COMILLAS
UNIVERSIDAD PONTIFICIA
ICAI

GRADO EN INGENIERÍA MATEMÁTICA E INTELIGENCIA ARTIFICIAL

TRABAJO FIN DE GRADO

CoGuide: Guía de Difusión Contrastiva para Problemas Inversos Espaciales

Author: Jorge Vančo Sampedro

Director: Jaime Pizarroso Gonzalo

Co-Director: Simón Rodríguez Santana

Madrid, May 2026

Declaration of originality

I declare under my responsibility that the Project presented with the title **CoGuide: Contrastive Diffusion Guidance for Spatial Inverse Problems** at the ICAI School of Engineering of the Comillas Pontifical University in the academic year 2025/2026 is of my authorship and has not been presented previously for other purposes. The Project is not plagiarised from any other, either totally or partially, and the information that has been taken from other documents is duly referenced.

Use of Artificial Intelligence[1]

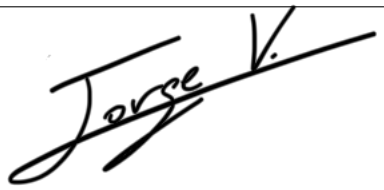
I declare under my responsibility (indicate the correct option):

☐ I have not used Artificial Intelligence in the preparation of this document.

☒ I have used Artificial Intelligence in the preparation of this document and/or Annex B under the conditions allowed by Comillas Pontifical University, i.e. applying Level 2 of the Perkins et al. (2024) Assessment Scale: *"AI can be used for pre-task activities such as brainstorming, description and initial research. This level focuses on the use of AI for planning, synthesising and generating ideas, but assessments should emphasise the ability to develop and refine these ideas independently"*. Specifically, Artificial Intelligence has been used to:

Artificial Intelligence has been used for brainstorming and broadening the search space for resources and materials. It has also been used for LaTeX formatting, table creation, grammar checking, and ensuring correctness.

[1]This declaration refers to the use of generative Artificial Intelligence to carry out the Project documents (Annex B and Memory). It does not apply to Projects where, by their nature, artificial intelligence must be used as part of them (application of machine learning techniques, neural networks, data analysis...).


Signature: Jorge Vančo Sampedro
Date: 29/05/2026

Authorisation for Project delivery

Thesis Supervisor	Thesis Deputy Supervisor (if applicable)
	
Signature: Jaime Pizarroso Gonzalo	Signature: Simón Rodríguez Santana
Date: 30/05/2026	Date: 30/05/2026

Acknowledgments

COGUIDE: GUÍA DE DIFUSIÓN CONTRASTIVA PARA PROBLEMAS INVERSOS ESPACIALES

Autor: Jorge Vančo Sampedro
Director: Jaime Pizarroso Gonzalo
Co-Director: Simón Rodríguez Santana

Resumen

Los problemas inversos buscan recuperar señales limpias a partir de medidas corruptas. Sin embargo, métodos actuales como el muestreo a posteriori por difusión (DPS) fallan drásticamente cuando el operador de degradación es desconocido. Este proyecto desarrolla un marco generativo robusto basado en aprendizaje contrastivo (CL-DPS) para resolver problemas inversos ciegos. Entrenando un modelo fundacional (MoCo) que proyecta múltiples degradaciones visuales en un espacio latente continuo, aproximamos la función de verosimilitud sin requerir parámetros matemáticos del operador. Los resultados demuestran que, frente a la inestabilidad del DPS ante operadores erróneos, CL-DPS mantiene una alta fidelidad perceptiva, logrando reconstrucciones estables e invariantes al desajuste paramétrico.

Palabras clave: Problemas Inversos Ciegos; Modelos de Difusión; Aprendizaje Contrastivo; Restauración de Imágenes; Sustituto de Verosimilitud.

Resumen ejecutivo

1 Introducción

En campos tan diversos como la imagen médica, la astrofísica o la robótica espacial, los investigadores rara vez observan un fenómeno de forma directa. En su lugar, se enfrentan a problemas inversos: el reto matemático de deducir una señal original a partir de una medida indirecta, ruidosa o degradada. Recientemente, los modelos de difusión [1] han surgido como potentes herramientas (prioris) para resolver estos problemas.

El método estándar actual, conocido como Muestreo a Posteriori por Difusión (DPS) [2], depende de un gradiente analítico para guiar la generación de la imagen paso a paso. Para que este gradiente se pueda calcular, se asume que el proceso de degradación (el "operador directo") es perfectamente conocido y diferenciable. Sin embargo, en el mundo real, parámetros como la severidad de un desenfoque o la trayectoria del movimiento de una cámara son desconocidos. Cuando se enfrenta a estos problemas inversos ciegos, el DPS estándar colapsa, introduciendo artefactos severos que destruyen la imagen.

2 Objetivos

El objetivo principal de esta tesis es validar **CL-DPS** (Contrastive Learning for Diffusion Posterior Sampling), un marco generativo capaz de resolver problemas inversos ciegos utilizando una verosimilitud basada en un espacio latente contrastivo [3], [4].

- **Formular el Problema Inverso Ciego:** Cuantificar la vulnerabilidad del DPS estándar ante el desajuste paramétrico en diversas familias de degradación visual (desenfoques gaussianos, direccionales y rotacionales).
- **Desarrollar el Mecanismo de Guía Contrastiva:** Diseñar y entrenar un modelo fundacional (codificador ResNet-50) utilizando la arquitectura *Momentum Contrast* (MoCo) [5] para mapear imágenes limpias y degradadas sintéticamente en un espacio latente unificado. Primero se entrenan modelos específicos para cada operador y después se valida el modelo entrenado en conjunto, además de entrenar un enrutador a cada experto.
- **Evaluación y Benchmarking:** Validar los resultados usando métricas informativas. Estudiar las diferencias en rendimiento de los distintos métodos y describir puntos fuertes y débiles del nuevo método.

3 Metodología

Para resolver esta vulnerabilidad, este proyecto implementa y optimiza el marco CL-DPS [4]. La solución elimina la dependencia de operadores matemáticos exactos dividiendo el problema en dos fases:

1. **Entrenamiento Offline de MoCo (Fig. 1 Phase 1):** En lugar de enseñar a una red a deshacer un desenfoque específico, entrenamos un codificador (ResNet-50) utilizando la arquitectura Momentum Contrast [5]. Se alimenta al modelo con imágenes limpias y sus correspondientes versiones degradadas sintéticamente (con desenfoques gaussianos, direccionales y rotacionales de intensidad aleatoria). Utilizando la pérdida InfoNCE [6], la red aprende a agrupar imágenes que comparten la misma identidad semántica, independientemente de cómo hayan sido degradadas. Para evitar pérdidas de color, se incluyó una cabecera de consistencia de color (CCH) guiada por el error cuadrático medio.
2. **Inferencia Guiada por Contraste (Fig. 1 Phase 2):** Durante el proceso de difusión inversa, en lugar de utilizar el error píxel a píxel, extraemos parches de la imagen generada en el instante t y los comparamos con la medida real en el espacio latente de MoCo. El gradiente de la similitud del coseno se utiliza para empujar la imagen generada hacia la identidad correcta.

Para evaluar la capacidad de generalización del marco en entornos completamente ciegos, se diseñaron dos arquitecturas de validación: un sistema dinámico basado en un enrutador que clasifica la degradación y asigna el problema a modelos MoCo "expertos" preentrenados en operadores específicos; y un codificador MoCo "Fundacional" (o universal), entrenado simultáneamente con múltiples familias de operadores para mapear cualquier degradación en un espacio latente unificado sin necesidad de clasificación previa.

Como mejora fundamental sobre la literatura existente, este trabajo identificó que guiar la difusión directamente sobre el estado ruidoso x_t genera artefactos estructurales catastróficos. La solución implementada consistió en evaluar la función de pérdida contrastiva sobre la predicción sin ruido de Tweedie ($\hat{x}_0$) [7], restaurando completamente la coherencia visual.

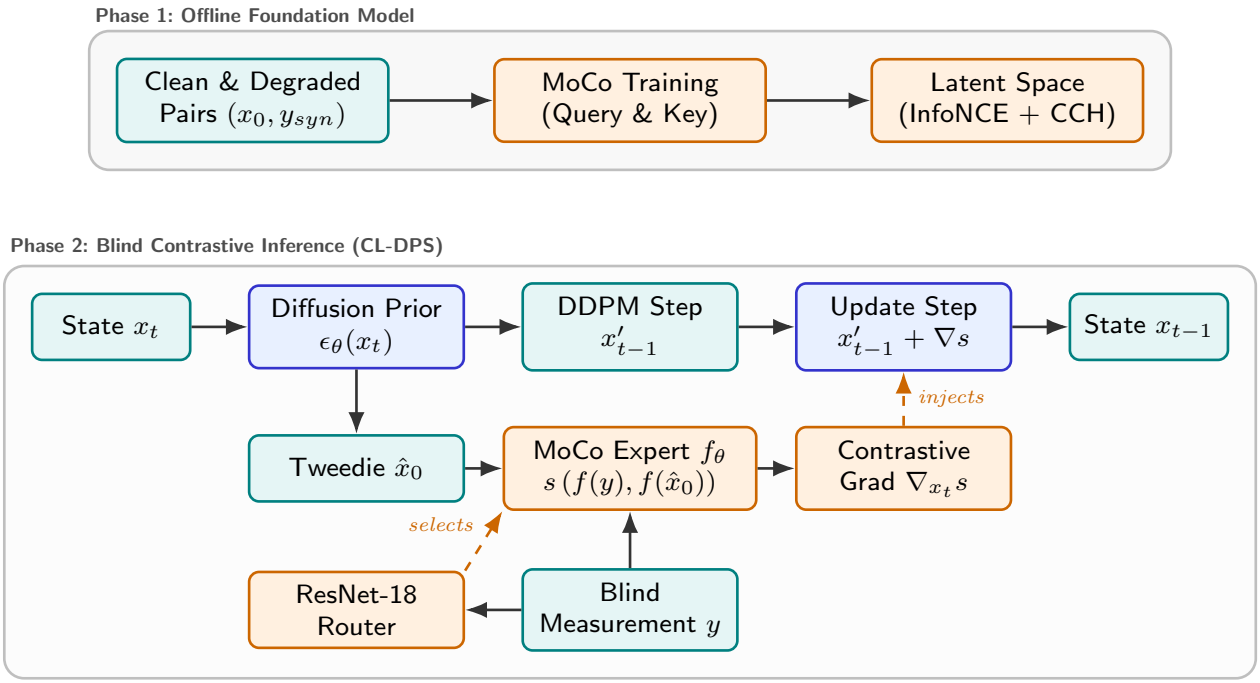


Figura 1: Arquitectura de la solución: fase offline de aprendizaje latente e inferencia guiada por $\hat{x}_0$.

4 Resultados

El sistema fue validado utilizando el dataset FFHQ frente a múltiples familias de degradaciones.

- **Sensibilidad al Desajuste de Operadores:** Al proporcionar a DPS un operador con parámetros incorrectos (ej. estimar mal la intensidad del desenfoque de movimiento), su rendimiento métrico (RMSE, LPIPS) sufre una degradación abrupta en forma de "V".

Por el contrario, CL-DPS mantiene un rendimiento completamente plano y estable a través de todo el espectro de errores paramétricos.

- **Clasificación y Enrutamiento Ciego:** Para escenarios completamente ciegos (donde ni siquiera se conoce el *tipo* de desenfoque), se entrenó un enrutador ResNet-18 que alcanzó un 99.4 % de precisión, superando con creces la extracción de características lineales del propio MoCo (37.5 %). Este enrutador selecciona dinámicamente al experto óptimo (Gaussiano, Movimiento o Rotación) durante la inferencia.
- **Modelo Universal:** Se entrenó con éxito un modelo MoCo Mixto capaz de manejar todas las degradaciones simultáneamente, logrando una robustez métrica comparable a los expertos especializados.

5 Conclusiones

Se ha demostrado que la inferencia guiada por contraste es superior en entornos ciegos. La aportación más importante es la metodología usada para comparar los métodos, así como optimización vía $\hat{x}_0$, el enrutador de expertos y el modelo universal MoCo. Sin embargo, queda hueco para mejoras del modelo contrastivo para reducir la diferencia en DPS y CL-DPS en calidad de reconstrucción.

6 Referencias

[1] J. Ho, A. Jain y P. Abbeel, "Denoising diffusion probabilistic models", *Advances in Neural Information Processing Systems*, vol. 33, págs. 6840-6851, 2020.

[2] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky y J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems", *arXiv preprint arXiv:2209.14687*, 2022.

[3] L. Weng, "Contrastive Representation Learning", *lilianweng.github.io*, mayo de 2021. dirección: `https://lilianweng.github.io/posts/2021-05-31-contrastive/`.

[4] L. Ye, S. M. Hamidi, M. Pilanci y K. N. Plataniotis, "CL-DPS: A Contrastive Learning Approach to Blind Nonlinear Inverse Problem Solving via Diffusion Posterior Sampling", en *The Fourteenth International Conference on Learning Representations*, 2026. dirección: `https://openreview.net/forum?id=KoLYNHJRBY`.

[5] K. He, H. Fan, Y. Wu, S. Xie y R. Girshick, *Momentum Contrast for Unsupervised Visual Representation Learning*, 2020. arXiv: `1911.05722 [cs.CV]`. dirección: `https://arxiv.org/abs/1911.05722`.

[6] A. v. d. Oord, Y. Li y O. Vinyals, "Representation learning with contrastive predictive coding", *arXiv preprint arXiv:1807.03748*, 2018.

[7] B. Efron, "Tweedie's Formula and Selection Bias", *Journal of the American Statistical Association*, vol. 106, n.º 496, págs. 1602-1614, 2011. DOI: 10.1198/jasa.2011.tm11181.

COGUIDE: CONTRASTIVE DIFFUSION GUIDANCE FOR SPATIAL INVERSE PROBLEMS

Author: Jorge Vančo Sampedro
Director: Jaime Pizarroso Gonzalo
Co-Director: Simón Rodríguez Santana

Abstract

Inverse problems seek to recover clean signals from corrupted measurements. However, current methods like Diffusion Posterior Sampling (DPS) fail drastically when the degradation operator is unknown. This project develops a robust generative framework based on contrastive learning (CL-DPS) to solve blind inverse problems. By training a foundational model (MoCo) that projects multiple visual degradations into a continuous latent space, we approximate the likelihood function without requiring mathematical parameters of the operator. Results demonstrate that, compared to the instability of DPS under incorrect operators, CL-DPS maintains high perceptual fidelity, achieving stable reconstructions that are invariant to parametric mismatch.

Keywords: Blind Inverse Problems; Diffusion Models; Contrastive Learning; Image Restoration; Likelihood Surrogate.

Executive Summary

1 Introduction

In fields as diverse as medical imaging, astrophysics, or space robotics, researchers rarely observe a phenomenon directly. Instead, they face inverse problems: the mathematical challenge of deducing an original signal from an indirect, noisy, or degraded measurement. Recently, diffusion models [1] have emerged as powerful tools (priors) to solve these problems.

The current standard method, known as Diffusion Posterior Sampling (DPS) [2], relies on an analytical gradient to guide the step-by-step generation of the image. For this gradient to be computed, it is assumed that the degradation process (the "forward operator") is perfectly known and differentiable. However, in the real world, parameters such as the severity of a blur or the motion trajectory of a camera are unknown. When faced with these blind inverse problems, standard DPS collapses, introducing severe artifacts that destroy the image.

2 Objectives

The main objective of this thesis is to validate **CL-DPS** (Contrastive Learning for Diffusion Posterior Sampling), a generative framework capable of solving blind inverse problems using

a likelihood based on a contrastive latent space [3], [4].

- **Formulate the Blind Inverse Problem:** Quantify the vulnerability of standard DPS to parametric mismatch across diverse visual degradation families (Gaussian, directional, and rotational blurs).
- **Develop the Contrastive Guidance Mechanism:** Design and train a foundational model (ResNet-50 encoder) using the *Momentum Contrast* (MoCo) architecture [5] to map clean and synthetically degraded images into a unified latent space. First, specific models for each operator are trained, and then a jointly trained model is validated, in addition to training a router to each expert.
- **Evaluation and Benchmarking:** Validate the results using informative metrics. Study the performance differences across the distinct methods and describe the strengths and weaknesses of the new method.

3 Methodology

To resolve this vulnerability, this project implements and optimizes the CL-DPS framework [4]. The solution eliminates the dependence on exact mathematical operators by splitting the problem into two phases:

1. **Offline MoCo Training (Fig. 1 Phase 1):** Instead of teaching a network to undo a specific blur, we train an encoder (ResNet-50) using the Momentum Contrast architecture [5]. The model is fed with clean images and their corresponding synthetically degraded versions (with Gaussian, directional, and rotational blurs of random intensity). Using the InfoNCE loss [6], the network learns to group images that share the same semantic identity, regardless of how they were degraded. To prevent color loss, a Color Consistency Head (CCH) guided by mean squared error was included.
2. **Contrastive-Guided Inference (Fig. 1 Phase 2):** During the reverse diffusion process, instead of using pixel-by-pixel error, we extract patches from the generated image at timestep t and compare them with the actual measurement in the MoCo latent space. The gradient of the cosine similarity is used to push the generated image toward the correct identity.

To evaluate the framework's generalization capability in fully blind environments, two validation architectures were designed: a dynamic system based on a router that classifies the degradation and assigns the problem to "expert" MoCo models pre-trained on specific operators; and a "Foundational" (or universal) MoCo encoder, trained simultaneously with multiple operator families to map any degradation into a unified latent space without the need for prior classification.

As a fundamental improvement over existing literature, this work identified that guiding diffusion directly on the noisy state x_t generates catastrophic structural artifacts. The implemented solution consisted of evaluating the contrastive loss function on Tweedie's denoised prediction ($\hat{x}_0$) [7], completely restoring visual coherence.

