure 6.2 for a detailed depiction of the methodology.

In this chapter, we initially present a baseline leveraging text-conditioned object detection, specifically Contrastive Language–Image Pre-training (CLIP) [41]. To assess this approach, we employ a recently introduced metric based on the Signature Transform, which accurately gauges summary quality compared to a uniform random sample. Our preliminary baseline effectively demonstrates the competitiveness of uniform random sampling [29]. Consequently, we introduce a technique utilizing prior knowledge of the Signature, specifi-

cally the element-wise mean comparison of the spectrum, to generate highly accurate random uniform samples for summarization. The Signature Transform allows for a design featuring an inherent link between the methodology, metric, and baseline. We first present a method for evaluation, followed by a set of metrics for assessment, and ultimately, we propose a state-of-the-art baseline that can function as an independent technique.

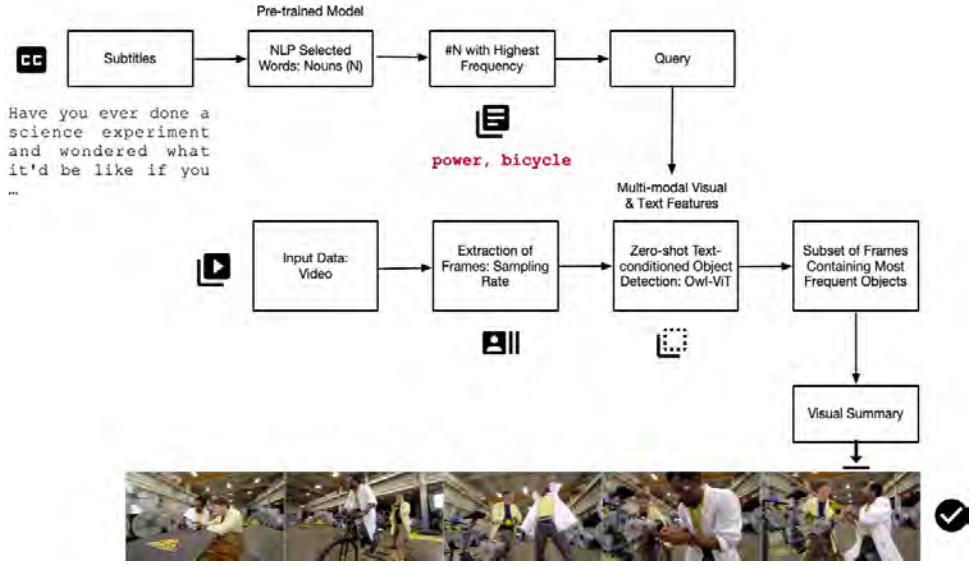


Figure 6.2: Video Summarization via Zero-shot Text-conditioned Object Detection.

6.4 Experiments: Dataset and Metrics

A dataset consisting of 28 videos about science experiments was sourced from Youtube, along with their automatic audio transcriptions, to evaluate the methodology and the proposed metrics. Table 6.1 provides a detailed description of the collected data and computed metrics, Figure 6.3 shows the distribution of selected frames using text-conditioned object detection over a subset of videos and the baselines based on the Signature Transform, Figure 6.4 depicts a visual comparison between methodologies, and Figures 6.5 and 6.6 visually elucidate the RMSE distribution for each video with mean and standard deviation.

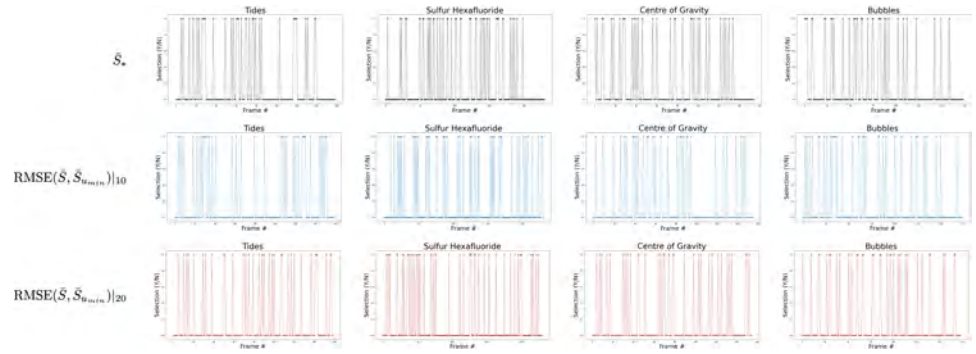


Figure 6.3: Comparison of distribution of selected frames for a subset of videos (Tides, Sulfur Hexafluoride, Centre of Gravity and Bubbles) using the method based on text-conditioned object detection and the baselines using the Signature Transform.

The dataset consists of science videos covering a wide range of experiments on several topics of interest; it has an average number of 264 frames per video (sampling rate $\frac{1}{4}$ s) and an average duration of 17 min 30 s.

Figure 6.3 depicts the selected frames when using our methodology for a subset of videos in the dataset. The selection coincides with the trigger of the zero-shot text-conditioned object detector by the 20 most frequent word code-phrase queries, which chooses a subset of the methodology that best explains the main factors of the argument. A comparison with the baselines based on the Signature Transform with 10 and 20 points is delivered.

In all experiments that involve the computation of the Signature Transform, we use the parameters proposed in [30] that were originally used to assess synthetic distributions generated with GANs; specifically, we employ truncated signatures of order 3 with a resized image size of 64×64 in grayscale.

RMSE $(\bar{S}, \bar{S}_*)$ computes the element-wise mean of the signatures of both the target summary to the score and a random uniform sample with the same number of frames, comparing their spectra with the use of the RMSE. Likewise, RMSE $(\bar{S}, \bar{S})$ computes the same measure between two random uniform samples with the same number of frames. The standard deviation of both results is compared to assess the quality of the summarized video concerning the present harmonic components. The preliminary technique based on text-conditioned object detection (see Table 6.1) achieves a zero-shot of 50% positive cases when compared against std (RMSE $(\bar{S}, \bar{S})$). The number of frames selected by the

methodology is consistent, and it automatically selects on average 20% of the total number of frames.

In this paragraph, we discuss the baseline based on the Signature Transform (see Table 6.1) in terms of the RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$. These techniques select a uniform random sample with minimum standard deviation in a set of 10 points and 20 points, respectively, and achieve 100% positive cases when compared to RMSE $(\bar{S}, \bar{S})$. Under the assumption that the summary can be approximated well by a random uniform sample, which holds true in many cases, the methodology finds a set of frames that maximizes the harmonic components relative to those present in the original video.

Table 6.1: Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ correspond to the baselines based on the Signature Transform using 10 and 20 random samples, respectively. Highlighted results in blue/brown correspond to values better than std (RMSE $(\bar{S}, \bar{S})$). Yellow values indicate when std (RMSE $(\bar{S}, \bar{S})$) is lower than std (RMSE $(\bar{S}, \bar{S}_*)$).

Descriptive Statistics			**Summary**	**RMSE $(\bar{S}, \bar{S}_*)$**		**RMSE $(\bar{S}, \bar{S})$**		**RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$**		**RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$**	
Video	**# Frames**	**Length**	**# Frames (%)**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**
Tides	159	10 m 29 s	35 (22%)	13,663	202,388	14,838	155,986	8859	157,455	7312	167,480
Sulfur Hexafluoride	230	15 m 12 s	47 (20%)	22,727	217,935	22,607	179,409	7194	161,995	7722	173,490
Centre of Gravity	155	10 m 14 s	33 (21%)	12,333	181,460	16,404	168,824	8481	160,779	12,416	175,971
Bubbles	174	11 m 30 s	35 (20%)	23,127	201,553	16,806	185,702	7461	194,993	5711	175,176
Airplanes	158	10 m 24 s	22 (14%)	19,964	215,688	23,591	231,539	8417	227,391	10,235	233,020
Protons	174	11 m 30 s	25 (14%)	29,853	252,224	20,186	262,434	12,835	251,907	11,542	250,512
Hydrophobic	168	11 m 06 s	29 (17%)	15,016	251,671	25,835	248,548	11,973	250,131	13,917	245,761
States of Matter	332	22 m 03 s	78 (23%)	16,249	156,408	9709	130,064	6630	115,454	5340	121,028
Spool Racer	332	22 m 02 s	90 (27%)	15,903	142,520	11,883	136,147	7054	137,621	8112	151,888
Paper Airplane	332	22 m 03 s	29 (9%)	20,642	235,639	11,829	221,220	5400	224,718	9385	177,448
Loudest Sound	332	22 m 01 s	93 (28%)	16,898	179,963	8304	148,885	7884	138,561	4355	147,016
Lightning	332	22 m 01 s	70 (21%)	15,237	169,338	21,862	162,849	9300	177,008	7494	153,797
Light Challenge	332	22 m 02 s	82 (25%)	12,566	152,488	10,546	126,117	5490	139,700	4874	129,044
Hot Air Balloon	332	22 m 01 s	98 (30%)	8620	150,366	5417	144,634	3516	137,141	4165	138,453
Hoop Glider	332	22 m 01 s	82 (25%)	6419	148,065	6752	132,544	4051	133,897	4966	133,894
Drag Race	332	22 m 03 s	73 (22%)	9384	135,228	8931	125,264	4375	122,615	4645	129,851
All about Balance	332	22 m 03 s	59 (18%)	14,023	182,063	14,238	182,179	7801	176,219	6914	167,727
Air Pressure	332	22 m 03 s	65 (20%)	10,123	166,342	18,314	151,664	6386	145,897	4602	148,232
Friction and Momentum	162	10 m 42 s	28 (17%)	18,754	217,403	22,443	218,203	13,348	202,288	12,238	205,680
Electricity	162	10 m 41 s	30 (19%)	24,376	298,238	22,885	279,820	16,889	268,932	10,263	270,619
Catapult	169	11 m 11 s	27 (16%)	26,413	271,643	31,265	214,727	15,158	203,290	10,222	188,008
Carbonation and More	165	10 m 53 s	40 (24%)	18,977	237,142	18,107	226,044	12,130	234,278	11,884	214,149
Carbon Dioxide	162	10 m 41 s	38 (23%)	25,862	245,415	18,806	217,270	13,838	207,828	7760	211,504
Bridge	164	10 m 51 s	21 (13%)	25,839	269,412	26,038	271,551	10,761	263,747	13,038	264,532
Bread Experiment	337	22 m 22 s	59 (18%)	15,099	189,086	8575	146,771	5542	153,224	5691	156,230
Balloon Power	337	22 m 22 s	53 (16%)	14,075	157,542	29,415	147,710	7741	128,920	7351	134,545
Attraction and Forces	654	43 m 30 s	81 (12%)	5955	107,097	7486	102,965	3701	96,266	2093	99,271
Puzzles	209	13 m 48 s	46 (22%)	11,258	185,502	19,012	196,762	14,620	199,556	14,622	197,064
Average	264	17 m 30 s	52 (20%)	14/28 (50%)				28/28 (100%)		28/28 (100%)	

Figure 6.4 displays examples of summaries using the baseline based on the Signature Transform compared to the summaries using text-conditioned object detection. The figure allows for a visual comparison of the results obtained using RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ and $\bar{S}_*$. The best summary among the three baselines according to the metric is highlighted (Table 6.1).

