ANaLOG: Anisotropic Native-Latent Operator Guidance for Solving Inverse Problems

Anonymous authors
Paper under double-blind review

Abstract

Native-latent guidance is a recent paradigm for solving inverse problems with latent diffusion models. It replaces repeated evaluations of the image-space forward model, each requiring a decoder pass, with efficient guidance computed using a learned latent-space surrogate. However, existing methods apply guidance uniformly across latent dimensions, ignoring that measurements are informative only along certain directions and that the reliability of model predictions varies across inputs and timesteps. We propose `ANaLOG`, a framework for efficient uncertainty-aware guidance with pretrained latent diffusion models. `ANaLOG` models uncertainty by learning an anisotropic, input- and time-dependent covariance that is integrated into the guidance mechanism to emphasize reliable directions and downweight uncertain ones. We theoretically analyze this framework in a linear model setting and prove that anisotropic, uncertainty-aware weighting is necessary for correct sampling, whereas isotropic guidance induces sampling errors. Experiments across five challenging inverse problems show that `ANaLOG` improves perceptual reconstruction quality over existing methods while preserving efficiency.

1 Introduction

Many image processing tasks, such as image deblurring and inpainting, can be formulated as inverse problems. The goal of these tasks is to recover a clean image $\mathbf{x}_0$ from a degraded measurement $\mathbf{y} \approx \mathcal{A}(\mathbf{x}_0)$ produced by a known forward operator $\mathcal{A}$, such as blurring or masking. A popular Bayesian approach is to sample from the posterior $p(\mathbf{x}_0 \mid \mathbf{y})$ using a data prior $p(\mathbf{x}_0)$ and a forward model $p(\mathbf{y} \mid \mathbf{x}_0)$. Pretrained diffusion priors have become the dominant approach for modeling $p(\mathbf{x}_0)$, where this Bayesian approach becomes akin to diffusion guidance methods (Chung et al., 2022; Daras et al., 2024).

Despite the successes of these methods, there remain unique challenges in developing accurate and efficient diffusion-based inverse problem solvers. One difficulty is computational: visual data are high-dimensional, motivating scalable priors and efficient guidance algorithms. Latent diffusion models (LDMs) provide one such scalable prior by performing diffusion in a compressed latent space rather than in pixel space (Rombach et al., 2022). Another difficulty is statistical: guidance should account for how much information the measurement provides at each timestep, and how uncertain the diffusion model is about the corresponding clean image features (Peng et al., 2024; Rissanen et al., 2024). For example, guidance at a given timestep should emphasize components that are both important to restore under the given forward operator and can be predicted reliably from the current noisy state. Capturing this timestep-, input-, and task-dependent uncertainty is therefore central to reliable guidance.

An ideal method should combine (1) a scalable diffusion prior, (2) efficient guidance, and (3) uncertainty-aware guidance. However, prior work addresses only subsets of these desiderata (Table 1), and no existing method achieves all three simultaneously. For instance, there is a mature line of work in developing uncertainty-aware guidance algorithms for pixel-space diffusion methods by modeling quantities such as $\mathrm{Cov}(\mathbf{y} \mid \mathbf{x}_t)$ at timestep t in the reverse diffusion process, exploiting linearity of the forward operator and approximations to the posterior covariance of $p(\mathbf{x}_0 \mid \mathbf{x}_t)$. However, these methods do not extend to LDM priors as the analogous uncertainty is $\mathrm{Cov}(\mathbf{y} \mid \mathbf{z}_t)$, and nonlinearity of the LDM decoder prevents any closed-form computation of this covariance. On the other hand, existing LDM inverse problem solvers fall into two categories: decode-and-guide methods and native-latent methods. Decode-and-guide methods map latents back to image space at each sampling

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

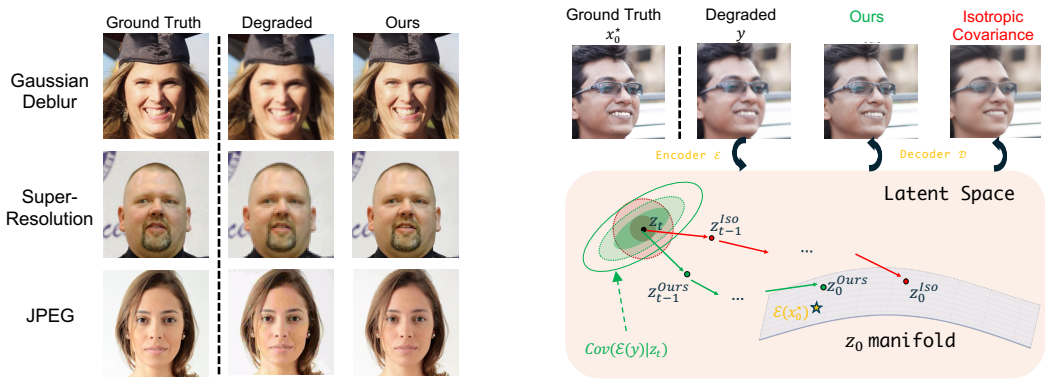


Figure 1: Left: Representative results of our method on the Gaussian deblurring, 8x super-resolution, and nonlinear JPEG decompression tasks. Right: ANaLOG models anisotropic covariance during guidance without expensive decoding steps. On a Gaussian deblurring task, ANaLOG preserves identity-relevant detail that isotropic approximations wash out.

step and apply guidance using the original forward model. This preserves the image-space forward operator, but sacrifices sampling efficiency through repeated decoder evaluations. Uncertainty-aware variants can further increase cost by requiring additional denoiser evaluations at each step (Rout et al., 2024). In contrast, native-latent methods improve sampling efficiency by applying guidance directly in latent space, for example by learning a latent forward operator H_θ that maps the denoiser-predicted clean latent $\mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t]$ to $\mathcal{E}(\mathbf{y})$ (Raphaeli et al., 2025). However, native-latent guidance was not derived under a probabilistic framework, leaving open the question of uncertainty estimation and its integration into guidance.

In this paper, we bridge this gap and introduce ANaLOG (**An**isotropic **Na**tive-**L**atent **O**perator **G**uidance), a framework that allows for efficient and uncertainty-aware guidance in the latent space. To develop our framework, we first reformulate existing work on native-latent guidance (Raphaeli et al., 2025) as approximate inference under a Gaussian latent measurement likelihood with a fixed covariance. We then extend to learn richer covariance structure without relying on covariance propagation through the LDM decoder. Specifically, our main contributions are:

Table 1: Comparison of posterior sampling methods across three desiderata.

	Efficient Diffusion	Efficient Guidance	Uncertainty-Aware
Pixel-space	✗	✓	✓
LDPS, PSLD	✓	✗	✗
STSL	✓	✗	✓
SILO	✓	✓	✗
ANaLOG (Ours)	✓	✓	✓

1. **Probabilistic Framework.** We develop ANaLOG, a probabilistic framework in which the latent measurement likelihood $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ is modeled as a Gaussian with learned mean and covariance. Viewed through this lens, existing work (Raphaeli et al., 2025) corresponds to the special case where the covariance is fixed to an isotropic, input-, and time-independent matrix. In contrast, ANaLOG learns an anisotropic, input-, and time-dependent covariance via a simple training procedure that augments the learning of the latent forward operator with a lightweight uncertainty head. At inference-time, we develop a guidance algorithm that steers using a Mahalanobis-weighted residual.

2. **Theoretical Results.** We theoretically isolate the role of covariance modeling in a tractable setting with Fourier-diagonal data and a linear LDM. We prove, via a lower bound, that even under perfect mean estimation, any isotropic covariance approximation incurs a strictly positive error relative to the true likelihood. In contrast, ANaLOG recovers the true likelihood at the standard parametric convergence rate.

3. **Empirical Evidence.** We empirically show that ANaLOG improves perceptual quality over existing LDM guidance algorithms across five inverse problems while remaining substantially faster than decode-and-guide methods. Furthermore, we show that visually, our learned uncertainty estimates closely match important input-, task-, and time-dependent structure.

2 Preliminaries and Related Work

Score-Based Generative Models. Diffusion models (Ho et al., 2020; Song et al., 2020) are powerful score-based generative models. They define a forward process that gradually corrupts data $\mathbf{x}_0 \sim p(\mathbf{x}_0)$ into noise, and train a neural network to reverse this process by learning the time-dependent score function $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$ (Efron, 2011). Sampling proceeds by solving the reverse-time SDE (Anderson, 1982):

$$\mathrm{d}\mathbf{x}_t = \left[f(t)\mathbf{x}_t - g(t)^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)\right] \mathrm{d}t + g(t)\, \mathrm{d}\bar{\mathbf{w}}, \tag{1}$$

where $f(t)$ and $g(t)$ are the drift and diffusion coefficients of the forward process and $\bar{\mathbf{w}}$ is a standard reverse-time Wiener process. Latent diffusion models (LDMs) (Rombach et al., 2022) apply this framework in the compressed latent space of a pretrained encoder $\mathcal{E} : \mathcal{X} \to \mathcal{Z}$, with a corresponding decoder $\mathcal{D} : \mathcal{Z} \to \mathcal{X}$. Here, we denote $\mathcal{X}$ as pixel space and $\mathcal{Z}$ as latent space (typically of dimension smaller than ambient image dimension). The forward process corrupts latents $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, and the score network is trained to approximate $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t)$.

Posterior Sampling. For inverse problems, we observe measurements $\mathbf{y} = \mathcal{A}(\mathbf{x}_0) + \sigma_y \boldsymbol{\epsilon}$, where $\mathcal{A}$ denotes the forward operator and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. We first describe the methodology for using pixel-space diffusion priors to solve inverse problems and then highlight key challenges in extending these approaches to LDMs. To solve inverse problems, we wish to sample from $p(\mathbf{x}_0 \mid \mathbf{y})$, which requires running the reverse SDE (1) with the *conditional* score. This decomposes via Bayes' rule as:

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{y}) = \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)}_{\text{unconditional score}} + \underbrace{\nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t)}_{\text{noisy likelihood score}}. \tag{2}$$

The unconditional score is provided by the pretrained diffusion prior. Conversely, the noisy likelihood factorizes as $\int p(\mathbf{y} \mid \mathbf{x}_0) p(\mathbf{x}_0 \mid \mathbf{x}_t)\, d\mathbf{x}_0$, which is intractable to analytically compute in general due to the unknown posterior denoising distribution (Daras et al., 2024). This has motivated the development of numerous approaches which focus on approximating $p(\mathbf{x}_0 \mid \mathbf{x}_t)$, e.g., as a Gaussian distribution with covariance approximated as a diagonal (Peng et al., 2024) or low-rank matrix (Meng et al., 2021) (see Appendix B for a more detailed history). Crucially, these approaches rely on linearity of the forward operator in order to compute a closed form of the integral above.

For LDMs, approximating the corresponding posterior covariance $p(\mathbf{z}_0 \mid \mathbf{z}_t)$ with a Gaussian is not sufficient to obtain a closed-form for the noisy likelihood since $p(\mathbf{y} \mid \mathbf{z}_0) = \mathcal{N}(\mathcal{A}(\mathcal{D}(\mathbf{z}_0)), \sigma_y^2 \mathbf{I})$, which depends on the nonlinear decoder $\mathcal{D}$. Thus, prior work has focused on two classes of solutions. Decode-and-guide methods such as LDPS (Rout et al., 2023) approximate $p(\mathbf{z}_0 \mid \mathbf{z}_t)$ as a Dirac delta around the posterior mean $\mathbf{z}_{0|t} \triangleq \mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t]$, so the noisy likelihood becomes

$$\log p(\mathbf{y} \mid \mathcal{D}(\mathbf{z}_{0|t})) \propto -\left\|\mathbf{y} - \mathcal{A}(\mathcal{D}(\mathbf{z}_{0|t}))\right\|_2^2. \tag{3}$$

Extensions iteratively optimize this residual (Song et al., 2023a) or add regularization to the above approximation, such as ensuring $\mathbf{z}_{0|t}$ is a fixed point of the autoencoder (Rout et al., 2023). Attempts to model the uncertainty in LDM guidance have focused on adding computationally expensive second-order correction terms to this guidance update (Rout et al., 2024). As is shown in Figure 2, decode-and-guide methods (red arrows) rely on decoder evaluations, eroding LDM efficiency gains.

To overcome this issue, native-latent methods such as SILO (Raphaeli et al., 2025) seek to replace the measurement operator $\mathcal{A}$, which maps the clean image $\mathbf{x}_0$ to the measurements $\mathbf{y}$, by a native-latent operator H that maps the latent image $\mathcal{E}(\mathbf{x}_0)$ to the latent measurements $\mathcal{E}(\mathbf{y})$. This is done by training a deep network H_θ to minimize a loss between $H_\theta(\mathbf{z}_{0|t})$ and latent measurements $\mathcal{E}(\mathbf{y})$. At inference time, the noisy likelihood score is approximated as in Equation 3, with $\mathcal{E}(\mathbf{y})$ replacing $\mathbf{y}$ and the learned H_θ replacing $\mathcal{A} \circ \mathcal{D}$. However, since the true noisy likelihood involves an expectation over the conditional distribution $p(\mathbf{z}_0 \mid \mathbf{z}_t)$, which is highly anisotropic and input-dependent, collapsing the noisy likelihood score into the gradient of a loss from a single deterministic mapping ignores its uncertainty structure. In contrast to pixel-space guidance algorithms, decode-and-guide methods, and native-latent algorithms, ANaLOG proposes an explicit probabilistic framework for approximating the noisy likelihood score and models the corresponding uncertainty.

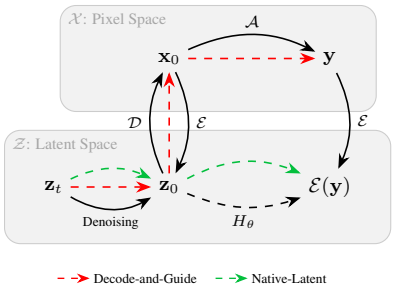


Figure 2: Guidance pathways at each sampling step.

3 ANISOTROPIC NATIVE-LATENT OPERATOR GUIDANCE

In this section, we present ANaLOG, our framework for efficient uncertainty-aware guidance for LDMs. Our framework is built on two precise approximations, described below, which we later prove generalize implicit assumptions made in prior work (Raphaeli et al., 2025).

Approximation 1 (Sufficiency of Encoded Measurements)**.** *We replace conditioning on the raw measurement* $\mathbf{y} \in \mathbb{R}^n$ *with conditioning on its encoding* $\mathcal{E}(\mathbf{y}) \in \mathbb{R}^k$*:*

$$p(\mathbf{z}_t \mid \mathbf{y}) \;\approx\; p(\mathbf{z}_t \mid \mathcal{E}(\mathbf{y})). \tag{4}$$

This replacement is exact when $\mathcal{E}(\mathbf{y})$ is a sufficient statistic for $\mathbf{y}$ with respect to the latent posterior. In Section 4, we prove in certain settings that common image processing tasks like deblurring and super-resolution satisfy Approximation 1. Our next approximation is of the latent forward model.

Approximation 2 (Gaussian Latent Forward Model)**.** *We approximate the intractable latent forward model with a learned Gaussian:*

$$p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \;\approx\; p_\theta(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \triangleq \mathcal{N}\big(\mathcal{E}(\mathbf{y});\; \mu_\theta(\mathbf{z}_t, t, \sigma_y),\; \Sigma_\theta(\mathbf{z}_t, t, \sigma_y)\big), \tag{5}$$

where μ_θ *and* Σ_θ *are learned mean and covariance functions*[1]*.*

While the SILO guidance rule was derived without an explicit probabilistic formulation (see Appendix A for details), we take an alternate route and directly model the latent likelihood $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$. We further show in Appendix A how these two approximations can be viewed as relaxations of two implicit assumptions made in SILO (Raphaeli et al., 2025), enabling a broader framework. Crucially, SILO corresponds to the special time and input-independent case $\Sigma_\theta \equiv \sigma^2\mathbf{I}$. This is a strong structural assumption. The latent likelihood $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ depends on a nonlinear encoder and a complex denoising process; different latent coordinates encode different spatial structures, so their uncertainties given $\mathbf{z}_t$ should vary across both coordinates and timesteps. Indeed, we show in Section 4 that even in an idealized linear Gaussian data model, the true covariance of $\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t$ is inherently anisotropic (Proposition 7, Part 2). Our framework generalizes SILO to allow for input-dependent, time-dependent, and anisotropic covariance.

Training and Inference Algorithms. Due to our explicit approximations, we can develop a simple, principled training objective and an uncertainty-aware guidance algorithm. Given the Gaussian model (5), we can use maximum likelihood estimation to learn the mean and covariance by minimizing the negative log-likelihood (NLL) (Bishop & Nasrabadi, 2006; Seitzer et al., 2022). The per-sample NLL is

$$\ell_\theta(\mathcal{E}(\mathbf{y}), \mathbf{z}_t, t, \sigma_y) \;=\; \frac{1}{2}\log|\Sigma_\theta| \;+\; \frac{1}{2}(\mathcal{E}(\mathbf{y}) - \mu_\theta)^\top \Sigma_\theta^{-1}(\mathcal{E}(\mathbf{y}) - \mu_\theta), \tag{6}$$

and the training objective is $\mathcal{L}_{\mathrm{NLL}}(\theta) = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},t,\mathbf{z}_t}[\ell_\theta(\mathcal{E}(\mathbf{y}), \mathbf{z}_t, t, \sigma_y)]$. At the optimum, the predicted covariance balances residual magnitude against uncertainty calibration, allowing the model to learn heterogeneous uncertainty estimates across coordinates, timesteps, and noise levels. In practice, we optimize $\mathcal{L}_{\mathrm{NLL}}$ in two stages: first learning a good mean prediction with a frozen isotropic covariance, and then learning covariance structure on top of a fixed mean prediction (details in Appendix E).

