Variational Flow Maps: Make Some Noise for One-Step Conditional Generation

Anonymous Authors[1]

Abstract

Flow maps enable high-quality image generation in a single forward pass. However, unlike iterative diffusion models, their lack of an explicit sampling trajectory impedes incorporating external constraints for conditional generation and solving inverse problems. We put forth *Variational Flow Maps*, a framework for conditional sampling that shifts the perspective of conditioning from "guiding a sampling path", to that of "learning the proper initial noise". Specifically, given an observation, we seek to learn a *noise adapter model* that outputs a noise distribution, so that after mapping to the data space via flow map, the samples respect the observation and data prior. To this end, we develop a principled variational objective that jointly trains the noise adapter and the flow map, improving noise-data alignment, such that sampling from complex data posterior is achieved with a simple adapter. Experiments on various inverse problems show that VFMs produce well-calibrated conditional samples in a single (or few) steps. For ImageNet, VFM attains competitive fidelity while accelerating the sampling by orders of magnitude compared to alternative iterative diffusion/flow models.

1. Introduction

Diffusion and flow-based methods have emerged as the dominant paradigm for high-fidelity generative modeling, achieving state-of-the-art results across images, audio, and video (Ho et al., 2020; Song & Ermon, 2020; Sohl-Dickstein et al., 2015; Karras et al., 2022; Lipman et al., 2022; Liu et al., 2022). These methods can be understood from the unified perspective of interpolating between two distributions; a simple noise distribution and a complex data distribution, and learning dynamics based on ordinary or stochastic differential equations (ODE/SDEs) that transport one to the other (Albergo et al., 2023). However, these share a fundamental limitation that generating a single sample requires dozens to hundreds of sequential function evaluations, creating high computational cost for real-time applications.

To address this issue, recent research have sought to dramatically reduce this sampling cost. Consistency models (Song et al., 2023b), for example, learn to map any point on the flow trajectory directly to the corresponding clean data, enabling few-step generation. Despite their promise, consistency models often suffer from training instabilities and frequently require re-noising steps for multi-step sampling to correct the drift trajectory, complicating the inference process (Geng et al., 2024). Flow maps (Boffi et al., 2024; 2025) offer an alternative framework that seeks to learn ODE flows directly, by training on the mathematical structure of such flows. For example, the state-of-the-art Mean Flow model (Geng et al., 2025) presents a particular parameterisation of flow maps based on *average velocities*, and trained on the so-called Eulerian condition satisfied by ODE flows.

While flow maps excel at unconditional few-steps generation, many applications require *conditional* generation to produce samples that satisfy external constraints. Inverse problems provide a canonical example: given a degraded observation $y = A(x) + \varepsilon$ (e.g., a blurred, masked, or noisy image), we seek to recover plausible original signals x consistent with both the observation and our learned prior $p(x)$. Iterative generative models naturally accommodate such conditioning through *guidance* mechanisms (Chung et al., 2022; 2024; Kawar et al., 2022; Song et al., 2023a), where the trajectory is iteratively nudged toward the conditional target. Flow maps, despite their efficiency, lack this iterative refinement mechanism: once the noise vector z is chosen, the generated sample $z \mapsto x$ is fixed; there is no intermediate state to guide, nor a trajectory to steer, hence there is no opportunity to incorporate measurement information during generation. This "guidance gap" has limited flow maps to unconditional settings, leaving their potential for conditional generation largely unexplored.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