Figure 6.4: Summarization of videos using the baseline based on the Signature Transform in comparison to the summarization using text-conditioned object detection. RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ and $\bar{S}_*$ summaries for two videos of the introduced dataset. The best summary among the three, according to the metric, is highlighted.

The selected frames are consistent and provide a good overall description of the original videos. Moreover, the metric based on the Signature Transform aligns well with our expectations of a high-quality summary, with better scores being assigned to summaries that effectively convey the content present in the original video.

Table 6.2 presents a qualitative analysis of the baseline based on the Signature Transform using 10 points, RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S})$ with a varying number of frames per summary. We observe that RMSE $(\bar{S}, \bar{S})$ reflects the variability of the harmonic components present; that is, it is preferable to work with lengths for which the variability among summaries is low, according to the standard deviation. RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ indicates the minimum standard deviation achieved in a set of 10 points, meaning that given a computational budget allowing us to select up to a specific number of frames, a good choice is to pick the length that yields the minimum RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ with low variability, as per RMSE $(\bar{S}, \bar{S})$.

RMSE $(\bar{S}, \bar{S}_*)$ (Figure 6.5) and RMSE $(\bar{S}, \bar{S})$ (Figure 6.6) show the respective distribution of RMSE values (10 points) with the mean and standard deviation. Low standard deviations, in comparison with the random uniform sample counterparts, indicate good summarization capabilities.

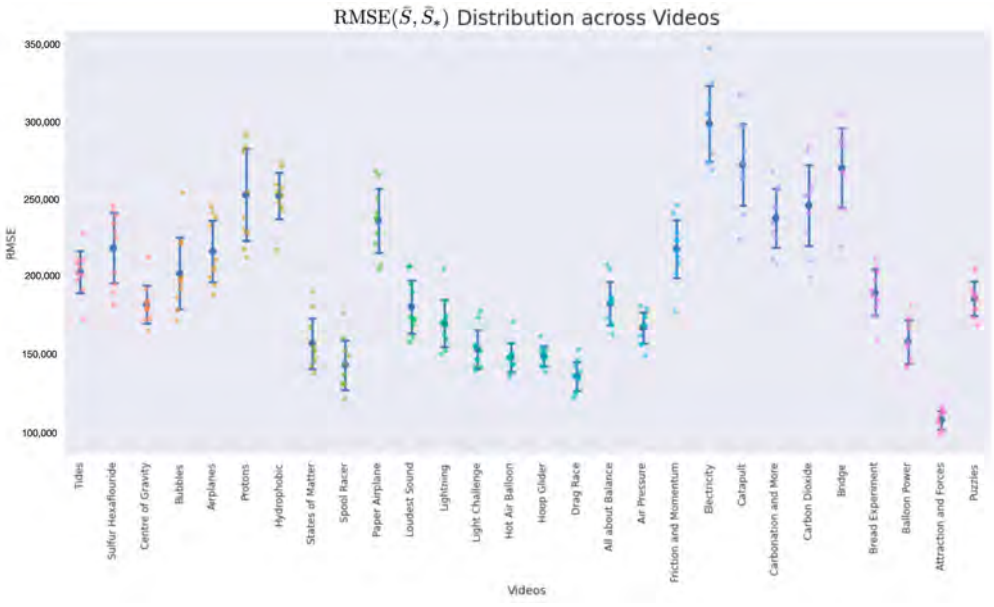


Figure 6.5: Plot with RMSE $(\bar{S}, \bar{S}_*)$ standard deviation and mean.

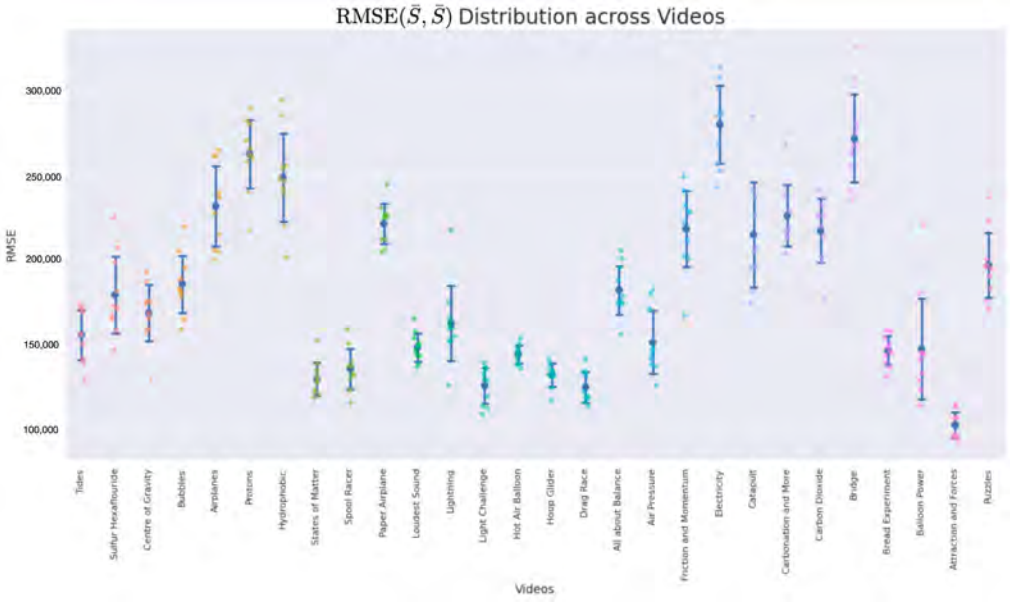
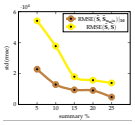
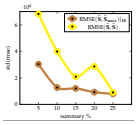
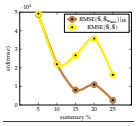
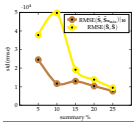


Figure 6.6: Plot with RMSE $(\bar{S}, \bar{S})$ standard deviation and mean.

Table 6.2: Descriptive statistics for a set of videos with varying numbers of frames per summary with RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ (brown) and RMSE $(\bar{S}, \bar{S})$ (yellow).

Dataset			RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$		RMSE $(\bar{S}, \bar{S})$		Visualization
Video	**# Frames**	**Summary (%)**	**Std**	**Mean**	**Std**	**Mean**	**Plot (Std,Std)**
Tides	159	8 (5%)	22,786	422,026	54,067	390,483	
		16 (10%)	12,851	254,984	37,713	263,881	
		24 (15%)	9423	202,925	17,935	224,797	
		32 (20%)	9074	183,933	15,700	186,621	
		40 (25%)	4782	158,183	13,903	159,452	
Sulfur Hexafluoride	230	12 (5%)	30,325	452,134	68,212	362,061	
		23 (10%)	12,701	281,425	39,872	246,967	
		35 (15%)	12,034	228,530	20,846	201,740	
		46 (20%)	9241	190,985	28,621	175,440	
		58 (25%)	7914	161,618	9021	152,310	
Centre of Gravity	155	8 (5%)	48,787	406,502	49,234	369,648	
		16 (10%)	22,163	252,841	21,974	276,366	
		24 (15%)	8050	212,893	26,776	229,959	
		31 (20%)	10,963	180,953	35,813	184,437	
		39 (25%)	2528	164,666	16,259	163,007	
Bubbles	174	9 (5%)	24,538	401,406	37,816	397,470	
		18 (10%)	11,669	272,430	49,740	276,152	
		27 (15%)	12,965	213,336	19,125	215,961	
		35 (20%)	10,331	190,639	13,792	183,984	
		44 (25%)	7625	173,009	9427	162,091	

6.4.1 Assessment of the Metrics

The metrics have been rigorously evaluated using the dataset in [1], which consists of short videos sourced from Youtube, and includes 5 annotated summaries per video for a total of 20. Tables 6.3 and 6.4 report the results, using a one-frame-per-second sampling rate. In this case, the average number of times that the human annotator outperforms uniform random sampling according to the proposed metric, std (RMSE $(\bar{S}, \bar{S})$), is 87%. Several observations emerge from these findings:

- The proposed metrics demonstrate that human evaluators can perform above average during the task, effectively capturing the dominant harmonic frequencies present in the video.
- Another crucial aspect to emphasize is that the metrics are able to evaluate human annotators with fair criteria and identify which subjects are creating competitive summaries.

- Moreover, the observations from this study indicate that the metrics serve as a reliable proxy for evaluating summaries without the need for annotated data, as they correlate strongly with human annotations.

Figure 6.7 shows the mean and standard deviation for each human-annotated summary (user 1 to user 5) for the subset of 20 videos from [1], using a sampling rate of 1 frame per second. For each video, a visual inspection of the error plot bar for each annotated summary provides an accurate estimate of the quality of the annotation compared to other users. Specifically:

- Annotations with lower standard deviations offer a better harmonic representation of the overall video;
- Annotations with higher standard deviations suggest that important harmonic components are missing from the given summary;
- The metrics make it simple to identify annotated summaries that may need to be relabeled for improved accuracy.

Furthermore, these metrics remain consistent when applied to various sampling rates.