At sampling time, the guidance signal is the score of the learned likelihood (5):

$$\nabla_{\mathbf{z}_t} \log p_\theta(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t, t, \sigma_y) \;=\; -\frac{1}{2}\nabla_{\mathbf{z}_t}\log|\Sigma_\theta| \;-\; \frac{1}{2}\nabla_{\mathbf{z}_t}\big[(\mathcal{E}(\mathbf{y}) - \mu_\theta)^\top \Sigma_\theta^{-1}(\mathcal{E}(\mathbf{y}) - \mu_\theta)\big]. \tag{7}$$

This is a *Mahalanobis-weighted* consistency signal: Σ_θ^{-1} acts as a precision matrix that steers each latent coordinate proportionally to the network's confidence in its prediction.

Practical Implementation. While our framework allows for expressive covariance parameterizations, storing and inverting large covariance matrices are often infeasible in practice (Rissanen et al., 2024). In addition, differentiating through the covariance head adds memory and computational cost and can destabilize guidance, similar to observations in prior work (Rissanen et al., 2024). Thus, we parameterize the covariance as a diagonal matrix (in a basis such as the Discrete Cosine Transform basis (Ahmed et al., 1974)) and use a stop-gradient approximation for Σ_θ, so gradients only flow through the mean approximation. This yields efficient training and inference, summarized in Algorithms 1 and 2 in Appendix E.3. We examine relaxing these approximations in Appendix F.2 and F.4.

[1]Henceforth, we will drop the function arguments $\mathbf{z}_t, t, \sigma_y$ for notational convenience

4 THEORETICAL RESULTS

In this section, we consider a setting where the data distribution follows a power law in the frequency domain, mimicking the spectral properties of natural image data (Schaaf & Hateren, 1996). We then demonstrate that in a linear model setting, deconvolution inverse problems satisfy Approximations 1 and 2 (Proposition 7). Further, we characterize the *anisotropic* and *time-dependent* structure of the latent likelihood covariance. Our main result is to show a provable separation between existing isotropic likelihood approximations (Raphaeli et al., 2025), and `ANaLOG`, which converges to the true likelihood at a standard parametric rate (Theorem 8). In Section 5, we show that these theoretical ideas extend to general settings, and modeling covariance structure is helpful for large-scale nonlinear LDMs beyond deconvolution problems. All proofs are given in Appendix C.

4.1 PROBLEM SETUP

Notation. Let $\mathbf{F}_d$ denote the d-dimensional Discrete Fourier Transform matrix. We will drop the subscript d when the dimension is clear from context. Assuming the DC frequency is at the center of the DFT spectrum, we index the Fourier-domain by centered frequencies over the index set

$$\Omega_d = \left\{-\tfrac{d-1}{2}, \ldots, -1, 0, 1, \ldots, \tfrac{d-1}{2}\right\}, \tag{8}$$

where d is assumed to be odd for notational convenience. We use $\circledast_{\downarrow M}$ to denote circular convolution with stride M and $\circledast_{\uparrow M}$ as transposed convolution with stride M. We let $\mathcal{E}_\sharp p$ denote the pushforward measure of p under the map $\mathcal{E}$. We consider the variance-preserving noising process for the latent diffusion model with T steps such that

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k). \tag{9}$$

Throughout this section, $\mathbf{z}_{0|t}$ denotes the exact posterior mean $\mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t]$ under the idealized Gaussian model analyzed here. In practice, LDMs are trained to predict this conditional mean (e.g., via Tweedie's formula (Efron, 2011)), and we do not study here the training and optimization error of learning. Next, we state our assumptions.

Assumption 3. *(Fourier-Diagonal Data Distribution) There exist constants $\alpha > 1$, $C > 0$ such that the clean image satisfies $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_f)$ where*

$$\boldsymbol{\Sigma}_f = \mathbf{F}_d^* \operatorname{diag}_{\omega \in \Omega_d}(\lambda_\omega)\, \mathbf{F}_d, \qquad \lambda_\omega = C(1+|\omega|)^{-\alpha}. \tag{10}$$

Assumption 3 is a standard assumption used in prior work (Thaker et al., 2025) and mimics the spectral properties of natural image data, which are observed to be approximately stationary in the frequency domain with a power law spectrum (Schaaf & Hateren, 1996). Next, inspired by the strided convolutional structure of pretrained LDM autoencoders (Rombach et al., 2022), we assume a similar convolutional structure for a linear LDM autoencoder.

Assumption 4. *(M-Strided Convolutional Autoencoder) Assume $d \bmod M = 0$ and let $k \triangleq d/M$ be the latent dimension of the autoencoder. Suppose the encoder $\mathcal{E} : \mathbb{R}^d \to \mathbb{R}^k$ and decoder $\mathcal{D} : \mathbb{R}^k \to \mathbb{R}^d$ are linear maps of the form:*

$$\mathcal{E}(\mathbf{x}) = \mathbf{W}(\mathbf{g} \circledast_{\downarrow M} \mathbf{x}) \tag{11}$$

$$\mathcal{D}(\mathbf{z}) = \mathbf{h} \circledast_{\uparrow M} (\mathbf{V}\mathbf{z}) \tag{12}$$

with $\mathbf{g}, \mathbf{h} \in \mathbb{R}^d$ and $\mathbf{W}, \mathbf{V} \in \mathbb{R}^{k \times k}$.

We study learning this autoencoder with a regularized autoencoder objective (Tolstikhin et al., 2017), similar to the objective studied in Rout et al. (2023):

$$\min_{\mathcal{E},\mathcal{D}} \mathcal{L} = \mathbb{E}\Big[\|\mathcal{D}(\mathcal{E}(\mathbf{x}_0)) - \mathbf{x}_0\|_2^2\Big] + \gamma\, \mathrm{KL}\big(\mathcal{E}_\sharp p \parallel \mathcal{N}(0, \mathbf{I}_k)\big). \tag{13}$$

Finally, we study deconvolution problems, which covers common image restoration tasks such as deblurring.

Assumption 5. *(Deconvolution Inverse Problem) Assume a measurement model $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \sigma_y \boldsymbol{\epsilon}$, where $\sigma_y > 0$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{A}\mathbf{x} = \boldsymbol{\psi} \circledast \mathbf{x}$ is a circular convolution. Let $a_\omega = diag(\mathbf{F}\mathbf{A}\mathbf{F}^*)_\omega$ be the frequency response of $\mathbf{A}$ at frequency ω.*

4.2 Verifying Approximation 1 and 2

We first show that Approximations 1 and 2 are exact under our model assumptions, i.e., there is no approximation error. Before studying the approximations, we characterize the structure of the latent space by characterizing the optimal encoder and decoder under the objective given in Equation (13).

Proposition 6. *The parameters of an M-strided convolutional autoencoder that achieve the global minimum of the loss in* (13) *are given by:*

$$(\mathbf{F}_d \mathbf{g})_\omega = (\mathbf{F}_d \mathbf{h})_\omega = \begin{cases} \sqrt{M} & \text{for } \omega \in \Omega_k, \\ 0 & \text{for } \omega \notin \Omega_k, \end{cases} \qquad \begin{aligned} \mathbf{W} &= \mathbf{F}_k \, \mathrm{diag}_{\omega \in \Omega_k}(\lambda_\omega^{-1/2}) \, \mathbf{F}_k^*, \\ \mathbf{V} &= \mathbf{F}_k \, \mathrm{diag}_{\omega \in \Omega_k}(\lambda_\omega^{1/2}) \, \mathbf{F}_k^*. \end{aligned} \tag{14}$$

This proposition states the optimal parameters of the convolutional encoder correspond to an ideal low-pass filter that retains the k lowest-magnitude centered frequencies Ω_k and discards all frequencies in $\Omega_d \setminus \Omega_k$. For the remainder of this section, we instantiate $(\mathcal{E}, \mathcal{D})$ with the optimal parameters characterized in Proposition 6. We now present the main result of this subsection.

Proposition 7 (Approximations 1 and 2 are exact)**.** *Under Assumptions 3–5 with the optimal autoencoder (Proposition 6), the following hold:*

1. ***(Sufficiency.)*** *$\mathcal{E}(\mathbf{y})$ is sufficient for predicting $\mathbf{z}_t$:*
$$p(\mathbf{z}_t \mid \mathbf{y}) = p(\mathbf{z}_t \mid \mathcal{E}(\mathbf{y})). \tag{15}$$
2. ***(Gaussian likelihood.)*** *The true $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_t$ and covariance $\mathbf{C}_t$ given by:*
$$\boldsymbol{\mu}_t = \mathbf{F}_k^* \, \mathrm{diag}_{\omega \in \Omega_k}(a_\omega) \, \mathbf{F}_k \mathbf{z}_{0|t}, \qquad \mathbf{C}_t = \mathbf{F}_k^* \, \mathrm{diag}_{\omega \in \Omega_k}\big(c_\omega(t)\big) \, \mathbf{F}_k, \tag{16}$$
where $\{a_\omega\}_{\omega \in \Omega_k}$ denotes the frequency response of $\mathbf{A}$ for the retained frequencies and
$$c_\omega(t) = (1 - \bar{\alpha}_t)|a_\omega|^2 + \frac{\sigma_y^2}{\lambda_\omega}, \qquad \omega \in \Omega_k. \tag{17}$$

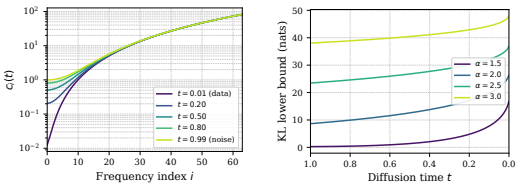
Part 1 proves that encoding the measurement preserves exactly the information needed for the latent posterior, motivating native-latent guidance. Part 2 shows the Gaussian structure of the latent likelihood induced by the data frequency structure. The proof of Part 2 draws a connection between native-latent guidance and multirate signal processing, specifically utilizing the Noble identities (Vaidyanathan, 2006) to obtain a closed form for the likelihood. The covariance eigenvalues $c_\omega(t)$ consist of two terms: $(1 - \bar{\alpha}_t)|a_\omega|^2$, which captures diffusion uncertainty propagated through the forward operator, and $\sigma_y^2/\lambda_\omega$, which captures measurement noise amplified by the encoder's whitening. Crucially, both terms are frequency-dependent, making the covariance inherently anisotropic.

4.3 Demonstrating Provable Need for Covariance Modeling

Our main result for our setting is that anisotropic, uncertainty-aware guidance is necessary for correct sampling, whereas isotropic guidance incurs a strictly positive error relative to the true likelihood.

Theorem 8 (Provable Separation of Isotropic vs. Anisotropic Covariance)**.** *Under Assumptions 3–5 with the optimal autoencoder (Proposition 6), let $\mathbf{C}_t$ be the true likelihood covariance from* (17) *with eigenvalues $c_\omega(t)$. Then, the following statements hold:*

1. ***(SILO: irreducible error.)*** *For any scalar $\eta > 0$, define the isotropic approximation $\tilde{p}(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \coloneqq \mathcal{N}(\boldsymbol{\mu}_t, \, \eta^2 \, \mathbf{I}_k)$. Then*
$$KL\big(p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \parallel \tilde{p}(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)\big) \geq \frac{k}{2} \log \frac{\mathrm{AM}\big(\{c_\omega(t)\}_{\omega \in \Omega_k}\big)}{\mathrm{GM}\big(\{c_\omega(t)\}_{\omega \in \Omega_k}\big)}, \tag{18}$$
where AM *and* GM *are the arithmetic and geometric means respectively. This lower bound is attained at $\eta^2 = \mathrm{AM}\big(\{c_\omega(t)\}_{\omega \in \Omega_k}\big)$ and under Assumptions 3–5, it is strictly positive with probability 1 for $t \sim \mathcal{U}[0, T]$.*
2. ***(ANaLOG: convergence to true likelihood.)*** *Let $\hat{p}$ denote the MLE of the Gaussian model* (5) *with linear mean parameterization and Fourier-diagonal covariance, trained on n i.i.d. samples at a fixed timestep t. Then*
$$\mathbb{E}\big[KL\big(p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \parallel \hat{p}(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)\big)\big] \; = \; O\left(\frac{k}{n}\right). \tag{19}$$



(a) Covariance eigenvalues $c_\omega(t)$ at different diffusion timesteps (α=2.5).

(b) KL lower bound (18) vs. diffusion time t for different spectral decays α.

Figure 3: Visualizing our theorem bounds.

Implications. Theorem 8 gives a clean separation within this model: an isotropic covariance family has an irreducible KL gap, while an anisotropic family recovers the true likelihood at the standard parametric rate. Figure 3 illustrates the qualitative behavior of the covariance eigenvalues and the KL lower bound. First, Figure 3a shows that the covariance spectrum is highly anisotropic, especially near the end of diffusion sampling. Second, Figure 3b shows the KL penalty grows with the spectral decay rate α. Crucially, the isotropic approximation error is largest at the end of the reverse process, where guidance precision matters most for preserving high-frequency details (Rissanen et al., 2022). This may provide theoretical evidence as to why SILO produces smoothed-out reconstructions (Figure 1). More analysis of this lower bound is given in Appendix D.1.

5 Experiments

5.1 Experimental Setup

Tasks, Datasets, and Models. We consider five inverse problems, grouped into three deconvolution tasks (Gaussian blur, 4x and 8x super-resolution) and two general inverse problems (box inpainting, JPEG decompression). The measurement noise for each task is $\sigma_y = 0.02$. Further task details are provided in Appendix E. We evaluate on a fixed 1000 image test set from the FFHQ dataset (Karras et al., 2019) and the ImageNet-1000 dataset (Deng et al., 2009), both at 512×512 resolution. We utilize the pretrained LDM Realistic Vision-v5.1 (Stable Diffusion API, 2023) from HuggingFace. For all methods, we utilize 100 sampling steps for the FFHQ dataset and 250 sampling steps for the more complex ImageNet dataset.

Baselines and Metrics. We compare to four leading LDM inverse problem solvers. The first class is decode-and-guide methods: LDPS (Rout et al., 2023), PSLD (Rout et al., 2023), and STSL (Rout et al., 2024). The other class is native-latent guidance, such as SILO (Raphaeli et al., 2025). These baselines serve different roles. SILO is the closest comparison because it is also a trained native-latent method. LDPS and PSLD are representative decode-and-guide posterior samplers that do not use the same native-latent operator, and STSL represents a more expensive uncertainty-weighted decode-and-guide approach. We report FID and LPIPS for perceptual quality (Zhang et al., 2018), and PSNR for distortion, each with subscripts denoting the standard error of the mean over the 1000 test examples. We also report sampling time in seconds per image.

Details of `ANaLOG`. We train a lightweight variance head alongside the mean prediction network, adding only $\approx 3.4\%$ additional parameters (480K out of 14.2M). The variance head takes inspiration from the Readout-Guidance architecture (Luo et al., 2024) and takes in as input aggregated features from the LDM UNet denoiser and a σ_y embedding. It then outputs a log-variance map in latent space, parametrized as diagonal both in spatial and DCT domain, with the latter inspired by our theoretical results. We test richer parameterizations of the covariance in Appendix F.2. We denote `ANaLOG` and `ANaLOG`-DCT as our method with the spatial and DCT covariances respectively. The exact architecture and training details are in Appendix E.

5.2 Main Results

Tables 2 and 3 show that, for both datasets, `ANaLOG` consistently improves perceptual metrics over the native-latent baseline, SILO, on most tasks while maintaining a nearly identical inference time. This highlights our main conceptual message: modeling covariance structure is critical for native-latent guidance. Modeling the covariance in the DCT domain rather than in spatial domain yields further improvements in most tasks, especially for JPEG decompression, where the forward operator also operates in the DCT domain (Wallace, 1991). Compared with decode-and-guide baselines, `ANaLOG` generally provides a better quality-efficiency tradeoff, matching or exceeding these baselines on the FFHQ dataset. Notably, `ANaLOG` is $18\times$ faster than STSL. On ImageNet, `ANaLOG` also yields substantial gains over SILO (e.g., $62.5 \rightarrow 26.9$ FID on Gaussian deblurring, $132.1 \rightarrow 71.1$ on $4\times$ super-resolution), confirming that covariance modeling transfers beyond face datasets. A gap to decode-and-guide baselines remains on super-resolution; since both native-latent methods show

degraded PSNR on ImageNet relative to FFHQ, we attribute this to difficulty of mean prediction rather than covariance. Improving μ_θ (e.g., by incorporating class labels as conditional inputs) is orthogonal to our contribution. Finally, results on inpainting and JPEG decompression (Table 3) suggest that, even when the exact sufficiency argument from Section 4 no longer applies, native-latent guidance remains empirically effective and continues to benefit from covariance-aware weighting.

Figure 4 illustrates the qualitative differences among methods. The isotropic covariance in SILO leads to over-smoothed reconstructions that lack high-frequency detail. STSL recovers finer detail but can also introduce some visual artifacts. In contrast, ANaLOG produces visually plausible and consistent reconstructions. Additional qualitative results are provided in Appendix F.1.