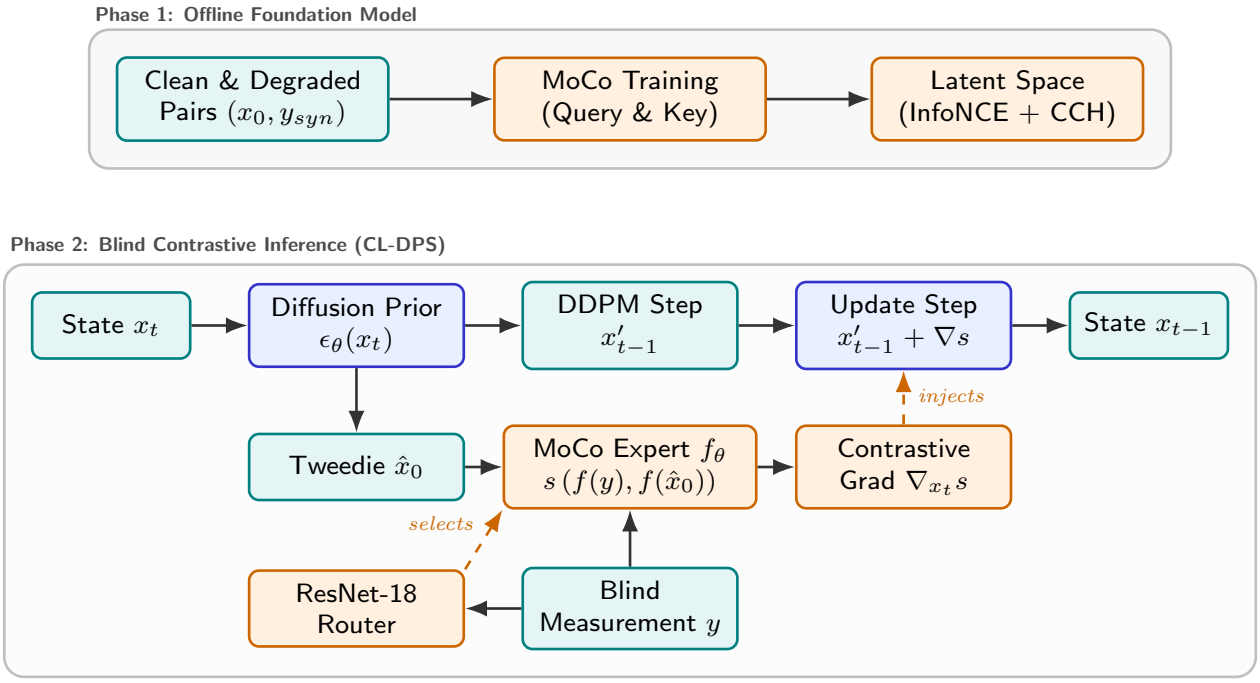


Figure 1: Design of the implemented CL-DPS solution: an offline phase learns an operator-invariant contrastive latent space, and inference uses this space as a likelihood surrogate to guide diffusion from the denoised estimate $\hat{x}_0$.

4 Results

The system was validated using the FFHQ dataset against multiple degradation families.

- **Sensitivity to Operator Mismatch:** When providing DPS with an operator with incorrect parameters (e.g., misestimating the intensity of motion blur), its metric performance (RMSE, LPIPS) suffers an abrupt "V-shaped" degradation. Conversely, CL-DPS maintains completely flat and stable performance across the entire spectrum of parametric errors.
- **Blind Classification and Routing:** For fully blind scenarios (where not even the *type* of blur is known), a ResNet-18 router was trained, achieving 99.4% accuracy, far surpassing the linear feature extraction of MoCo itself (37.5%). This router dynamically selects the optimal expert (Gaussian, Motion, or Rotation) during inference.

- **Universal Model:** A Mixed MoCo model capable of handling all degradations simultaneously was successfully trained, achieving a metric robustness comparable to the specialized experts.

5 Conclusions

It has been demonstrated that contrastive-guided inference is superior in blind environments. The most important contribution is the methodology used to compare the methods, as well as the optimization via $\hat{x}_0$, the expert router, and the universal MoCo model. However, there remains room for improvement in the contrastive model to narrow the gap between DPS and CL-DPS in terms of reconstruction quality.

6 References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[2] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems", *arXiv preprint arXiv:2209.14687*, 2022.

[3] L. Weng, "Contrastive representation learning", *lilianweng.github.io*, May 2021. [Online]. Available: `https://lilianweng.github.io/posts/2021-05-31-contrastive/`.

[4] L. Ye, S. M. Hamidi, M. Pilanci, and K. N. Plataniotis, "CL-DPS: A contrastive learning approach to blind nonlinear inverse problem solving via diffusion posterior sampling", in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: `https://openreview.net/forum?id=KoLYNHJRBY`.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, *Momentum contrast for unsupervised visual representation learning*, 2020. arXiv: `1911.05722 [cs.CV]`. [Online]. Available: `https://arxiv.org/abs/1911.05722`.

[6] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding", *arXiv preprint arXiv:1807.03748*, 2018.

[7] B. Efron, "Tweedie's formula and selection bias", *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011. DOI: `10.1198/jasa.2011.tm11181`.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Inverse problems (IP) are a fundamental challenge in engineering and science, seeking to recover unknown signals from indirect, partial, and often noisy measurements. The unknown signal x and the measurement y are typically related via a forward process $y = \mathcal{A}(x) + n$, where $\mathcal{A}$ is a forward operator and n denotes noise. While remarkable progress has been made using generative models for differentiable operators (e.g., image deblurring), a significant class of problems involves operators that are non-linear, non-differentiable, and highly complex.

This project focuses on one such challenging spatial inverse problem: **reconstructing the original signal from the corrupted input without notion of the corruption operator**. In this scenario, the forward operator $\mathcal{A}(x)$ use to distort the original data is not known.

Direct inversion of this problem is ill-posed, as multiple floorplans can explain the same trajectory. Furthermore, standard diffusion-based inverse solvers fail because the they require calculating the gradients of the forward operator, which is not known in this setting.

This project analyzes the behaviour of SOTA inverse problem solving with a newer paradigm that uses **Contrastive Learning** to learn a smooth embedding space. By projecting both the input and the corruption into a shared latent space, we derive a valid surrogate likelihood score that guides a pre-trained diffusion model toward the correct posterior distribution. This approach enables the reconstruction of the original signal without requiring the use of the forward operator. However, the forward operator is still required during the training of the Contrastive Learning model. This project focuses on expanding this idea to the Blind Inverse Problem setting, where the forward operator is unknown, by training a foundational Contrastive Learning Model.

1.1 Motivation

The mathematical framework of inverse problems (recovering a hidden true signal from a corrupted or indirect measurement) is a fundamental pillar of numerous critical industries. While the forward process (cause to effect) is often physically well-understood, the inverse process (effect to cause) is inherently ill-posed and highly sensitive to noise.

The primary motivation for this work stems from the limitations of current state-of-the-art analytical solvers. Methods like Diffusion Posterior Sampling (DPS) strictly assume that the mathematical degradation process (the forward operator, $\mathcal{A}$) is perfectly known and differentiable. However, real-world scenarios predominantly present **blind inverse problems**, where these parameters are unknown, inexact, or mathematically unstable. Solving these blind problems is essential for several key domains explored in this project:

- **Spatial Planning and Robotics:** As highlighted by recent advancements in contrastive guidance [1], inverse problems extend into discrete spatial reasoning. Reconstructing a building's hidden floorplan based solely on sparse human walking trajectories requires inverting complex, non-differentiable path-planning algorithms.

- **Medical Imaging and Computational Photography:** In clinical settings, unexpected patient movement during an MRI introduces an unknown, non-linear motion blur. Similarly, in forensic photography, images are frequently degraded by unquantified camera shake or defocus. Reconstructing clear, artifact-free images in these scenarios is impossible with solvers that require exact mathematical parameters.

Because standard analytical solvers collapse catastrophically when the measurement operator is misparameterized or entirely unknown, closing this gap is critical. Developing robust, operator-agnostic generative models, such as the contrastive latent framework explored in this project, is necessary to advance the reliability of these technologies in unpredictable, real-world environments.

1.2 Objectives

The **main objective** of this Bachelor's Thesis is to develop, implement, and validate a robust generative framework capable of solving blind inverse problems without requiring prior knowledge of the underlying mathematical degradation operator. By leveraging a contrastive learning latent space as a fully differentiable likelihood surrogate, this project aims to overcome the catastrophic brittleness of standard analytical Diffusion Posterior Sampling (DPS).

To systematically achieve this main goal, the project pursues the following **secondary objectives**:

- **Quantification of Operator Mismatch Vulnerability:** To systematically benchmark and quantify the sensitivity of standard DPS to mis-specified parameters across diverse spatial degradations (Gaussian, linear motion, non-linear motion, and rotation blurs), establishing a baseline of failure for analytical solvers.
- **Architectural and Inference Optimization:** To evaluate and establish an optimal contrastive guidance architecture by analyzing the effects of different network backbones (e.g., ResNet-18 vs. ResNet-50) and determining the most mathematically stable diffusion guidance state (x_t vs. the Tweedie prediction $\hat{x}_0$) to maximize perceptual fidelity.
- **Extension to Fully Blind Scenarios:** To expand the framework into a strictly zero-shot blind setting by designing, training, and comparing two distinct strategies: a Universal (Mixed) MoCo encoder capable of generalizing across multiple degradation families, and a dynamic routing mechanism that automatically classifies the corruption to deploy a targeted specialist expert.

1.3 SDG Alignment

This project aligns with the United Nations Sustainable Development Goals (SDGs) by advancing the foundational algorithms necessary for resilient technological and social infrastructure. Specifically, the developments within this thesis contribute to:

- **SDG 9: Industry, Innovation, and Infrastructure.** This goal emphasizes upgrading technological capabilities and fostering scientific research. The proposed contrastive framework presents a significant methodological innovation in signal processing and computational photography. By providing a stable, operator-agnostic method for recovering corrupted sensor data, it directly enhances the reliability of automated industrial systems, remote sensing infrastructure, and autonomous robotics operating in unpredictable environments.
- **SDG 3: Good Health and Well-being.** Medical imaging modalities (such as MRI, CT, and Ultrasound) fundamentally rely on solving inverse problems. A prevalent challenge in clinical settings is patient movement during scans, which introduces an unknown, non-linear motion blur into the measurement. By improving the robustness of blind inverse solvers, the methodologies explored in this project can aid in recovering diagnostic-quality images from corrupted scans. This reduces the need for repeated patient testing, thereby lowering healthcare costs, minimizing patient anxiety, and directly reducing unnecessary exposure to ionizing radiation.

Chapter 2 State of the Art

2.1 Bayesian Inference for Inverse Problems

Inverse problems constitute a fundamental class of challenges across various scientific disciplines, from astrophysics to medical imaging. Broadly defined, an inverse problem involves estimating the hidden state or unobserved causal variables of a system from a set of noisy, indirect observations.

The most principled framework to tackle these challenges is Bayesian inference. Instead of seeking a single deterministic solution, which is often impossible because multiple hidden states might produce the exact same observation, the Bayesian approach models all unknown variables probabilistically.

Formally, let x represent the unknown target signal and y represent the constrained or corrupted measurement. They are typically related via a forward measurement process:

$$y = \mathcal{A}(x) + \eta \tag{2.1}$$

where $\mathcal{A}$ is the forward operator (e.g., a blurring kernel, a downsampling matrix, or a path planner) and η represents stochastic measurement noise.

To recover x from y, we rely on Bayes' theorem, which formally describes how to compute the probability of a specific state given the noisy observation:

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \tag{2.2}$$

This theorem introduces the core terminology of probabilistic inference:

- **Prior** $p(x)$: Our baseline assumption about the underlying signal before observing any data. It defines the manifold of valid solutions (e.g., the statistical distribution of natural images or structurally sound floorplans).
- **Likelihood** $p(y|x)$: The probability of observing y given a specific signal x. This term explicitly enforces *data consistency* by evaluating how well a proposed signal matches the measurement through the forward operator $\mathcal{A}$ and the noise η.
- **Evidence** $p(y)$: The marginal probability of the measurement. Because it is independent of x, it acts merely as a normalizing constant during optimization.
- **Posterior** $p(x|y)$: The ultimate objective of the inverse problem. It represents our updated, refined belief about the true signal x after incorporating both our prior knowledge and the observed evidence.

By formulating inverse problems through this lens, the task of reconstruction becomes the task of *posterior sampling*: drawing valid samples from $p(x|y)$ that satisfy both the unconditional prior and the measurement likelihood.

2.2 Diffusion Models for Posterior Sampling

Recently, diffusion models [2] have emerged as highly expressive architectures capable of learning the unconditional data prior $p(x)$. Rather than training a task-specific network to map y to x end-to-end, methods like **Diffusion Posterior Sampling (DPS)** [3] leverage a pre-trained, unconditional diffusion model to traverse the reverse generative process, conditioning it on y at inference time.

This conditional sampling is achieved by applying Bayes' theorem to the score function at each intermediate timestep t. We begin with the standard Bayes' rule adapted for the noisy diffusion state x_t:

$$p(x_t|y) = \frac{p(y|x_t)p(x_t)}{p(y)} \tag{2.3}$$

Taking the natural logarithm of both sides yields:

$$\log p(x_t|y) = \log p(y|x_t) + \log p(x_t) - \log p(y) \tag{2.4}$$

To obtain the score function, we take the gradient with respect to the intermediate noisy state x_t. Because the marginal probability of the measurement, $p(y)$, is completely independent of x_t, its gradient evaluates to zero ($\nabla_{x_t} \log p(y) = 0$). This exactly recovers the decomposition into the unconditional prior score and the likelihood score:

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y|x_t) \tag{2.5}$$

The first term, the prior score, is provided by the pre-trained diffusion model. The second term, the likelihood score, enforces measurement consistency. Because the exact likelihood $p(y|x_t)$ at intermediate noise levels is analytically intractable, DPS utilizes Tweedie's formula [4] to estimate the denoised signal $\hat{x}_0(x_t)$.

By assuming the measurement noise is Gaussian, $\eta \sim \mathcal{N}(0, \sigma^2 I)$, the likelihood function for the measurement takes the explicit form:

$$p(y|\hat{x}_0) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} \exp\left[-\frac{||y - \mathcal{A}(\hat{x}_0)||_2^2}{2\sigma^2}\right] \tag{2.6}$$

By taking the natural logarithm of this likelihood and substituting the approximation $p(y|x_t) \approx p(y|\hat{x}_0(x_t))$, the constant scaling factors vanish when differentiating with respect to x_t. This allows the guidance step to be analytically approximated via the gradient of the measurement error:

$$\nabla_{x_t} \log p(y|x_t) \approx -\zeta \nabla_{x_t} ||y - \mathcal{A}(\hat{x}_0(x_t))||_2^2 \tag{2.7}$$

where $\zeta \triangleq 1/\sigma^2$ acts as a scaling hyperparameter (or step size) that balances the data consistency.

Crucially, this formulation strictly dictates that the forward operator $\mathcal{A}$ must be known and fully differentiable, as gradients must backpropagate through $\mathcal{A}$ to update the intermediate diffusion state x_t.

2.3 Principles of Contrastive Learning and MoCo

At its core, contrastive representation learning can be formulated as a dynamic dictionary look-up problem. The goal is to learn an embedding space where similar sample pairs (a query and its positive key) are mapped close to each other, while dissimilar pairs (the query and negative keys) are pushed far apart [5].

This objective is predominantly optimized using the InfoNCE (Noise Contrastive Estimation) loss [6]. Given an encoded query q, a positive key k_+, and a set of K negative keys $\{k_i\}_{i=1}^K$, the InfoNCE loss acts as a categorical cross-entropy objective that attempts to correctly classify the positive sample among many noise samples:

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\exp(q \cdot k_+/\tau) + \sum_{i=1}^K \exp(q \cdot k_i/\tau)} \tag{2.8}$$

where τ is a temperature hyperparameter that scales the similarity dot products.

Intuitively, minimizing this loss function balances two opposing forces within the embedding space. The numerator drives the network to maximize the dot product $q \cdot k_+$, effectively pulling the query and its corresponding positive key together. On the other hand, the denominator acts as a repulsive force: to minimize the overall loss, the network must also minimize the sum of the exponential similarities with all negative keys. This forces the dot products $q \cdot k_i$ to become as small as possible, pushing the negative samples away from the query. The temperature parameter τ dictates the strictness of this separation, concentrating the model's penalty on the "hardest" negatives, those that are incorrectly clustered too close to the query.

For contrastive learning to yield highly expressive and robust features, the dictionary of negative samples must satisfy two critical conditions: it must be **large** (to densely sample the underlying high-dimensional continuous space) and **consistent** (the keys must be encoded by identical or smoothly evolving network parameters) [7].

Traditional end-to-end contrastive mechanisms fail to meet both criteria simultaneously. Because both the query and key encoders are updated via back-propagation, the dictionary size is strictly coupled to the mini-batch size, which is heavily constrained by GPU memory limits [7]. Conversely, alternative approaches like Memory Banks allow for large dictionaries by storing representations of the entire dataset, but suffer from severe key inconsistency because the stored features are computed at vastly different stages of the training epoch [7].

Momentum Contrast (MoCo) [7] resolves this bottleneck and is widely considered a state-of-the-art framework for contrastive optimization. MoCo decouples the dictionary size from the mini-batch size by maintaining the dictionary as a dynamic **queue** of data samples. During training, the current mini-batch of encoded representations is enqueued, and the oldest mini-batch is dequeued. This allows the network to process massive dictionaries of negative samples efficiently.