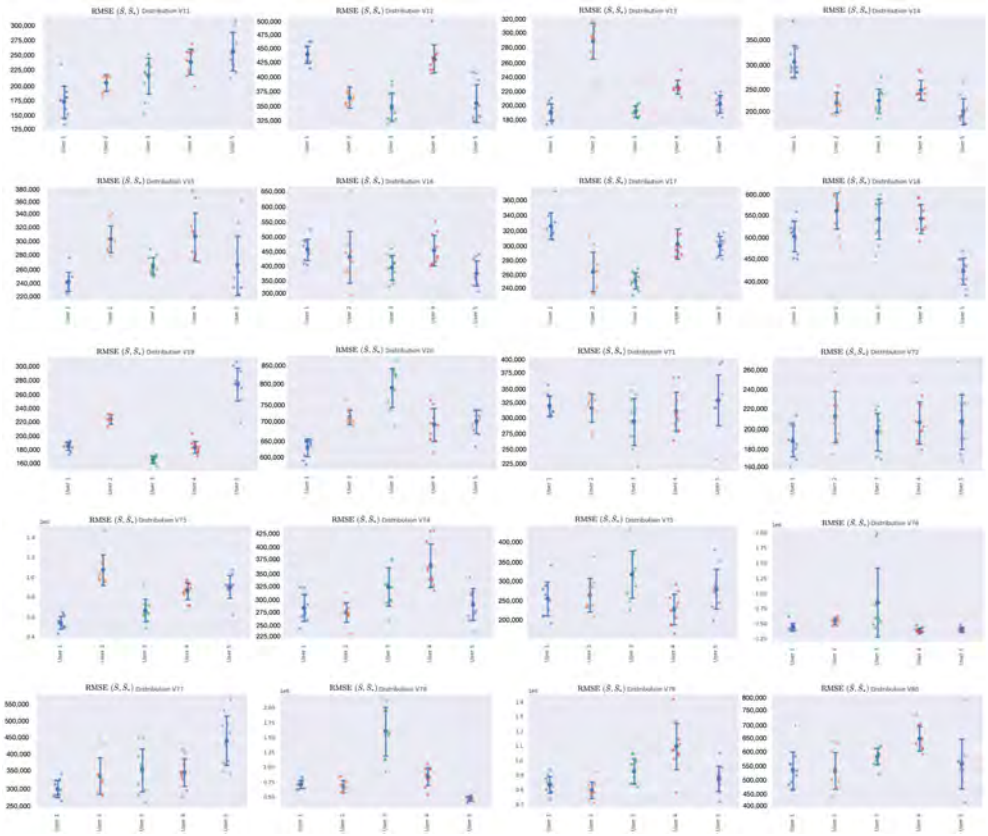


Figure 6.7: Error bar plot with mean and standard deviation for each human-annotated summary of the subset of 20 videos from [1]. Sampling rate: 1 frame per second.

That being said, there are several standard measures that are commonly used for video summarization, such as F1 score, precision, recall, and Mean Opinion Score (MOS). Each of these measures has its own strengths and weaknesses. Compared to these standard measures, the proposed benchmark based on the Signature Transform has several potential advantages. Here are a few reasons for this:

- Content based: the Signature Transform is a content-based approach that captures the salient features of the video data. This means that the proposed measure is not reliant on manual annotations or subjective human ratings, which can be time consuming and prone to biases.

- Robustness: the Signature Transform is a robust feature extraction technique that can handle different types of data, including videos with varying frame rates, resolutions, and durations. This means that the proposed measure can be applied to a wide range of video datasets without the need for pre-processing or normalization.
- Efficiency: the Signature Transform is a computationally efficient approach that can be applied to large-scale datasets. This means that the proposed measure can be used to evaluate the effectiveness of visual summaries quickly and accurately.
- Flexibility: the Signature Transform can be applied to different types of visual summaries, including keyframe-based and shot-based summaries. This means that the proposed measure can be used to evaluate different types of visual summaries and compare their effectiveness.

Overall, the proposed measure based on the Signature Transform has the potential to provide a more accurate and comprehensive assessment of the standard of visual summaries compared to the preceding measures used in video summarization.

Table 6.3: Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). Lower is better. Sampling rate: 1 frame per second. Dataset in [1], videos from V11 to V20. Highlighted results in blue/yellow correspond to the lowest values, either std (RMSE $(\bar{S}, \bar{S}_*)$) or std (RMSE $(\bar{S}, \bar{S})$), respectively.

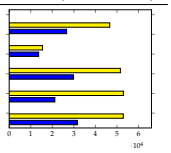
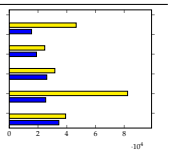
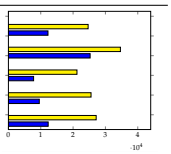
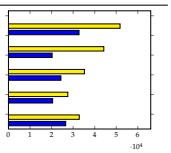
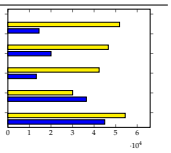
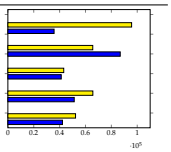
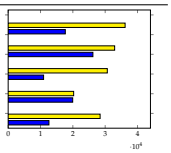
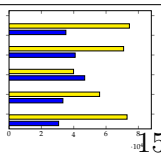
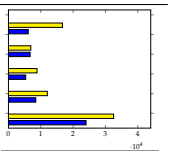
Youtube, Dataset				RMSE $(\bar{S}, \bar{S}_*)$		RMSE $(\bar{S}, \bar{S})$		Visualization
Video	**# Frames**	**User**	**# Frames User**	**Std**	**Mean**	**Std**	**Mean**	**Plot (Std,Std)**
V11	48	1	10	26,644	171,106	46,655	151,483	
		2	12	13,673	202,172	15,479	155,481	
		3	10	29,857	213,880	51,590	182,327	
		4	9	21,192	236,959	52,982	196,303	
		5	8	31,627	254,336	52,925	193,520	
V12	59	1	11	15,497	436,723	46,551	252,142	
		2	17	18,927	359,562	24,665	177,286	
		3	15	26,071	342,161	31,703	180,066	
		4	11	25,330	429,272	82,323	242,627	
		5	14	34,479	348,834	39,199	188,417	
V13	59	1	19	12,238	187,001	24,649	114,155	
		2	9	25,267	287,479	34,635	166,495	
		3	18	7790	187,346	21,203	126,432	
		4	14	9544	222,496	25,553	140,508	
		5	18	12,298	198,349	27,138	124,386	
V14	59	1	9	32,739	302,118	51,770	183,978	
		2	16	20,249	219,068	44,235	141,927	
		3	17	24,345	222,559	35,235	113,806	
		4	10	20,498	244,509	27,548	155,515	
		5	16	26,561	200,139	32,840	143,384	
V15	57	1	12	14,454	237,551	51,812	207,845	
		2	11	20,018	301,650	46,590	209,491	
		3	13	13,192	261,014	42,337	171,810	
		4	13	36,408	305,376	30,041	179,442	
		5	14	44,931	261,859	54,428	180,145	
V16	70	1	9	35,722	449,758	95,662	376,411	
		2	9	86,863	425,107	65,626	328,563	
		3	12	41,260	388,869	43,186	340,133	
		4	9	51,299	447,523	65,698	375,162	
		5	13	42,200	369,517	52,316	302,677	
V17	59	1	12	17,668	324,562	36,166	242,235	
		2	13	26,203	262,895	32,930	243,366	
		3	18	10,957	250,543	30,660	177,779	
		4	12	19,956	300,390	20,252	223,791	
		5	16	12,611	297,707	28,433	207,258	
V18	50	1	13	35,152	501,230	74,454	260,574	
		2	14	40,896	559,244	70,863	274,572	
		3	14	46,791	540,747	39,899	246,964	
		4	10	33,309	541,490	56,012	329,343	
		5	14	30,663	420,924	72,998	308,756	
V19	65	1	15	6114	186,893	16,695	119,136	
		2	20	6701	225,075	6899	103,517	
		3	20	5339	167,085	8834	103,752	
		4	13	8462	185,452	12,020	129,608	
		5	6	23,992	275,155	32,512	208,629	

Table 6.3: *Cont.*

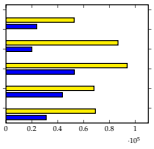
Youtube, Dataset				RMSE $(\bar{S}, \bar{S}_*)$		RMSE $(\bar{S}, \bar{S})$		Visualization
Video	**# Frames**	**User**	**# Frames User**	**Std**	**Mean**	**Std**	**Mean**	**Plot (Std,Std)**
V20	61	1	15	23,716	627,121	52,711	540,857	
		2	12	19,933	707,823	86,586	609,589	
		3	9	52,818	787,188	93,656	747,199	
		4	11	43,598	688,065	68,016	617,091	
		5	11	31,058	695,905	69,077	618,156	

Table 6.4: Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). Lower is better. Sampling rate: 1 frame per second. Dataset in [1], videos from V71 to V80. Highlighted values correspond to the lowest standard deviation.

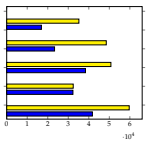
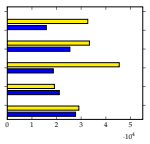
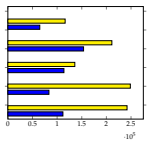
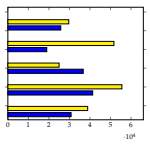
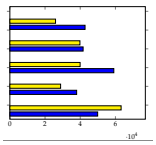
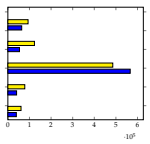
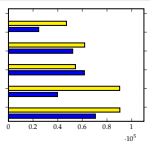
Youtube, Dataset				RMSE $(\bar{S}, \bar{S}_*)$		RMSE $(\bar{S}, \bar{S})$		Visualization
Video	**# Frames**	**User**	**# Frames User**	**Std**	**Mean**	**Std**	**Mean**	**Plot (Std,Std)**
V71	277	1	18	16,916	319,975	35,173	330,114	
		2	18	23,314	315,996	48,511	339,793	
		3	20	38,384	293,853	50,766	345,021	
		4	17	32,270	310,193	32,411	359,049	
		5	18	41,753	329,353	59,688	334,337	
V72	536	1	18	15,842	187,019	32,676	194,820	
		2	16	25,427	211,466	33,363	202,442	
		3	16	18,684	196,149	45,453	217,699	
		4	18	21,112	205,421	19,122	177,117	
		5	18	27,718	206,335	29,057	205,808	
V73	201	1	11	64,802	538,239	116,284	484,970	
		2	7	153,682	106,8305	211,124	704,655	
		3	8	113,805	661,992	135,899	653,041	
		4	8	83,387	856,406	248,619	689,301	
		5	7	111,767	899,150	241,947	794,828	
V74	293	1	17	25,780	282,200	29,674	309,051	
		2	16	18,954	273,776	51,670	331,322	
		3	15	36,714	322,833	24,961	335,618	
		4	13	41,327	363,665	55,543	369,875	
		5	16	30,798	289,135	38,881	353,928	
V75	383	1	14	42,736	254,385	25,959	282,877	
		2	13	41,632	263,431	39,826	337,124	
		3	10	59,083	315,531	39,925	330,766	
		4	17	37,954	227,411	28,843	250,314	
		5	12	49,908	278,966	63,236	312,366	
V76	89	1	6	64,097	440,825	93,524	422,565	
		2	4	53,727	536,138	123,009	464,922	
		3	1	566,208	843,799	485,614	878,793	
		4	6	40,356	382,643	78,354	424,418	
		5	6	39,194	395,906	60,916	401,751	
V77	168	1	12	24,546	302,076	47,095	366,748	
		2	9	52,176	339,285	61,880	385,056	
		3	9	61,623	355,883	54,390	413,118	
		4	10	39,765	349,207	90,313	400,379	
		5	7	70,562	440,656	90,468	451,833	