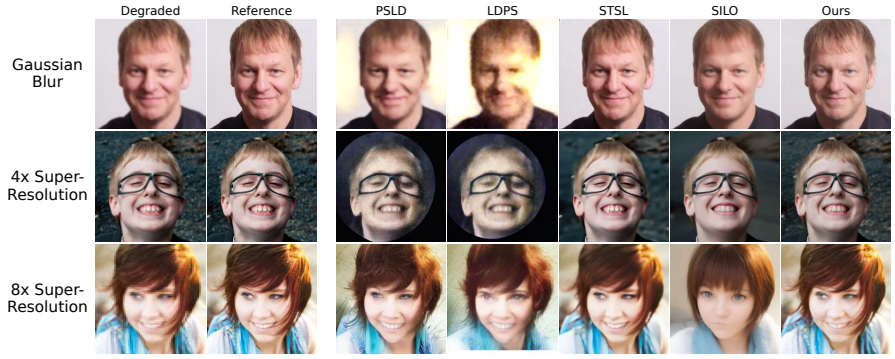


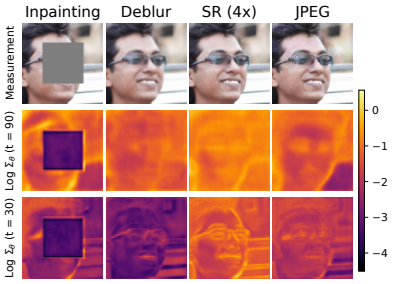
Figure 4: Qualitative comparison on FFHQ deconvolution tasks. Best viewed zoomed in.

Figure 5a visualizes the learned log-variance spatial maps for a single input across four tasks at two different timesteps. It highlights that the learned covariances are anisotropic, input- and time-dependent. Across different tasks, the variance captures operator-specific structure; e.g., in super-resolution it concentrates along edges and fine textures, where high-frequency information is lost. At early timesteps, the variance is relatively uniform, whereas at later timesteps, spatial structure emerges and becomes more pronounced. The learned covariance is also well-calibrated on held-out data across all tasks as shown in Appendix F.3.

		Method	Time (s)	Gaussian Blur FID↓	Gaussian Blur LPIPS↓	Gaussian Blur PSNR↑	4x Super-Resolution FID↓	4x Super-Resolution LPIPS↓	4x Super-Resolution PSNR↑	8x Super-Resolution FID↓	8x Super-Resolution LPIPS↓	8x Super-Resolution PSNR↑
FFHQ	*Decode*	LDPS	52.2	84.8	$0.564_{(.004)}$	$21.48_{(.16)}$	68.6	$0.554_{(.003)}$	$20.70_{(.11)}$	55.5	$0.536_{(.003)}$	$21.24_{(.08)}$
		PSLD	62.2	70.8	$0.535_{(.003)}$	$22.65_{(.14)}$	53.5	$0.505_{(.003)}$	$22.52_{(.10)}$	54.0	$0.513_{(.003)}$	$21.68_{(.08)}$
		STSL	433	41.7	$0.419_{(.002)}$	$\mathbf{27.36}_{(.07)}$	39.7	$0.415_{(.002)}$	$26.88_{(.10)}$	38.6	$0.439_{(.002)}$	$25.58_{(.08)}$
	Native	SILO	**24.0**	47.3	$0.367_{(.002)}$	$25.03_{(.06)}$	38.8	$0.331_{(.002)}$	$26.26_{(.06)}$	35.0	$0.356_{(.002)}$	$25.81_{(.06)}$
		ANaLOG	**23.7**	**29.8**	$\mathbf{0.326}_{(.002)}$	$\mathbf{27.34}_{(.09)}$	**27.0**	$\mathbf{0.303}_{(.002)}$	$\mathbf{27.56}_{(.08)}$	38.8	$0.357_{(.002)}$	$25.39_{(.05)}$
		ANaLOG-DCT	**24.1**	30.9	$0.338_{(.002)}$	$26.36_{(.10)}$	**26.4**	$\mathbf{0.302}_{(.002)}$	$26.92_{(.09)}$	**31.1**	$\mathbf{0.341}_{(.002)}$	$\mathbf{26.37}_{(.07)}$
ImageNet	*Decode*	LDPS	137.9	55.9	$0.463_{(.003)}$	$24.99_{(.13)}$	111.7	$0.548_{(.003)}$	$23.26_{(.12)}$	149.3	$0.576_{(.003)}$	$21.58_{(.12)}$
		PSLD	164.8	53.4	$0.455_{(.003)}$	$25.05_{(.13)}$	95.8	$0.546_{(.003)}$	$23.28_{(.12)}$	146.9	$0.577_{(.003)}$	$21.38_{(.11)}$
		STSL	1155	39.8	$0.404_{(.003)}$	$\mathbf{25.95}_{(.13)}$	**38.2**	$\mathbf{0.422}_{(.003)}$	$\mathbf{25.95}_{(.13)}$	**69.6**	$\mathbf{0.491}_{(.003)}$	$\mathbf{23.41}_{(.12)}$
	Native	SILO	**58.6**	62.5	$0.450_{(.003)}$	$22.94_{(.11)}$	132.1	$0.541_{(.004)}$	$20.73_{(.10)}$	129.1	$0.544_{(.004)}$	$19.41_{(.09)}$
		ANaLOG	**57.8**	28.6	$\mathbf{0.370}_{(.003)}$	$24.77_{(.14)}$	85.8	$0.477_{(.003)}$	$21.71_{(.09)}$	119.2	$0.524_{(.003)}$	$19.68_{(.08)}$
		ANaLOG-DCT	**59.1**	**26.9**	$0.403_{(.002)}$	$23.74_{(.10)}$	71.1	$0.440_{(.002)}$	$22.44_{(.08)}$	105.2	$0.509_{(.003)}$	$19.83_{(.09)}$

Table 2: Quantitative results on deconvolution tasks on the FFHQ and ImageNet datasets. Bolded entries indicate best performance, underlined entries indicate second best performance, and subscripts denote standard error of mean. We bold multiple rows in the event of approximately tied results. Gaussian deblurring is the representative task for timings since timings do not vary significantly across the different forward operators.

Ablation Studies. We now ablate the covariance structure on a Gaussian deblurring task on the FFHQ dataset to support the need for anisotropic, input- and time-dependent covariance. We compare four variants: (i) *isotropic covariance*, which corresponds to SILO and uses a scalar variance; (ii) *time-independent covariance*, which computes the variance only at the first timestep and reuses it



(a) Learned spatial variance maps. Variance is averaged across the 4 latent channels.

Variant	FID↓	LPIPS↓	PSNR↑
Isotropic (SILO)	47.3	$0.367_{(.002)}$	$25.03_{(.06)}$
Time-independent	<u>41.09</u>	$0.378_{(.002)}$	$25.48_{(.09)}$
Theory-predicted	48.87	<u>0.353</u>$_{(.001)}$	<u>25.86</u>$_{(.05)}$
ANaLOG	**29.79**	**0.326**$_{(.002)}$	**27.34**$_{(.09)}$

(b) Ablation of covariance structure on Gaussian deblurring (FFHQ).

Figure 5: Learned variance maps and ablation of covariance structure. **(a)** Task-dependent variance maps show the model learns distinct spatial patterns for each inverse problem. **(b)** Progressively richer covariance models improve reconstruction quality.

		Method	Time (s)	Box Inpainting FID↓	Box Inpainting LPIPS↓	Box Inpainting PSNR↑	JPEG FID↓	JPEG LPIPS↓	JPEG PSNR↑
FFHQ	*Decode*	LDPS	<u>52.2</u>	68.2	$0.205_{(.001)}$	$16.61_{(.12)}$	101.0	$0.612_{(.003)}$	$17.83_{(.14)}$
		PSLD	62.2	62.7	$0.196_{(.001)}$	$17.09_{(.12)}$	N/A	N/A	N/A
		STSL	433	63.1	$0.191_{(.001)}$	$18.19_{(.13)}$	63.5	$0.502_{(.003)}$	$22.09_{(.17)}$
	Native	SILO	**24.0**	<u>32.8</u>	<u>0.142</u>$_{(.001)}$	**22.59**$_{(.08)}$	68.3	$0.471_{(.002)}$	$20.87_{(.06)}$
		ANaLOG	**23.7**	**22.8**	**0.135**$_{(.001)}$	$21.31_{(.11)}$	<u>46.3</u>	<u>0.393</u>$_{(.002)}$	<u>23.91</u>$_{(.06)}$
		ANaLOG-DCT	**24.1**	**22.1**	**0.132**$_{(.001)}$	<u>22.08</u>$_{(.10)}$	**27.3**	**0.326**$_{(.002)}$	**26.26**$_{(.09)}$

Table 3: Quantitative results on center box inpainting and JPEG decompression on the FFHQ dataset. Note PSLD does not apply to nonlinear inverse problems.

throughout sampling; (iii) *theory-predicted covariance*, which uses the diagonal covariance predicted by Proposition 7, i.e., the optimal covariance under our assumptions; and (iv) our model with time- and input-dependent covariance. Figure 5b shows that the theory-predicted covariance already provides improvements in perceptual quality, suggesting that the spectral covariance structure is beneficial. As we further incorporate input and time dependence, both perceptual and distortion metrics improve, highlighting the importance of modeling anisotropic, input- and time-dependent covariance. We prove additional ablation studies on varying guidance scale as well as considering richer covariance structure in Appendix F.4.

6 Conclusion

We introduced ANaLOG, a framework for efficient uncertainty-aware guidance with pretrained latent diffusion models. By interpreting existing native-latent guidance as approximate inference under a Gaussian latent likelihood, we identify covariance modeling as the *missing degree of freedom* in native-latent guidance methods. In a Fourier-diagonal data model, we proved that the true latent likelihood is inherently anisotropic, and any isotropic approximation incurs an irreducible KL gap, whereas ANaLOG recovers the true likelihood at the standard parametric rate. Empirically, modeling the mean and covariance together yields consistent improvements in perceptual reconstruction quality, even with a lightweight diagonal covariance head.

Limitations and Future Work. ANaLOG inherits some limitations of native-latent guidance, as discussed in detail in Appendix G. Nonetheless, we hope the probabilistic formulation of native-latent guidance can inspire work on uncertainty-aware native-latent guidance, possibly considering richer covariance parameterizations.

AI use statement

In this work, we used generative AI tools for tasks: design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, formulate

mathematical claims, assist in the writing of proofs, and interpret results. We have not used generative AI tools for tasks: propose or refine hypotheses and help develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims . Tasks such as: generate synthetic data sets, assist with translation, clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools for creating/modifying scientific figures, writing software code, brainstorming, and identifying relevant literature. We have reviewed all AI-assisted work. We checked LLM-generated code, verified all references and proofs by hand. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

Our work improves the quality of posterior sampling with pretrained latent diffusion models for image inverse problems. However, any improvement to diffusion-based image generation carries dual-use risk: more accurate posterior samplers could be repurposed to generate realistic synthetic imagery for disinformation or identity fraud. We emphasize that our method operates on top of existing pretrained LDMs and does not introduce new generative capabilities beyond what these models already provide. Nonetheless, we encourage the continued development of deepfake detection methods, provenance tracking, and watermarking techniques as necessary complements to generative modeling research.

Reproducibility statement

We have provided all experimental details in Appendix E, and will release the code upon acceptance.

References

Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. Discrete cosine transform. *IEEE transactions on Computers*, 100(1):90–93, 1974.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12 (3):313–326, 1982.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *arXiv preprint arXiv:2310.06721*, 2023.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Giannis Daras, Hyungjin Chung, Chieh-Hsin Lai, Yuki Mitsufuji, Jong Chul Ye, Peyman Milanfar, Alexandros G Dimakis, and Mauricio Delbracio. A survey on diffusion models for inverse problems. *arXiv preprint arXiv:2410.00083*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3): 211–218, 1936.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496): 1602–1614, 2011.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Tero Karras, S. Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

Erich Leo Lehmann and George Casella. *Theory of point estimation*. Springer, 1998.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8217–8227, 2024.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34:25359–25369, 2021.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994.

Xinyu Peng, Ziyang Zheng, Wenrui Dai, Nuoqian Xiao, Chenglin Li, Junni Zou, and Hongkai Xiong. Improving diffusion models for inverse problems using optimal posterior covariance. *arXiv preprint arXiv:2402.02149*, 2024.

Ron Raphaeli, Sean Man, and Michael Elad. Silo: Solving inverse problems with latent operators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10570–10580, 2025.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*, 2022.

Severi Rissanen, Markus Heinonen, and Arno Solin. Free hunch: Denoiser covariance estimation for diffusion models without extra costs. *arXiv preprint arXiv:2410.11149*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alex Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:49960–49990, 2023.

Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9472–9481, 2024.

A. van der Schaaf and J. H. van Hateren. Modelling the Power Spectra of Natural Images: Statistics and Information. *Vision Research*, 36(17):2759–2770, 1996. ISSN 0042-6989. doi: https://doi.org/10.1016/0042-6989(96)00002-8. URL https://www.sciencedirect.com/science/article/pii/0042698996000028.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022.

Ivan Skorokhodov, Sharath Girish, Benran Hu, Willi Menapace, Yanyu Li, Rameen Abdal, Sergey Tulyakov, and Aliaksandr Siarohin. Improving the diffusability of autoencoders. *arXiv preprint arXiv:2502.14831*, 2025.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. *arXiv preprint arXiv:2307.08123*, 2023a.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023b.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Stable Diffusion API. Realistic vision v5.1. https://huggingface.co/stablediffusionapi/realistic-vision-v51, 2023. Hugging Face model card.

Darshan Thaker, Abhishek Goyal, and René Vidal. Frequency-guided posterior sampling for diffusion-based image restoration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12873–12882, 2025.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

Parishwad P Vaidyanathan. *Multirate systems and filter banks*. Pearson Education India, 2006.

Gregory K Wallace. The jpeg still picture compression standard. *Communications of the ACM*, 34(4):30–44, 1991.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Implicit Assumptions Behind SILO **13**

A.1 ANaLOG as a Relaxation of SILO 14

B Extended Related Work **14**

C Proofs from Section 4 **15**

C.1 Proof of Proposition 6 (Optimal Autoencoder) 15

C.2 Proof of Proposition 7 16

C.3 Proof of Theorem 8 (Isotropic vs. Anisotropic Covariance Approximation Error) . 18

D Discussion of Theoretical Analysis **19**

D.1 Regimes Where Isotropic Approximation Error Is Large 19

D.2 Additional Visualizations 20

E Experimental Details **21**

E.1 Training Details 21

E.2 Inference Details and Baselines 23

E.3 Full Training and Inference Algorithms 23

F Extended Experiments **25**

F.1 Extended Qualitative Results 25

F.2 Extending to Richer Covariance Parameterizations 31

F.3 Probing Calibration of Uncertainty Head 31

F.4 Further Ablation Studies 32

G Limitations and Broader Impact **33**

A Implicit Assumptions Behind SILO

The central object needed for posterior sampling with latent diffusion models is the noisy likelihood score $\nabla_{\mathbf{z}_t} \log p(\mathbf{y} \mid \mathbf{z}_t)$. The main challenge is that $\mathbf{y}$ is obtained by a forward model operating on images, whereas $\mathbf{z}_t$ is in the latent space of the LDM encoder. Decode-and-guide methods (Rout et al., 2023; 2024) approximate this noisy likelihood score by mapping latents back to pixel space at each step, which inevitably sacrifices the efficiency of operating in the latent domain. SILO (Raphaeli et al., 2025) takes a different approach: it encodes the measurement via $\mathcal{E}(\mathbf{y})$ and directly learns a latent forward operator H_θ. This latent operator is trained to map denoised latent estimates $\mathbf{z}_{0|t}$ directly to $\mathcal{E}(\mathbf{y})$ by minimizing a residual between $H_\theta(\mathbf{z}_{0|t})$ and $\mathcal{E}(\mathbf{y})$. At inference time, the noisy likelihood score is approximated as the negative gradient of an ℓ_2 loss:

$$-\nabla_{\mathbf{z}_t} \left\| \mathcal{E}(\mathbf{y}) - H_\theta(\mathbf{z}_{0|t}) \right\|_2^2, \tag{A.1}$$

While this yields fast and effective guidance, the distributional assumptions that justify the training objective and justify (A.1) as the correct guidance signal are left implicit. For completeness, we repeat the derivation of SILO native-latent guidance below. Assuming we have a trained network H_θ that approximates the forward operator $\mathcal{A}$ in latent space (as shown in Figure 2), the starting point of the derivation is the pixel-space residual

$$\|\mathbf{y} - \mathcal{A}(\mathbf{x}_0)\|_2^2 \approx \|\mathcal{D}(\mathcal{E}(\mathbf{y})) - \mathcal{D}(\mathcal{E}(\mathcal{A}(\mathbf{x}_0)))\|_2^2 \tag{A.2}$$

$$\approx \|\mathcal{D}(\mathcal{E}(\mathbf{y})) - \mathcal{D}(H_\theta(\mathbf{z}_0))\|_2^2 \tag{A.3}$$

$$\leq C_{\mathcal{D}}^2 \|\mathcal{E}(\mathbf{y}) - H_\theta(\mathbf{z}_0)\|_2^2 \tag{A.4}$$

Above, $C_{\mathcal{D}}$ denotes the Lipschitz constant of the LDM decoder. The final equation is then used as a proxy to $\nabla_{\mathbf{z}_t} \log p(\mathbf{y} \mid \mathbf{z}_t)$.