To ensure key consistency without the intractable computational cost of back-propagating through the entire queue, MoCo updates the parameters of the key encoder (θ_k) using a

momentum-based moving average of the query encoder's parameters (θ_q):

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{2.9}$$

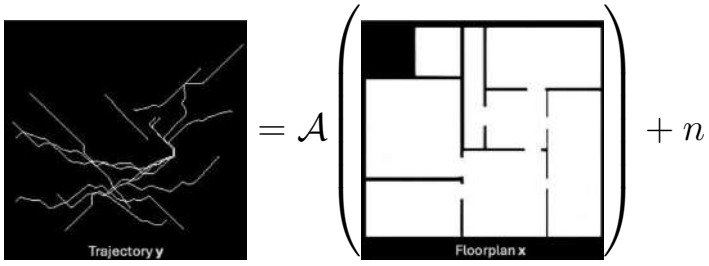
where $m \in [0, 1)$ is a momentum coefficient (typically set to a high value, e.g., $m = 0.999$). This momentum update ensures that the key encoder evolves extremely slowly, keeping the queued negative keys highly consistent with current queries.

Because of its ability to maintain a massive, consistent dictionary of negative samples, MoCo provides an exceptionally stable and scalable embedding space. As detailed in the following subsections, this stability is exactly what enables modern frameworks to utilize contrastive distances as a differentiable, robust surrogate for intractable likelihoods in diffusion posterior sampling.

2.4 Contrastive Learning as a Likelihood Surrogate

To overcome the instability of the forward operator, this project leverages **Contrastive Learning** (InfoNCE) [6]. It has been theoretically shown that the optimal contrastive classifier recovers the density ratio between joint and marginal distributions.

CoGuide [1] builds upon this theoretical foundation to address spatial inverse problems, such as reconstructing a floorplan from human walking trajectories. When the forward operator $\mathcal{A}$ is a path planner:



$$= \mathcal{A}\left(\quad\right) + n$$

Direct likelihood-based guidance becomes severely unstable due to the non-differentiable and highly non-linear nature of path generation. Tiny topological changes in a floorplan can drastically alter a planned path, rendering standard gradient-based optimization ineffective. Fig. 1 shows two different paths obtained by opening two small pathways in the wall. The difference is just a couple of pixels but the resulting paths are drastically different.

In the context of standard DPS, the measurement guidance step (Eq. (2.7)) inherently relies on backpropagating the error through the forward operator. By the chain rule, this computation depends directly on the Jacobian of the operator, $\mathcal{J}_\mathcal{A} = \partial\mathcal{A}(x)/\partial x$. Because discrete path planning algorithms are highly sensitive and discontinuous, $\mathcal{J}_\mathcal{A}$ is fundamentally unstable. As illustrated above, a minimal perturbation in the predicted spatial layout $\hat{x}_0$ triggers a catastrophic, non-linear shift in the resulting trajectory. Consequently, the gradient signal derived from $\mathcal{J}_\mathcal{A}$ becomes chaotic and unreliable, causing standard DPS to either diverge or become trapped in poor local optima.

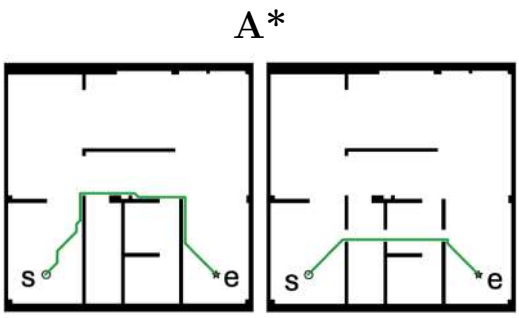


Figure 1: A* algorithm sensitivity. Changing a couple of pixels can lead to greatly different paths.

To resolve this, CoGuide reformulates the likelihood score in a smoother, continuous embedding space. By training a floorplan encoder f_θ and a trajectory encoder g_ϕ to pull compatible (floorplan, trajectory) pairs together and push mismatched pairs apart, we construct a smooth joint embedding space (Fig. 2). The distance in this space serves as a stable, differentiable surrogate for the intractable likelihood, effectively bypassing the need to differentiate through the complex path-planning algorithm itself.

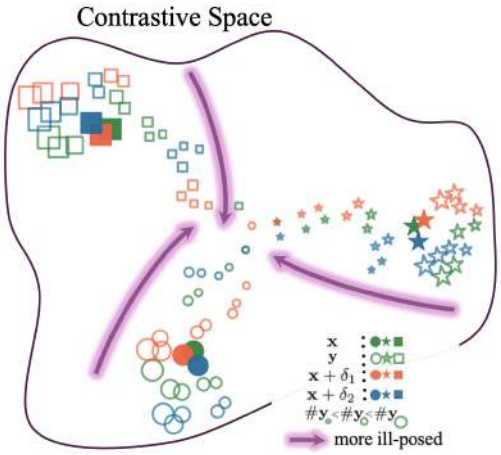


Figure 2: t-SNE embeddings from CoGuide for 3 floorplans (solid green) and two perturbed variants (solid orange/blue). Trajectories from these floorplans are shown as hollow shapes; larger hollow markers indicate higher trajectory density.

2.4.1 Mathematical Formulation of Contrastive Guidance

Under the InfoNCE framework, the intractable conditional likelihood $p(y|x_t)$ is approximated using the exponential similarity between the encoded noisy floorplan x_t and the trajectory

measurement y. Given a temperature parameter τ, the similarity score is defined as:

$$s(x_t, y) = \frac{\langle f_\theta(x_t), g_\phi(y) \rangle}{\tau} \tag{2.10}$$

The contrastive likelihood surrogate is then formulated as a softmax distribution over a dictionary of negative trajectory samples $\mathcal{Y}$:

$$p(y|x_t) \approx \frac{\exp(s(x_t, y))}{\sum_{\tilde{y} \in \mathcal{Y}} \exp(s(x_t, \tilde{y}))} \tag{2.11}$$

During diffusion posterior sampling, we require the gradient of the log-likelihood to guide the reverse process. By taking the gradient of this surrogate, CoGuide derives a highly stable guidance signal. In practice, isolating the numerator-only update is often sufficient and computationally efficient to steer the denoising trajectory:

$$\nabla_{x_t} \log p(y|x_t) \approx \nabla_{x_t} s(x_t, y) = \frac{1}{\tau} \nabla_{x_t} \langle f_\theta(x_t), g_\phi(y) \rangle \tag{2.12}$$

2.4.2 CoGuide Sampling Algorithm

This contrastive gradient seamlessly integrates into the standard reverse diffusion process. The trajectory is encoded only once, while the noisy floorplan is encoded at each timestep to provide continuous geometric guidance. In this scenario, there is an additional intersection penalty $\mathcal{L} = \lambda_{int} \nabla_{\mathbf{x}_t} \|\mathbf{y} \odot (1 - \hat{\mathbf{x}}_0)\|_1$ specific to the floorplans scenario that we disregard in other tasks.

2.5 Contrastive Learning to solve Blind Inverse Problems (CL-DPS)

As established in Section 2, standard DPS relies on the analytical gradient of the measurement error (Eq. (2.7)), which strictly requires the forward operator $\mathcal{A}$ to be known and differentiable. However, in **blind inverse problems**, such as restoring a photograph degraded by unquantified camera shake, the exact parameters of $\mathcal{A}$ are entirely unknown at inference time. Under these conditions, the standard DPS gradient $\nabla_{x_t} \log p(y|x_t)$ becomes theoretically intractable, making it impossible to backpropagate the error and causing traditional solvers to fail catastrophically.

Building upon the foundational principles of contrastive likelihood estimation used in CoGuide, recent work [8] adapts this framework to bypass the need for explicit operator parameters. To resolve the intractable likelihood, CL-DPS estimates the measurement consistency numerically using an auxiliary encoder f.

Algorithm 1 CoGuide: Contrastive Diffusion Guidance for Spatial Inverse Problems

Require: T timesteps; trajectory $\mathbf{y}$; step sizes $\{\zeta_t\}$; Adam base LR η_0, γ_1, γ_2, ε, noise scales $\{\tilde{\sigma}_t\}$; diffusion params $\{\alpha_t, \bar{\alpha}_t\}$; score s_θ; encoders f_φ, g_ψ; temperature τ, intersection weight λ_{int}.

Require: Annealing & gating: start t_s, end t_e, min-LR ρ, stop t_{stop}

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; Adam: $m \leftarrow 0$, $v \leftarrow 0$

for $t = T-1$ **down to** 0 **do**

DDIM: $\hat{s} \leftarrow s_\theta(\mathbf{x}_t, t)$; $\hat{\mathbf{x}}_0 \leftarrow \bar{\alpha}_t^{-1/2}(\mathbf{x}_t + (1-\bar{\alpha}_t)\hat{s})$; ▷ score model and one step denoising

$\hat{\epsilon} \leftarrow \frac{\mathbf{x}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0}{\sqrt{1-\bar{\alpha}_t}}$; $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. $\sigma_t \leftarrow \tilde{\sigma}_t$.

$\mathbf{x}'_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\hat{\epsilon} + \sigma_t \mathbf{z}$. ▷ ddim step

CoGuide: $G_t \leftarrow -\frac{1}{2\tau}\nabla_{\mathbf{x}_t}\|g_\psi(y) - f_\varphi(\hat{\mathbf{x}}_0)\|_2^2 + \lambda_{int}\nabla_{\mathbf{x}_t}\|\mathbf{y} \odot (1 - \hat{\mathbf{x}}_0)\|_1$. ▷ contrastive likelihood score

Adam: $\eta_t \leftarrow \text{AnnealLR}(\eta_0, \rho, t; t_s, t_e)$; **if** t_{stop} set **and** $t \geq t_{\text{stop}}$ **then** $\eta_t \leftarrow 0$.

$\mathbf{x}_{t-1} \leftarrow \text{Adam}(\mathbf{x}'_{t-1}, G_t; \eta_t, \gamma_1, \gamma_2, \varepsilon)$. *(Optional SGD:* $\mathbf{x}_{t-1} \leftarrow \mathbf{x}'_{t-1} + \zeta_t G_t$*.)*

end for

return $\hat{\mathbf{x}}_0$

2.5.1 Operator-Agnostic Contrastive Surrogate

To handle the blind setting without training task-specific diffusion models, CL-DPS trains a single auxiliary encoder f offline using synthetic degradations generated from a broad prior of operator families (e.g., Gaussian, motion, and rotation blurs). By pulling together positive pairs (x_t, y) drawn from these simulated measurement processes and pushing apart negatives sampled from a MoCo queue [7], the encoder learns a generalized, operator-agnostic energy landscape.

2.5.2 Overlapping Patch-Wise Inference & Color Consistency

One distinct challenge in image-based blind restoration is the preservation of low-level structural details, which standard convolutional encoders often compress or discard. To mitigate this, CL-DPS introduces an overlapping patch-wise inference strategy. By partitioning the image into U overlapping patches and stacking the features, the mutual information $I(x; f(\{p_j^x\}))$ between the clean signal and the encoded features is theoretically guaranteed to be preserved [8].

Furthermore, pure contrastive embeddings can become insensitive to global color statistics, leading to hue or brightness shifts in the reconstructed image. To prevent this, a lightweight Color-Consistency Head (CCH) is optimized alongside the encoder during the offline training phase to explicitly capture global color information.

Algorithm 2 CL-DPS

Input number of steps T, measurement $\boldsymbol{y}$, noise schedule $\{\tilde{\sigma}_t\}$, pretrained encoder $f(\cdot)$, step size $\eta > 0$, number of overlapping patches U.
$\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. // initialize with Gaussian noise
Extract U overlapping patches from $\boldsymbol{y}$ once.
// cache measurement features
$\{\boldsymbol{p}_j^{\boldsymbol{y}}\}_{j\in[U]} \leftarrow \boldsymbol{y}$.
for $t = T-1 \dots 0$ **do**
  $\hat{\boldsymbol{s}} \leftarrow \boldsymbol{s}_\theta(\boldsymbol{x}_t, t)$. // score model estimate of $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t)$
  $\tilde{\boldsymbol{x}}_0 \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} (\boldsymbol{x}_t + (1-\bar{\alpha}_t)\hat{\boldsymbol{s}})$. // Tweedie posterior mean
  $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
  $\boldsymbol{x}'_{t-1} \leftarrow \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} \boldsymbol{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \tilde{\boldsymbol{x}}_0 + \tilde{\sigma}_t \boldsymbol{z}$.
  // DDPM update
  Extract U overlapping patches from $\boldsymbol{x}_t$.
  $\{\boldsymbol{p}_j^{\boldsymbol{x}_t}\}_{j\in[U]} \leftarrow \boldsymbol{x}_t$.
  $\boldsymbol{x}_{t-1} \leftarrow \boldsymbol{x}'_{t-1} - \eta \nabla_{\boldsymbol{x}_t} \langle f(\{\boldsymbol{p}_j^{\boldsymbol{x}_t}\}_{j\in[U]}), f(\{\boldsymbol{p}_j^{\boldsymbol{y}}\}_{j\in[U]}) \rangle$.
  // contrastive guidance
end for
Output $\boldsymbol{x}_0$.

2.5.3 Diffusion Posterior Guidance

When solving blind inverse problems, the exact parameters of the measurement operator remain unknown. We make use of the pretrained auxiliary encoder f to evaluate the extracted patches from the diffusion state x_t and the actual measurement y, leaving us with embeddings from the smooth latent space designed to help guide the diffusion process.

Analogous to the CoGuide formulation, CL-DPS avoids computing the full, computationally heavy denominator-aware gradient. Instead, it leverages the highly efficient numerator-only energy guidance step to steer the reverse diffusion update. This contrastive guidance is applied directly as an additive gradient step at each timestep t:

$$\boldsymbol{x}_{t-1} \leftarrow \boldsymbol{x}'_{t-1} - \eta \nabla_{\boldsymbol{x}_t} \langle f(\{\boldsymbol{p}_j^{\boldsymbol{x}_t}\}_{j\in[U]}), f(\{\boldsymbol{p}_j^{\boldsymbol{y}}\}_{j\in[U]}) \rangle \tag{2.13}$$

The full procedure, which seamlessly plugs this contrastive gradient into a standard diffusion sampler, is outlined in Algorithm 2.

Chapter 3 Methodology

The core premise of this project is that traditional Diffusion Posterior Sampling (DPS) is highly fragile when the forward measurement operator is unstable, non-differentiable, or inaccurately specified. While our ultimate motivation lies in complex spatial planning problems (as discussed in Section 2), evaluating operator sensitivity in those domains is obscured by the inherent non-differentiability of path planners.

To rigorously isolate and quantify how DPS performance degrades under operator mismatch, we formulate our methodology within the visual domain using the FFHQ dataset. This domain provides a stable, well-understood environment where standard DPS excels when the operator is perfectly known. By using continuous visual degradations, we can systematically control the mathematical divergence between the true corruption and the assumed guidance operator, allowing for a precise comparison against Contrastive Learning for DPS (CL-DPS).

3.1 Problem Formulation and Operator Mismatch

In standard inverse problems, the goal is to recover a clean signal $\boldsymbol{x}_0$ from a measurement $\boldsymbol{y}$. The true measurement is generated by a specific, true forward operator $\mathcal{A}_{true}$:

$$\boldsymbol{y} = \mathcal{A}_{true}(\boldsymbol{x}_0) + \boldsymbol{n} \tag{3.1}$$

In standard DPS, the reverse diffusion process is guided by the analytical gradient of the measurement consistency loss. During inference, DPS requires the explicit definition of a guidance operator, $\mathcal{A}_{guide}$. If the problem is strictly non-blind, $\mathcal{A}_{guide} \equiv \mathcal{A}_{true}$, and the guidance step is optimal, we obtain from Eq. (2.7):

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y}|\boldsymbol{x}_t) \approx -\zeta \nabla_{\boldsymbol{x}_t} \|\boldsymbol{y} - \mathcal{A}_{guide}(\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t))\|_2^2 \tag{3.2}$$

where ζ is the guidance scale parameter.

However, in real-world or blind scenarios, the exact measurement process is often partially unknown. We formally define an **operator mismatch** scenario where $\mathcal{A}_{guide} \neq \mathcal{A}_{true}$. For instance, the true operator might be a Gaussian blur with a standard deviation $\sigma_{true} = 3.0$, while the guidance operator assumes $\sigma_{guide} = 1.5$. We hypothesize that as the mathematical distance between $\mathcal{A}_{true}$ and $\mathcal{A}_{guide}$ increases, the analytical gradients will actively push the denoising trajectory off the natural data manifold, leading to severe artifacts.

3.2 Forward Measurement Operators

To systematically evaluate the robustness of standard DPS and to train our universal contrastive encoder, we define a diverse set of forward measurement operators. In real-world imaging, these operators correspond to physical camera artifacts such as defocus blur, camera shake, and sensor movement. Mathematically, these spatial degradations are modeled as convolution operations or geometric transformations applied to the clean image $\boldsymbol{x}_0$.

Below, we detail the theoretical formulation of the four distinct degradation families used in our experiments. Fig. 3 illustrates the visual effect of each operator on a clean target image.

(a) Clean Image (b) Gaussian (c) Linear Motion (d) Non-linear (e) Rotation Blur

Figure 3: Visual comparison of the forward measurement operators used to degrade the clean images. From left to right: the original clean image, Gaussian blur, linear motion blur, non-linear (random walk) motion blur, and rotation blur.

3.2.1 Gaussian Blur

Gaussian blur is a fundamental isotropic degradation, simulating an out-of-focus camera lens or atmospheric scattering. The operation is defined as a 2D discrete convolution of the image with a Gaussian kernel $\boldsymbol{k}_{gauss}$. The kernel weights are determined by a standard deviation σ, which dictates the severity of the blur:

$$\boldsymbol{k}_{gauss}(u, v) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{u^2 + v^2}{2\sigma^2}\right) \tag{3.3}$$

where (u, v) are the spatial coordinates relative to the kernel center. Because the blur spreads energy uniformly in all directions, it acts as a severe low-pass filter, permanently destroying high-frequency edge information.