Table 6.4: *Cont.*

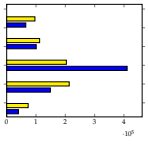
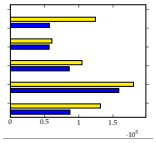
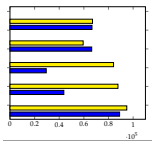
Youtube, Dataset				RMSE $(\bar{S}, \bar{S}_*)$		RMSE $(\bar{S}, \bar{S})$		Visualization
Video	**# Frames**	**User**	**# Frames User**	**Std**	**Mean**	**Std**	**Mean**	**Plot (Std,Std)**
V78	310	1	13	65,238	706,978	96,368	770,000	
		2	14	100,771	672,121	112,412	807,250	
		3	3	410,792	159,3229	203,589	188,2757	
		4	9	149,063	839,743	213,286	106,1204	
		5	23	40,178	466,571	73,228	614,140	
V79	49	1	7	56,918	831,057	124,249	835,575	
		2	8	56,569	793,831	60,657	859,241	
		3	6	85,973	925,025	104,621	990,479	
		4	5	158,480	109,3141	179,902	109,9105	
		5	6	87,104	873,950	131,597	895,318	
V80	159	1	18	66,585	529,875	67,019	572,836	
		2	17	66,367	527,930	59,432	602,819	
		3	13	29,459	579,078	84,101	726,883	
		4	12	43,740	643,016	87,688	685,117	
		5	14	89,016	553,274	94,849	649,317	

Figure 6.8 shows a summary that is well annotated by all users, demonstrating that the metrics can accurately indicate when human annotators have effectively summarized the information present in the video.

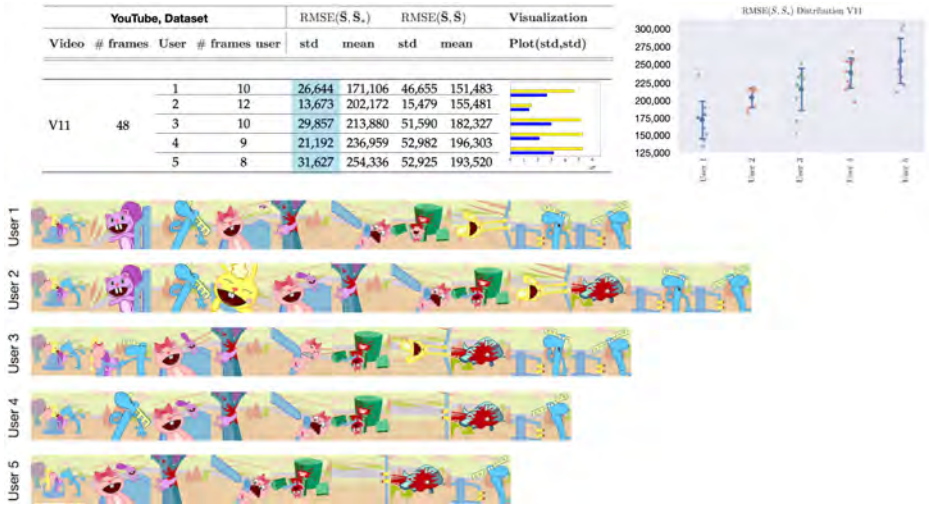


Figure 6.8: Visual depiction of human annotated summaries together with RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$ of video V11, Table 6.3. Sampling rate: 1 frame per second. Highlighted values on the table correspond to the lowest standard deviation.

To illustrate how these metrics can help improve annotations, Figure 6.9 displays the metrics along with the annotated summaries of users 1 to 5. We observe that selecting the frames highlighted by users 1–4 can increase the performance if user 5 is asked to relabel its summary.

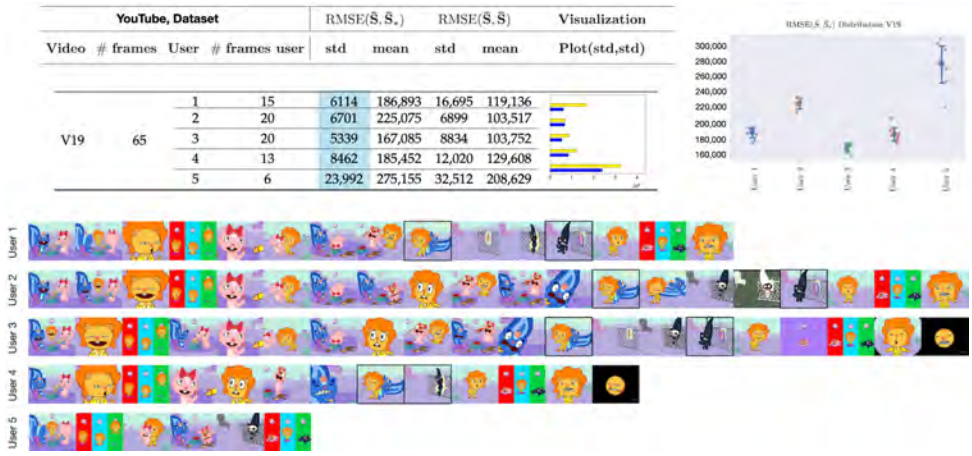
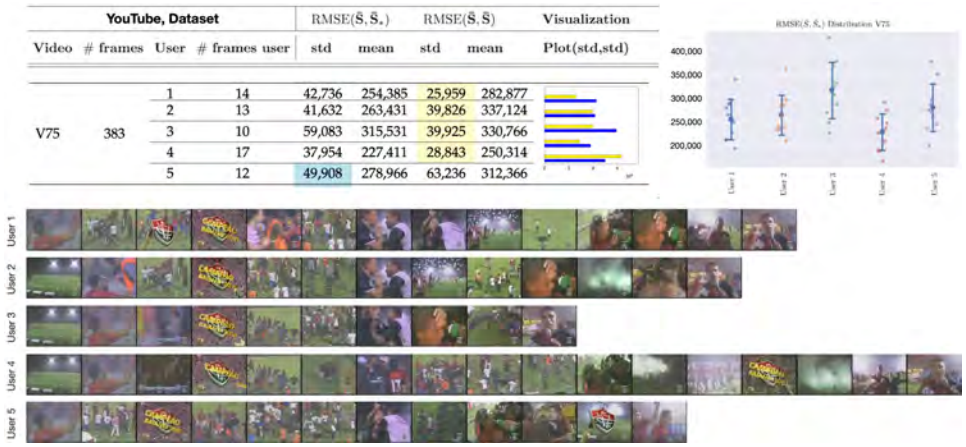


Figure 6.9: Visual depiction of human annotated summaries together with RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$ of video V19, Table 6.3. Sampling rate: 1 frame per second. Highlighted frames can increase the accuracy of the annotated summary by user 5. Highlighted values on the table correspond to the lowest standard deviation.

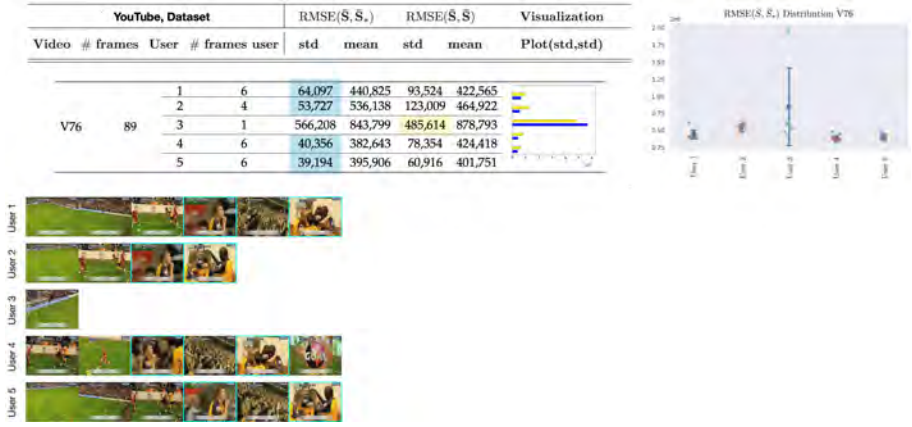
Figure 6.10 showcases an example in which random uniform sampling outperforms the majority of human annotators. This occurs because the visual information is uniformly distributed throughout the video. In this case, user 5 performs the best, scoring slightly higher than std (RMSE $(\bar{S}, \bar{S})$. Highlighted values on the table correspond to the lowest standard deviation.).



YouTube, Dataset				RMSE($\bar{S}, \bar{S}_*$)		RMSE($\bar{S}, \bar{S}$)		Visualization
Video	# frames	User	# frames user	std	mean	std	mean	Plot(std,std)
V75	383	1	14	42,736	254,385	25,959	282,877	
		2	13	41,632	263,431	39,826	337,124	
		3	10	59,083	315,531	39,925	330,766	
		4	17	37,954	227,411	28,843	250,314	
		5	12	49,908	278,966	63,236	312,366	

Figure 6.10: Visual depiction of human annotated summaries, together with RMSE ($\bar{S}, \bar{S}_*$) and RMSE ($\bar{S}, \bar{S}$) of video V75, Table 6.4. Sampling rate: 1 frame per second. Highlighted values on the table correspond to the lowest standard deviation.

Similarly, Figure 6.11 presents an example in which incorporating the highlighted frames improves the accuracy of the annotated summary by user 3, which is currently performing worse than uniform random sampling, according to the metrics.



YouTube, Dataset				RMSE($\bar{S}, \bar{S}_*$)		RMSE($\bar{S}, \bar{S}$)		Visualization
Video	# frames	User	# frames user	std	mean	std	mean	Plot(std,std)
V76	89	1	6	64,097	440,825	93,524	422,565	
		2	4	53,727	536,138	123,009	464,922	
		3	1	566,208	843,799	485,614	878,793	
		4	6	40,356	382,643	78,354	424,418	
		5	6	39,194	395,906	60,916	401,751	

Figure 6.11: Visual depiction of human annotated summaries together with RMSE ($\bar{S}, \bar{S}_*$) and RMSE ($\bar{S}, \bar{S}$) of video V76, Table 6.4. Sampling rate: 1 frame per second. Highlighted frames can increase the accuracy of the annotated summary by user 3. Highlighted values on the table correspond to the lowest standard deviation.