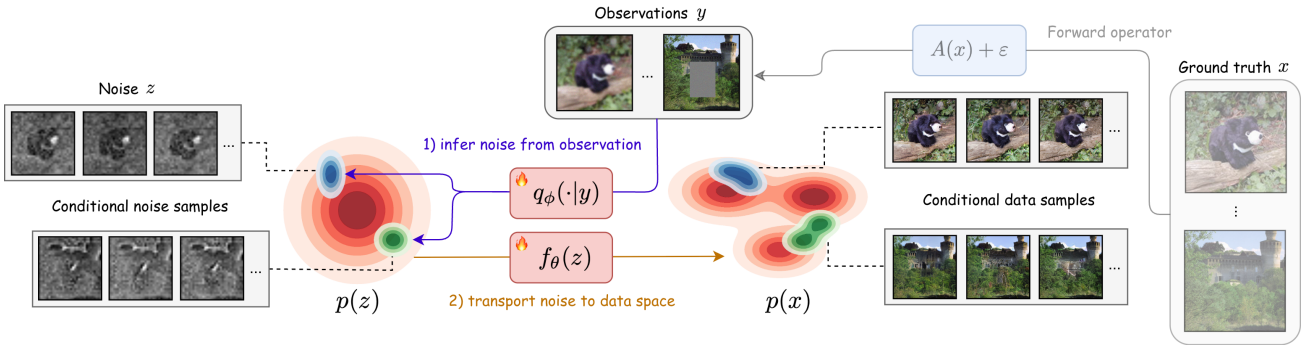


Figure 1. One-step conditional generation with Variational Flow Maps (VFM). Given an observation y, VFM learns a noise adapter network $q_\phi(z|y)$, which approximates the noise space posterior $p(z|y)$ via amortized variational inference. Conditional noise samples $z \sim q_\phi(z|y)$ are then mapped to data space in a single step via a learned flow map $x = f_\theta(z)$, producing conditional samples that approximate $p(x|y)$. In VFM, the networks q_ϕ and f_θ are trained jointly by extending the variational autoencoder framework to learn the correspondence between the triple (x, y, z). By jointly training, f_θ learns to compensate for the simple Gaussian assumption on q_ϕ.

To fill this "guidance gap", we introduce *Variational Flow Maps* (VFMs), a framework for conditional sampling that is compatible with one/few-step generation using flow maps. Our approach is based on the following perspective: rather than steer the generation process itself, we can *find the noise z to generate from*, as each z deterministically maps to a data $x = f_\theta(z)$ (see Figure 1). Specifically, given an observation y, we seek to produce a distribution of z's, such that each $x = f_\theta(z)$ is a candidate data that produced y. Formulating this as a Bayesian inverse problem, we can derive a principled variational training objective to jointly learn the flow map f_θ and a *noise adapter model* q_ϕ that produces appropriate noise z from observations y.

We note the resemblance to variational autoencoders (VAEs) (Kingma & Welling, 2013), where q_ϕ plays the role of an encoder that takes y to a latent z, and f_θ acts as a decoder from z to data x. Our key innovation is in learning the alignment of all three variables (x, y, z) *simultaneously*, allowing updates to q_ϕ to reshape the noise-to-data coupling by f_θ and vice versa. Notably, we observe that joint training can compensate for limited adapter expressivity by learning a noise-to-data coupling that makes the conditional posterior easier to represent in latent space.

Altogether, our contributions can be summarized as follows:

- We introduce Variational Flow Maps (VFMs), a new paradigm enabling **one and few-step conditional generation** with flow maps by learning an observation-dependent noise sampler.
- We derive a **principled variational objective** for joint adapter/flow map training, linking the mean flow loss to likelihood bounds.
- We demonstrate empirically and theoretically that joint training yields **better noise-data coupling** to fit complex posteriors in data space using **simple** variational posteriors in noise space.

2. Background

We review essential backgrounds on flow maps for few-step generation, the Bayesian formulation of inverse problems, and variational inference with amortization.

2.1. Flow-based Generative Models and Flow Maps

Flow-based generative models learn to transport samples from a prior distribution $p_1(z) = \mathcal{N}(0, I)$ to the data distribution $p_0(x) = p_{\text{data}}(x)$ via an ODE:

$$\frac{dx_t}{dt} = v_t(x_t), \quad t \in [0, 1], \tag{1}$$

where v_t is a time-dependent velocity field. Flow matching (Lipman et al., 2022; Liu et al., 2022; Albergo et al., 2023) provides a training objective to learn v_t: given $x_0 \sim p_{\text{data}}$ and $x_1 \sim \mathcal{N}(0, I)$, we construct a linear interpolant $x_t = (1-t)x_0 + tx_1$ with conditional velocity $v_t = x_1 - x_0$. Then, $v_\theta(x_t, t) \approx v_t(x_t)$ is trained via:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{x_0, x_1, t}\left[\|v_\theta(x_t, t) - (x_1 - x_0)\|^2\right]. \tag{2}$$

At inference time, samples are generated by integrating the ODE backwards from $t = 1$ to $t = 0$, typically requiring 50–250 function evaluations.

To accelerate sample generation, *flow maps* (Boffi et al., 2024; 2025) directly learn the solution operator of the ODE, instead of the instantaneous velocity v_t. Denoting by $\phi_{t,s} : x_t \mapsto x_s$ the backward flow of the ODE, the *two-time flow map* $f_\theta(x_t, s, t)$ learns to approximate $\phi_{t,s}(x_t)$ for any $0 \le s < t \le 1$. This enables generation with an arbitrary number of steps chosen post-training, e.g. a single evaluation $f_\theta(x_1, 0, 1)$ produces a one-step sample, while intermediate evaluations can be composed for multi-step refinement.