From this derivation, we can deduce two implicit assumptions embedded in SILO's procedure.

(Implicit Assumption A1) Encoder Reconstruction Quality. The guidance signal (A.1) works entirely with $\mathcal{E}(\mathbf{y})$ instead of $\mathbf{y}$. SILO justifies this by showing that the autoencoder reconstructs measurements well (high PSNR of $\mathcal{D}(\mathcal{E}(\mathbf{y}))$ vs $\mathbf{y}$) for common image processing tasks.

(Implicit Assumption A2) Isotropic Gaussian Likelihood. Denoting $\mu_\theta(\mathbf{z}_{0|t}) = H_\theta(\mathbf{z}_{0|t})$, SILO's guidance (A.1) takes the form $-\nabla_{\mathbf{z}_t} \left\|\mathcal{E}(\mathbf{y}) - \mu_\theta(\mathbf{z}_{0|t})\right\|_2^2$. We observe that this is, up to a positive scalar, the score of an isotropic Gaussian:

Proposition 9. *Suppose* $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) = \mathcal{N}(\mathcal{E}(\mathbf{y}); \mu_\theta(\mathbf{z}_{0|t}), \sigma^2\mathbf{I})$ *for some fixed* $\sigma > 0$*. Then, the score is* $\nabla_{\mathbf{z}_t} \log p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) = \frac{1}{\sigma^2} \left(\nabla_{\mathbf{z}_t} \mu_\theta(\mathbf{z}_{0|t})\right)^\top (\mathcal{E}(\mathbf{y}) - \mu_\theta(\mathbf{z}_{0|t}))$*, recovering SILO's guidance* (A.1) *with scale* $\eta = 1/(2\sigma^2)$.

SILO is therefore implicitly modeling $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ as a Gaussian with isotropic, input-independent covariance.

A.1 ANaLOG AS A RELAXATION OF SILO

We can view ANaLOG as a relaxation of the two implicit assumptions made in SILO.

Specifically, considering Implicit Assumption A1, we notice that reconstruction quality of the autoencoder on $\mathbf{y}$ is not the same as statistical sufficiency for posterior sampling, which is what Approximation 1 requires. For example, a lossy encoder can achieve high reconstruction PSNR yet still discard information relevant to $\mathbf{z}_t$. This perspective is important for future work that may consider new encoding maps that allow for generalization to tasks beyond image-space inverse problems, e.g., phase retrieval.

Regarding Implicit Assumption A2, as we demonstrated in Proposition 9, SILO corresponds to the special time and input-independent case $\Sigma_\theta \equiv \sigma^2\mathbf{I}$. As we demonstrate in Sections 4 and 5, the lifting of this assumption to include richer covariance structure is both provably necessary in certain settings and empirically beneficial. Further, SILO is derived from the lens that for efficiency, we can learn a forward operator directly operating in the latent space: from this perspective, there is no need for the latent forward operator to be time-dependent. In fact, SILO also experiments with a time-independent CNN architecture for H_θ, but adopts a time-dependent Readout-Guidance (Luo et al., 2024) architecture in general. Our framework exposes that time-dependency is not a curious artifact of the architecture, but *necessary* if we aim to approximate the likelihood $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ which is a time-varying likelihood. This is further illustrated in Figure 5b, which shows that using a time-independent covariance harms performance in practice.

B EXTENDED RELATED WORK

Uncertainty in Pixel-Space Diffusion Models. We review here in more detail the rich line of work of exploiting uncertainty estimates for guidance in pixel-space diffusion models, and we contrast this with the more challenging LDM setting. Recall from Section 2 that pixel-space diffusion models consider a diffusion process directly in pixel space. The corresponding conditional score for solving inverse problems then becomes

$$\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{y}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + \nabla_{\mathbf{x}_t} \log p(\mathbf{y} \mid \mathbf{x}_t), \tag{B.1}$$

where the noisy likelihood score requires approximation. This distribution factorizes as

$$p(\mathbf{y} \mid \mathbf{x}_t) = \int p(\mathbf{y} \mid \mathbf{x}_0) p(\mathbf{x}_0 \mid \mathbf{x}_t)\, d\mathbf{x}_0. \tag{B.2}$$

As we know the forward model $p(\mathbf{y} \mid \mathbf{x}_0)$, the main challenge is in approximating the posterior denoising distribution. Early methods such as DPS (Chung et al., 2022) approximate $p(\mathbf{x}_0 \mid \mathbf{x}_t)$ as a Dirac delta around the posterior mean $\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t]$, and Pi-GDM (Song et al., 2023b) extends to use a Gaussian distribution centered at the posterior mean with isotropic covariance. This was extended to richer covariance models, leveraging higher-order Tweedie's formulas as in TMPD (Boys et al., 2023) or approximating via diagonal matrices as in Peng (Peng et al., 2024) or FreeHunch (Rissanen et al., 2024). Certain variants also proposed to learn the posterior covariance by parameterizing the covariance by a deep network (Meng et al., 2021). Crucially, these more complex covariance models such as TMPD, Peng, and FreeHunch are only applicable when the forward operator $\mathcal{A}$ is linear, otherwise the integral in B.2 is still intractable.

This issue is precisely why these more complex posterior covariance models do not aid in modeling the noisy likelihood score for LDMs . Namely, for LDMs, Equation B.2 becomes

$$p(\mathbf{y} \mid \mathbf{z}_t) = \int p(\mathbf{y} \mid \mathbf{z}_0) p(\mathbf{z}_0 \mid \mathbf{z}_t)\, d\mathbf{z}_0. \tag{B.3}$$

Even with Gaussian parameterizations of $p(\mathbf{z}_0 \mid \mathbf{z}_t)$, the likelihood $p(\mathbf{y} \mid \mathbf{z}_0)$ is $\mathcal{N}(\mathcal{A}(\mathcal{D}(\mathbf{z}_0)), \sigma_y^2 \mathbf{I})$. Because the decoder is nonlinear, this integral no longer maintains any closed form. As such, prior works are mostly limited to considering approximations to the noisy likelihood evaluated at the posterior mean only. Attempts to include uncertainty have focused on using surrogate guidance terms that include regularizations on second-order estimates of uncertainty. Specifically, STSL (Rout et al., 2024) uses a surrogate guidance that targets uncertainty by the second-order Tweedie formula, which links denoising covariance to the Jacobian of the denoiser. The norm of this Jacobian, estimated via finite difference evaluations of the denoiser in a perturbation set, is then added as a regularizer. However, these finite difference evaluations greatly increase sampling time, increasing denoiser evaluations by five to ten fold during every guidance step. Our work avoids these closed form computations and additional regularizations by directly targeting the likelihood in the latent space $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$, which permits a straightforward learning rule for the covariance.

Uncertainty in Classification and Regression. Our work draws on a mature line of research on heteroscedastic uncertainty estimation in supervised learning. The foundational idea of training a neural network to predict both a mean and an input-dependent variance dates back to Nix & Weigend (1994), who proposed minimizing the Gaussian negative log-likelihood (NLL) so that the network learns to output larger variance on inherently noisy inputs. Kendall & Gal (2017) popularized this approach in deep learning for computer vision, distinguishing *aleatoric* uncertainty (irreducible noise inherent to the data) from *epistemic* uncertainty (model uncertainty reducible with more data), and showing that modeling aleatoric uncertainty via a learned heteroscedastic variance improves both regression and classification. More recently, Seitzer et al. (2022) identified optimization pitfalls in heteroscedastic NLL training—specifically, that the network can exploit the learned variance to explain away large residuals rather than fitting the mean—and proposed β-NLL, which reweights the loss to mitigate this failure mode.

Our framework can be viewed as an adaptation of this heteroscedastic paradigm from supervised prediction to diffusion posterior sampling. In the classification/regression setting, one trains a network to output $(\mu_\theta(\mathbf{x}), \sigma_\theta^2(\mathbf{x}))$ and minimizes $\frac{1}{2}\log\sigma_\theta^2 + \frac{1}{2\sigma_\theta^2}\|y - \mu_\theta\|^2$. Analogously, `ANaLOG` trains a latent forward operator to output a mean $\mu_\theta(\mathbf{z}_t, t, \sigma_y)$ and a diagonal covariance $\Sigma_\theta(\mathbf{z}_t, t, \sigma_y)$ and minimizes the multivariate Gaussian NLL (Equation 6). The key difference is the downstream use: rather than producing calibrated predictive intervals, the learned covariance is used to construct a Mahalanobis-weighted guidance signal that steers diffusion sampling. In this sense, `ANaLOG` transfers the benefits of heteroscedastic uncertainty estimation—emphasizing reliable predictions and downweighting uncertain ones—from discriminative tasks to generative posterior sampling with diffusion models. We also note that in our two-stage training procedure, we first train the mean with a frozen isotropic covariance before learning the covariance, which sidesteps the optimization pitfalls identified by Seitzer et al. (2022), as the mean network is not incentivized to exploit the variance head during its own training.

C Proofs from Section 4

C.1 Proof of Proposition 6 (Optimal Autoencoder)

Proof. We prove global optimality by direct substitution and a lower bound argument.

First, recall the encoder and decoder are of the form:

$$\mathcal{E}(\mathbf{x}) = \mathbf{W}(\mathbf{g} \circledast_{\downarrow M} \mathbf{x}) \tag{C.1}$$

$$\mathcal{D}(\mathbf{z}) = \mathbf{h} \circledast_{\uparrow M} (\mathbf{V}\mathbf{z}) \tag{C.2}$$

Suppose the weights of these maps satisfy the conditions of the proposition, so

$$(\mathbf{F}_d \mathbf{g})_\omega = \begin{cases} \sqrt{M}, & \omega \in \Omega_k, \\ 0, & \omega \notin \Omega_k, \end{cases} \qquad (\mathbf{F}_d \mathbf{h})_\omega = \begin{cases} \sqrt{M}, & \omega \in \Omega_k, \\ 0, & \omega \notin \Omega_k, \end{cases} \tag{C.3}$$

and

$$\mathbf{W} = \mathbf{F}_k \, \mathrm{diag}_{\omega \in \Omega_k}(\lambda_\omega^{-1/2}) \, \mathbf{F}_k^*, \tag{C.4}$$

$$\mathbf{V} = \mathbf{F}_k \, \mathrm{diag}_{\omega \in \Omega_k}(\lambda_\omega^{1/2}) \, \mathbf{F}_k^*. \tag{C.5}$$

Then, the encoder acts an ideal low-pass filter followed by whitening of the retained Fourier coefficients, and the decoder is an inverse of the encoder on the subspace spanned by the first k (centered) frequencies. Thus, the composition $\mathcal{D} \circ \mathcal{E}$ is the orthogonal projector P_{Ω_k} onto the subspace spanned by the frequencies in Ω_k. Therefore the reconstruction error is

$$\mathbb{E}\|\mathcal{D}(\mathcal{E}(\mathbf{x}_0)) - \mathbf{x}_0\|_2^2 = \sum_{\omega \in \Omega_d \setminus \Omega_k} \lambda_\omega. \tag{C.6}$$

Further, because of the whitening, the pushforward $\mathcal{E}_\sharp p$ is exactly a standard k-dimensional Gaussian, so the KL-term in the objective is exactly 0.

We now show that the above setting of the encoder and decoder is a global minimizer of (13). To show this, we will demonstrate that (C.6) matches a lower bound on the reconstruction error of any encoder/decoder pair. This fact combined with the non-negativity of the KL divergence will prove that the above setting must be a global minimizer of (13).

The lower bound of reconstruction error follows from the Eckart-Young-Mirsky theorem (Eckart & Young, 1936). Specifically, because $\mathcal{E}$ is a linear map with codomain $\mathbb{R}^k$ where $k < d$, the composition $\mathcal{D} \circ \mathcal{E}$ has rank at most k and thus the reconstruction error is minimized by retaining the k largest-variance principal components of the data. By Assumption 3, due to the power law distribution of the data in frequency domain, the top principal components are precisely the lowest-magnitude centered frequencies (the set Ω_k). Since the setting of the encoder and decoder given above precisely performs this projection, it must be a global minimizer of the reconstruction loss and therefore also a global minimizer of (13) by non-negativity of the KL loss. □

Remark (Role of the isotropic latent prior)**.** *At the optimal autoencoder, $\mathcal{E}_\sharp p = \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, so the latent marginal $p(\mathbf{z}_t)$ is standard Gaussian for all t and the unconditional diffusion process is trivial (the score is simply $-\mathbf{z}_t$). This is by design: the purpose of our theoretical model is not to study unconditional generation, but to characterize the* likelihood *$p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ and its covariance structure, which is the object that governs guidance. The covariance anisotropy (Proposition 7, Part 2) arises entirely from the interaction of the forward operator, the encoder's whitening, and the diffusion timestep—none of which depend on the latent prior being non-trivial. The isotropic prior is an analytical convenience that makes the posterior mean available in closed form via Tweedie's formula; in practice, LDM encoders do not achieve $\mathrm{KL} = 0$ and the latent distribution retains further structure.*

C.2 Proof of Proposition 7

We prove Proposition 7 by establishing two lemmas, one for each part, and then combining them.

Lemma 10 (Sufficiency of encoded measurements)**.** *Under Assumptions 3–5 with the optimal autoencoder (Proposition 6), $\mathcal{E}(\mathbf{y})$ is a sufficient statistic for $\mathbf{z}_t$ given $\mathbf{y}$:*

$$p(\mathbf{z}_t \mid \mathbf{y}) = p(\mathbf{z}_t \mid \mathcal{E}(\mathbf{y})). \tag{C.7}$$

Proof. Since $\mathbf{A}$ is a circular convolution as per Assumption 5, the measurement decomposes component-wise in centered Fourier coordinates:

$$(\mathbf{F}_d \mathbf{y})_\omega = a_\omega (\mathbf{F}_d \mathbf{x}_0)_\omega + \sigma_y (\mathbf{F}_d \boldsymbol{\epsilon})_\omega, \qquad \omega \in \Omega_d, \tag{C.8}$$

where a_ω is the Fourier response of $\mathbf{A}$ at index ω. We denote $\mathbf{y}^{\mathrm{lo}}$ for the Fourier coefficients indexed by Ω_k, and $\mathbf{y}^{\mathrm{hi}}$ for the remaining coefficients indexed by $\Omega_d \setminus \Omega_k$.

Under the optimal autoencoder from Proposition 6, $\mathcal{E}(\mathbf{y})$ is an invertible whitening transform of $\mathbf{y}^{\mathrm{lo}}$ (since $\mathbf{W}$ is invertible), so conditioning on $\mathcal{E}(\mathbf{y})$ is equivalent to conditioning on $\mathbf{y}^{\mathrm{lo}}$.

The latent variable $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$ depends only on the coefficients of $\mathbf{x}_0$ in Ω_k, and $\mathbf{z}_t$ depends on $\mathbf{x}_0$ only through $\mathbf{z}_0$. Under Assumption 3, the Fourier coefficients of $\mathbf{x}_0$ are independent across frequencies, and the measurement noise is independent across frequencies. Therefore the high-frequency measurement $\mathbf{y}^{\mathrm{hi}}$ is conditionally independent of $\mathbf{z}_t$ given $\mathbf{y}^{\mathrm{lo}}$:

$$\mathbf{z}_t \perp \mathbf{y}^{\mathrm{hi}} \mid \mathbf{y}^{\mathrm{lo}}. \tag{C.9}$$

Hence

$$p(\mathbf{z}_t \mid \mathbf{y}) = p(\mathbf{z}_t \mid \mathbf{y}^{\mathrm{lo}}) = p(\mathbf{z}_t \mid \mathcal{E}(\mathbf{y})), \tag{C.10}$$

which proves sufficiency. □

Lemma 11 (Gaussian latent likelihood)**.** *Under Assumptions 3–5 with the optimal autoencoder (Proposition 6), $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_t$ and covariance $\mathbf{C}_t$ given by:*

$$\boldsymbol{\mu}_t = \mathbf{F}_k^* \operatorname{diag}_{\omega\in\Omega_k}(a_\omega)\, \mathbf{F}_k \mathbf{z}_{0|t}, \qquad \mathbf{C}_t = \mathbf{F}_k^* \operatorname{diag}_{\omega\in\Omega_k}\big(c_\omega(t)\big)\, \mathbf{F}_k, \tag{C.11}$$

where $c_\omega(t) = (1-\bar\alpha_t)|a_\omega|^2 + \sigma_y^2/\lambda_\omega$ for $\omega \in \Omega_k$.

Proof. We observe that the latent likelihood factorizes as

$$p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) = \int p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_0) p(\mathbf{z}_0 \mid \mathbf{z}_t)\, d\mathbf{z}_0 \tag{C.12}$$

Thus, the strategy for deriving $p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)$ is to (i) express $\mathcal{E}(\mathbf{y})$ as a linear function of $\mathbf{z}_0$ plus noise (the *latent measurement model*), and then (ii) marginalize out $\mathbf{z}_0$ using the posterior $p(\mathbf{z}_0 \mid \mathbf{z}_t)$.