6.4.2 Evaluation

In this section, we evaluate the baselines and metrics compared to VSUMM [1], a methodology based on handcrafted techniques that performs particularly well on this dataset. Table 6.5 displays the comparison between the standard deviation of RMSE $(\bar{S}, \bar{S}_*)$ and RMSE $(\bar{S}, \bar{S})$, as well as against the baselines based on the Signature Transform,
RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$, with 10 and 20 points, respectively.

We can observe how the metrics effectively capture the quality of the visual summaries and how the introduced methodology based on the Signature Transform achieves state-of-the-art results with both 10 and 20 points. The advantages of using a technique that operates on the spectrum of the signal, compared to other state-of-the-art systems, is that it can generate visual summaries without fine-tuning the methodology. In other words, there is no need to train on a subset of the target distribution of videos, but rather, compelling summaries can be generated at once for any dataset. Moreover, this approach is highly efficient, as computation is performed on the CPU and consists only of calculating the Signature Transform, element-wise mean, and RMSE. These operations can be further optimized for rapid on-device processing or for deploying in parallel at the tera-scale level.

Table 6.5: VSUMM [1] comparison against baseline based on the Signature Transform for the first 20 videos of the dataset crawled from Youtube. Descriptive statistics with RMSE $(\bar{S}, \bar{S}_*)$ (target summary against random uniform sample) and RMSE $(\bar{S}, \bar{S})$ (random uniform sample against random uniform sample). RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{10}$ and RMSE $(\bar{S}, \bar{S}_{u_{min}})|_{20}$ correspond to the baselines based on the Signature Transform using 10 and 20 random samples, respectively. Highlighted results are better than std (RMSE $(\bar{S}, \bar{S})$). Sampling rate: 1 frame per second. Highlighted results correspond to lowest standard deviation as described in Table 6.1.

Descriptive Statistics		**VSUMM**	**RMSE $(\bar{S}, \bar{S}_*)$**		**RMSE $(\bar{S}, \bar{S})$**		**RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{10}$**		**RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{20}$**	
Video	**# Frames**	**# Frames**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**
V11	48	11	25,981	185,959	37,907	175,031	16,343	148,128	18,343	159,157
V12	59	13	56,274	313,156	41,613	205,004	17,770	181,533	11,665	206,951
V13	59	19	7018	184,865	15,319	120,307	10,578	110,258	6655	134,846
V14	59	8	21,415	281,969	39,412	171,935	19,069	157,531	10,104	180,199
V15	57	10	20,159	271,197	46,041	219,182	27,536	192,667	27,765	218,787

Table 6.5: *Cont.*

Descriptive Statistics		VSUMM	RMSE $(\bar{S}, \bar{S}_*)$		RMSE $(\bar{S}, \bar{S})$		RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{10}$		RMSE $(\bar{S}, \bar{S}_{u_{min}})\|_{20}$	
Video	**# Frames**	**# Frames**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**
V16	70	9	65,997	513,440	84,667	428,025	38,088	283,324	30,235	446,068
V17	59	15	10,697	255,666	41,831	197,136	17,625	197,944	19,102	227,646
V18	50	14	42,731	449,324	51,635	230,695	33,525	261,288	30,179	242,746
V19	65	16	3891	235,797	5739	121,766	5883	116,245	4582	111,766
V20	61	9	43,864	796,448	39,035	733,547	28,460	684,546	39,414	644,681
V71	277	17	20,840	383,945	43,176	341,779	14,908	352,365	20,657	327,732
V72	536	12	61,886	233,649	48,603	252,688	17,604	276,631	18,966	248,489
V73	201	10	40,261	717,107	156,051	533,457	64,344	681,064	38,361	711,039
V74	293	17	26,274	270,374	36,674	334,265	17,622	354,621	17,486	330,606
V75	383	10	37,516	272,804	38,026	366,510	23,163	339,078	21,295	360,216
V76	89	7	36,084	353,323	114,266	377,699	31,131	335,958	34,724	405,954
V77	168	9	26,653	361,516	67,134	422,612	33,214	407,085	27,562	480,795
V78	310	13	95,305	831,043	127,705	823,938	33,903	980,397	36,361	951,784
V79	49	7	67,052	965,267	101,325	878,917	42,513	818,629	47,401	885,023
V80	159	15	48,115	613,702	118,428	644,529	43,411	589,256	37,487	808,984
Average	153	12	17/20 (85%)				19/20 (95%)		19/20 (95%)	

6.5 Conclusions and Future Work

In this thesis, we propose a benchmark based on the Signature Transform to evaluate visual summaries. For this purpose, we introduce a dataset consisting of videos obtained from Youtube related to science experiments with automatic audio transcriptions. A baseline, based on zero-shot text-conditioned object detection, is used as a preliminary technique in the study to evaluate the metrics. Subsequently, we present an accurate baseline built on the prior knowledge that the Signature provides. Furthermore, we conduct rigorous comparison against human-annotated summaries to demonstrate the high correlation between the measures and the human notion of a good summary.

One of the main contributions of this work is that techniques based on the Signature Transform can be integrated with any state-of-the-art method in the form of a gate that activates when the method performs worse than the metric, std (RMSE $(\bar{S}, \bar{S}_*)$) > std (RMSE $(\bar{S}, \bar{S})$).

The experiments conducted in this work lead to the following conclusion: if a method for delivering a summarization technique is proposed that involves complex computation (e.g., DNN techniques or Foundation Models), it must provide better summarization capabilities than the baselines based on the Signature Transform, which serve as lower bounds for uniform random samples. If not, there is no need to use a more sophisticated technique that would in-

volve greater computational and memory overhead and possibly require training data. The only exception to this would be when additional constraints are present in the problem, such as when summarization must be performed by leveraging audio transcriptions (as in the technique based on text-conditioned object detection) or any other type of multimodal data.

That being said, the methodology proposed based on the Signature Transform, although accurate and effective, is built on the overall representation of harmonic components of the signal. Videlicet, under certain circumstances, can provide summaries in which frames are selected due to low-level representations of the signal, such as color and image intensity, rather than the storyline. Moreover, it assumes that, in general, uniform random sampling can provide good summarization capabilities, which is supported by the literature. However, this assumption is not fulfilled in all circumstances. Therefore, in subsequent works, it would be desirable to develop techniques that perform exceptionally well according to the metrics while simultaneously bestowing a level of intelligence similar to the methodology based on Foundation Models. This would take into account factors such as the human concept of detected objects, leading to more context-aware and meaningful summarization.

Bibliography

[1] de Avila, S.E.F.; Lopes, A.; da Luz, A., Jr.; de Albuquerque Araújo, A. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognit. Lett.* **2011**, *32*, 56–68. [CrossRef]

[2] Gygli, M.; Grabner, H.; Gool, L.V. Video summarization by learning submodular mixtures of objectives. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.

[3] Gygli, M.; Grabner, H.; Riemenschneider, H.; Van Gool, L. (2014). Creating summaries from user videos. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September; Springer: Berlin/Heidelberg, Germany, 2014.

[4] Kanehira, A.; Gool, L.V.; Ushiku, Y.; Harada, T. Viewpoint-aware video summarization. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018.

[5] Liang, G.; Lv, Y.; Li, S.; Zhang, S.; Zhang, Y. Video summarization with a convolutional attentive adversarial network. *Pattern Recognit.* **2022**, *131*, 108840. [CrossRef]

[6] Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A. TVSum: Summarizing web videos using titles. In Proceedings of the 2015 IEEE Conference on

Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 5179–5187.

[7] Zhu, W.; Lu, J.; Han, Y.; Zhou, J. Learning multiscale hierarchical attention for video summarization. *Pattern Recognit.* **2022**, *122*, 108312. [CrossRef]

[8] Ngo, C.-W.; Ma, Y.-F.; Zhang, H.-J. Automatic video summarization by graph modeling. In Proceedings of the Ninth IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003.

[9] Fajtl, J.; Sokeh, H.S.; Argyriou, V.; Monekosso, D.; Remagnino, P. Summarizing videos with attention. In Proceedings of the Computer Vision—ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, 2–6 December; Springer: Berlin/Heidelberg, Germany, 2018.

[10] Zhu, W.; Lu, J.; Li, J.; Zhou, J. DSNet: A flexible detect-to-summarize network for video summarization. *IEEE Trans. Image Process.* **2020**, *30*, 948–962. [CrossRef] [PubMed]

[11] He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.

[12] Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]

[13] Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.

[14] Dwivedi, K.; Bonner, M.F.; Cichy, R.M.; Roig, G. Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Comput. Biol.* **2021**, *17*, e100926. [CrossRef]

[15] Dwivedi, K.; Roig, G.; Kembhavi, A.; Mottaghi, R. What do navigation agents learn about their environment? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 10276–10285.

[16] Rakshit, S.; Tamboli, D.; Meshram, P.S.; Banerjee, B.; Roig, G.; Chaudhuri, S. Multi-source open-set deep adversarial domain adaptation. In

Proceedings of the Computer Vision—ECCV: 16th European Conference, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 735–750.

[17] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI*; Springer: Cham, Switzerland, 2015.

[18] Tan, M.; Chen, B.; Pang, R.; Vasudevan, V.; Le, Q.V. Mnasnet: Platform-aware neural architecture search for mobile. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

[19] Thao, H.; Balamurali, B.; Herremans, D.; Roig, G. Attendaffectnet: Self-attention based networks for predicting affective responses from movies. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 8719–8726.

[20] Mahasseni, B.; Lam, M.; Todorovic, S. Unsupervised video summarization with adversarial lstm networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

[21] Zhang, K.; Chao, W.-L.; Sha, F.; Grauman, K. Video summarization with long short-term memory. In Proceedings of the Computer Vision–ECCV: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016.

[22] Zhao, B.; Li, X.; Lu, X. Hierarchical recurrent neural network for video summarization. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017.

[23] Rochan, M.; Ye, L.; Wang, Y. Video summarization using fully convolutional sequence networks. In Proceedings of the Computer Vision–ECCV: 15th European Conference, Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.

[24] Yuan, L.; Tay, F.E.; Li, P.; Zhou, L.; Feng, J. Cycle-sum: Cycle-consistent adversarial lstm networks for unsupervised video summarization. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.

[25] Zhang, K.; Grauman, K.; Sha, F. Retrospective encoders for video summarization. In Proceedings of the Computer Vision–ECCV: 15th European Conference, Munich, Germany, 8–14 September 2018; Springer: Cham, Switzerland, 2018.

[26] Zhou, K.; Qiao, Y.; Xiang, T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In Proceedings of the Association for the Advancement of Artificial Intelligence Conference (AAAI), New Orleans, LA, USA, 2–7 February 2018.