One such approach to learn flow maps is *mean flows* (Geng et al., 2025), which introduce the *average velocity* as an

alternative characterization:

$$u(x_t, r, t) := \frac{1}{t-r} \int_r^t v_s(\phi_{t,s}(x_t))\, ds. \tag{3}$$

The average velocity satisfies $x_r = x_t - (t-r) \cdot u(x_t, r, t)$, enabling one-step generation via $x_0 = x_1 - u(x_1, 0, 1)$. Thus the corresponding flow map is given by $f_\theta(x_t, r, t) = x_t - (t-r) \cdot u_\theta(x_t, r, t)$. For simplicity, we denote the one-step flow map as $f_\theta(z) := z - u_\theta(z, 0, 1)$, mapping noise $z \sim \mathcal{N}(0, I)$ directly to data $x = f_\theta(z)$.

2.2. Inverse Problems

Inverse problem seeks to recover an unknown signal $x \in \mathbb{R}^d$ from noisy observations, given by

$$y = A(x) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{4}$$

where $A : \mathbb{R}^d \to \mathbb{R}^m$ is a known forward operator and $\sigma > 0$ is the noise level. Given a prior $p(x)$ over signals, the Bayesian formulation seeks the posterior distribution:

$$p(x|y) \propto \exp\left(-\frac{\|y - A(x)\|^2}{2\sigma^2}\right) p(x). \tag{5}$$

When $p(x)$ is defined implicitly by a generative model, guidance-based methods (Chung et al., 2024; Song et al., 2023a) approximate posterior sampling by incorporating likelihood gradients $\nabla_x \log p(y|x)$ at each denoising step. While effective, these methods inherently require iterative refinement and cannot be applied to one-step flow maps.

2.3. Variational Inference and Data Amortization

Variational inference seeks to approximate an intractable posterior $p(z|x)$ with a tractable disribution $q(z|x)$ by minimizing the Kullback-Leibler (KL) divergence:

$$\mathrm{KL}(q(z|x) \| p(z|x)) := \mathbb{E}_{q_\phi}[\log q(z|x) - \log p(z|x)]. \tag{6}$$

Extending this, amortized inference uses a neural network to directly predict the variational distribution from the conditioning variable x, rather than optimizing separately for each instance. For example, if we choose the variational family to be Gaussians with diagonal covariance, then amortized inference learns a neural network $x \mapsto (\mu_\phi(x), \sigma_\phi(x))$ with parameter ϕ, such that $q_\phi(z|x) = \mathcal{N}(z|\mu_\phi(x), \mathtt{diag}(\sigma_\phi^2(x)))$ is close to $p(z|x)$ under the KL divergence.

A prototypical example is the *Variational Autoencoder (VAE)* (Kingma & Welling, 2013), which learns both an encoder $q_\phi(z|x)$ and a decoder $p_\theta(x|z)$ by optimizing the VAE objective $\mathcal{L}_{\text{VAE}}(\theta, \phi) = \mathbb{E}_{p(x)}[\ell(\theta, \phi; x)]$, where

$$\ell(\theta, \phi; x) := -\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] + \mathrm{KL}(q_\phi(z|x) \| p(z)), \tag{7}$$

is the negative evidence lower bound (ELBO), yielding $q_\phi(z|x) \approx p_\theta(z|x) \propto p_\theta(x|z)p(z)$ for any $x \sim p(x)$. Probabilistically, the VAE objective can be derived from the KL divergence between two representations of the *joint distribution* of (x, z), i.e., $\mathrm{KL}(q_\phi(z, x) || p_\theta(z, x))$, where $q_\phi(z, x) = q_\phi(z|x)p(x)$ and $p_\theta(z, x) = p_\theta(x|z)p(z)$. This perspective will be useful in the derivation of our loss later.

3. Variational Flow Maps (VFMs)

Our proposed method for one-step conditional generation, which we term *Variational Flow Maps (VFMs)*, is based on reformulating the inverse problem (5) in noise space. To motivate our methodology, we begin with a simple "strawman" approach that is intuitively sound but ultimately insufficient for our task: Let $x = f_\theta(z)$ denote a pretrained flow map. Then the posterior over latent noise variables induced by the inverse problem can be written as

$$p(z|y) \propto \exp\left(-\frac{\|y - A(f_\theta(z))\|^2}{2\sigma^2}\right) p(z). \tag{8}$$