To derive the latent measurement model, we would like to find a commuting diagram (Figure 6) such that $\mathcal{E}(\mathbf{A}\mathbf{x}_0) = \mathbf{H}\,\mathcal{E}(\mathbf{x}_0) = \mathbf{H}\mathbf{z}_0$. While this motivates our learning objective in general, in our setting, we can exploit spectral properties of the autoencoder and the forward operator to find a closed form of $\mathbf{H}$. Specifically, we use the Noble identities from multirate signal processing (Vaidyanathan, 2006). The Noble identities show that filtering followed by downsampling commutes with downsampling followed by an appropriately aliased filter. In our setting, since the encoder's ideal low-pass filter eliminates all frequencies outside Ω_k before downsampling, and $\mathbf{A}$ acts independently on each frequency, the two operations commute exactly: $\mathcal{E} \circ \mathbf{A} = \mathbf{H} \circ \mathcal{E}$, where $\mathbf{H}$ retains only the action of $\mathbf{A}$ on the frequencies in Ω_k. Expressed mathematically,

$$\mathbf{H} = \mathbf{F}_k^* \operatorname{diag}_{\omega\in\Omega_k}(a_\omega)\, \mathbf{F}_k. \tag{C.13}$$

For the noise term, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ has flat spectrum, and the encoder whitens each retained frequency by $\lambda_\omega^{-1/2}$, amplifying the noise non-uniformly. Combining both terms gives the latent measurement model:

$$\mathcal{E}(\mathbf{y}) = \mathbf{H}\mathbf{z}_0 + \boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}\big(\mathbf{0},\, \sigma_y^2\, \mathbf{F}_k^* \operatorname{diag}_{\omega\in\Omega_k}(\lambda_\omega^{-1})\, \mathbf{F}_k\big), \quad \boldsymbol{\eta} \perp \mathbf{z}_0. \tag{C.14}$$

This is a standard linear-Gaussian observation model in the latent space, with a diagonal operator and anisotropic noise whose amplification $\sigma_y^2/\lambda_\omega$ grows for higher frequencies.

Next, we can marginalize using (C.12). Under the variance-preserving noising process, the posterior is $p(\mathbf{z}_0 \mid \mathbf{z}_t) = \mathcal{N}(\hat{\mathbf{z}}_{0|t},\, (1-\bar\alpha_t)\, \mathbf{I}_k)$, where $\hat{\mathbf{z}}_{0|t} := \mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t] = \sqrt{\bar\alpha_t}\, \mathbf{z}_t$. The last equality follows from Tweedie's formula: since $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$, $p(\mathbf{z}_t)$ is also standard Gaussian and $\nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t) = -\mathbf{z}_t$, giving $\mathbb{E}[\mathbf{z}_0 \mid \mathbf{z}_t] = \frac{1}{\sqrt{\bar\alpha_t}}\big(\mathbf{z}_t + (1-\bar\alpha_t)(-\mathbf{z}_t)\big) = \sqrt{\bar\alpha_t}\, \mathbf{z}_t$. Conditioning on $\mathbf{z}_t$, the latent measurement model (C.14) becomes a linear function of the Gaussian $\mathbf{z}_0 \mid \mathbf{z}_t$ plus independent Gaussian noise $\boldsymbol{\eta}$. By standard Gaussian marginalization,

$$p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) = \mathcal{N}(\boldsymbol{\mu}_t,\ \mathbf{C}_t)\,, \tag{C.15}$$

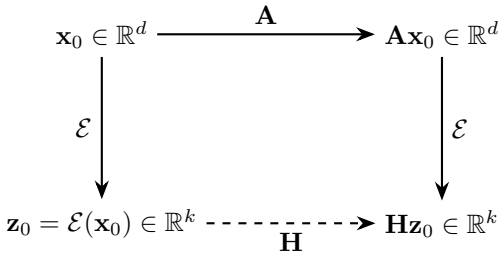


Figure 6: Commuting diagram for the latent forward operator. The Noble identity guarantees that encoding after degradation ($\mathcal{E} \circ \mathbf{A}$, right then down) equals latent degradation after encoding ($\mathbf{H} \circ \mathcal{E}$, down then right), so the diagram commutes: $\mathcal{E}(\mathbf{A}\mathbf{x}_0) = \mathbf{H}\,\mathcal{E}(\mathbf{x}_0) = \mathbf{H}\mathbf{z}_0$.

where the mean is $\boldsymbol{\mu}_t \coloneqq \mathbf{H}\hat{\mathbf{z}}_{0|t}$ and the covariance collects two sources of uncertainty:

$$\mathbf{C}_t = \underbrace{(1-\bar{\alpha}_t)\,\mathbf{H}\mathbf{H}^*}_{\text{diffusion uncertainty through } \mathbf{A}} + \underbrace{\sigma_y^2\,\mathbf{F}_k^*\,\mathrm{diag}_{\omega\in\Omega_k}(\lambda_\omega^{-1})\,\mathbf{F}_k}_{\text{measurement noise amplified by whitening}}. \tag{C.16}$$

Since $\mathbf{H}$ is diagonal in the latent Fourier basis, the covariance eigenvalues are

$$c_\omega(t) = (1-\bar{\alpha}_t)\,|a_\omega|^2 + \frac{\sigma_y^2}{\lambda_\omega}, \qquad \omega \in \Omega_k. \tag{C.17}$$

Both terms vary across frequencies, making the covariance inherently anisotropic. □

Proof of Proposition 7. Part 1 follows directly from Lemma 10, and Part 2 follows directly from Lemma 11. □

C.3 Proof of Theorem 8 (Isotropic vs. Anisotropic Covariance Approximation Error)

Proof. **Part 1: Irreducible error of isotropic covariance.** The isotropic approximation replaces the true covariance $\mathbf{C}_t$ with $\eta^2\,\mathbf{I}_k$ for a scalar $\eta > 0$, while keeping the mean $\boldsymbol{\mu}_t$ unchanged. We want to show that the KL divergence between the true and approximate likelihoods is lower bounded by a strictly positive quantity for any choice of η.

Since both the true and approximate likelihoods are Gaussians with the same mean, the KL divergence depends only on the covariances. Recall that for two Gaussians with shared mean,

$$\mathrm{KL}\big(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_1) \,\|\, \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_2)\big) = \frac{1}{2}\big[\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + \log\det\boldsymbol{\Sigma}_2 - \log\det\boldsymbol{\Sigma}_1 - k\big]. \tag{C.18}$$

We set $\boldsymbol{\Sigma}_1 = \mathbf{C}_t$ and $\boldsymbol{\Sigma}_2 = \eta^2\,\mathbf{I}_k$. Since both are diagonalized by the Fourier basis $\mathbf{F}_k$, we can evaluate each term using the eigenvalues $\{c_\omega(t)\}_{\omega\in\Omega_k}$ directly:

$$\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) = \frac{1}{\eta^2}\sum_{\omega\in\Omega_k} c_\omega, \tag{C.19}$$

$$\log\det\boldsymbol{\Sigma}_2 = k\log\eta^2, \tag{C.20}$$

$$\log\det\boldsymbol{\Sigma}_1 = \sum_{\omega\in\Omega_k}\log c_\omega. \tag{C.21}$$

Substituting, the KL divergence is

$$\mathrm{KL}\big(p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \,\|\, \mathcal{N}(\boldsymbol{\mu}_t,\, \eta^2\,\mathbf{I}_k)\big) = \frac{1}{2}\left[\frac{1}{\eta^2}\sum_{\omega\in\Omega_k} c_\omega + k\log\eta^2 - \sum_{\omega\in\Omega_k}\log c_\omega - k\right]. \tag{C.22}$$

We now minimize over η^2 to find the tightest possible lower bound. Differentiating (C.22) with respect to η^2 and setting the derivative to zero:

$$\frac{\partial}{\partial \eta^2}\left[\frac{1}{\eta^2}\sum_{\omega} c_\omega + k\log\eta^2\right] = -\frac{1}{\eta^4}\sum_{\omega} c_\omega + \frac{k}{\eta^2} = 0 \tag{C.23}$$

$$\implies \quad \eta^{*2} = \frac{1}{k}\sum_{\omega\in\Omega_k} c_\omega(t) = \mathrm{AM}\big(\{c_\omega(t)\}_{\omega\in\Omega_k}\big). \tag{C.24}$$

The second derivative $\frac{2}{\eta^6}\sum_\omega c_\omega - \frac{k}{\eta^4}$ is positive at η^{*2}, confirming this is a minimum. Substituting η^{*2} back into (C.22), the trace term becomes $\frac{1}{\eta^{*2}}\sum_\omega c_\omega = k$, which cancels the $-k$ term. The remaining expression is

$$\frac{k}{2}\log \mathrm{AM}(\{c_\omega\}_{\omega\in\Omega_k}) - \frac{1}{2}\sum_{\omega\in\Omega_k}\log c_\omega = \frac{k}{2}\log \mathrm{AM}(\{c_\omega\}_{\omega\in\Omega_k}) - \frac{k}{2}\log \mathrm{GM}(\{c_\omega\}_{\omega\in\Omega_k}) \tag{C.25}$$

$$= \frac{k}{2}\log\frac{\mathrm{AM}(\{c_\omega\}_{\omega\in\Omega_k})}{\mathrm{GM}(\{c_\omega\}_{\omega\in\Omega_k})}, \tag{C.26}$$

where we used the identity $\sum_\omega \log c_\omega = k\log \mathrm{GM}(\{c_\omega\}_{\omega\in\Omega_k})$. By the AM–GM inequality, $\mathrm{AM} \geq \mathrm{GM}$ with equality if and only if all c_ω are equal, establishing the bound (18).

It remains to show strict positivity under Assumptions 3–5. The bound (18) equals zero if and only if all eigenvalues $c_\omega(t)$ are identical. We show this cannot happen.

At $t = 0$ ($\bar\alpha_0 = 1$), the diffusion term vanishes and $c_\omega(0) = \sigma_y^2/\lambda_\omega$. Since $\lambda_\omega = C(1+|\omega|)^{-\alpha}$ with $\alpha > 1$, this equals $(\sigma_y^2/C)(1+|\omega|)^\alpha$, which is strictly increasing in $|\omega|$ and therefore non-constant (using $\sigma_y > 0$ from Assumption 5).

For $t > 0$ ($\bar\alpha_t < 1$), the eigenvalues are $c_\omega(t) = (1-\bar\alpha_t)\,|a_\omega|^2 + \sigma_y^2/\lambda_\omega$. For these to be constant across ω, we need $|a_\omega|^2 = (K - \sigma_y^2/\lambda_\omega)/(1-\bar\alpha_t)$ for some constant K. Since $|a_\omega|^2$ is fixed while $\bar\alpha_t$ varies with t, this can hold for at most one $t \in (0,T)$, so under $t \sim \mathcal{U}[0,T]$, this is a measure-zero event.

Therefore the $\{c_\omega(t)\}_{\omega\in\Omega_k}$ are non-constant with probability 1 when $t \sim \mathcal{U}[0,T]$, and the KL divergence (18) is strictly positive.

Part 2: Convergence of diagonal covariance model. In the Fourier basis, the true likelihood decomposes into k independent 1D Gaussians: the ω-th component has mean $a_\omega\,(\mathbf{F}_k\mathbf{z}_{0|t})_\omega$ and variance $c_\omega(t)$. A diagonal Gaussian model with linear mean and free diagonal covariance matches this structure exactly, so the model is well-specified.

For each frequency ω, the MLE estimates the coefficient a_ω via ordinary least squares and the variance $c_\omega(t)$ as the sample variance of residuals. Standard results for Gaussian MLE on a well-specified model (Lehmann & Casella, 1998) give, for each component, $\mathbb{E}[\mathrm{KL}_\omega] = O(1/n)$, where the implicit constant depends on the true parameters but is bounded for each fixed t. Summing over k independent components yields $\mathbb{E}\big[\mathrm{KL}\big(p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t) \parallel \hat p(\mathcal{E}(\mathbf{y}) \mid \mathbf{z}_t)\big)\big] = O(k/n)$. □

D Discussion of Theoretical Analysis

D.1 Regimes Where Isotropic Approximation Error Is Large

The KL lower bound (18) equals $\frac{k}{2}\log(\mathrm{AM}/\mathrm{GM})$, which measures the *multiplicative spread* of the eigenvalues $\{c_\omega(t)\}_{\omega\in\Omega_k}$. The discussion below identifies regimes in which Theorem 8 predicts this quantity to be $\Omega(1)$ or larger.

To understand the scaling of the lower bound, recall that $c_\omega(t) = (1-\bar\alpha_t)\,|a_\omega|^2 + \sigma_y^2/\lambda_\omega$. Since $\lambda_\omega \propto (1+|\omega|)^{-\alpha}$, the second term scales as $\sigma_y^2/\lambda_\omega \propto (1+|\omega|)^\alpha$. Let $\omega_{\max} \in \Omega_k$ denote a retained frequency with largest magnitude, $|\omega_{\max}| = (k-1)/2$. The ratio between the largest and smallest eigenvalues satisfies

$$\frac{c_{\omega_{\max}}(t)}{c_0(t)} \;\geq\; \frac{\sigma_y^2/\lambda_{\omega_{\max}}}{\sigma_y^2/\lambda_0 + (1-\bar\alpha_t)\,\|a\|_\infty^2} \;=\; \frac{(1+|\omega_{\max}|)^\alpha}{1 + (1-\bar\alpha_t)\,\|a\|_\infty^2\,\lambda_0/\sigma_y^2}, \tag{D.1}$$

where $\|a\|_\infty = \max_\omega |a_\omega|$. This tells us that the eigenvalue ratio grows exponentially in the spectral decay α. Since the lower bound is the logarithm of this quantity, the lower bound grows linear in the spectral decay.

Near the data end of the reverse process ($t \to 0$, $\bar{\alpha}_t \approx 1$). The diffusion noise contribution $(1 - \bar{\alpha}_t)\,|a_\omega|^2$ is small, so the eigenvalue spread is dominated by $\sigma_y^2/\lambda_\omega$ and the isotropic error is largest. As the reverse process moves towards the data (t decreases, $\bar{\alpha}_t \to 1$), the diffusion term $(1 - \bar{\alpha}_t)\,|a_\omega|^2$ shrinks, leaving the eigenvalue spread increasingly dominated by the measurement noise term $\sigma_y^2/\lambda_\omega$, which varies strongly across frequencies. This *increases* the isotropic penalty near the data end, consistent with Fig. 3b.

Beyond convolutional operators. Our theoretical analysis focuses on deconvolution, where the convolutional structure of the forward operator and autoencoder enables exact characterization. In practice, native-latent guidance also applies to non-convolutional inverse problems such as inpainting and JPEG restoration, where the sufficiency and likelihood structure may not admit closed-form expressions. We verify empirically in Section 5 that the anisotropic covariance structure improves performance on these tasks as well; extending the theoretical analysis to non-convolutional operators is an interesting direction for future work.

D.2 Additional Visualizations

We provide additional visualizations that complement the analysis of the covariance eigenvalues $c_\omega(t)$ from Proposition 7 and the isotropic KL lower bound from Theorem 8. All plots use data dimension $d = 256$, latent dimension $k = 64$, $\sigma_y = 0.05$, $\alpha = 2.5$ (unless varied), and the Gaussian blur operator from the main text.

Effect of measurement noise σ_y. Figure 7 shows the KL lower bound as a function of σ_y at a fixed near-data timestep ($t = 0.05$). Recall that $c_\omega(t) = (1 - \bar{\alpha}_t)\,|a_\omega|^2 + \sigma_y^2/\lambda_\omega$. At very low σ_y, the noise-amplification term $\sigma_y^2/\lambda_\omega$ is negligible and the eigenvalue spread is determined entirely by the frequency response $|a_\omega|^2$ of the forward operator. For mild spectral decay ($\alpha = 1.5$), the blur's frequency response is sufficiently non-flat to produce a sizeable KL even at low noise. As σ_y increases, the noise-amplification term dominates and the eigenvalue spread grows as $(1 + |\omega|)^\alpha$, so the bound increases monotonically for larger α. For intermediate α values, the two regimes create a non-monotonic profile: the KL first dips as the diffusion-noise term and the measurement-noise term partially balance, then rises as the noise amplification takes over. This suggests that the optimal guidance strategy depends on the noise level in a non-trivial way that a scalar η cannot capture.

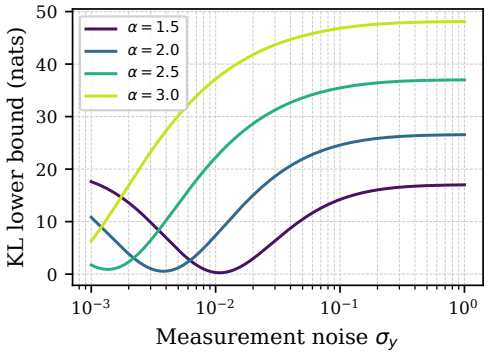


Figure 7: KL lower bound (18) vs. measurement noise σ_y at $t = 0.05$ for different spectral decay rates α. The isotropic penalty depends non-trivially on σ_y: at low noise the bound is set by the operator's frequency response, while at high noise the encoder's whitening ($\sigma_y^2/\lambda_\omega$) dominates.

Dependence on the forward operator. Figure 8 compares the covariance spectrum $c_\omega(t)$ at a fixed timestep for four different forward operators: the identity ($|a_\omega| = 1$ for all ω), a mild Gaussian blur (bandwidth $= 30$), a strong Gaussian blur (bandwidth $= 8$), and $4\times$ super-resolution ($|a_\omega| = 1$ for $|\omega| < k/4$, $|a_\omega| = 0$ otherwise). To isolate the effect of the operator, we use a low measurement noise $\sigma_y = 0.001$ so that the diffusion term $(1 - \bar{\alpha}_t)|a_\omega|^2$ shapes the spectrum rather than the noise floor $\sigma_y^2/\lambda_\omega$.