3.2.2 Linear Motion Blur

Linear motion blur simulates the effect of a camera or subject moving at a constant velocity along a straight line during the camera sensor's exposure time. This directional blur is modeled using a 1D line kernel parameterized by its length L (in pixels) and its angle θ. The kernel is defined as:

$$\boldsymbol{k}_{linear}(u, v) = \begin{cases} \frac{1}{L} & \text{if } u\cos\theta + v\sin\theta = 0 \text{ and } \sqrt{u^2 + v^2} \leq \frac{L}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

Unlike Gaussian blur, linear motion blur is highly anisotropic. It destroys structural details along the axis of motion but preserves edges that are parallel to the direction of movement.

3.2.3 General (Non-Linear) Motion Blur

While linear motion blur assumes a perfectly straight trajectory, real-world camera shake (e.g., caused by hand tremors) is highly erratic and non-linear. To simulate this, we utilize a general motion blur operator that generates a continuous, random-walk trajectory over a 2D plane.

The kernel is constructed by integrating a sub-pixel point spread function along a randomized spline or Markovian path, resulting in complex, overlapping exposure patterns. Let $\boldsymbol{p}(t) = [u(t), v(t)]$ define the camera's position at time t during the exposure interval $t \in [0, 1]$. The resulting kernel is the temporal integration of the path:

$$\boldsymbol{k}_{motion}(u, v) = \int_0^1 \delta(u - u(t), v - v(t))\, dt \tag{3.5}$$

This operator is exceptionally challenging for analytical inverse solvers like DPS, as the random trajectory creates highly unpredictable phase shifts and zeroes in the frequency domain.

3.2.4 Rotation Blur

Rotation blur occurs when the camera rotates around its optical axis (or another fixed center point) during exposure. Rather than a standard convolution kernel, this degradation is a geometric integration. Let $\mathcal{R}_\phi(\boldsymbol{x}_0)$ denote the image rotated by an angle ϕ. The blurred measurement is the average of all rotated states uniformly distributed within a maximum angular range $[-\phi_{max}, \phi_{max}]$:

$$\mathcal{A}_{rot}(\boldsymbol{x}_0) = \frac{1}{2\phi_{max}} \int_{-\phi_{max}}^{\phi_{max}} \mathcal{R}_\phi(\boldsymbol{x}_0)\, d\phi \tag{3.6}$$

Because the linear displacement caused by rotation increases with the distance from the center of rotation, this operator produces a spatially variant blur. The center of the image remains relatively sharp, while the periphery suffers from severe rotational smearing.

3.3 Contrastive Foundation Model for Blind Inverse Problems

To overcome the brittleness of explicit operator matching, we propose training a unified contrastive foundation model capable of solving blind inverse problems. Unlike standard DPS, which requires $\mathcal{A}_{guide}$ at inference time, our approach relies on an auxiliary encoder f_θ trained offline using a Momentum Contrast (MoCo) framework.

3.3.1 Offline Representation Learning

The objective is to train a single, universal encoder over a highly diverse mixture of degradation families $\mathcal{O} = \{\text{Gaussian, Linear Motion, Non-Linear Motion, Rotation}\}$. During the

offline training phase, we sample a clean image $\boldsymbol{x}_0$, randomly select an operator family $o \sim \mathcal{O}$, and sample its parameters ψ from a broad uniform prior to generate a synthetic measurement:

$$\boldsymbol{y}_{syn} = \mathcal{A}_{o,\psi}(\boldsymbol{x}_0) + \boldsymbol{n} \tag{3.7}$$

By pulling together clean images $\boldsymbol{x}_0$ with their corresponding $\boldsymbol{y}_{syn}$ across this massive variety of perturbations, and pushing apart semantically unrelated images via the InfoNCE loss, the encoder learns a generalized, operator-agnostic latent space.

3.3.2 Color Consistency Preservation

A known limitation of purely structural contrastive learning is its tendency to become invariant to global color statistics, which can cause hue or brightness shifts during image reconstruction. To mitigate this issue, [8] incorporates a lightweight auxiliary Color Consistency Head (CCH), denoted by H_c, on top of the auxiliary encoder.

Formally, let the input be $x_t \in \mathbb{R}^{C\times N_1\times N_2}$. We define its spatial average across the height N_1 and width N_2 for each color channel c as:

$$[\mathrm{AP}(x_t)]_c = \frac{1}{N_1 N_2}\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} x_{t\,cij} \tag{3.8}$$

The CCH, $\mathrm{H}_c(x_t) \in \mathbb{R}^C$, is implemented as a two-layer convolutional module with global pooling followed by a sigmoid activation. It is trained to directly predict these global color statistics from the encoded representation. We define the color-consistency loss using the Mean Squared Error between the CCH prediction and the pooled input, which forces the encoder to retain global color fidelity within its learned manifold:

$$\mathcal{L}_{CC}(x_t) = ||\mathrm{H}_c(x_t) - \mathrm{AP}(x_t)||_2^2 \tag{3.9}$$

We add this term to our primary objective, leaving us with the following combined loss function:

$$\mathcal{L}_{\text{CL-DPS}} = \mathcal{L}_{p(y_{\text{syn}}|x_t)} + \lambda\mathcal{L}_{CC}(x_t) \tag{3.10}$$

where $\lambda > 0$ is a parameter that balances the primary likelihood estimation and color preservation. Because the CCH is utilized solely to guide the encoder's learning process during training, it is entirely discarded at inference.

3.4 Contrastive-Guided Inference

During the inference phase, the system must solve the inverse problem without knowing the exact parameters of the true measurement operator, $\mathcal{A}_{true}$. To achieve this, our methodology guides the diffusion model using learned latent representations rather than raw pixel comparisons.

Traditional methods (such as DPS, see Eq. (2.7)) guide the diffusion process by calculating the L_2 distance between the reconstructed image and the measurement in pixel space. Because $\mathcal{A}_{true}$ is unknown in blind settings, this direct comparison is impossible. Instead, we approximate the likelihood solely via the contrastive feature distance in the latent space (Eq. (2.12)).

Relying on global features can be problematic, as global pooling operations often discard fine, high-frequency structural details. To preserve this crucial information, the inference methodology in [8] utilizes an overlapping patch-wise extraction strategy. Specifically, the input image is divided into N overlapping patches, each having a spatial dimension of $M \times M$ and extracted using a stride of S.

To preserve fine structural details that global pooling might discard, the inference methodology utilizes an overlapping patch-wise extraction strategy: the input image is divided into N overlapping patches with size $M \times M$ and stride S.

Using these extracted patches, denoted as $\{\boldsymbol{p}_j\}$, the generalized gradient step used to guide the diffusion model is defined as:

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{y}|\boldsymbol{x}_t) \approx \frac{1}{\tau} \nabla_{\boldsymbol{x}_t} \langle f_\theta(\{\boldsymbol{p}_j^{\hat{\boldsymbol{x}}_0}\}), f_\theta(\{\boldsymbol{p}_j^{\boldsymbol{y}}\}) \rangle \tag{3.11}$$

where τ is the contrastive temperature and f_θ is the pretrained auxiliary encoder.

Because the encoder f_θ relies solely on the contrastive similarities between the extracted patch representations and is never explicitly given the parameters of $\mathcal{A}_{true}$, it functions as a highly robust, zero-shot blind solver.

3.5 Experimental Design

To systematically validate the robustness of this contrastive framework, our methodology dictates two primary evaluation tracks:

- **Track 1: Sensitivity to Parametric Mismatch.** We compare standard DPS against CL-DPS under varying degrees of parametric error. We corrupt images using a fixed true operator (e.g., motion blur with length 30.0). For standard DPS, we sweep the parameters of $\mathcal{A}_{guide}$ from an exact match to extreme mismatch, measuring the degradation in image quality. For CL-DPS, we provide the exact same corrupted image $\boldsymbol{y}$ and evaluate its ability to maintain high-fidelity reconstructions using solely the latent surrogate.

- **Track 2: Out-of-Distribution (OOD) Generalization.** A true foundation model must maintain stability when exposed to severities beyond its training data. We evaluate the contrastive encoder on corruption intensities strictly outside the uniform priors used during the MoCo offline training phase. This measures the capacity of the embedding space to extrapolate the concept of "measurement consistency" beyond its defined boundaries without catastrophic failure.

Chapter 4 Experimental Results

4.1 Offline Training of the Contrastive Likelihood Surrogate

Before executing diffusion posterior sampling, we train the auxiliary encoder f offline to learn the contrastive likelihood surrogate. Following the CL-DPS framework, the model is trained using a Momentum Contrast (MoCo) architecture. To thoroughly evaluate the method's capability to handle blind inverse problems, we train models under two distinct regimes: **Specialist Models** (trained on a single degradation family) and a **Universal Model** (trained on a mixture of degradations).

4.1.1 Dataset and Preprocessing

The encoders are trained using the FFHQ dataset. For training, we utilize the first 69,000 images, reserving 1,000 separate images for periodic validation. During the data loading pipeline, images are first resized to 128×128. We then apply a Joint Random Resized Crop to extract 64×64 patches (with a scale of $[0.20, 0.30]$ and aspect ratio of $[0.5, 2]$) applied consistently to both the clean query and the distorted key images. Finally, random horizontal flipping ($p = 0.5$) and standard FFHQ normalization are applied.

4.1.2 Network Architecture and Objective

We employ a ResNet-50 backbone for both the query encoder and the momentum-updated key encoder. The standard fully connected layer is replaced with a custom Projection Head. This head serves a dual purpose:

1. **Contrastive Feature Branch:** It outputs a 128-dimensional embedding vector used to compute the InfoNCE contrastive loss.

2. **Color Consistency Branch:** It outputs a 3-channel spatial prediction to compute an auxiliary Color Consistency loss ($\mathcal{L}_{CC}$). This loss is computed as the Mean Squared Error (MSE) between the color prediction and a detached 2×2 Adaptive Average Pooling of the clean query image, ensuring the encoder preserves global color statistics.

To expose the network to varying noise levels during training, we generate $NNS = 10$ noisy variants for each distorted key image using a randomized diffusion noise schedule. To prevent the model from finding trivial solutions via intra-batch communication, we utilize Shuffled Batch Normalization (ShuffleBN) across GPUs for the key encoder.

4.1.3 Hyperparameters

The hyperparameters used for training the contrastive models are extracted directly from the CL-DPS paper [8]. We enumerate the core training configuration below:

- **Dictionary Queue Size (K):** $65,536$
- **InfoNCE Temperature (τ):** 0.07
- **MoCo Momentum (m):** 0.996
- **Batch Size:** 256
- **Feature Dimension:** 128
- **Optimizer:** SGD with a momentum of 0.9 and weight decay of 10^{-4}
- **Learning Rate:** 0.03 (with a linear warmup for the first 5 epochs, followed by a Cosine Annealing schedule)
- **Maximum Epochs:** 400. We then select the model with lowest validation loss.
- **Measurement Noise (σ_n):** Sampled uniformly from $\mathcal{U}(0.005, 0.03)$ and added to the distorted key views.

4.1.4 Specialist Contrastive Models

To evaluate performance when the degradation operator family is known but its exact parameters remain blind, we train distinct "Specialist" encoders for individual distortion types. The training sets for these models apply a single forward operator with randomized parameters per step:

- **Gaussian Blur:** The measurement y is corrupted using a Gaussian kernel of size 61×61, with the standard deviation uniformly sampled as $\sigma \sim \mathcal{U}(0.6, 2.4)$.
- **Linear Motion Blur:** We simulate camera shake or object motion using a linear motion blur. The length of the motion is sampled uniformly $l \sim \mathcal{U}(15.0, 60.0)$ pixels, and the angle is sampled uniformly from $\theta \sim \mathcal{U}(0°, 180°)$.
- **Rotation Blur:** To simulate non-linear rotational camera shake, the image is rotated around a randomly chosen center point. The maximum rotation angle is uniformly sampled as $\phi \sim \mathcal{U}(10°, 30°)$, and the blur is constructed by averaging 21 rotated views distributed within this angular range.

4.1.5 Universal (Mixed) Contrastive Model

To address truly blind inverse problems where even the family of the degradation operator is unknown at inference time, we train a "Universal" (or mixed) MoCo model.

During the training of the Universal model, the forward operator is dynamically selected on the fly for each sample. For every training iteration, the dataloader uniformly randomly selects between Gaussian, Linear Motion, or Rotation blur. Once the operator family is selected, its continuous parameters (such as σ, motion length, or rotation angle) are sampled from the exact same ranges defined for the specialist models. By pulling together clean queries and keys corrupted by this highly varied mixture of degradations, the Universal encoder learns a generalized, robust energy landscape capable of providing stable likelihood guidance regardless of the underlying distortion type.

4.2 Evaluation Metrics

To quantitatively assess the performance of the proposed CL-DPS framework against the standard DPS baseline, we employ a comprehensive suite of metrics. Because generative models inherently involve a trade-off between strict spatial adherence and high-level perceptual realism, we utilize a combination of distortion-based (pixel-wise) and perceptual metrics.

- **Root Mean Square Error (RMSE):** A standard measure of the average magnitude of the reconstruction error. It quantifies the strict pixel-wise spatial difference between the ground truth image $\boldsymbol{x}_0$ and the reconstructed image $\hat{\boldsymbol{x}}_0$:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{0,i} - \hat{x}_{0,i})^2} \tag{4.1}$$

  where N is the total number of pixels. Lower RMSE values indicate better structural alignment at the pixel level.

- **Peak Signal-to-Noise Ratio (PSNR):** A traditional distortion metric widely used to measure reconstruction quality. It is defined via the Mean Squared Error (MSE) and the maximum possible pixel fluctuation (MAX_I, which is typically 255 for 8-bit images or 1.0 for normalized tensors):

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{\text{MSE}}\right) \tag{4.2}$$

  While higher PSNR values generally correlate with better reconstruction, PSNR is highly sensitive to microscopic spatial shifts (e.g., a one-pixel translation) and often fails to accurately reflect human visual perception.

- **Structural Similarity Index Measure (SSIM):** To overcome the limitations of absolute error metrics, SSIM [9] measures the perceived change in structural information, independently evaluating luminance, contrast, and structure. It provides a value between -1 and 1, where 1 indicates perfect structural identity.

  The formula for SSIM is:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4.3}$$

  Variables:

  μ_x and μ_y: The average pixel values (mean luminance) of images x and y.

  σ_x^2 and σ_y^2: The variance of x and y (measuring contrast).

  σ_{xy}: The covariance of x and y (measuring structural correlation).

  C_1 and C_2: Constants used to stabilize the division when the denominator is very close to zero. They are typically defined as $C_1 = (k_1 L)^2$ and $C_2 = (k_2 L)^2$, where L is the dynamic range of the pixel values (e.g., 255 for 8-bit grayscale images), $k_1 = 0.01$, and $k_2 = 0.03$.

- **Learned Perceptual Image Patch Similarity (LPIPS):** Standard distortion metrics penalize generative models for synthesizing highly realistic but non-identical high-frequency textures (e.g., hallucinating a plausible, but slightly different, strand of hair). LPIPS [10] resolves this by measuring the distance between image patches in the latent feature space of a pre-trained deep neural network (such as VGG or AlexNet). Because it mimics human visual judgment, LPIPS serves as our primary indicator of generative success under severe operator mismatch. Lower values indicate higher perceptual similarity.

  Given a reference image x and a distorted image x_0, the LPIPS distance $d(x, x_0)$ is calculated as:

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|w_l \odot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})\|_2^2 \tag{4.4}$$

  Variables:

  l: A specific layer within the pre-trained neural network.

  $\hat{y}^l$ and $\hat{y}^l_0$: The feature maps extracted from layer l for images x and x_0. These feature activations are scaled and unit-normalized in the channel dimension.

  H_l and W_l: The spatial dimensions (height and width) of the feature map at layer l.

  w_l: A vector of learned weights used to scale the active channels. It dictates how "important" a specific channel in layer l is to human perception.

  In simple terms, LPIPS pushes both images through a neural network, extracts their "features" at various layers, scales those features by human-calibrated weights, and

calculates the squared distance between them. Lower LPIPS values indicate that the images look more similar to a human eye.

A Note on Fréchet Inception Distance (FID)

In the generative modeling literature, the Fréchet Inception Distance (FID) [11] is frequently used as a gold-standard metric for visual quality. Conceptually, FID is highly similar to LPIPS: both leverage the deep feature space of pre-trained image classification networks (InceptionV3 for FID, typically VGG for LPIPS) to evaluate perceptual quality rather than relying on shallow, pixel-wise comparisons.

However, there is a fundamental difference in their statistical application: while LPIPS computes a pairwise distance between two specific images (image-to-image), FID computes the Fréchet distance (also known as the Wasserstein-2 distance) between two multivariate Gaussian distributions fitted to the network features of two entire datasets (distribution-to-distribution).

Consequently, FID is exceptionally sensitive to sample size. To accurately estimate the high-dimensional covariance matrices required by the FID formulation, it is standard practice to use a minimum of $10,000$ to $50,000$ samples. In the scope of our evaluation phase, generating such a massive volume of reconstructions across multiple operator sweeps and diffusion timesteps is computationally prohibitive. Because our test sets evaluate 25 samples per specific degradation scenario, calculating FID would yield highly unstable, biased, and statistically meaningless covariance estimates. Therefore, we exclude FID from our quantitative analysis, relying entirely on LPIPS to capture deep perceptual fidelity at the individual sample level.

4.3 MoCo Model Validation and Architecture Selection

We performed preliminary experiments to verify the convergence and representational quality of the trained MoCo models, as well as to determine the optimal backbone architecture. Fig. 4 compares the validation loss during training for two different backbone architectures (ResNet-18 and ResNet-50) using Gaussian-blurred images. As expected for the offline contrastive task, the deeper ResNet-50 architecture provides a higher capacity for complex feature extraction and converges to a notably lower validation loss.