[27] Narasimhan, M.; Rohrbach, A.; Darrell, T. Clip-it! Language-Guided Video Summarization. *Adv. Neural Inf. Process. Syst.* **2021**; *34*, 13988–14000.

[28] Plummer, B.A.; Brown, M.; Lazebnik, S. Enhancing video summarization via vision-language embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.

[29] Otani, M.; Nakashima, Y.; Rahtu, E.; Heikkilä, J. Rethinking the evaluation of video summaries. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.

Type = Article

[30] de Curtò, J.; de Zarzà, I.; Yan, H.; Calafate, C.T. Signature and Logsignature for the Study of Empirical Distributions Generated with GANs. *arXiv* **2022**, arXiv:2203.03226.

[31] Lyons, T. Rough paths, signatures and the modelling of functions on streams. *arXiv* **2014**, arXiv:1405.4537.

[32] Bonnier, P.; Kidger, P.; Arribas, I.P.; Salvi, C.; Lyons, T. Deep signature transforms. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December2019; Volume 32.

[33] Chevyrev, I.; Kormilitzin, A. A primer on the signature method in machine learning. *arXiv* **2016**, arXiv:1603.03788.

[34] Kidger, P.; Lyons, T. Signatory: Differentiable computations of the signature and logsignature transforms, on both CPU and GPU. *arXiv* **2020**, arXiv:2001.00706.

[35] Liao, S.; Lyons, T.J.; Yang, W.; Ni, H. Learning stochastic differential equations using RNN with log signature features. *arXiv* **2019**, arXiv:1908.0828.

[36] Morrill, J.; Kidger, P.; Salvi, C.; Foster, J.; Lyons, T.J. Neural CDEs for long time series via the log-ode method. *arXiv* **2021**, arXiv:2009.08295.

[37] Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.

[38] Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2022**, arXiv:2104.13921.

[39] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[40] de Curtò, J.; de Zarzà, I.; Calafate, C.T. Semantic scene understanding with large language models on unmanned aerial vehicles. *Drones* **2023**, *7*, 114. [CrossRef]

[41] Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J. et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020.

[42] Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the 38th International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8821–8831.

[43] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.

[44] Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Lopes, R.G.; et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv* **2022**, arXiv:2205.11487.

[45] Cui, Y.; Niekum, S.; Gupta, A.; Kumar, V.; Rajeswaran, A. Can foundation models perform zero-shot task specification for robot manipulation? In Proceedings of the Learning for Dynamics and Control Conference, Palo Alto, CA, USA, 23–24 June 2022 .

[46] Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A universal visual representation for robot manipulation. *arXiv* **2022**, arXiv:2203.12601.

[47] Zeng, A.; Florence, P.; Tompson, J.; Welker, S.; Chien, J.; Attarian, M.; Armstrong, T.; Krasin, I.; Duong, D.; Wahid, A.; et al. Transporter networks: Rearranging the visual world for robotic manipulation. In Proceedings of the Conference on Robot Learning, Online, 15–18 November 2020.

[48] Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. *arXiv* **2022**, arXiv:2201.07207.

[49] Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv* **2022**, arXiv:2204.00598.

[50] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 ×16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.

[51] Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. Simple open-vocabulary object detection with vision transformers. *arXiv* **2022**, arXiv:2205.06230.

[52] Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python*, 1st ed.; O'Reilly Media, Inc.: Sebastopol, CA, USA, 2009.

Chapter 7

Discussion

The chapters of this thesis build upon a common thread - the utilisation of LLMs and related techniques to solve multifaceted problems [23, 24, 25], ranging from scene understanding, decision optimization under uncertainty, and GANs evaluation to video summarization. The body of research within this thesis elucidates the transformative capacity of LLMs and AI in reshaping the paradigms of these diverse domains.

In Chapter 3, we ventured into the realm of semantic scene understanding with UAVs. The sophisticated amalgamation of LLMs with state-of-the-art object detection algorithms and a RIZE Tello drone yielded insightful scene descriptions in real-time, lending potential applications to surveillance, search and rescue, and environmental monitoring. However, this research also paves the way for further enhancements, such as the inclusion of additional sensors, refining captioning algorithms, and optimising the drone's trajectory for improved scene descriptions. Furthermore, real-time analysis could amplify the system's effectiveness for immediate alerts and updates. By achieving a balance between the drone's autonomy, LLMs' language understanding capabilities, and other potential enhancements, the goal is to equip autonomous systems like UAVs and self-driving cars with human-like literary capabilities.

Chapter 4 tackled the complex problem of non-stationary multi-armed bandits [26] using the predictive prowess of LLMs. The use of LLMs to augment traditional RL strategies presented a novel and dynamic approach to decision-making [27, 28]. Nonetheless, further research is warranted to refine the strategy recommendation process and potentially integrate other RL strategies within the LLM-informed framework. There also lies potential in extending this LLM-informed strategy to real-world domains like personalized healthcare and financial trading systems, albeit with unique challenges such as ensuring the interpretability of the decision-making process and effectively handling domain-specific information [29].

In Chapter 5, the focus shifted to assessing GAN convergence and the goodness of fit using the Signature Transform. This novel methodology introduces a reliable, consistent, and efficient measure to evaluate GAN performance, and also any other generative model capable of producing high-fidelity imagery such as Stable Diffusion [30]. Despite offering a robust alternative to existing GAN evaluation methods, this approach is not without limitations. Future work could focus on addressing these, such as improving sensitivity to fine-grained image details, optimizing computation for large datasets or high-resolution images, and extending the approach to handle non-stationary data behavior. The assessment of GANs using the Signature Transform presents an exciting prospect for future research, aiming to further optimize GAN evaluation and extend the proposed metrics to other applications.

Finally, Chapter 6 presented a new benchmark based on the Signature Transform for evaluating visual summaries, while at the same time proposed an innovative technique based on LLMs and VLMs for summarization of videos. While it demonstrated high correlation with the human notion of a good summary, it also highlighted a challenge - achieving a balance between accuracy, computational efficiency, and the generation of context-aware, meaningful summaries. Future research could address this, potentially by developing techniques that combine the efficiency of Signature Transform-based metrics with the intelligence of methodologies based on Foundation Models [31]. This approach would foster more context-aware and meaningful summarization, taking into account factors like human-identified objects and storyline consistency.

Overall, the four chapters showcase the transformative power of LLMs across various domains. While significant strides have been made [32], each chapter also uncovers new avenues for future research, offering an exciting prospect for

further exploration in this rapidly evolving field. Each of these research works provides us with not just a wealth of understanding but also leaves us with thought-provoking questions and future directions to explore - a testament to the ever-evolving nature of AI.

The research presented in this doctoral thesis demonstrates a broad spectrum of advanced methodologies and their implications in various domains. A comparative view helps in understanding the essence of each chapter, their methodologies, main contributions, potential applications, and prospective directions for future research. Table 7.1 encapsulates this comparative analysis for Chapters 3 to 6.

To sum up, Chapter 3 introduced the concept of UAVs, like the RIZE Tello drone, narrating real-time literary stories, combining aviation, computer vision, and AI. This development hints at a future where autonomous vehicles not only perceive but also narrate their observations in human-friendly formats. Chapter 4 emphasized the potential of LLMs in decision-making, such as the capabilities of GPT-3.5-turbo. Despite their advantages, there's a need for ethical application and maintaining a human touch in decisions. Chapter 5 delved into GAN evaluation, presenting a new method to assess GAN convergence while acknowledging the need for ongoing adaptability given the rising quality of synthetic images. Chapter 6 examined evaluating visual summaries using the Signature Transform, setting benchmarks and guiding future research towards innovative summarization methods that bring tangible value.

Table 7.1: Comparative Analysis of Chapters 3 to 6

Chapter	**Techniques & Methodologies Used**	**Main Contributions/Findings**	**Potential Applications**	**Future Research Directions**
Chapter 3	RIZE Tello drone, LLMs, YOLOv7, GPT-3	Zero-shot UAV literary storytelling with state-of-the-art accuracy	Surveillance, search & rescue, environmental monitoring	Drone trajectory optimization, integration of more sensors, fine-tuning algorithms, domain-specific applications
Chapter 4	Integration of LLMs in decision-making	Enhanced strategy recommendation with LLMs, bridging traditional RL strategies	Healthcare, financial systems, decision-making in dynamic situations	Improving strategy recommendation, integration of other RL strategies, application in specific domains
Chapter 5	Signature Transform for GAN evaluation	Efficient GAN Synthetic image quality assessment with reduced computation	GAN evaluation, image quality assessment	Enhance descriptor complexity, use in other tasks, integrate into training loops
Chapter 6	Signature Transform for video summaries	Effective video summarization capabilities; methods for evaluating summarization	Video summarization, visual content analysis	Develop smarter techniques considering storyline, optimize for non-uniform content

7.1 Contributions

To best encapsulate the rich tapestry of contributions derived from this doctoral work, we must step back and recognize the unifying metric, algorithmic, and methodological advancements it has heralded. The exploration into the UAV domain witnessed a pioneering intersection of LLMs and VLMs with real-world UAV applications. The groundbreaking outcome was the literary scene descriptions, which registered a GUNNING Fog median grade level spanning 7-12, effectively humanizing the mechanical interpretations of UAV-captured scenes. Transcending into the terrain of decision-making, we unraveled a novel nexus between LLMs and the MAB problem in non-stationary environments. Experimental evaluations reflected the dexterity of our LLM-informed strategy in navigating the capricious nature of the problem. Advancing to our journey in understanding GANs, we leveraged the mathematical prowess of the Signature Transform, facilitating a powerful metric—through RMSE and MAE Signature—to diagnose GAN convergence with efficacy rivalling conventional GPU-intensive methods, but with a remarkable tilt towards computational efficiency. Notably, the discriminative potential of our method was evidenced by PCA and t-SNE visual representations. Lastly, our strides in video summarization converged with the Signature Transform, introducing a new benchmark for visual summary assessment sans human annotation. Through a systematic approach, we authenticated the Signature Transform's resonance with human perceptions of summary quality.

Evidently, each chapter is not a mere siloed contribution; they collectively echo the paramount potential of LLMs, their synergies with existing computational paradigms, and the myriad of applications waiting to be enhanced through such interdisciplinary mergers.

Chapter 8

Conclusions

8.1 Concluding Remarks

The intricate interplay of AI and LLMs in complex decision-making, scene understanding, and summarization tasks, as demonstrated in this thesis, marks a significant advancement in the field of AI. The journey through each chapter of this thesis underscores the transformative power of LLMs, serving as a testament to the pivotal role they play in pushing the frontiers of research.