The identity operator produces a nearly flat covariance spectrum, since $|a_\omega| = 1$ for all ω and the only variation comes from the small noise term. As the blur strengthens, the covariance dips at frequencies where $|a_\omega| \approx 0$: the forward operator destroys these components, so the only remaining uncertainty is the noise floor. Super-resolution produces the most dramatic structure: a sharp cliff at $|\omega| = k/4$ where the operator's frequency response drops to zero. For $|\omega| < k/4$, the covariance is dominated by diffusion uncertainty; for $|\omega| \geq k/4$, it drops to the noise floor.

This visualization illustrates that different inverse problems induce qualitatively different covariance landscapes. An isotropic approximation applies the same precision to all coordinates regardless of these structural differences, while the learned covariance in our framework adapts to each task's spectral profile.

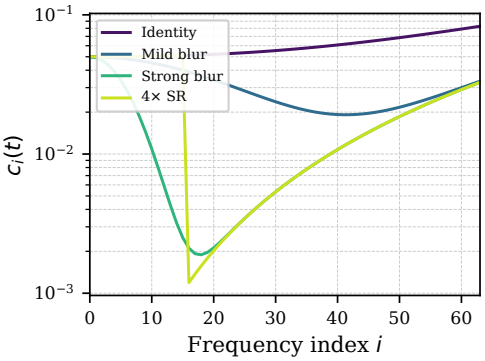


Figure 8: Covariance eigenvalues $c_\omega(t)$ at $t = 0.05$ for different forward operators ($\alpha = 2.5$, $\sigma_y = 0.001$). Eigenvalues are plotted against nonnegative frequency magnitude $|\omega|$. Each operator produces a distinct spectral profile. The identity is nearly flat; blurs suppress high frequencies progressively; $4\times$ SR introduces a sharp cutoff. An isotropic covariance cannot adapt to these task-dependent structures.

E Experimental Details

Datasets, resolution, and prompts. Both training and evaluation use 512×512 images. On FFHQ (Karras et al., 2019) (CC BY-NC-SA 4.0) we hold out 1,000 images for evaluation and use the remainder for training; on ImageNet (Deng et al., 2009) (distributed under ImageNet Terms of Access for non-commercial research and educational purposes) we train on the standard train split and evaluate on 1,000 randomly selected validation images. Following SILO (Raphaeli et al., 2025), the FFHQ runs use the fixed text prompt "*A high quality photo of a face*" during sampling, whereas the ImageNet runs use an empty prompt. The negative prompt is empty for all runs. We do not use Classifier-Free Guidance (CFG) (Ho & Salimans, 2022) for any of the sampling results.

Inverse Problem Task Descriptions. For Gaussian deblurring, the forward operator uses a kernel of size 61×61 with standard deviation 3.0. For super-resolution task, the forward operator uses downsampling by bicubic interpolation with antialiasing. For inpainting, we mask out the center 256×256 box; note this is a particularly challenging inpainting task, especially for face datasets where this box usually occludes most of the face. Thus, we cannot expect exact reconstruction and perceptual quality of reconstructions is a more faithful metric. For JPEG decompression, the forward operator compresses by a quality factor of 10.

E.1 Training Details

To foster reproducibility, we detail all architectural and training details below. We did not attempt to optimize these training hyperparameters or architectures.

Mean head architecture and UNet feature aggregation. The mean head and variance head share a common backbone of aggregated UNet features, following the Readout-Guidance design (Luo et al., 2024). We attach hooks to all twelve ResNet blocks in the four upsampling stages of the SD-1.5 UNet (the up-block channel widths are $[1280, 1280, 640, 320]$ with three ResNet blocks each), giving a per-block feature with channel dimensions

[1280,1280,1280,1280,1280,1280, 640,640,640,320,320,320]. Each block's feature map is bilinearly resampled to 64×64 (matching the latent spatial resolution), concatenated with the broadcast sigma embedding (see below), and passed through a per-block bottleneck projecting to 384 channels (a $1\times1\to3\times3\to1\times1$ ResNet block with timestep conditioning); the twelve projected features are then summed with learned softmax mixing weights w_ℓ (one per block, also trained). The resulting 384-channel aggregated feature $\mathbf{f}_{\text{agg}} \in \mathbb{R}^{384\times64\times64}$ is concatenated with the 1-channel sigma embedding map and fed into the mean head, whose architecture is $\text{Conv}(385 \to 128, 3\times3) \to \text{SiLU} \to \text{Conv}(128\to32, 3\times3) \to \text{SiLU} \to \text{Conv}(32\to4, 1\times1)$. The output is the per-pixel mean $\mu_\theta \in \mathbb{R}^{4\times64\times64}$; no output activation is applied. The total trainable parameter count is 14,227,540 for the mean head. The variance head h_ϕ adds another 480,676 parameters and is detailed in Table 4; it consumes the same pre-head feature $[\mathbf{f}_{\text{agg}} \,\|\, \mathbf{e}_\sigma]$. The raw output of the variance head is passed through a positivity-preserving transform to yield a numerically stable variance:

$$\sigma^2 = \text{softplus}(\mathbf{r}) + \varepsilon_{\min}, \tag{E.1}$$

$$\log\sigma^2 = \text{clip}\big(\log\sigma^2,\ \ell_{\min},\ \ell_{\max}\big), \tag{E.2}$$

$$\sigma^2 \leftarrow \exp(\log\sigma^2), \tag{E.3}$$

with variance floor $\varepsilon_{\min} = 10^{-4}$ and clipping range $[\ell_{\min}, \ell_{\max}] = [-6, 4]$. We further initialize the final convolutional layer with zero weights. The bias is chosen such that the variance is isotropic unit variance at the beginning of training after passing through the softplus activation. For our two-stage training objective, we train using the mean-fitting objective for 75000 steps and then train the variance-fitting objective for 10000 steps.

Table 4: Architecture of the log-variance head h_ϕ. Input is the pre-head feature $\mathbf{f} \in \mathbb{R}^{B\times385\times64\times64}$ (aggregated UNet features $\mathbf{f}_{\text{agg}} \in \mathbb{R}^{384\times64\times64}$ concatenated with the sigma embedding map $\mathbf{e}_\sigma \in \mathbb{R}^{1\times64\times64}$). Output is per-pixel $\log\sigma^2 \in \mathbb{R}^{B\times4\times64\times64}$.

Layer	Operation	Output shape	#Params
Input	$\mathbf{f} = [\mathbf{f}_{\text{agg}} \,\|\, \mathbf{e}_\sigma]$	$385 \times 64 \times 64$	—
Conv-1	$\text{Conv2d}(385\to128,\ 3\times3,\ \text{pad} = 1)$	$128 \times 64 \times 64$	443,648
	SiLU	$128 \times 64 \times 64$	—
Conv-2	$\text{Conv2d}(128\to32,\ 3\times3,\ \text{pad} = 1)$	$32 \times 64 \times 64$	36,896
	SiLU	$32 \times 64 \times 64$	—
Conv-3	$\text{Conv2d}(32\to4,\ 1\times1)$	$4 \times 64 \times 64$	132
		Total	**480,676**

DCT-Domain Training. To train the DCT-domain variant, we reuse the spatial-diagonal architecture above unchanged, but interpret the head's output as per-bin diagonal log-variance in the DCT domain. The variance head, however, is a small-kernel translation-equivariant CNN, whose smoothing prior makes sense for spatial-domain inputs since adjacent input positions are spatial neighbors. But the head's output positions index DCT frequency bins, not pixels, and the input features remain spatial; this mismatch causes the conv to incorrectly smooth across DCT-bin neighbors that have no reason to share variance (e.g., the DC bin vs. its neighbors). We make a simple adjustment: apply a 2D DCT to the variance head's input feature map before the first conv. The head then operates on DCT-domain features, with input-position (i, j) corresponding to DCT bin (i, j), so the conv's local-smoothing prior is applied across neighboring frequencies, a meaningful prior in DCT space, and the output's per-bin semantics match the input's per-bin semantics.

Variance regularization. We add a small quadratic penalty $\lambda_{\text{var}} \cdot \mathbb{E}_{\mathbf{z}_t,t,\sigma_y}[\,\|\log\Sigma_\theta\|_2^2\,]$ to the training loss, with $\lambda_{\text{var}} = 10^{-4}$. This penalty discourages $\log\Sigma_\theta$ from drifting toward the clipping endpoints $\ell_{\min}$ or $\ell_{\max}$ on coordinates that receive little gradient signal, and it has negligible effect on calibrated bins where $|\log\Sigma_\theta| \ll 1$.

Measurement noise embedding. The measurement-noise level σ_y is encoded by re-using the LDM's sinusoidal timestep embedding: we map σ_y to a pseudo-timestep $\tilde{t} = (\sigma_y/0.1) \cdot 999$ and pass $\tilde{t}$ through the SD-1.5 timestep MLP, producing a 1280-d vector. A learned linear layer projects this vector to $64\times64 = 4096$ dimensions, which we reshape to a single-channel spatial map

$\mathbf{e}_\sigma \in \mathbb{R}^{1\times 64\times 64}$ and apply LayerNorm. This map is broadcast-concatenated to every block's feature before bottlenecking, so both the aggregation backbone and the heads are conditioned on σ_y.

Timestep sampling during training. At every training step, we sample the diffusion timestep uniformly from $t \sim \mathrm{Uniform}\{1, 2, \ldots, T-1\}$ with $T = 1000$ (matching the DDPM noise schedule of the base LDM), and form $\mathbf{z}_t$ via the variance-preserving forward process. We use a single random timestep per sample rather than averaging over t.

Measurement-noise sampling during training. For every training step we sample $\sigma_y \sim \mathrm{Uniform}[0, 0.1]$ and apply $\mathbf{y} = \mathcal{A}(\mathbf{x}) + \sigma_y\, \boldsymbol{\epsilon}_y$ with $\boldsymbol{\epsilon}_y \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The same sampled σ_y is fed to the network via the sigma embedding above.

Optimizer and hyperparameters. Both stages use AdamW with the PyTorch defaults (β_1=0.9, β_2=0.999, ε=10^{-8}), a constant learning rate of 2×10^{-4}, and weight decay 0. Training is in fp32 with no mixed precision. We use a per-GPU batch size of 8 (no gradient accumulation) and clip the global parameter gradient norm to 1.0 before each optimizer step.

Hardware and timing. All training, sampling, and reported timings were run on NVIDIA RTX A5000 (24 GB) GPUs. For FFHQ dataset, we utilize pretrained Stage 1 checkpoints from the SILO codebase. For ImageNet, we pretrain Stage 1, which takes roughly 28 GPU hours for an A5000 GPU. Stage 2 training takes roughly 4 GPU hours on the same GPU.

E.2 Inference Details and Baselines

Inference step size and guidance clipping. We follow SILO (Raphaeli et al., 2025) and normalize the guidance objective by the ℓ_2 norm of the residual $\|\mathbf{r}\|_2$, treated as a constant (i.e., detached from the computation graph so that gradients do not flow through the normalization). Following SILO, we additionally clip the per-coordinate guidance gradient to $[-1, 1]$ for stability against isolated low-variance outliers.

DCT-domain inference details. For the DCT variant of `ANaLOG` the variance head predicts a diagonal covariance $\Sigma_\theta = \mathrm{diag}(v)$ in the 2D DCT basis, so guidance applies a per-frequency precision weighting $\hat{\mathbf{r}} \mapsto \hat{\mathbf{r}}/v$ on the DCT residual $\hat{\mathbf{r}} = \mathrm{DCT}(\mathcal{E}(\mathbf{y}) - \mu_\theta)$ before mapping back to latent space via the inverse DCT. The remainder of the inference algorithm is identical to the spatial-domain variant.

Hyperparameters. To ensure fairness, we first find the optimal hyperparameters for the SILO baselines via cross-validation and then use the same hyperparameters for our spatial covariance baseline. This directly measures the impact of varying covariance parameterizations while keeping all else fixed. For our DCT covariance parameterization, the scales are tuned separately as the covariance has different magnitudes than in spatial domain and the optimal scales for SILO are not the optimal scales for `ANaLOG`-DCT. We now summarize these hyperparameters here. On FFHQ, the guidance scale is set to 2.0 for Gaussian blur, 3.0 for super-resolution (4× and 8×), 2.5 for inpainting, and 1.0 for JPEG decompression. For the DCT covariance variants, these become guidance scale 1.0 for Gaussian blur, 3.0 and 2.0 for super-resolution 4× and 8× respectively, 3.0 for JPEG decompression, and 1.5 for inpainting. For ImageNet, the spatial covariance variant uses 1.0 for all tasks. For the DCT variant, we use scale 0.5 for deblurring and 0.7 for super-resolution tasks. See Appendix F.4 for details on this hyperparameter search.

Baselines. For the LDPS (Rout et al., 2023) and PSLD (Rout et al., 2023) baselines, we use the code from `https://github.com/LituRout/PSLD` (CC-BY-4.0). We use the default hyperparameters for all tasks, given in their codebase. The only task which they do not evaluate on is JPEG decompression, for which we use the same hyperparameters as the Gaussian deblurring task. As the STSL (Rout et al., 2024) baseline does not provide code, we re-implement within the PSLD codebase, utilizing the same hyperparameters from their paper for all tasks. For the SILO (Raphaeli et al., 2025), we use the code from `https://github.com/ronraphaeli/SILO` (CC-BY-4.0).

E.3 Full Training and Inference Algorithms

Algorithms 1 and 2 summarize the full procedures.

Algorithm 1 ANaLOG two-stage training: Gaussian NLL on encoded measurements

Require: Pretrained LDM (encoder $\mathcal{E}$, denoiser ϵ_ϕ, scheduler $\{\bar{\alpha}_t\}$); mean head μ_θ; variance head h_ϕ; forward model $\mathcal{A}$; training set $\{\mathbf{x}^{(i)}\}$; optimizer; noise range $[\sigma_{\min}, \sigma_{\max}]$; regularization weight λ_{var}; gradient-norm clip $g_{\max}$; stage-1 steps $S_1 = 75{,}000$; stage-2 steps $S_2 = 10{,}000$.

Stage 1: train μ_θ with $\Sigma_\theta = \mathbf{I}$ (equivalent to SILO).

for $s = 1, \ldots, S_1$ **do**
  Sample $\mathbf{x}, \sigma_y, t$; $\mathbf{y} \leftarrow \mathcal{A}(\mathbf{x}) + \sigma_y \boldsymbol{\epsilon}_y$; $\mathbf{z}_0 \leftarrow \mathcal{E}(\mathbf{x})$; $\mathbf{w} \leftarrow \mathcal{E}(\mathbf{y})$
  $\mathbf{z}_t \leftarrow \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$
  $\mathbf{f} \leftarrow$ aggregated UNet features at $(\mathbf{z}_t, t, \sigma_y)$; $\mu_\theta \leftarrow$ mean head$(\mathbf{f})$
  $\ell_\theta^{(1)} \leftarrow \frac{1}{2}\|\mathbf{w} - \mu_\theta\|_2^2$ ▷ Equation (6) with $\Sigma_\theta = \mathbf{I}$
  Backpropagate $\ell_\theta^{(1)}$ through μ_θ and the backbone; clip global grad norm to $g_{\max}$; optimizer step.
end for

Stage 2: freeze μ_θ and the backbone; train only h_ϕ.

for $s = 1, \ldots, S_2$ **do**
  Sample $\mathbf{x}, \sigma_y, t$; $\mathbf{y} \leftarrow \mathcal{A}(\mathbf{x}) + \sigma_y \boldsymbol{\epsilon}_y$; $\mathbf{z}_0 \leftarrow \mathcal{E}(\mathbf{x})$; $\mathbf{w} \leftarrow \mathcal{E}(\mathbf{y})$
  $\mathbf{z}_t \leftarrow \sqrt{\bar{\alpha}_t}\,\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}$
  With `no_grad`: compute aggregated features $\mathbf{f}$ and $\mu_\theta \leftarrow$ mean head$(\mathbf{f})$
  $\Sigma_\phi \leftarrow h_\phi(\mathbf{f})$ ▷ diagonal in spatial or DCT basis
  $\mathbf{r} \leftarrow \mathbf{w} - \mu_\theta$ ▷ stationary target since μ_θ frozen
  if DCT variant **then**
    $\hat{\mathbf{r}} \leftarrow \mathrm{DCT}(\mathbf{r})$ ▷ transform residual to DCT domain
    $\ell_\phi^{(2)} \leftarrow \frac{1}{2}\hat{\mathbf{r}}^\top \Sigma_\phi^{-1}\hat{\mathbf{r}} + \frac{1}{2}\log|\Sigma_\phi|$
  else
    $\ell_\phi^{(2)} \leftarrow \frac{1}{2}\mathbf{r}^\top \Sigma_\phi^{-1}\mathbf{r} + \frac{1}{2}\log|\Sigma_\phi|$
  end if ▷ Equation (6)
  $\mathcal{L} \leftarrow \ell_\phi^{(2)} + \lambda_{\text{var}} \|\log \Sigma_\phi\|_2^2$
  Backpropagate $\mathcal{L}$ through h_ϕ only; clip $\|\nabla_\phi \mathcal{L}\|_2$ to $g_{\max}$; optimizer step on ϕ.
end for

Algorithm 2 ANaLOG inference: covariance-aware native-latent guidance

Require: Encoded measurement $\mathbf{w} = \mathcal{E}(\mathbf{y})$; noise level σ_y; trained mean head μ_θ and variance head h_ϕ; pretrained LDM denoiser; sampler with schedule $\{\bar{\alpha}_t\}_{t=1}^T$; step size η; per-coordinate gradient clip c.