However, offline training convergence does not strictly guarantee superior performance during diffusion guidance. We hypothesized that a shallower network (ResNet-18) might allow gradients to propagate more directly to the input image space, potentially steering the reverse diffusion process more effectively than a deep network. To test this, we compared the inference performance of both architectures across a guidance scale sweep (Fig. 5).

While ResNet-18 achieves competitive, and occasionally slightly superior, performance on raw pixel-wise metrics (RMSE and PSNR), ResNet-50 demonstrates a clear advantage on structural and perceptual metrics (SSIM and LPIPS). The success of ResNet-50 in propagating stable, high-quality guidance gradients without degradation can be attributed to

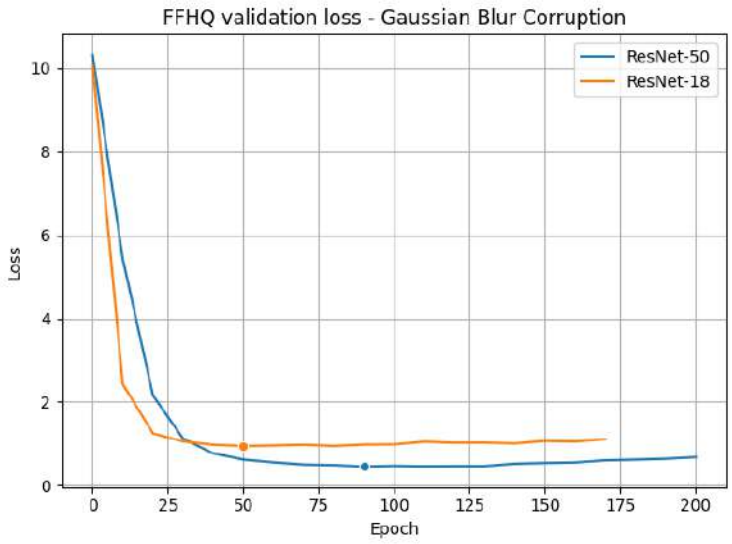


Figure 4: Validation loss comparison between ResNet-18 and ResNet-50. The deeper ResNet-50 architecture converges to a notably lower validation loss during offline MoCo training.

its residual connections, which inherently prevent gradient vanishing [12]. Because it excels at preserving perceptual fidelity, which is the primary goal of the contrastive surrogate, we selected ResNet-50 as the default backbone for all subsequent experiments.

Beyond training convergence, it is critical to validate that the encoder successfully learned a smooth similarity landscape rather than merely memorizing pixel-level distributions. To this end, we evaluated the cosine similarity in the latent space between clean target images and their progressively corrupted counterparts.

As illustrated in Fig. 6, the model effectively separates semantically related and unrelated images within the latent space. Positive pairs (differently corrupted views of the same source image) exhibit a heavily skewed, high cosine similarity, demonstrating the encoder's robustness to measurement degradation. Conversely, the embeddings for negative samples are driven apart, yielding a similarity distribution centered near zero. This distinct separation confirms the efficacy of the contrastive objective.

To further illustrate this, Fig. 7 visualizes the qualitative similarity scores across varying degrees of rotation blur. As the maximum rotation angle (ϕ) increases from 10° to 40°, the positive pairs (the clean reference versus its corrupted views) maintain high cosine similarities, consistently scoring above 0.92. In contrast, the similarity scores for negative pairs (the clean reference versus entirely different identities) remain significantly lower, rarely exceeding 0.5 and often hovering near zero. This substantial margin confirms the encoder's ability to reliably match true identities under severe, non-linear degradations without collapsing the representation space.

Ultimately, these results confirm the representational quality and correctness of the trained MoCo encoders. Rather than merely memorizing structural artifacts, the network

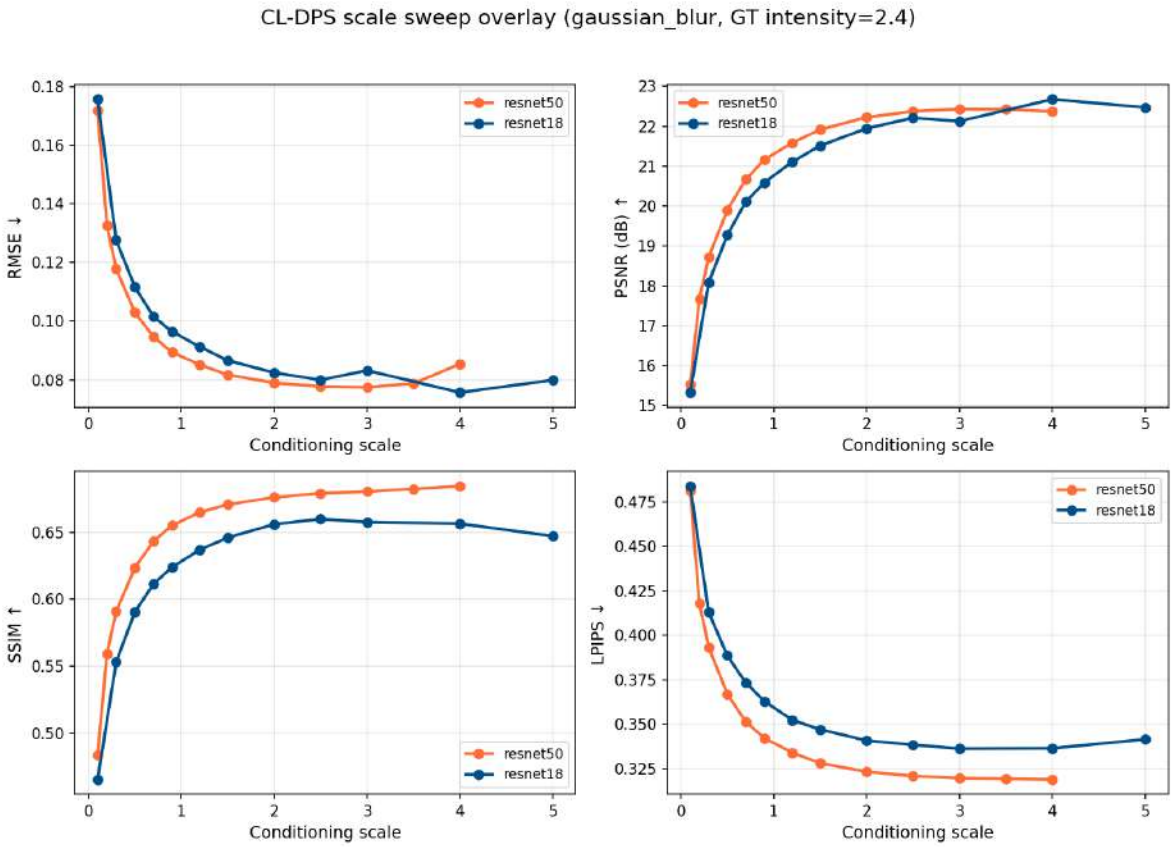


Figure 5: Guidance scale sweep comparing ResNet-18 and ResNet-50 during diffusion inference (Gaussian blur). While both models yield similar RMSE and PSNR, ResNet-50 significantly outperforms ResNet-18 in structural (SSIM) and perceptual (LPIPS) fidelity.

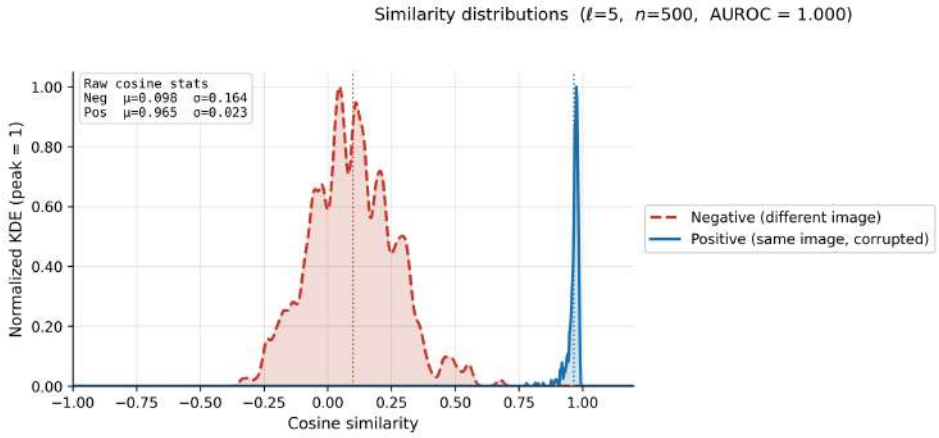


Figure 6: Distribution of Cosine Similarities between positive and negative samples

has successfully constructed a smooth, continuous energy landscape. In this latent space, differently corrupted views of the same identity map to a tight, high-similarity cluster, while unrelated identities are orthogonalized with uncorrelated embeddings.

Crucially, this continuous decay in similarity with respect to degradation severity is exactly what is required for effective diffusion guidance. Because the landscape is smooth, the gradient of the similarity score, $\nabla_{\boldsymbol{x}_t}\langle f(\boldsymbol{x}_t), f(\boldsymbol{y})\rangle$, provides a stable and highly informative directional signal. It effectively points the intermediate diffusion state $\boldsymbol{x}_t$ toward the underlying clean identity of the measurement $\boldsymbol{y}$. By leveraging this validated embedding space as a fully differentiable likelihood surrogate for $p(\boldsymbol{y}|\boldsymbol{x}_t)$, we entirely bypass the need to explicitly model, estimate, or invert the unknown non-linear forward operators. With this robust surrogate established, the following sections evaluate its integration into the reverse diffusion process to perform blind image restoration.

4.4 Standard DPS Baseline and Sanity Checks

Before comparing our approach to the standard Diffusion Posterior Sampling (DPS) baseline, we established a rigorous testing environment to ensure that the standard DPS was operating optimally. Initial preliminary runs exhibited unexpected instability, prompting a thorough hyperparameter validation to confirm the baseline mechanics.

First, we evaluated the sensitivity of the DPS guidance scale parameter (ζ), which dictates the step size of the likelihood gradient during the reverse SDE. Sweeping across a broad range of values (Fig. 8) demonstrated that $\zeta = 0.9$ optimally balances measurement consistency with the unconditional diffusion prior. Additional sweeps were performed to ensure that the DPS algorithm is acting optimally across operators.

Interestingly, as illustrated in Fig. 9, the visual outputs generated across this stable window of scale values are practically identical, with only marginal variations in fine, high-frequency textures such as hair. This confirms that within a reasonable range, the baseline

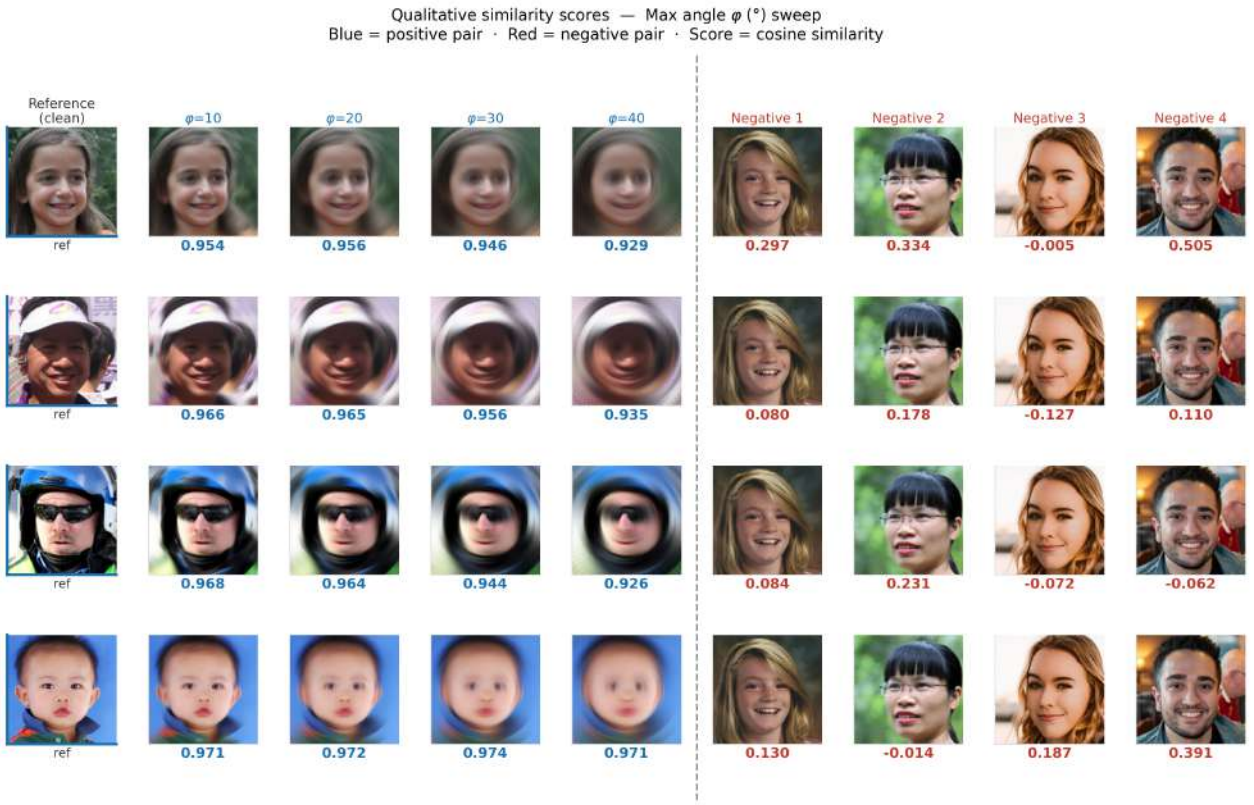


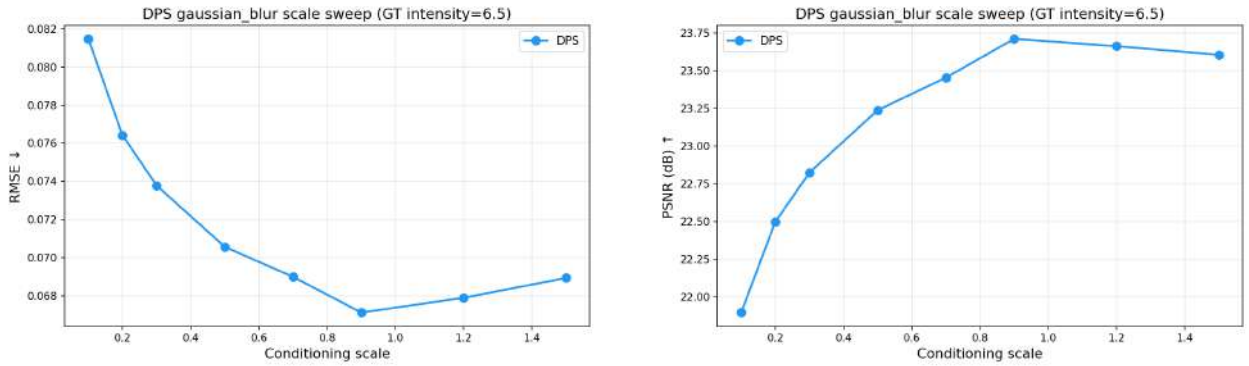
Figure 7: Qualitative cosine similarity scores for positive and negative pairs under varying rotation blur severities. Positive pairs (blue scores) exhibit robust similarity (> 0.92) even at extreme rotation angles ($\phi = 40°$). Conversely, negative pairs depicting different identities (red scores) yield substantially lower similarity scores, demonstrating the encoder's strong discriminative capacity.

is robust to the scale parameter, isolating operator mismatch as the primary source of failure in blind settings.

To further verify the internal mechanics of the DPS baseline, we tracked the L_2 measurement reconstruction loss, $||y - \mathcal{A}(\hat{x}_0(x_t))||_2^2$, at each step of the reverse diffusion trajectory (Fig. 10). Naturally, the loss begins extremely high due to the pure Gaussian noise initialization at x_T. However, as the analytical gradient of the perfectly known operator guides the diffusion SDE, the loss reliably converges toward zero across different inference samples. This diagnostic confirms that the standard analytical guidance functions correctly under ideal, non-blind conditions.

4.5 DPS Sensitivity to Operator Mismatch

Having validated the baseline under ideal conditions, we designed a sensitivity test to expose the fundamental brittleness of standard DPS when confronted with blind or mischaracterized



(a) RMSE evaluation.

(b) PSNR evaluation.

Figure 8: Sweep over values for the guidance scale parameter (ζ) in DPS, comparing RMSE and PSNR metrics.

operators. We corrupted input images using a ground-truth Gaussian blur kernel with a fixed standard deviation of $\sigma_{true} = 6.5$. During inference, we intentionally supplied the DPS algorithm with an incorrect guidance operator, sweeping its assumed intensity across a wide spectrum.

The quantitative results, evaluated across three distinct operators, are shown in Fig. 11. Linear motion blur exhibits a sharp, parabolic "V-shaped" performance curve, which is what we expected. Optimal reconstruction predictably occurs when the guidance operator precisely matches the ground truth, and result in catastrophic reconstructions when the operator varies slightly. Gaussian blur and Rotation blur operators show some tolerance when the operator *underestimates* (i.e. the operator used to guide the inverse diffusion process has a lower corruption capability than the originally used) the corruption, resulting in blurrier but coherent images. Nevertheless, performance deteriorates catastrophically when the operator *overestimates* the blur. In such cases, the guidance gradient forces the diffusion model to hallucinate extreme high-frequency details to compensate for a massive blur that does not exist in the measurement, leading to severe artifacts.

A qualitative example of this behavior is illustrated in Fig. 12, which displays DPS reconstructions of an input image originally corrupted by a Linear Motion blur kernel with a length of 30.0. The figure demonstrates the effect of varying the intensity of the Linear Motion blur guidance operator: lower guidance intensities yield overly blurry outputs, whereas higher intensities introduce visible artifacts and darker borders. Only the image reconstructed with the correct guidance operator (with length 30.0) matches the original with minimal differences. Examples for the different operators can be seen in Appendix A.