In conclusion, this body of work establishes LLMs as powerful tools in addressing complex tasks across diverse domains, pushing the boundaries of conventional AI applications. While it unravels groundbreaking pathways in leveraging LLMs, it also leaves us with thought-provoking questions and future directions. The implications of this research are far-reaching, setting the stage for continued exploration and innovation in the field. It underscores the importance of further research in this domain and paves the way for deeper insights, innovative applications, and novel solutions to complex problems. As we continue to tread on the path of AI research, the findings of this thesis serve as a beacon, guiding us towards unexplored terrains and promising possibilities.

In this opus of interdisciplinary exploration, the confluence of mathematical formalism and state-of-the-art computational models has elucidated paths previously obscured in the domain of AI and its applications. The mathematical elegance of tools, be it the Signature Transform's differential geometric roots or the statistical sophistication embedded in MAB theory, is emblematic of a deeper truth: AI's future is inherently intertwined with the profound axioms of mathematical sciences. As evidenced in our exploration of LLMs with UAVs, the harmony achieved between rigorous object detection algorithms, represented as $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the input image space and $\mathcal{Y}$ the semantic description, and nuanced linguistic models, testifies to this profound union. Similarly, the MAB's strategic deployment, harnessing LLMs, can be envisaged as a dynamical system with reward feedback loops, modeled by the stochastic differential equation $dX_t = b(X_t, t)dt + \sigma(X_t, t)dW_t$. Here, X_t represents the bandit's state, while b and σ encapsulate the drift and volatility contingent on LLM guidance, and W_t is a Brownian motion. The confluence with GANs and video summarization further consolidates this sentiment, indicating a mathematical underpinning as the fulcrum for AI's next paradigm shift. As this narrative unfolds, it beckons a future where mathematical rigor and AI innovation walk hand-in-hand, perpetually pushing the boundaries of what machines can comprehend, elucidate, and innovate. With this groundwork, the horizon promises an era where foundational models not only evolve to mirror human cognition but do so grounded in the immutable laws of mathematics, forging a potent scaffold for significant advancements.

8.2 Publications and Related Works

The extensive research conducted during this doctoral study is manifested in a collection of articles published in reputable journals and conferences. The following is a list of these scholarly contributions, organized by their direct relation to the topics explored in the thesis and related works that bear relevance to the overarching themes.

8.2.1 Publications Included in the Thesis:

1. [33] de Curtò, J., de Zarzà, I., & Calafate, C. T. (2023). "Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles." Drones, vol(7), 114. DOI: 10.3390/drones7020114. IF 2022: 4.8; SCImago-SJR: 1st Quartile.

2. [34] de Curtò, J., de Zarzà, I., Roig, G., Manzoni, P., & Calafate, C. T. (2023). "LLM- Informed Multi-Armed Bandit Strategies for Non-Stationary Environments." Electronics, vol(12), 2814. DOI: 10.3390/electronics12132814. IF 2022: 2.9; JCR: 2nd Quartile.

3. [35] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. (2023). "Signature and Log-Signature for the Study of Empirical Distributions Generated with GANs." Electronics, vol(12), 2192. DOI: 10.3390/electronics12102192. IF 2022: 2.9; JCR: 2nd Quartile.

4. [36] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. (2023). "Summarization of Videos with the Signature Transform." Electronics, vol(12), 1735. DOI: 10.3390/electronics12071735. IF 2022: 2.9; JCR: 2nd Quartile.

8.2.2 Related Publications:

In addition to these publications, other works closely related to the thesis's themes were also conducted. They provide supplementary insights into the main thesis topics, contributing to a more comprehensive understanding of the applied methodologies.

1. [37] de Curtò, J., de Zarzà, I., & Calafate, C. T. (2023). "UWB and MB-OFDM for Lunar Rover Navigation and Communication" Mathematics, vol(11), 3835. DOI: 10.3390/math11183835. IF 2022: 2.4; JCR: 1st Quartile.

2. [38] de Zarzà, I., de Curtò, J., Cano, J. C., & Calafate, C. T. (2023) "Drone-Based Decentralized Truck Platooning with UWB Sensing and Control" Mathematics, vol(11), 4627. DOI: 10.3390/math11224627. IF 2022: 2.4; JCR: 1st Quartile.

3. [39] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. (2023). "LLM Adaptive PID Control for B5G Truck Platooning Systems" Sensors, vol(23), 5899. DOI: 10.3390/s23135899. IF 2022: 3.9; SCImago-SJR: 1st Quartile.

4. [40] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. (2023). "LLM Multimodal Traffic Accident Forecasting" Sensors, vol(23), 9225. DOI: 10.3390/s23229225. IF 2022: 3.9; SCImago-SJR: 1st Quartile.

5. [41] de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P., & Calafate, C. T. (2023). "Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs." Electronics, vol(12), 2722. DOI: 10.3390/electronics12122722. IF 2022: 2.9; JCR: 2nd Quartile.

6. [42] de Zarzà, I., de Curtò, J., Hernández-Orallo, E., & Calafate, C. T. (2023). "Cascading and Ensemble Techniques in Deep Learning." Electronics, vol(12), 3354. DOI: 10.3390/electronics12153354. IF 2022: 2.9; JCR: 2nd Quartile.

7. [43] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "Optimizing Neural Networks for Imbalanced Data." Electronics, vol(12), 2674. DOI: 10.3390/electronics12122674. IF 2022: 2.9; JCR: 2nd Quartile.

8. [44] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2022). "Detection of glaucoma using three-stage training with EfficientNet." Intelligent Systems with Applications, vol(16), 200140.
DOI: 10.1016/j.iswa.2022.200140. SCImago-SJR: 1st Quartile.

9. [45] de Curtò, J., de Zarzà, Yan, H., & Calafate, C. T. (2022). "On the applicability of the Hadamard as an input modulator for problems of classification." Software Impacts, vol(13), 100325.
DOI: 10.1016/j.simpa.2022.100325 IF 2022: 2.1; JCR: 3rd Quartile.

8.2.3 Conference Papers:

Furthermore, a series of conference papers have been contributed, further enriching the research presented in this thesis:

1. [46] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "Socratic Video Understanding on Unmanned Aerial Vehicles." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pending assignment. CORE B.

2. [47] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "Area Estimation of Forest Fires using TabNet with Transformers." 27th International Conference on Knowledge Based and Intelligent information and Engineering

Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pending assignment. CORE B.

3. [48] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "UMAP for Geospatial Data Visualization." 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023), Athens, Greece, 6–8 September, 2023. DOI: pending assignment. CORE B.

8.2.4 Workshop Papers:

Additionally, the research also included the presentation of key findings in international workshops, emphasizing the practical implications of the methods and models developed throughout the thesis.

1. [49] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "Decentralized Platooning Optimization for Trucks: A MILP and ADMM-based Convex Approach to Minimize Latency and Energy Consumption" 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023), Hong Kong. July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00031. CORE A.

2. [50] de Zarzà, I., de Curtò, J., & Calafate, C. T. (2023). "Decentralized Planning of Platoons in Road Transport using Reinforcement Learning" 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023), Hong Kong. July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00030. CORE A.

These publications and conference papers constitute the breadth of research conducted during this study, highlighting the multifaceted aspects of LLMs and their applications.

8.2.5 Future Work

The research journey outlined in this thesis has brought forth several promising opportunities for future investigation. The multidimensional versatility and scope of LLMs, their interplay with different computational approaches, and their potential to redefine various fields form an exciting frontier for future research. Following are some prospective directions:

- **Integration of other modalities with LLMs:** Future work could explore the combination of LLMs with more complex sensor data, such as LiDAR, RADAR, and other multispectral data. This would allow for a more detailed and richer understanding of the environment, significantly improving autonomous decision-making capabilities in complex scenarios.
- **Enhancing LLM-informed strategies:** Refinement of the MAB approach by integrating more advanced RL techniques, as well as refining the interpretation of LLM advice, could lead to more robust and efficient decision-making systems.
- **Domain-specific LLM applications:** Exploring domain-specific applications, like healthcare, financial trading, and climate modeling, where LLMs can provide enhanced decision-making capacities, can revolutionize these fields and create impactful solutions.
- **Improving the Signature Transform:** Research on enhancing the Signature Transform's capabilities to capture the fine-grained details of images could lead to a more accurate assessment of GANs and other generative models able to produce high-fidelity images, pushing the boundaries of the present state of the art.
- **Novel metrics for video summarization:** The pursuit of more intelligent summarization techniques that not only perform well according to proposed metrics but also account for the human concept of a good summary could lead to groundbreaking methodologies in the field of video summarization.
- **Further explorations in AI ethics and bias:** As LLMs continue to gain prominence, further investigations into the ethical considerations, transparency, and potential biases within these models are crucial. Strategies to ensure fair, unbiased AI systems need to be devised and continually refined.

While the trajectory of this research has already yielded considerable results, the scope of possibilities is immense. The inherent complexities and continual advancements in the field ensure that the road ahead will be challenging but, without a doubt, intellectually rewarding.

8.3 Synthesis of Contributions

The collective contributions of this thesis, as manifested in the peer-reviewed articles published in respected journals, present a coherent narrative of innovation and advancement within the field of AI, particularly in the application and enhancement of LLMs. This body of work not only highlights the capabilities of LLMs in a variety of complex tasks, but also illustrates the transformative potential of AI when applied with precision and creativity.

8.3.1 Integrated Contributions

- The integration of LLMs with UAV technology, as detailed in the studies, has not only advanced the field of autonomous aerial surveillance but has also set a precedent for real-time, context-aware scene understanding and description.
- The novel strategies informed by LLMs for decision-making in non-stationary environments have provided a framework that could redefine strategic optimization in dynamic contexts, potentially influencing sectors as diverse as finance, logistics, and network security.
- The application of the Signature Transform to assess GANs has introduced an efficient and robust methodology that could revolutionize the evaluation of generative models, impacting areas such as synthetic data generation and multimedia content creation.
- The development of new benchmarks for video summarization based on the Signature Transform has charted a course for future research in efficient data compression and retrieval, with implications for surveillance, entertainment, and education.

8.3.2 Broader Implications

The implications of these integrated contributions are profound, both from a theoretical and a practical standpoint. Theoretically, this work has extended the understanding of how LLMs can be synergized with other computational models to address tasks that require a deep fusion of linguistic and visual information processing. Practically, the applications demonstrated in these studies not only offer immediate benefits but also serve as a foundation for future developments that could significantly alter the technological landscape.