Sample $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
for $t = T, T-1, \ldots, 1$ **do**
  Run one unconditional sampler step from $\mathbf{z}_t$ to obtain $\mathbf{z}'_{t-1}$ and the Tweedie estimate $\mathbf{z}_{0|t}$.
  Compute features $\mathbf{f}$ at $(\mathbf{z}_t, t, \sigma_y)$ via the UNet; obtain μ_θ from the mean head and $\Sigma_\theta = \mathrm{diag}(v)$ from $h_\phi(\mathbf{f})$.
  $v \leftarrow \text{stop-grad}(v)$ ▷ detach covariance for guidance
  $\mathbf{r} \leftarrow \mathbf{w} - \mu_\theta$
  if DCT variant **then**
    $\hat{\mathbf{r}} \leftarrow \mathrm{DCT}(\mathbf{r})$
    $\mathbf{u} \leftarrow \mathrm{DCT}^{-1}\big(\hat{\mathbf{r}} \,/\, v\big)$ ▷ per-frequency precision weighting
    $\mathcal{J}_t \leftarrow \frac{1}{\text{stop-grad}(\|\mathbf{r}\|_2)}\langle \mathbf{u}, \mathbf{r}\rangle$ ▷ detach denominator
  else
    $\mathcal{J}_t \leftarrow \frac{1}{\text{stop-grad}(\|\mathbf{r}\|_2)}\big\|\, \mathbf{r} \,/\, \sqrt{v}\,\big\|_2^2$ ▷ normalize by $\|\mathbf{r}\|_2$
  end if
  $\mathbf{g}_t \leftarrow \nabla_{\mathbf{z}_t} \mathcal{J}_t$ ▷ gradient flows through μ_θ only
  $\mathbf{z}_{t-1} \leftarrow \mathbf{z}'_{t-1} - \eta \cdot \mathrm{clip}(\mathbf{g}_t, -c, c)$
end for
return $\mathcal{D}(\mathbf{z}_0)$.

F Extended Experiments

F.1 Extended Qualitative Results

The following figures demonstrate extended qualitative results of our method on the FFHQ dataset. We observe that ANaLOG consistently recovers crucial high-frequency detail that SILO misses, highlighting the importance of modeling uncertainty in guidance. STSL on average also retains some of these high-frequency details at the expense of significantly increased time. We observe that STSL often introduces certain artifacts, which may be due to the Jacobian of the decoder, as pointed out in Raphaeli et al. (2025). Note that the center box inpainting task we have on FFHQ is particularly challenging as the entire face is often occluded. This stress tests the need for powerful pretrained priors coupled with accurate guidance algorithms. We observe that baselines often do not preserve simple facial structure such has having two eyes, a nose, and a mouth. ANaLOG usually preserves this, but sometimes struggles with smooth boundaries between the masked and unmasked region. On the JPEG decompression task, we clearly see the benefit of modeling covariance in the DCT domain, which gives high-quality reconstructions and drastically improves upon the spatial variant of ANaLOG.

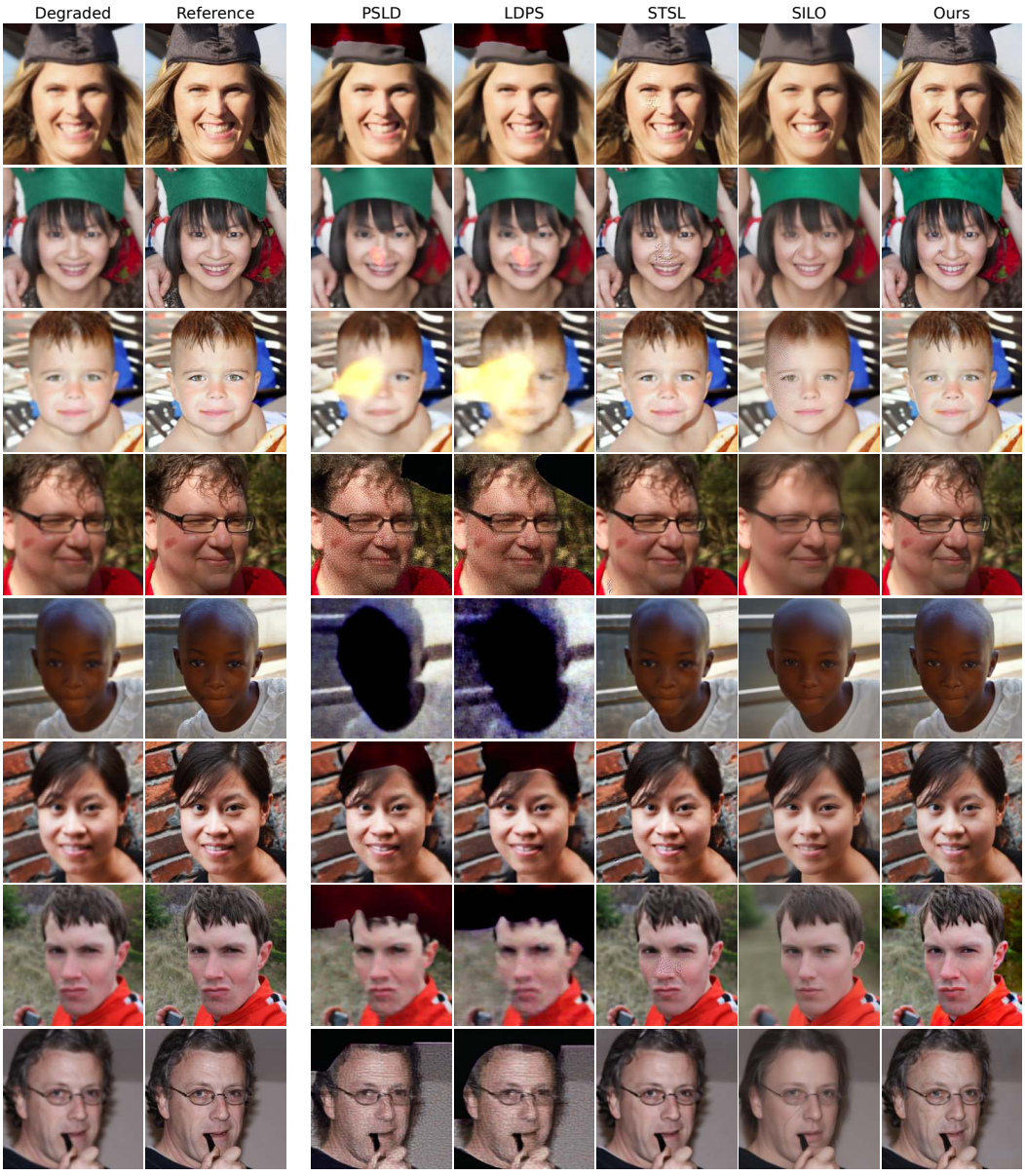


Figure 9: Gaussian Blur task on FFHQ dataset.

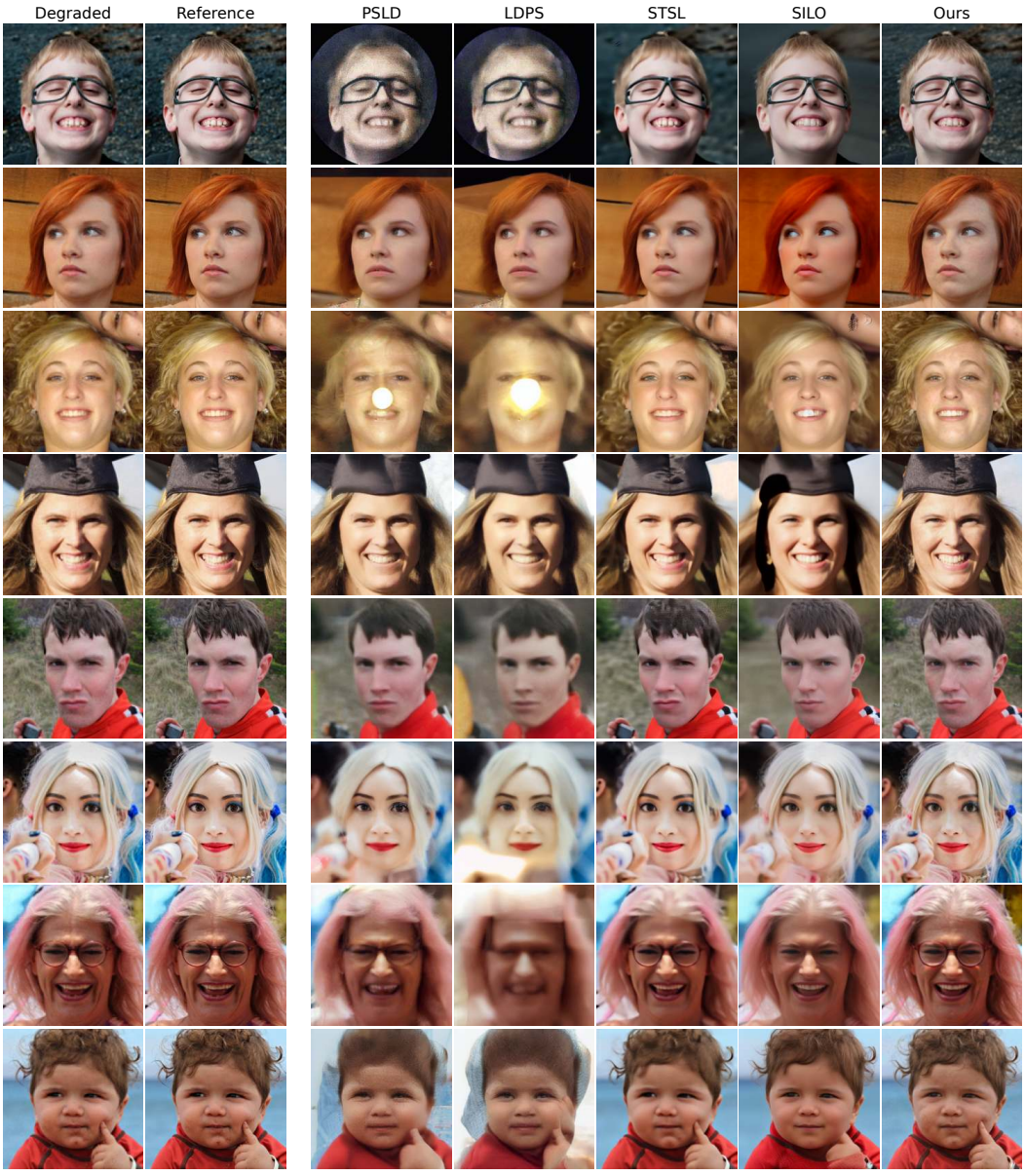


Figure 10: 4x Super-Resolution task on FFHQ dataset.

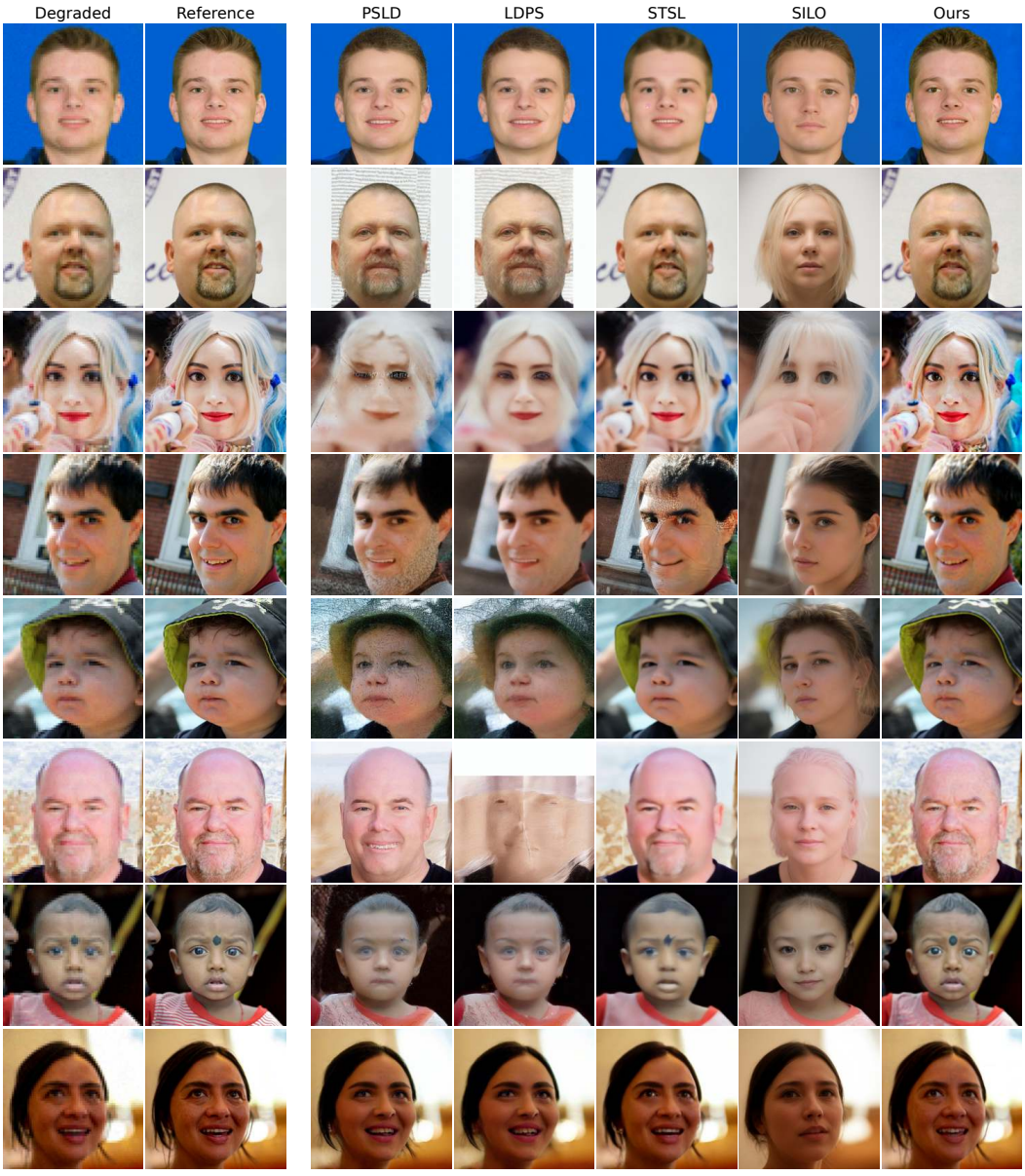


Figure 11: 8x Super-Resolution task on FFHQ dataset.

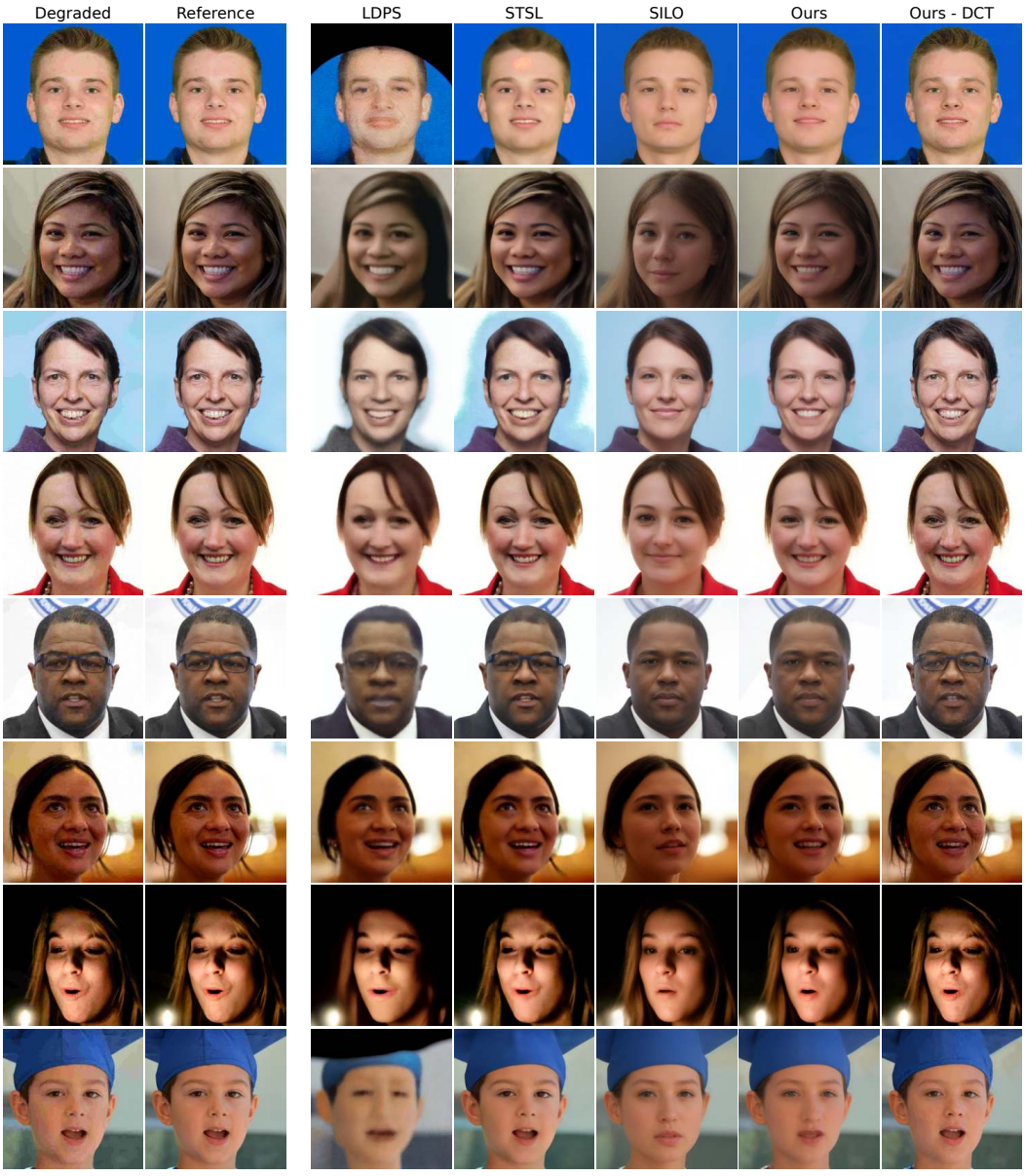


Figure 12: JPEG Decompression task on FFHQ dataset.