4.6 Analysis of the Guidance State: x_t vs. $\hat{x}_0$

During the implementation and reproduction of the CL-DPS framework, we encountered a critical divergence in the choice of the guidance state compared to standard DPS and our

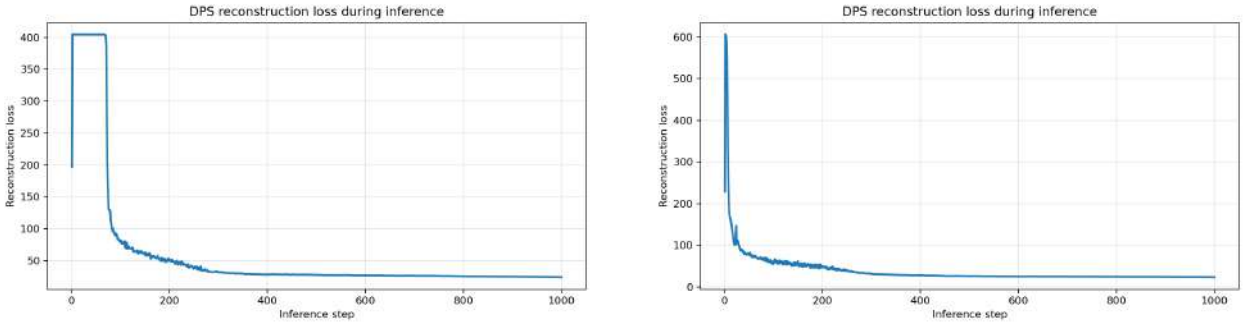
Figure 9: Reconstructed images over the scale parameter sweep

prior work, CoGuide.

Algorithm 2 of the original CL-DPS framework dictates that the contrastive guidance gradient should be computed with respect to the intermediate noisy diffusion state x_t. However, our empirical evaluations revealed that guiding directly on x_t produced highly unstable and unsatisfactory reconstructions. Despite extensive efforts to stabilize the gradients, including applying various normalizations to the encoder inputs, outputs, and the gradients themselves, the resulting images suffered from severe degradation.

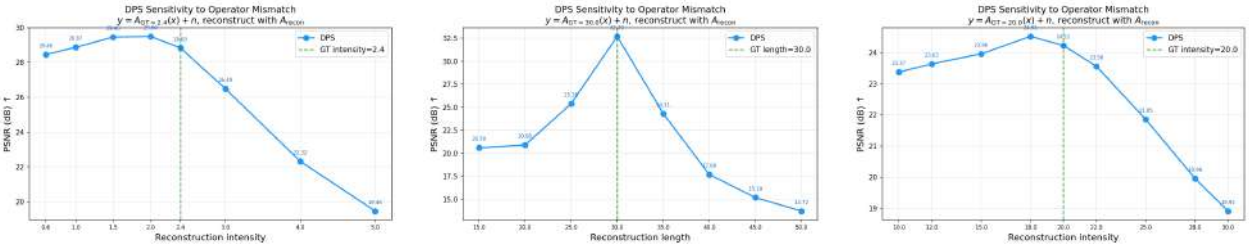
As seen in Fig. 13 (third column), when guiding on x_t, the overall semantic identity of the person is roughly guided towards the correct manifold. However, the image is plagued by severe color shifts, high-frequency noise, and distinct grid-like patch artifacts. This instability likely occurs because the overlapping patch-wise encoder struggles to extract coherent structural features from the heavily noised x_t states at early timesteps, leading to chaotic and destructive gradient signals.

Drawing inspiration from the standard DPS framework and our previous formulation in CoGuide (Algorithm 1), we modified the guidance step to instead evaluate the contrastive loss on the Tweedie's posterior mean estimate, $\hat{x}_0(x_t)$. This estimate provides a denoised prediction of the original signal at every timestep. By feeding $\hat{x}_0$ into the contrastive encoder rather than the noisy x_t, the reconstruction quality improved dramatically. The gradient signal became significantly more stable, the patch-like artifacts vanished entirely, and global color consistency was fully restored. Consequently, we adopted the $\hat{x}_0$ guidance state for our optimal CL-DPS implementation.



(a) Sample 00005 (b) Sample 00004

Figure 10: Comparison of DPS reconstruction loss across different inference runs.



(a) Gaussian Blur (b) Linear Motion Blur (c) Rotation Blur

Figure 11: PSNR sensitivity analysis across different degradation operators. For all three corruption types (Gaussian, linear motion, and rotation blur), DPS exhibits a sharp drop in reconstruction quality when the assumed inference operator deviates from the ground-truth parameters, highlighting the method's brittleness to operator mismatch.

4.7 CL-DPS Inference Dynamics and Limitations

With the optimal $\hat{x}_0$ guidance state established, we fully deployed our contrastive model (CL-DPS) to the blind inverse tasks. To thoroughly understand the internal behavior of this contrastive guidance during generation, we simultaneously tracked the pixel-wise reconstruction loss and the contrastive cosine similarity over the 1000 reverse diffusion timesteps (Fig. 14).

As expected, the cosine similarity steadily increases, reliably converging toward 1.0 across the evaluated samples. This confirms that the reverse diffusion process successfully navigates the contrastive energy landscape, effectively pulling the intermediate generated state $\boldsymbol{x}_t$ toward the semantic identity of the corrupted measurement. However, we observed an interesting phenomenon regarding the explicit L_2 pixel-wise reconstruction loss: while it initially drops alongside the cosine similarity, it occasionally plateaus or even diverges slightly in the final stages of inference ($t \to 0$).

This divergence highlights a fundamental characteristic and trade-off of purely latent-guided generative models. Because the MoCo encoder is explicitly trained via data augmen-

Figure 12: DPS reconstruction results across a range of guidance intensities for an input corrupted by Linear Motion blur (length 30.0).

tation to be invariant to severe structural perturbations (which is precisely what grants it robustness against unknown blur operators), it inherently prioritizes high-level perceptual and semantic features over strict, pixel-perfect spatial adherence. Consequently, the final generated image may drift slightly in raw pixel space, hallucinating plausible high-frequency textures or minor structural shifts, even while it perfectly minimizes the contrastive energy score. We perform an analysis of this phenomenom at Section 4.8.

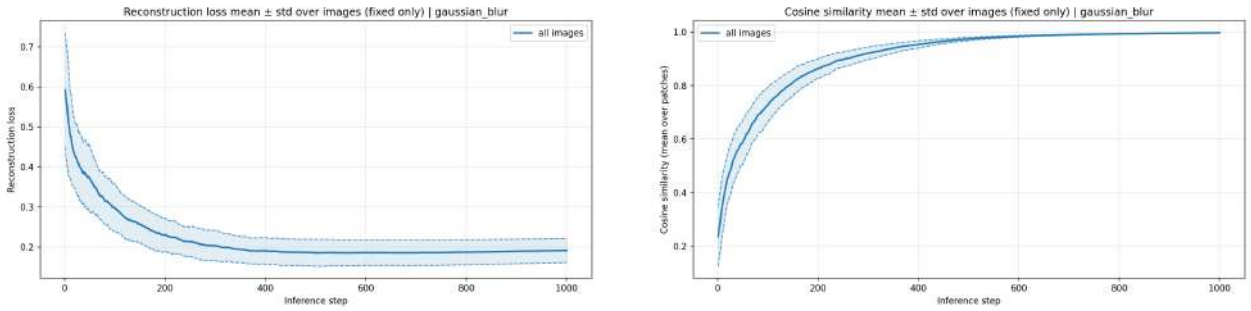
We also performed a sweep over the guidance scale parameter (ζ). As illustrated in Fig. 15, the optimal value for the contrastive guidance was found to be 3.0 for Gaussian blur. Sweeps were also performed for the other operators to ensure optimal performance from the algorithm.

We also performed a sweep over the guidance scale parameter (ζ). As illustrated in Fig. 15, the optimal value for the contrastive guidance was found to be $\zeta = 3.0$ for Gaussian blur. This magnitude is notably higher than the optimal scale found for standard DPS ($\zeta = 0.9$), which is theoretically consistent: because CL-DPS derives its guidance from a bounded cosine similarity within a normalized latent space, a larger scalar multiplier is required to match the influence of standard DPS gradients computed directly in the unbounded pixel domain. Sweeps were similarly performed for the other operators to ensure optimal algorithmic performance.

Furthermore, Fig. 16 investigates the Out-of-Distribution (OOD) generalization of CL-DPS by sweeping Gaussian blur intensities from $\sigma = 0.0$ to 3.0. This range was deliberately chosen to evaluate intensities strictly inside the MoCo model's training distribution ($[0.6, 2.4]$) as well as outside of it (< 0.6 and > 2.4). Theoretically, one might expect a sharp deterioration in performance beyond these boundaries, as the encoder had never seen those specific perturbations. Instead, the observed results deviate from this expectation: the reconstruction quality improves as the intensity increases and finally it degrades, without

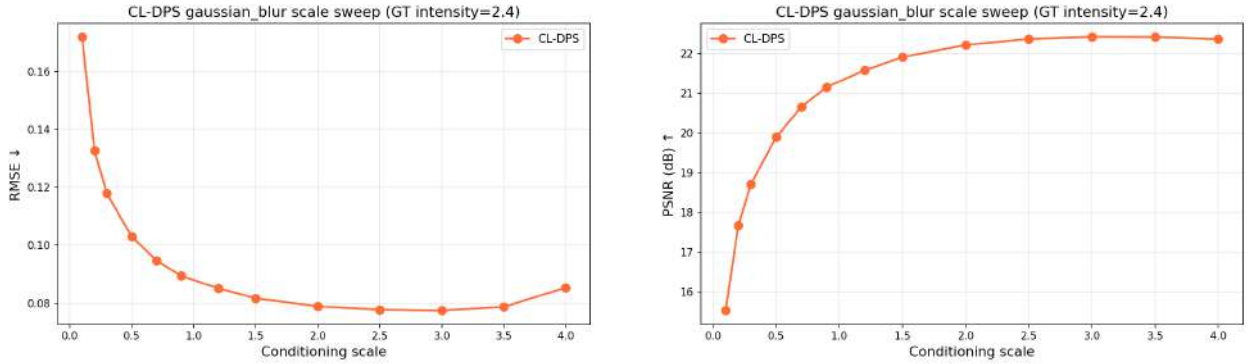
(a) Ground Truth (b) Measurement (c) Guidance on x_t (d) Guidance on $\hat{x}_0$

Figure 13: Qualitative comparison of the diffusion guidance state. Guiding the reverse process directly on the noisy state x_t (as originally proposed in CL-DPS) forces the encoder to process heavy noise, resulting in severe grid artifacts and color distortions. Modifying the algorithm to evaluate the contrastive surrogate on the denoised Tweedie prediction $\hat{x}_0$ yields a clean, high-fidelity reconstruction.



(a) Reconstruction Loss (b) Cosine Similarity

Figure 14: Evolution of metrics over 1000 steps in inference, showing mean and standard deviation for Reconstruction Loss and Cosine Similarity.



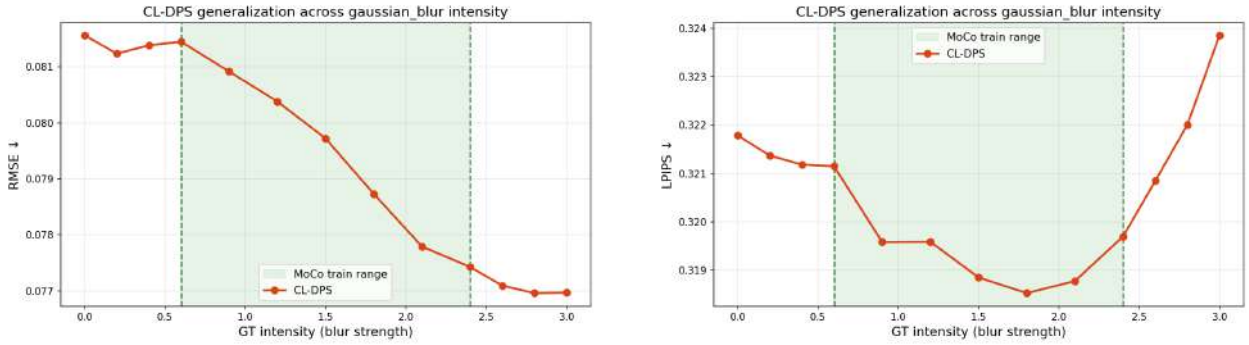
(a) RMSE vs. Guidance Scale (ζ) (b) PSNR vs. Guidance Scale (ζ)

Figure 15: Guidance scale parameter (ζ) sweep for CL-DPS using Gaussian Blur, evaluated via RMSE (left) and PSNR (right).

any sharp boundaries at 0.6 or 2.4. While PSNR and RMSE decay predictably at higher intensities, the perceptual LPIPS metric forms a distinct valley aligning with the training range. This suggests that while macroscopic structure is preserved OOD, subtle perceptual artifacts emerge that perturb the embeddings obtained through VGG. The lack of a catastrophic failure boundary implies that these moderate OOD magnitudes are insufficient to completely break the learned contrastive representations.

The qualitative reconstructions across this sweep are displayed in Fig. 17. The generated images consistently maintain high fidelity to the ground truth, even when slightly exceeding the training range. To test the limits of the system, we additionally evaluated an extreme OOD corruption with an intensity of $\sigma = 6.5$. At this level, the reconstruction significantly deviates from the ground truth. However, the nature of this failure is highly revealing: rather than injecting severe high-frequency ringing or structural artifacts, as standard DPS did under operator mismatch, CL-DPS defaults to the unconditional diffusion prior.

Because the heavily corrupted measurement provides an insufficient identity signal, the model hallucinates a plausible, photorealistic human face that simply lacks the specific identity of the ground truth. This confirms that the contrastive guidance of the MoCo model preserves the original data distribution of the generative process, ensuring visually coherent outputs even when the underlying identity is entirely lost. This is a clear effect of the smoothness of the latent space, as unknown perturbations are projected into a more general embedding that minimizes the effect of the surrogate likelihood. Expanding the training distribution of the contrastive model to contain these extreme intensities would likely restore identity-matching capabilities in these edge cases.



(a) RMSE evaluation.

(b) LPIPS evaluation.

Figure 16: CL-DPS performance across varying intensities of Gaussian blur corruption, measured in LPIPS and RMSE.

4.8 Impact of Guidance Truncation (Scale Cutoff)

As mentioned before, the reconstruction loss during inference seemed to diverge at the end of the inverse diffusion process. This suggested that the contrastive learning guidance might be

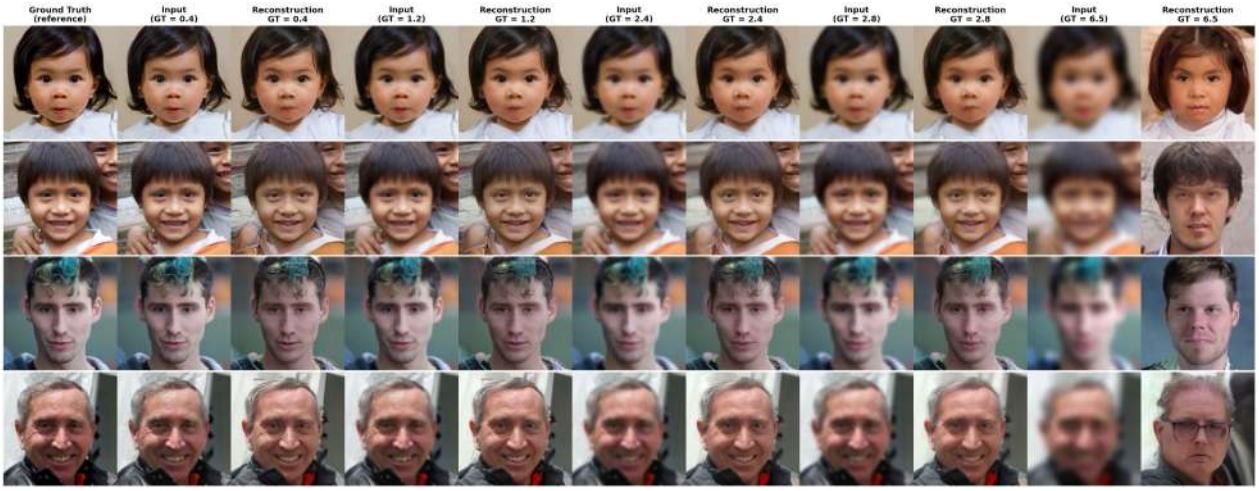


Figure 17: Generated images using CL-DPS accross varying intensities of Gaussian blur Corruption.

counterproductive in the last stages of inference, where the image is mostly generated, and that extra guidance might lead to some artifacts being generated that improve the cosine similarity at the cost of worse reconstruction loss.

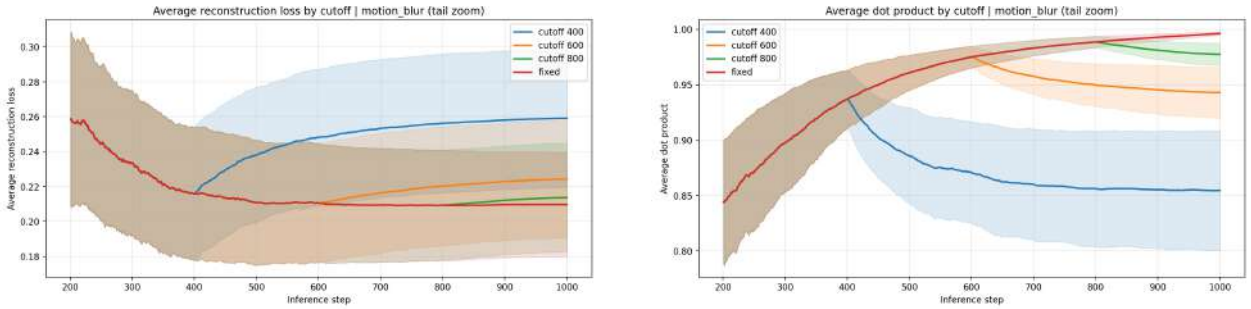
To investigate the temporal importance of contrastive guidance, we conducted an experiment where the guidance scale ζ is set to zero after a specific number of diffusion steps (Fig. 18). We evaluated cutoffs at $t = 400$, $t = 600$, and $t = 800$ steps, comparing them against a baseline where guidance is maintained throughout the entire 1000-step trajectory. As expected, the cosine similarity (dot product) plateaus or begins to diverge the moment the scale is truncated, indicating that the model stops aligning the generation with the measurement features.