8.3.3 Collective Impact

The collective impact of the studies conducted as part of this thesis transcends the sum of its parts. When viewed together, they represent a significant stride towards creating autonomous systems that are not only more intelligent and efficient but also more aligned with human ways of understanding and interacting with the world. This alignment, achieved through the interpretability and descriptiveness afforded by LLMs, is a step towards more naturalistic and accessible AI systems.

In essence, the research presented in this thesis, validated by the scientific community through peer-review, stands as a beacon of innovation. It provides valuable insights into the capabilities of modern AI and lays down a roadmap for future explorations that will undoubtedly expand the reach and efficacy of autonomous systems.

Acronyms

AI	Artificial Intelligence
AMT	Amazon Mechanical Turk
CLIP	Contrastive Language-Image Pre-training
CPU	Central Processing Unit
DNN	Deep Neural Networks
DL	Deep Learning
FID	Fréchet Inception Distance
GAN	Generative Adversarial Networks
GPT	Generative Pre-training Transformer
GPU	Graphics Processing Unit
KL	Kullback–Leibler
LiDAR	Light Detection And Ranging
LLM	Large Language Models
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAB	Multi-Armed Bandit
ML	Machine Learning
MOS	Mean Opinion Score
MS-SSIM	Structural Similarity Index Measure
NLP	Natural Language Processing
PCA	Principal Component Analysis
PSNR	Peak Signal-to-Noise Ratio
QLO	Quantized Low-Rank Adapters
RADAR	Radio Detection And Ranging
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
t-SNE	t-Distributed Stochastic Neighbor Embedding
UCB	Upper Confidence Bound
VLM	Visual Language Models
YOLO	You Only Look Once

Bibliography

[1] Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2021**, arXiv:2010.11929.

[2] Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, arXiv:2206.07682.

[3] Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021.

[4] Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. Flamingo: A visual language model for few-shot learning. *arXiv* **2022**, arXiv:2204.14198.

[5] Gu, X.; Lin, T.-Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv* **2022**, arXiv:2104.13921.

[6] Nair, S.; Rajeswaran, A.; Kumar, V.; Finn, C.; Gupta, A. R3M: A universal visual representation for robot manipulation. *arXiv* **2022**, arXiv:2203.12601.

[7] Huang, W.; Abbeel, P.; Pathak, D.; Mordatch, I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MA, USA, 17–23 July 2022.

[8] Saharia, C.; Chan, W.; Chang, H.; Lee, C.; Ho, J.; Salimans, T.; Fleet, D.; Norouzi, M. Palette: Image-to-image diffusion models. In Proceedings of the ACM SIGGRAPH 2022, Vancouver, BC, Canada, 7–11 August 2022; pp. 1–10.

[9] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; Volume 27.

[10] Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 6629–6640.

[11] Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.

[12] Chen, K.-T. Iterated path integrals. *Bull. Am. Math. Soc.* **1977**, *83*, 831–879. [CrossRef]

[13] Lyons, T. Rough paths, signatures and the modelling of functions on streams. In Proceedings of the International Congress of Mathematicians, Madrid, Spain, 22–30 August 2014.

[14] Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; et al. Simple open-vocabulary object detection with vision transformers. *arXiv* **2022**, arXiv:2205.06230.

[15] Gygli, M.; Grabner, H.; Riemenschneider, H.; Van Gool, L. (2014). Creating summaries from user videos. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September; Springer: Berlin/Heidelberg, Germany, 2014.

[16] Mundur, P.; Rao, Y.; Yesha, Y. Keyframe-based video summarization using delaunay clustering. *International journal on digital libraries* **2006**, *6*, 219–232.

[17] Taskiran, C. M.; Pizlo, Z.; Amir, A.; Ponceleon, D.; Delp, E. J. Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia* **2006**, *8*(4), 775–791.

[18] Liang, G.; Lv, Y.; Li, S.; Zhang, S.; Zhang, Y. Video summarization with a convolutional attentive adversarial network. *Pattern Recognit.* **2022**, *131*, 108840. [CrossRef]

[19] Zhu, W.; Lu, J.; Han, Y.; Zhou, J. Learning multiscale hierarchical attention for video summarization. *Pattern Recognit.* **2022**, *122*, 108312. [CrossRef]

[20] Zeng, A.; Florence, P.; Tompson, J.; Welker, S.; Chien, J.; Attarian, M.; Armstrong, T.; Krasin, I.; Duong, D.; Wahid, A.; et al. Transporter networks: Rearranging the visual world for robotic manipulation. *arXiv* **2022**, arXiv:2010.14406.

[21] Zeng, A.; Attarian, M.; Ichter, B.; Choromanski, K.; Wong, A.; Welker, S.; Tombari, F.; Purohit, A.; Ryoo, M.; Sindhwani, V.; et al. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv* **2022**, arXiv:2204.00598.

[22] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, and others. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *arXiv:2307.15818*, 2023.

[23] Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Visual Language Maps for Robot Navigation. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023.

[24] Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685.

[25] Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv* **2023**, arXiv:2305.14314.

[26] Cavenaghi, E.; Sottocornola, G.; Stella, F.; Zanker, M. Non stationary multi-armed bandit: Empirical evaluation of a new concept drift-aware algorithm. *Entropy* **2021**, *23*, 380. [CrossRef]

[27] Cesa-Bianchi, N.; Lugosi, G. *Prediction, Learning, and Games*; Cambridge University Press: Cambridge, UK, 2017.

[28] Oroojlooy, A.; Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. *arXiv* **2022**, arXiv:1908.03963.

[29] Wei, J.; Bosma, M.; Zhao, V.; Guu, K.; Yu, A.W.; Lester, B.; Du, N.; Dai, A.M.; Le, Q.V. Finetuned Language Models are Zero-Shot Learners. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.

[30] Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.

[31] Wortsman, M.; Dettmers, T.; Zettlemoyer, L.; Morcos, A.; Farhadi, A.; Schmidt, L. Stable and low-precision training for large-scale vision-language models. *arXiv* **2023**, arXiv:2304.13013.

[32] Shah, D.; Osiński, B.; ichter, B.H.; Levine, S. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In Proceedings of the 6th Conference on Robot Learning, Proceedings of Machine Learning Research, PMLR, Atlanta, GA, USA, 6–9 November 2023; Volume 205, pp. 492–504.

[33] de Curtò, J., de Zarzà, I., & Calafate, C. T. Semantic Scene Understanding with Large Language Models on Unmanned Aerial Vehicles. *Drones*, 7:114, 2023. DOI: 10.3390/drones7020114.

[34] de Curtò, J., de Zarzà, I., Roig, G., Manzoni, P., & Calafate, C. T. LLM-Informed Multi-Armed Bandit Strategies for Non-Stationary Environments. *Electronics*, 12:2814, 2023. DOI: 10.3390/electronics12132814.

[35] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. Signature and Log-Signature for the Study of Empirical Distributions Generated with GANs. *Electronics*, 12:2192, 2023. DOI: 10.3390/electronics12102192.

[36] de Curtò, J., de Zarzà, I., Roig, G., & Calafate, C. T. Summarization of Videos with the Signature Transform. *Electronics*, 12:1735, 2023. DOI: 10.3390/electronics12071735.

[37] de Curtò, J., de Zarzà, I., & Calafate, C. T. UWB and MB-OFDM for Lunar Rover Navigation and Communication *Mathematics*, 11:3835, 2023. DOI: 10.3390/math11183835.

[38] de Zarzà, I.; de Curtò, J.; Cano, J.C.; Calafate, C.T. Drone-Based Decentralized Truck Platooning with UWB Sensing and Control. *Mathematics*, 11:4627, 2023. DOI: 10.3390/math11224627.

[39] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. LLM Adaptive PID Control for B5G Truck Platooning Systems. *Sensors*, 23:5899, 2023. DOI: 10.3390/s23135899.

[40] de Zarzà, I., de Curtò, J., Roig, G., & Calafate, C. T. LLM Multimodal Traffic Accident Forecasting. *Sensors*, 23:9225, 2023. DOI: 10.3390/s23229225.

[41] de Zarzà, I., de Curtò, J., Roig, G., Manzoni, P., & Calafate, C. T. Emergent Cooperation and Strategy Adaptation in Multi-Agent Systems: An Extended Coevolutionary Theory with LLMs. *Electronics*, 12:2722, 2023. DOI: 10.3390/electronics12122722.

[42] de Zarzà, I., de Curtò, J., Hernández-Orallo, E., & Calafate, C. T. Cascading and Ensemble Techniques in Deep Learning. *Electronics*, 12:3354, 2023. DOI: 10.3390/electronics12153354.

[43] de Zarzà, I., de Curtò, J., & Calafate, C. T. Optimizing Neural Networks for Imbalanced Data. *Electronics*, 12:2674, 2023. DOI: 10.3390/electronics12122674.

[44] de Zarzà, I., de Curtò, J., & Calafate, C. T. Detection of glaucoma using three-stage training with EfficientNet. *Intelligent Systems with Applications*, 16:200140, 2022. DOI: 10.1016/j.iswa.2022.200140.

[45] de Curtò, J., de Zarzà, I., Yan, H., & Calafate, C. T. On the applicability of the Hadamard as an input modulator for problems of classification. *Software Impacts*, 13:100325, 2022. DOI: 10.1016/j.simpa.2022.100325.

[46] de Zarzà, I., de Curtò, J., & Calafate, C. T. Socratic Video Understanding on Unmanned Aerial Vehicles. In *Proceedings of the 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023)*, Athens, Greece, 6–8 September, 2023.

[47] de Zarzà, I., de Curtò, J., & Calafate, C. T. Area Estimation of Forest Fires using TabNet with Transformers. In *Proceedings of the 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023)*, Athens, Greece, 6–8 September, 2023.

[48] de Zarzà, I., de Curtò, J., & Calafate, C. T. UMAP for Geospatial Data Visualization. In *Proceedings of the 27th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2023)*, Athens, Greece, 6–8 September, 2023.

[49] de Zarzà, I., de Curtò, J., & Calafate, C. T. Decentralized Platooning Optimization for Trucks: A MILP and ADMM-based Convex Approach to Minimize Latency and Energy Consumption. In *Proceedings of the 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023)*, Hong Kong, July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00031.

[50] de Zarzà, I., de Curtò, J., & Calafate, C. T. Decentralized Planning of Platoons in Road Transport using Reinforcement Learning. In *Proceedings of the 6th International Workshop on Vehicular Networking and Intelligent Transportation Systems (VENITS 2023)*, Hong Kong, July 18, 2023. Held in conjunction with the 43rd IEEE International Conference on Distributed Computing Systems (ICDCS), Hong Kong, 18–21 July, 2023. DOI: 10.1109/ICDCSW60045.2023.00030.