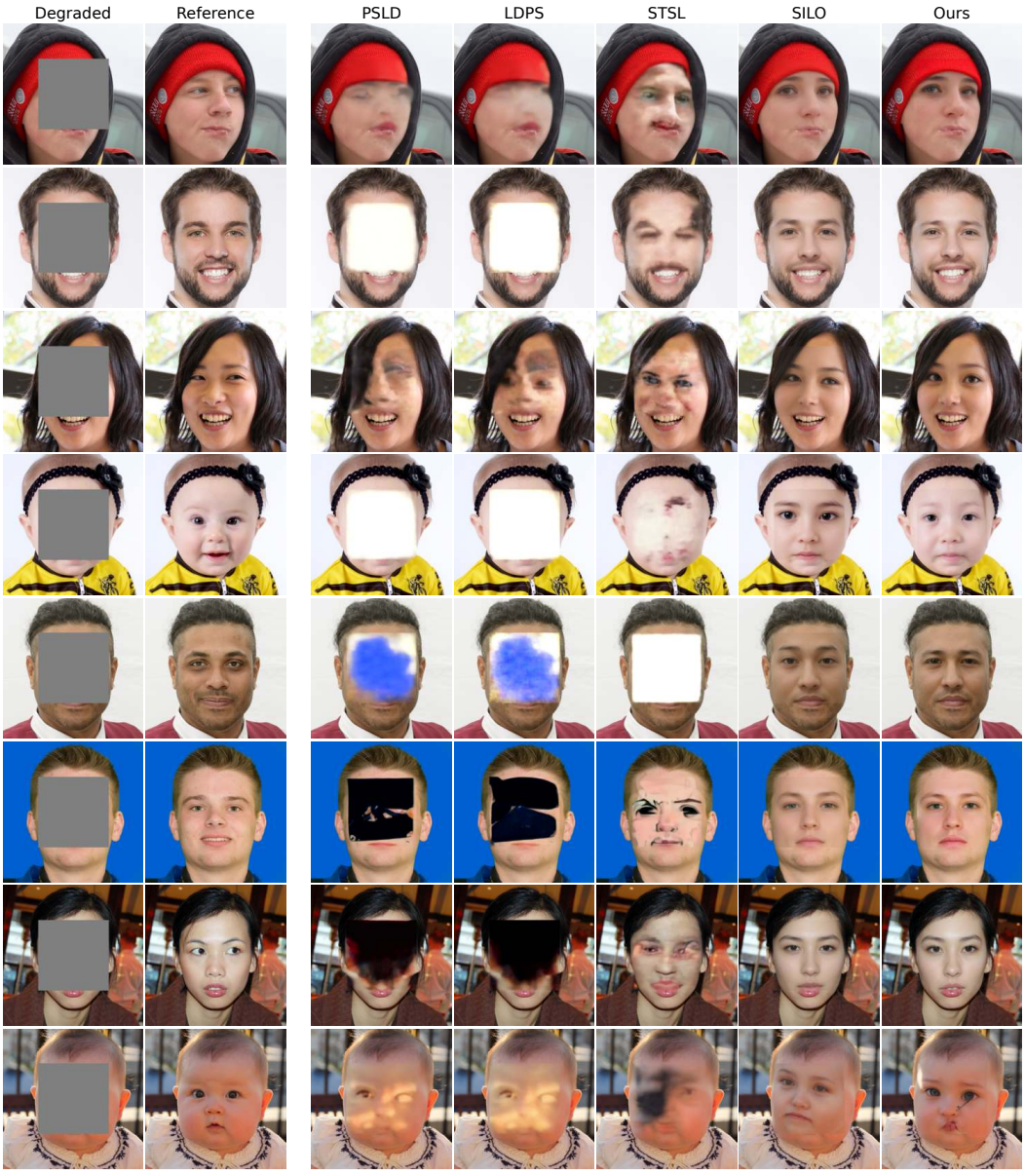


Figure 13: Box Inpainting task on FFHQ dataset.

F.2 Extending to Richer Covariance Parameterizations

Our probabilistic framework in principle permits arbitrary covariance models. In the main text, spatial- and DCT-diagonal covariances were chosen to show even simple covariance models can substantially improve upon baselines. Nonetheless, to test a richer parameterization, in this section, we learn a dense 4×4 covariance across latent channels at each DCT frequency (u, v). We parameterize each covariance block through its Cholesky factor to guarantee positive definiteness.

As shown in Table 5, the block-DCT covariance can further improve over the DCT-diagonal model. These results demonstrate that explicitly modeling cross-channel correlations provides further gains without sacrificing the efficiency advantage of the native-latent formulation. We leave still richer covariance parameterizations to future work.

Table 5: **Richer covariance parameterizations improve restoration quality.**

Task	Method	FID ↓	LPIPS ↓	PSNR ↑
Gaussian deblurring	SILO	47.3	0.367	25.03
	ANaLOG	29.8	0.326	**27.34**
	ANaLOG-DCT	30.9	0.338	26.36
	ANaLOG-DCT-block	**22.0**	**0.292**	26.48
JPEG decompression	SILO	68.3	0.471	20.87
	ANaLOG	46.3	0.393	23.91
	ANaLOG-DCT	27.3	0.326	26.26
	ANaLOG-DCT-block	**24.1**	**0.303**	**27.14**

F.3 Probing Calibration of Uncertainty Head

This section examines whether the learned diagonal covariance Σ_θ obtained by minimizing the negative log-likelihood in Eq. 6 produces well-calibrated uncertainty estimates on held-out data, both on average and at the level of individual latent coordinates.

Testing Setup. For each held-out measurement $\mathbf{y}$ we compute the standardized residual

$$z_i \;=\; \frac{\mathcal{E}(\mathbf{y})_i \;-\; \mu_\theta(\mathbf{z}_t,\, t,\, \sigma_y)_i}{\sqrt{\Sigma_\theta(\mathbf{z}_t,\, t,\, \sigma_y)_{ii}}}, \tag{F.1}$$

indexed by latent coordinate $i \in \{1, \ldots, 4{\times}64{\times}64\}$. Under a calibrated diagonal Gaussian likelihood, $z_i \sim \mathcal{N}(0, 1)$ at every coordinate, so $\mathbb{E}[z_i^2] = 1$ and the empirical variance of z_i across an i.i.d. batch of measurements is also 1. We therefore report two complementary quantities. First, the *aggregate* $\mathbb{E}[z^2]$ pooled over all coordinates—a necessary condition for calibration. Second, the *per-coordinate* empirical variance $\mathrm{Var}_b[z_i]$ taken across the held-out batch and studied as a distribution over i. The per-coordinate quantity probes calibration at the level of individual latent coordinates; uniform calibration requires this distribution to be concentrated at 1 for every i.

For each task we evaluate the trained ANaLOG checkpoint on $N = 256$ held-out FFHQ images. For every test image we apply the task-specific operator, add measurement noise with σ_y drawn from $\mathcal{U}(0, 0.1)$, encode both the clean image and the noisy measurement through the latent autoencoder, and condition the covariance head on $(\mathbf{z}_t,\, t,\, \sigma_y)$. We then compute z from Eq. F.1 on the resulting tensor of shape $(N,\, 4,\, 64,\, 64)$. With $N = 256$ samples, the sampling-noise floor on a per-coordinate variance estimate around the true value of 1 is $\sqrt{2/(N-1)} \approx 0.089$, which sets an irreducible lower bound on the cross-coordinate spread of $\mathrm{Var}_b[z_i]$ that any calibration test at this N can resolve.

Calibration holds on all five tasks. Table 6 summarises the results. The aggregate $\mathbb{E}[z^2]$ lands within a few percent of 1 on every task. The per-coordinate empirical variance $\mathrm{Var}_b[z_i]$ is also tightly distributed around 1: the cross-coordinate *mean* ranges from 0.94 to 1.03, the *median* from 0.95 to 1.01, and on every task at least 99.5% of the 16,384 latent coordinates fall within a factor of two of perfect calibration. Under the stricter window $\mathrm{Var}_b[z_i] \in [0.8,\, 1.25]$ (within $\pm 25\%$ of the calibrated value), 87%–97% of coordinates qualify. Because the sampling-noise floor at $N = 256$ is already ≈ 0.09, the observed cross-coordinate spread of 0.10–0.20 leaves only modest room for systematic miscalibration.

Figure 14(a) overlays the distribution of $\mathrm{Var}_b[z_i]$ over the 16,384 latent coordinates for all five tasks. Every distribution is sharply peaked at the calibrated value $\mathrm{Var}_b[z_i] = 1$ with no heavy tails on either

Table 6: Per-coordinate calibration of the learned diagonal covariance Σ_θ on FFHQ ($N = 256$ held-out images per task). $\mathbb{E}[z^2]$, $\overline{z}$ and $\mathrm{std}(z)$ are pooled over all latent coordinates. The remaining columns characterise the distribution of $\mathrm{Var}_b[z_i]$ over the $4 \times 64 \times 64 = 16{,}384$ latent coordinates, and the per-coordinate residual bias $|\overline{z_i}|$ averaged over coordinates. The sampling-noise floor on $\mathrm{Var}_b[z_i]$ at $N = 256$ is ≈ 0.089.

Task	$\mathbb{E}[z^2]$	$\overline{z}$	$\mathrm{std}(z)$	$\mathrm{Var}_b[z_i]$ across coords			$\Pr[\,\mathrm{Var}_b[z_i] \in [0.8, 1.25]\,]$	$\lvert\overline{z_i}\rvert$
				mean	median	std		avg over coords
`gb`	1.033	+0.024	1.016	1.030	1.010	0.199	93.7%	0.068
`sr4`	1.069	−0.139	1.024	0.943	0.947	0.125	87.4%	0.326
`sr8`	1.015	−0.074	1.005	0.958	0.964	0.109	93.1%	0.213
`ip`	0.993	+0.020	0.996	0.992	0.981	0.137	93.6%	0.058
`jpeg`	0.999	−0.030	0.999	0.997	0.992	0.102	97.0%	0.059

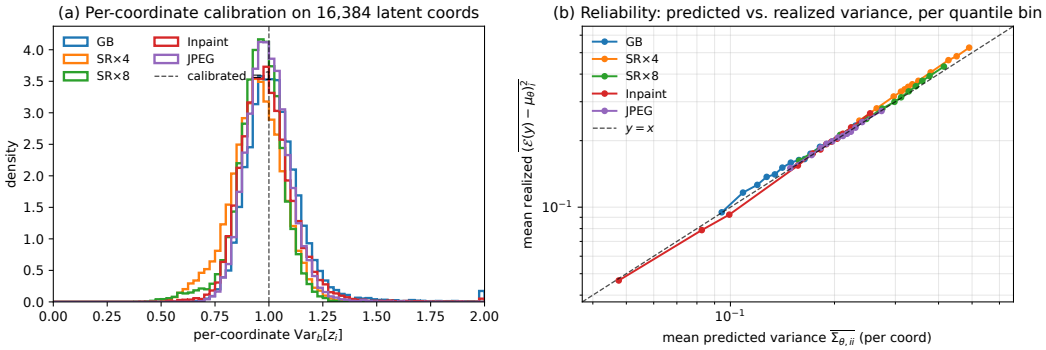


Figure 14: Calibration of the learned diagonal covariance Σ_θ on FFHQ, all five tasks overlaid (`gb`, `sr4`, `sr8`, `ip`, `jpeg`; $N = 256$ held-out images each). **(a)** Distribution of the per-coordinate empirical variance $\mathrm{Var}_b[z_i]$ over the 16,384 latent coordinates; the dashed line marks the calibrated value 1. All five tasks are sharply peaked at 1. **(b)** Reliability diagram on log-log axes: coordinates binned into 12 quantiles by mean predicted variance; the dashed line is $y = x$. All five tasks track $y = x$ across roughly two orders of magnitude of predicted variance, indicating that the variance head is calibrated not only on average but conditional on its own predictions.

side; the visible width is dominated by the $N = 256$ sampling-noise floor. Figure 14(b) is a reliability diagram: latent coordinates are binned by predicted variance quantile, and the mean predicted variance in each bin is plotted against the mean realised squared residual on log-log axes. All five tasks track the perfect-calibration line $y = x$ across roughly two orders of magnitude of predicted variance, indicating that the variance head is calibrated not only on average but also *conditionally on its own predictions*: the coordinates the model says it is uncertain about are exactly the coordinates whose realised residuals are large, and vice versa.

The cross-coordinate concentration of $\mathrm{Var}_b[z_i]$ around 1 together with the alignment of the reliability diagram with $y = x$ provide empirical evidence that the learned diagonal covariance Σ_θ is calibrated on held-out data, supporting the use of Σ_θ^{-1} as a precision matrix in the guidance term of Eq. 7.

F.4 Further Ablation Studies

Guidance scale sensitivity. On a held-out validation set of 100 images, Figure 15 shows the sensitivity of `ANaLOG` to the guidance scale hyperparameter on FFHQ. We see that there exists a broad range of scales for which all metrics remain high. Our results were optimized for perceptual quality (LPIPS).

Stop-gradient approximation. To evaluate the effect of stopping gradients through the learned covariance on reconstruction quality, we compare our implementation against guidance using the full gradient of the learned likelihood on 100 FFHQ images on the Gaussian deblurring task. Results are shown in Table 7.

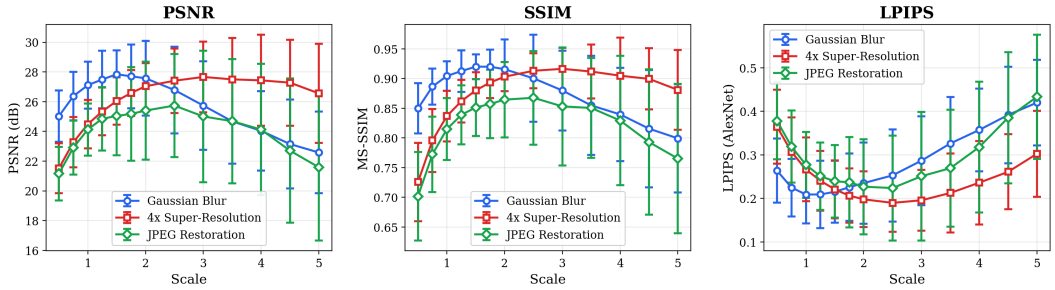


Figure 15: Sensitivity to the guidance scale hyperparameter on FFHQ dataset.

Table 7: **Effect of stopping gradients through the learned covariance.**

Guidance	FID ↓	LPIPS ↓	PSNR ↑
Stop-gradient (ours)	61.7	**0.317**	**27.47**
Full covariance gradient	**60.9**	0.325	27.44

Stop-gradient preserves essentially the same reconstruction quality while providing substantially greater optimization stability. With the full covariance gradient, we often observe large gradient-norm spikes that destabilize guidance, resulting in a much narrower range of usable step sizes. In contrast, stop-gradient remains stable across a broad range of guidance scales, as shown in Figure 15.

G Limitations and Broader Impact

We discuss below some limitations of our work.

The first set of limitations are of the broader framework of native-latent guidance. `ANaLOG` inherits the limitation of SILO in that sufficiency can only hold when the measurement itself lies in pixel-space; for example, for tasks like phase retrieval, encoding the measurement via a pretrained image-space LDM will not work and clearly in this case, $\mathcal{E}(\mathbf{y})$ is not a sufficient statistic for the latent posterior. We hope that our characterization of sufficiency will inspire future work on learning the proper measurement transformation. Secondly, native-latent guidance trades off inference time with task-specific training. Our method does not directly address these limitations, but considers an orthogonal improvement: formulating a principled probabilistic framework for native-latent guidance and uncovering the role of covariance structure.

In practice, to maintain efficiency and for computational tractability, we restrict the learned covariance to be diagonal in spatial or DCT domain. We hope that future work may consider richer covariance models, such as low-rank approximations.

On the theoretical front, we do not claim that our theoretical setting is perfectly indicative of real-world data. Rather, we aim to use the theoretical results as a motivating example for studying covariance structure in native-latent guidance. We point out an interesting connection between our autoencoder latent structure as a low-pass filter and recent work that studies the frequency properties of nonlinear LDMs (Skorokhodov et al., 2025). It is an interesting direction to theoretically formalize the frequency structure of LDM latent spaces in nonlinear architectures.