Regarding the reconstruction loss, we observe a significant divergence when the cutoff is applied early at $t = 400$, as the unconditional prior lacks sufficient structural information to complete the reconstruction accurately. In contrast, for the $t = 600$ and $t = 800$ cutoffs, the loss maintains a trajectory similar to the baseline, suggesting that much of the fundamental layout is established by the midpoint of the inference process. Ultimately, maintaining a fixed scale throughout the entire trajectory yields the lowest final reconstruction loss, although the marginal gains after $t = 800$ appear to be minimal.

As there is no clear advantage in cutting off the guidance from the contrastive learning model, in addition to another hyperparameter adding complexity to the system, we decided to leave the guidance throughout the whole inverse diffusion process.

4.9 CL-DPS vs DPS

The final phase of our evaluation directly compares the standard DPS baseline against CL-DPS under conditions of operator mismatch. Fig. 19 plots the RMSE of both methods as



(a) Motion Blur: Reconstruction Loss

(b) Motion Blur: Dot Product Similarity

Figure 18: Temporal sensitivity of contrastive guidance for Gaussian and Motion blur. The plots track average reconstruction loss and cosine similarity (dot product) when the guidance scale is truncated at different stages of the reverse diffusion process.

the assumed inference operator deviates from the ground truth parameters ($\Delta = 0$).

This experiment encapsulates the core thesis of our project. As expected, standard DPS (plotted in blue) achieves its peak performance, and slightly edges out CL-DPS, strictly when its analytical guidance operator perfectly matches the ground truth. However, the moment the assumed parameters deviate, the analytical gradients mislead the diffusion process. This causes a distinct "V-shaped" escalation in RMSE. Motion blur exhibits the most catastrophic divergence, while Gaussian blur shows a slightly more forgiving curve when the blur is underestimated, aligning with our earlier findings in Fig. 11.

Conversely, CL-DPS (plotted in orange) plots a perfectly flat, horizontal line across all mismatch values. This is the direct result of the blind framework's design: CL-DPS does not require, nor does it use, explicit operator parameters (such as intensity, length, or angle) during inference. It relies solely on the corrupted measurement y and its learned contrastive surrogate. Consequently, its reconstruction process is completely invariant to parameter misspecification. While its baseline error is marginally higher than a perfectly tuned, non-blind DPS oracle (an RMSE difference of less than 0.1), its resilience effectively eliminates the catastrophic failures associated with blind or inaccurately specified inverse problems.

This quantitative stability translates into dramatic qualitative differences, as depicted in Fig. 20. In this example, the input image is corrupted by a motion blur with an intensity of 0.5. When standard DPS (third row) is fed this exact parameter, it successfully recovers the image. However, when provided with mismatched intensities (e.g., 0.1, 0.3, 0.7, or 0.9), the analytical gradients violently disrupt the diffusion process, resulting in severe artifacts.

In contrast, CL-DPS (bottom row) consistently recovers a high-quality, artifact-free image. Because it bypasses the explicit mathematical operator entirely, its output remains identical and stable across every column. This demonstrates the practical superiority of utilizing a contrastive likelihood surrogate in blind inverse scenarios where true degradation parameters are unavailable.

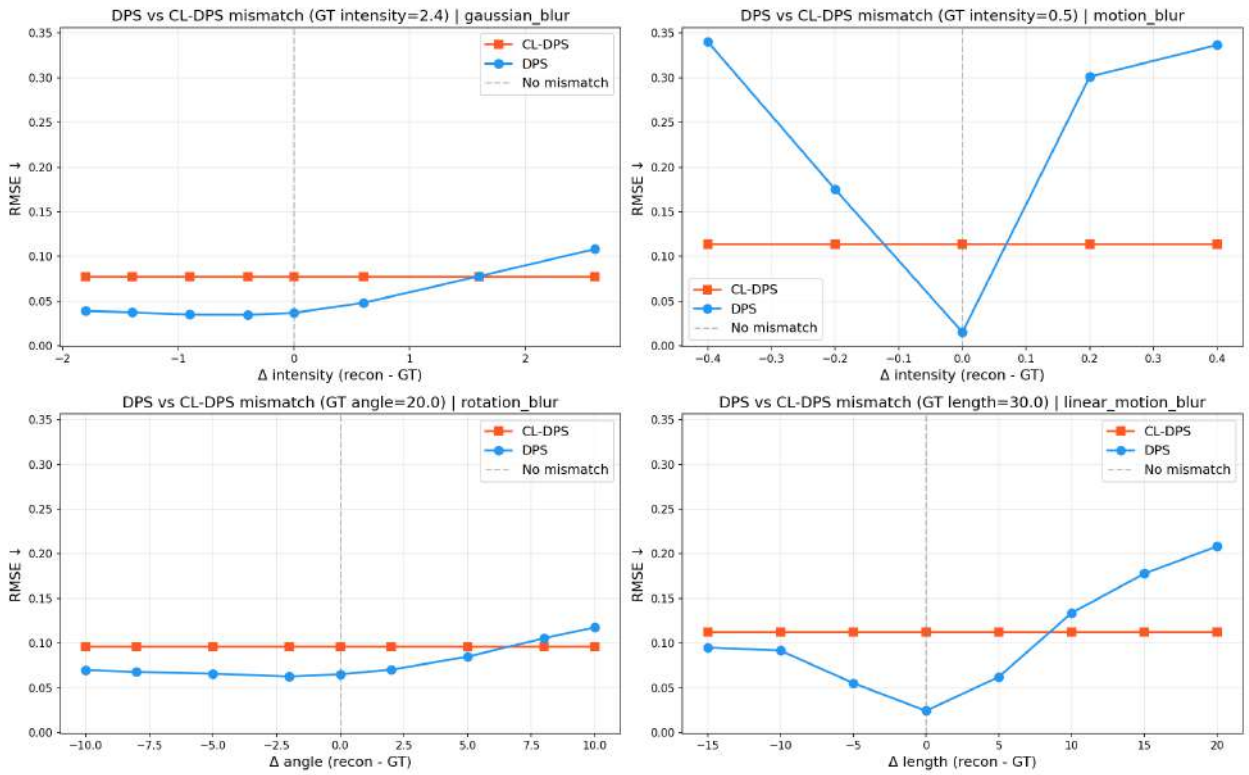


Figure 19: Quantitative comparison of DPS vs. CL-DPS under operator mismatch. Standard DPS degrades sharply when the assumed operator parameters deviate from the ground truth. CL-DPS, relying solely on the observed image, maintains a flat, highly stable performance across the entire spectrum.

Figure 20: Qualitative comparison demonstrating the robustness of CL-DPS to operator mismatch. The top two rows display the ground truth and the fixed input image, corrupted by a motion blur of intensity 0.5. The third row shows standard DPS reconstructions using assumed blur intensities ranging from 0.1 to 0.9; slight variations in the guidance operator introduce severe, destructive artifacts. In contrast, the bottom row demonstrates that CL-DPS consistently recovers a high-quality image regardless of the assumed parameter, remaining completely stable.

4.10 Expanding to the Fully Blind Setting: Mixed MoCo and Expert Routing

While the specialist MoCo models perform well when confronted with their exact respective corruption families, a fully blind scenario implies that not only the parameters but also the *family* of the degradation operator is completely unknown. To counter this, two distinct strategies were explored: deploying a single "Universal" Mixed MoCo encoder, and training an external "Router" to classify the degradation and select the optimal specialist expert.

4.10.1 Expert Routing: Classifier Training

The routing mechanism attempts to identify the underlying corruption applied to the input measurement y and dynamically routes it to the correct specialist MoCo checkpoint (e.g., Gaussian, Motion, or Rotation). We formulated this as a supervised image classification problem on progressively corrupted images.

A crucial design choice was selecting the router's backbone architecture. We compared extracting the embeddings from a pretrained and frozen Universal MoCo encoder (pairing it with a simple linear probe) against training a supervised ResNet-18 from scratch. The intuition behind the MoCo linear probe was that the contrastive latent space might already organically disentangle degradation families.

However, as demonstrated in Table 1, the frozen MoCo features failed to accurately classify the degradation, yielding a top-1 validation accuracy of a mere 37.5% for the Mixed model and 35.0% for the Gaussian-specialist model. Conversely, the supervised ResNet-18 model, trained from scratch, achieved near-perfect classification accuracy (99.4%). This contrast highlights an essential property of the MoCo embeddings: because they are explicitly trained via data augmentation to be invariant to massive structural corruptions, they actively discard the low-level noise signatures required to classify the degradation itself. By design, the contrastive latent space collapses differently corrupted views of the same underlying image into a single identity cluster, rendering the extracted features inherently blind to the specific degradation operator.

Router Model	Training Policy	Top-1 Accuracy	Top-2 Accuracy
ResNet-18	Supervised (Scratch)	**99.4%**	**100.0%**
MoCo Mixed	Frozen MoCo Features	37.5%	69.9%
MoCo Gaussian	Frozen MoCo Features	35.0%	69.6%

Table 1: Validation classification performance for different Router architectures. Training a network from scratch significantly outperforms the frozen MoCo representation features.

4.10.2 Routed CL-DPS vs. Mixed MoCo Performance

With an effective router established, we successfully integrated it into the CL-DPS inference pipeline. For each input image, the router predicts the degradation class, dynamically loads the corresponding specialist encoder, and provides the downstream contrastive likelihood guidance. We compared this routing strategy against the Universal Mixed MoCo model and fixed individual oracle experts.

Table 2 summarizes the quantitative reconstruction quality for inputs afflicted by different random degradations. The results reveal that the router approach successfully matches the performance of the oracle experts (the correct fixed specialist) by efficiently identifying the corruption. While the Mixed MoCo model remains robust, it falls slightly behind the

individually routed experts. Constructing a single smooth latent plane that accommodates multiple wildly different degradation physics naturally diffuses and reduces the specificity of the generation gradients compared to a specialized network.

	Gaussian Corruption		Rotation Corruption		Linear Motion Corruption	
Guidance Model	**RMSE ↓**	**LPIPS ↓**	**RMSE ↓**	**LPIPS ↓**	**RMSE ↓**	**LPIPS ↓**
Fixed Gaussian Expert	**0.077 ± 0.017**	**0.320 ± 0.046**	0.103 ± 0.022	**0.403 ± 0.049**	0.119 ± 0.025	0.482 ± 0.041
Fixed Motion Expert	0.115 ± 0.037	0.475 ± 0.107	0.119 ± 0.018	0.494 ± 0.079	0.128 ± 0.018	0.548 ± 0.096
Fixed Rotation Expert	0.089 ± 0.014	0.373 ± 0.048	0.108 ± 0.019	0.430 ± 0.059	0.125 ± 0.018	0.519 ± 0.039
Router-Selected	**0.077 ± 0.017**	**0.320 ± 0.046**	0.108 ± 0.019	0.430 ± 0.059	0.126 ± 0.020	0.542 ± 0.087
Universal Mixed MoCo	0.098 ± 0.018	0.422 ± 0.074	**0.097 ± 0.018**	0.420 ± 0.075	**0.096 ± 0.017**	**0.421 ± 0.083**

Table 2: Quantitative evaluation of CL-DPS reconstruction under various corruption scenarios using distinct contrastive guidance strategies. Results are reported as mean ± standard deviation. Notably, the Universal Mixed MoCo model achieves highly competitive metrics across multiple degradations, while the Fixed Gaussian expert demonstrates unexpected cross-degradation robustness, often surpassing specialized experts.

Interestingly, the Fixed Gaussian expert demonstrates strong cross-degradation generalization. While it naturally excels on Gaussian corruption, it also achieves highly competitive metrics on rotation and linear motion blur, outperforming the dedicated experts for those specific degradations. This suggests that training on isotropic Gaussian degradation may encourage the model to learn highly robust, generalizable latent representations. However, as illustrated in Figure 21, quantitative superiority does not perfectly correlate with perceptual quality. Despite the Gaussian expert's strong numerical performance on rotation blur, qualitative inspection reveals that it fails to fully resolve the specific directional artifacts compared to the dedicated Fixed Rotation expert. Finally, the Universal Mixed MoCo model yields unexpectedly strong quantitative results, securing the best performance in three of the six evaluated metric categories, notably dominating linear motion corruption. This indicates that the universal model successfully maps an effective and versatile latent space, maintaining high fidelity even when tasked with resolving multiple, disparate operators simultaneously.

Figure 21: Qualitative comparison of reconstruction models under rotation blur corruption. While the Fixed Gaussian expert achieves competitive quantitative metrics (see Table 2), visual inspection reveals residual directional artifacts (visible as streaks). In contrast, the dedicated Fixed Rotation expert, the Router-Selected approach, and the Universal Mixed MoCo successfully resolve these specific degradation patterns, highlighting the necessity of qualitative assessment alongside numerical metrics.

Chapter 5 Conclusions and Future Work

The primary objective of this Bachelor's Thesis was to develop, implement, and validate a robust generative framework capable of solving blind inverse problems without relying on explicit, exact mathematical degradation operators. Standard analytical methods, such as Diffusion Posterior Sampling (DPS), strictly require perfect knowledge of the forward operator. The work presented in this document conclusively demonstrates that relying on analytical gradients is fundamentally brittle in uncertain environments, and successfully validates the Contrastive Learning for Diffusion Posterior Sampling (CL-DPS) framework as a highly stable alternative.

5.1 Conclusions

By systematically evaluating both baseline and contrastive methodologies, this project has fulfilled its established objectives and yielded several key insights into blind inverse problems. First, our systematic quantification of operator mismatch vulnerability confirmed the initial hypothesis regarding standard Diffusion Posterior Sampling (DPS). While standard DPS excels as an oracle when provided with the perfect forward operator, it exhibits a catastrophic, "V-shaped" degradation in performance the moment the assumed operator parameters deviate from the ground truth. This failure is particularly severe for directional and non-linear motion blurs, rendering traditional DPS unusable in real-world, blind scenarios.

To address this vulnerability, the implemented CL-DPS framework successfully provided a stable likelihood surrogate. During the architectural optimization of this contrastive sur-

rogate, ResNet-50 was proven to capture a significantly richer and more perceptually aligned latent space than shallower networks. More importantly, this work identified a critical flaw in applying contrastive guidance directly to the intermediate noisy diffusion state x_t, which introduced severe high-frequency noise and grid-like artifacts. By modifying the inference pipeline to guide the diffusion process using Tweedie's denoised prediction ($\hat{x}_0$), structural coherence and global color consistency were completely restored. As a result, CL-DPS maintained a perfectly flat, highly stable reconstruction performance across the entire spectrum of operator mismatches. Instead of forcing destructive high-frequency artifacts, the contrastive guidance effectively decoupled the generation process from explicit parameter estimation.

Finally, the project successfully expanded the framework to handle strictly zero-shot blind settings, where even the degradation family is unknown. We demonstrated that a dedicated ResNet-18 router could classify structural degradations with 99.4% accuracy, successfully deploying targeted MoCo experts dynamically. Furthermore, the Universal (Mixed) MoCo encoder proved that a single foundational model can map multiple different degradation physics into a unified, robust latent plane, offering highly competitive performance without any routing mechanisms. Ultimately, these findings confirm that mapping diverse visual degradations into a unified contrastive latent space provides a vastly superior, robust, and mismatch-invariant guidance signal for diffusion models facing uncertain inverse problems.

5.2 Future Work

While the CL-DPS framework has demonstrated remarkable robustness, this study highlights several promising avenues for future research, particularly regarding algorithmic refinement and domain expansion. A primary focus for future work should be closing the baseline reconstruction error gap between CL-DPS and a perfectly tuned, non-blind DPS oracle. Employing advanced representation learning techniques, such as hard negative mining, integrating Vision Transformer (ViT) backbones, or utilizing multi-scale contrastive loss formulations, could yield a more precisely disentangled embedding space. This would allow the likelihood surrogate to preserve identity with greater pixel-wise fidelity, matching the raw metric performance of analytical gradients without sacrificing operator invariance.

Additionally, expanding the out-of-distribution (OOD) prior is critical. As observed in extreme intensity sweeps (e.g., Gaussian blur at $\sigma = 6.5$), when the measurement corruption vastly exceeds the MoCo training distribution, the contrastive signal weakens. This causes the diffusion model to default to its unconditional prior and hallucinate a different identity. Broadening the augmentation boundaries during the offline MoCo training phase will help construct a more inclusive energy landscape capable of handling these extreme physical corruptions.

Beyond algorithmic improvements, the practical application of this framework presents exciting opportunities. The current methodology was validated using synthetic degradations on the FFHQ dataset. To fully align with the outlined Sustainable Development Goals (SDG 3 and SDG 9), the next logical step is to deploy this blind framework on real-world medical imaging datasets, such as MRI or CT scans suffering from actual patient movement, as

well as physical camera shake datasets. Furthermore, building upon the foundational ideas of CoGuide, the insights gained regarding $\hat{x}_0$ stabilization and Universal MoCo encoders should be ported back to discrete spatial planning problems and extended to audio signal processing, where blind deconvolution and unknown room acoustics present mathematically identical challenges.

Finally, addressing the computational bottleneck of pure latent-guided diffusion is essential for practical deployment. Because the current methodology requires hundreds of reverse sampling steps, integrating accelerated inference techniques, such as Denoising Diffusion Implicit Models (DDIM) or exploring Consistency Models under contrastive guidance, would dramatically enhance the real-time utility of the framework in industrial settings.

Chapter 6 References

[1] S. Basu, C. Amballa, Z. Xu, J. V. Sampedro, S. Nelakuditi, and R. R. Choudhury, *Contrastive diffusion guidance for spatial inverse problems*, 2026. arXiv: 2509.26489 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2509.26489.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models", *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[3] H. Chung, J. Kim, M. T. Mccann, M. L. Klasky, and J. C. Ye, "Diffusion posterior sampling for general noisy inverse problems", *arXiv preprint arXiv:2209.14687*, 2022.

[4] B. Efron, "Tweedie's formula and selection bias", *Journal of the American Statistical Association*, vol. 106, no. 496, pp. 1602–1614, 2011. DOI: 10.1198/jasa.2011.tm11181.

[5] L. Weng, "Contrastive representation learning", *lilianweng.github.io*, May 2021. [Online]. Available: https://lilianweng.github.io/posts/2021-05-31-contrastive/.

[6] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding", *arXiv preprint arXiv:1807.03748*, 2018.

[7] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, *Momentum contrast for unsupervised visual representation learning*, 2020. arXiv: 1911.05722 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1911.05722.

[8] L. Ye, S. M. Hamidi, M. Pilanci, and K. N. Plataniotis, "CL-DPS: A contrastive learning approach to blind nonlinear inverse problem solving via diffusion posterior sampling", in *The Fourteenth International Conference on Learning Representations*, 2026. [Online]. Available: https://openreview.net/forum?id=KoLYNHJRBY.

[9] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity", *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[10] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[11] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, *Gans trained by a two time-scale update rule converge to a local nash equilibrium*, 2017.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition", *CoRR*, vol. abs/1512.03385, 2015. arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385.

Chapter A More examples

Figure 22: DPS reconstruction results across a range of guidance intensities for an input corrupted by Gaussian blur (intensity 2.4).

Figure 23: DPS reconstruction results across a range of guidance intensities for an input corrupted by Rotation blur (intensity 20.0).

Figure 24: DPS reconstruction results across a range of guidance intensities for an input corrupted by nonlinear Motion blur (intensity 0.5).

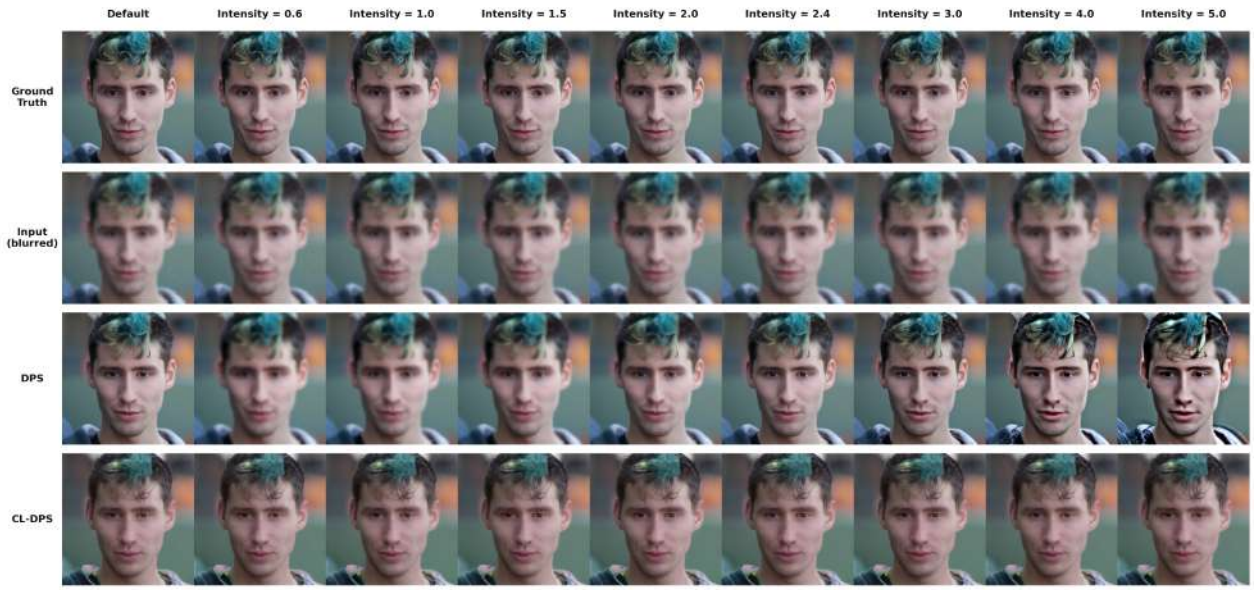
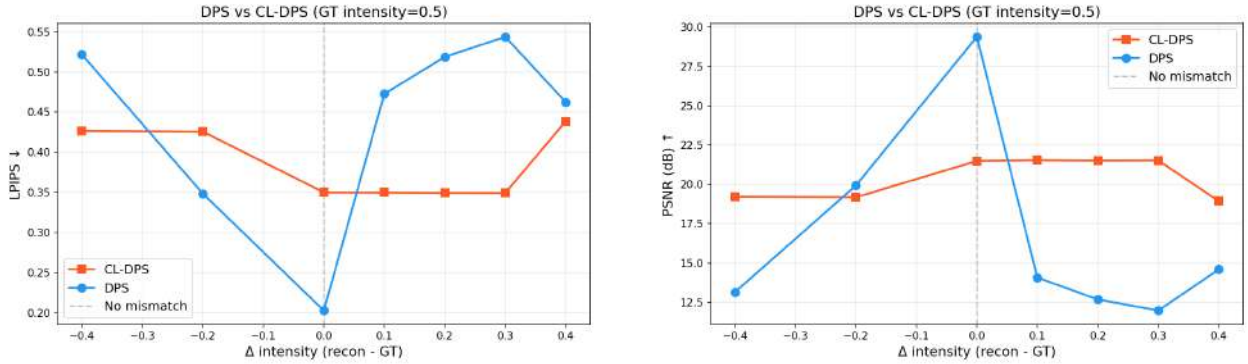
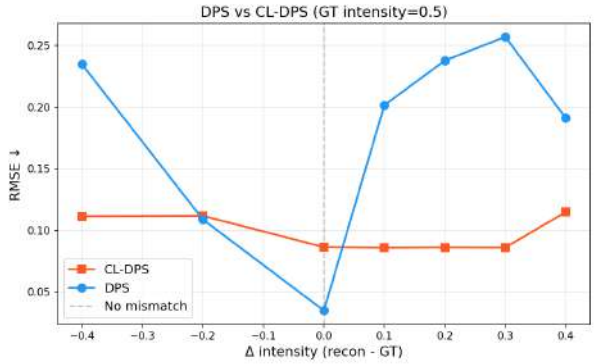


Figure 25: Qualitative comparison demonstrating the robustness of CL-DPS to operator mismatch. The top two rows display the ground truth and the fixed input image, which was corrupted by a Gaussian blur of intensity 2.4. The third row shows standard DPS reconstructions using assumed blur intensities ranging from 0.6 to 5.0. DPS is highly sensitive to this parameter: underestimating the intensity leaves residual blur, while overestimating introduces severe over-sharpening artifacts. In contrast, the bottom row demonstrates that CL-DPS consistently recovers a high-quality image regardless of the mismatched intensity parameter provided to the reconstruction operator.

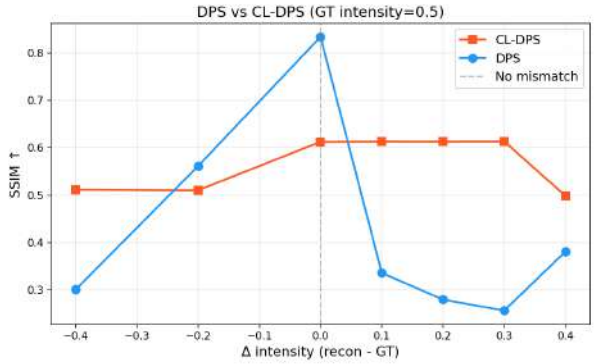


(a) LPIPS Comparison (↓)

(b) PSNR Comparison (↑)

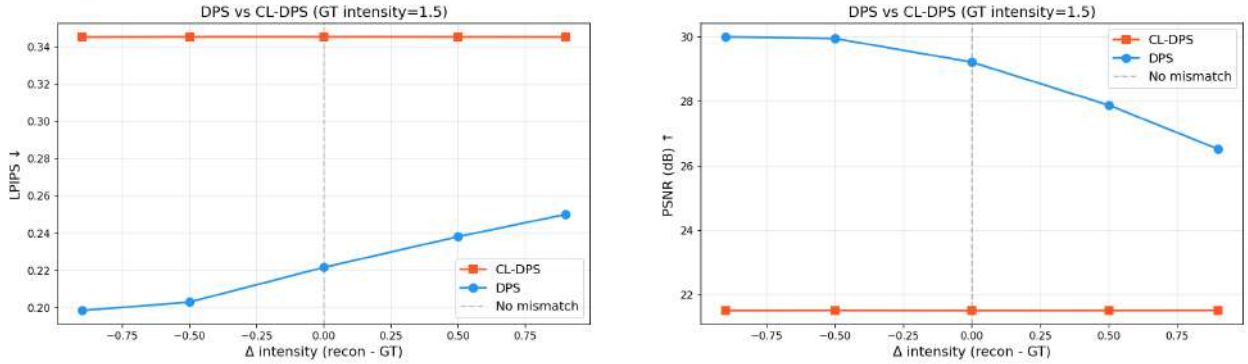


(c) RMSE Comparison (↓)



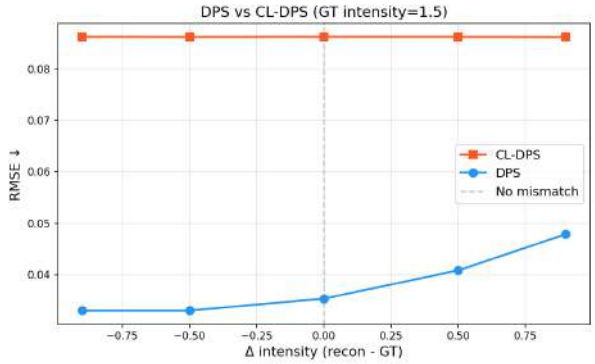
(d) SSIM Comparison (↑)

Figure 26: Comparison of DPS vs. CL-DPS under motion blur corruption. Standard DPS (blue) shows extreme sensitivity to operator mismatch, while CL-DPS (orange) demonstrates robust performance regardless of intensity delta.

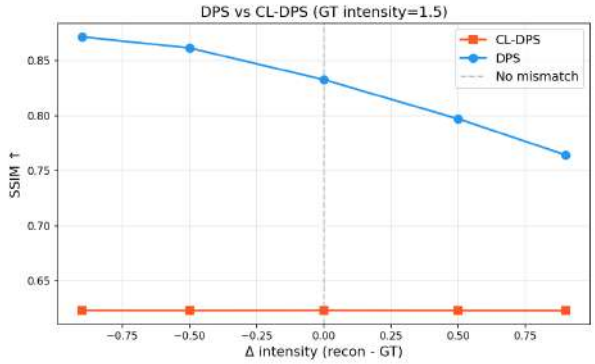


(a) LPIPS Comparison (↓)

(b) PSNR Comparison (↑)



(c) RMSE Comparison (↓)



(d) SSIM Comparison (↑)

Figure 27: Comparison of DPS vs. CL-DPS under Gaussian blur corruption ($GT\sigma = 1.5$). In this regime, standard DPS demonstrates higher stability compared to the motion blur experiments, consistently outperforming CL-DPS across the intensity sweep.

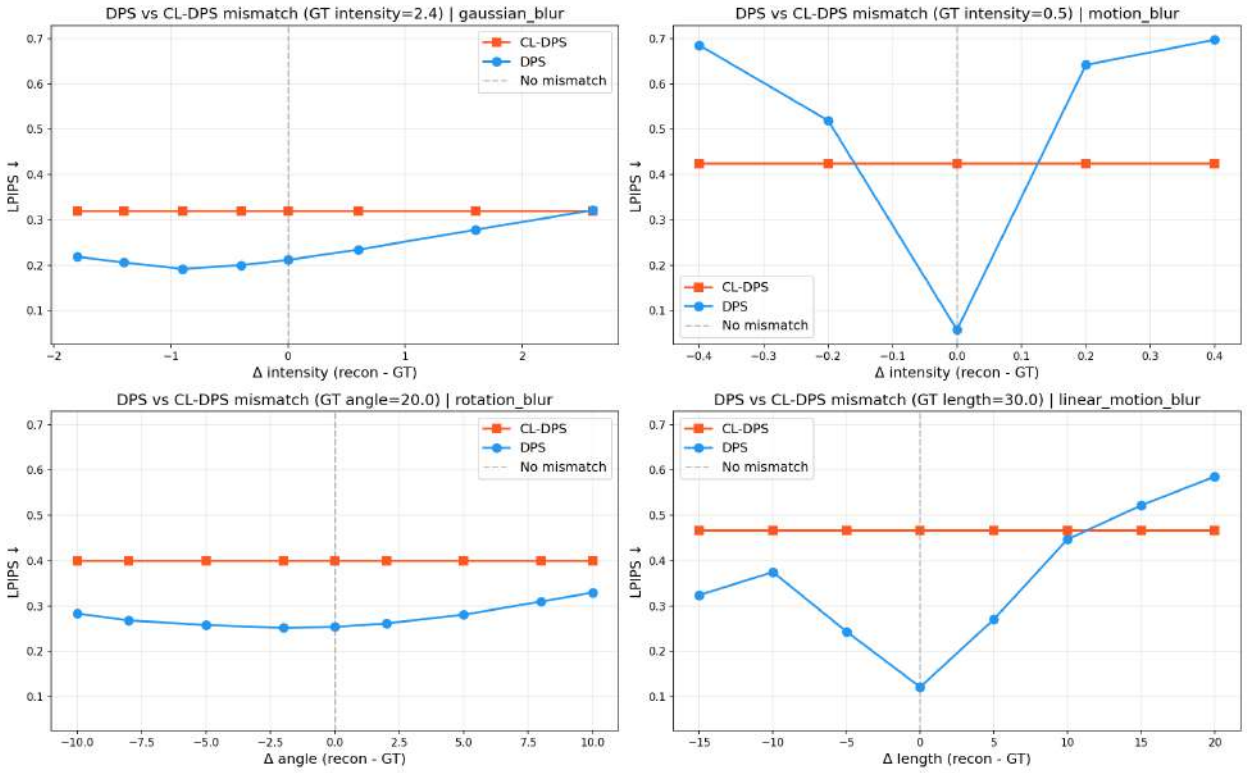


Figure 28: Comparison of DPS vs. CL-DPS

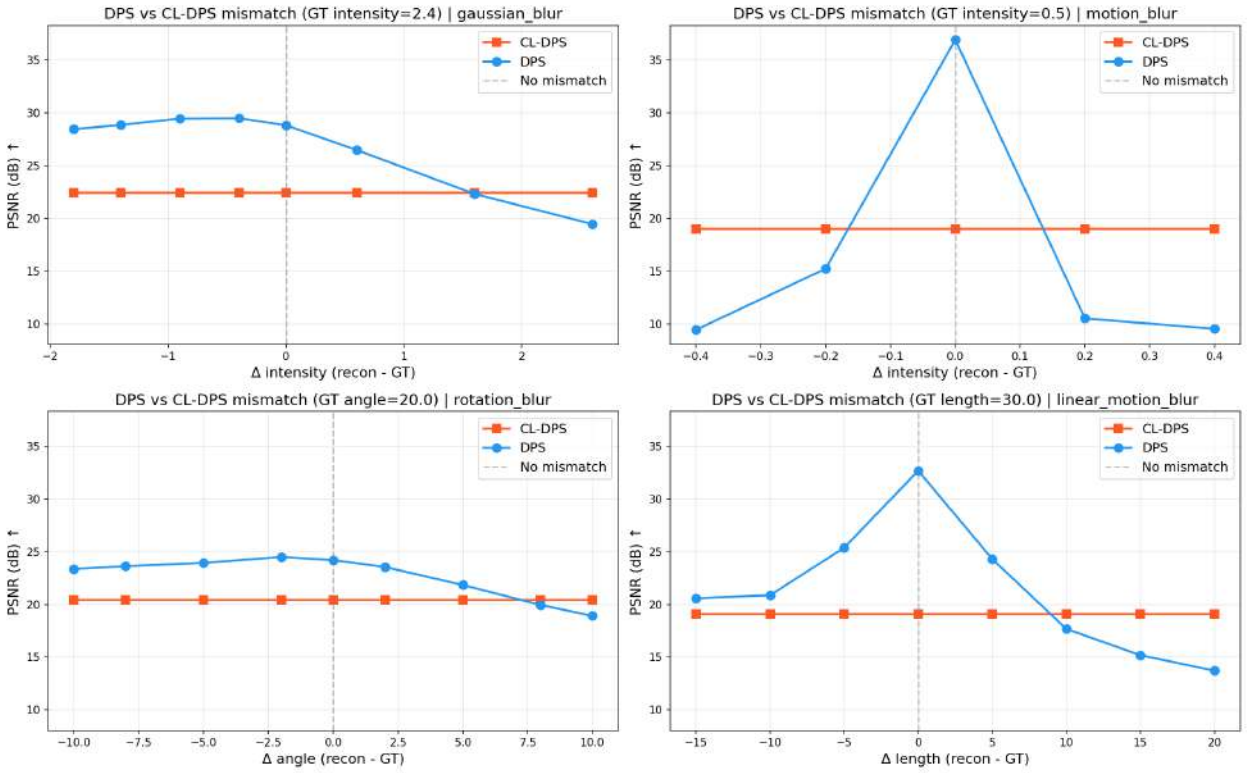


Figure 29: Comparison of DPS vs. CL-DPS