Although the posterior (8) is intractable, we can approximate it in the same spirit as VAEs. In particular, introducing a variational posterior $q_\phi(z|y) \approx p(z|y)$, we minimize the objective $\mathcal{L}_{\text{VAE}}(\theta, \phi) = \mathbb{E}_{p(y)}[\ell(\theta, \phi; y)]$, where,

$$\ell(\theta, \phi; y) := -\mathbb{E}_{q_\phi(z|y)}[\log p_\theta(y|z)] + \mathrm{KL}(q_\phi(z|y) \| p(z)), \tag{9}$$

and $p_\theta(y|z) := \mathcal{N}(y|A(f_\theta(z)), \sigma^2 I)$, the likelihood in noise space. A key advantage of working in the noise space rather than the original data space is that the noise prior $p(z)$ is simple and tractable (commonly $\mathcal{N}(0, I)$, which we assume hereafter). Thus, imposing a conjugate variational posterior, such as $q_\phi(z|y) = \mathcal{N}(z|\mu_\phi(y), \mathtt{diag}(\sigma_\phi^2(y)))$, makes the computation of the KL term in (9) tractable.

However, the objective (9) has two major limitations in our setting. First, it does not impose structural properties of flow maps, such as the semi-group property (Boffi et al., 2025), known to be crucial for learning said maps. Second, when the flow map f_θ is pretrained and held fixed, a Gaussian variational posterior $q_\phi(z|y)$ may not be expressive enough to approximate the true posterior $p(z|y)$ accurately.

Motivated by this observation, we pursue training the parameters θ and ϕ *jointly*. By adapting the map $f_\theta : z \mapsto x$ alongside learning the variational posterior q_ϕ, we can compensate for the limited expressibility of $q_\phi(z|y)$ by reshaping the correspondence between noise and data. In the next section, we formalize this idea by deriving a modified objective that enables joint training of (θ, ϕ) while explicitly incorporating additional structural constraints to the flow.

3.1. Joint Training of the Flow Map and Noise Adapter

We now propose a joint training strategy that simultaneously aligns the data variable x, the observation y, and the latent noise variable z. Following the probabilistic perspective underlying VAEs (see Section 2.3), we achieve this by matching the following two factorizations of $p(x, y, z)$:

$$q_\phi(z|y)p(y|x)p(x) \approx p_\theta(x, y|z)p(z). \tag{10}$$

For simplicity, we assume a Gaussian decoder of the form

$$p_\theta(x, y|z) = \mathcal{N}(x|f_\theta(z), \tau^2 I)\, \mathcal{N}(y|A(f_\theta(z)), \sigma^2 I), \tag{11}$$

where we introduce a new hyperparameter $\tau > 0$ that relaxes the correspondence between x and z. Taking the KL divergence between the two representations in (10) yields

$$\begin{aligned} &\mathrm{KL}(q_\phi(z|y)p(y|x)p(x)\,||\,p_\theta(x, y|z)p(z)) \\ &\le \frac{1}{2\tau^2}\mathcal{L}_{\text{data}}(\theta, \phi) + \frac{1}{2\sigma^2}\mathcal{L}_{\text{obs}}(\theta, \phi) + \mathcal{L}_{\text{KL}}(\phi), \end{aligned} \tag{12}$$

(see Appendix A.1 for details), where

$$\mathcal{L}_{\text{data}}(\theta, \phi) = \mathbb{E}_{q_\phi(z|y)p(y|x)p(x)}\left[\|x - f_\theta(z)\|^2\right], \tag{13}$$

$$\mathcal{L}_{\text{obs}}(\theta, \phi) = \mathbb{E}_{q_\phi(z|y)p(y)}\left[\|y - A(f_\theta(z))\|^2\right], \tag{14}$$

$$\mathcal{L}_{\text{KL}}(\phi) = \mathbb{E}_{p(y)}\left[\mathrm{KL}\left(q_\phi(z|y)\,||\,p(z)\right)\right]. \tag{15}$$

We note that relative to (9), this formulation gives rise to an additional term $\mathcal{L}_{\text{data}}(\theta, \phi)$ that measures closeness of the reconstructed state $f_\theta(z)$ and the ground-truth data x, where noise z is drawn from the noise adapter $q_\phi(z|y)$, with observation y taken from x. This term couples the adapter model and flow map more tightly, encouraging the samples $\{f_\theta(z)\}_{z\sim q_\phi(z|y)}$ to remain consistent with data manifold.

In the following result, we identify a concrete benefit of jointly learning f_θ and q_ϕ to target the true posterior $p(x|y)$, under a simple Gaussian setting. While this does not claim that the distribution of samples $\{f_\theta(z)\}_{z\sim q_\phi(z|y)}$ matches $p(x|y)$ exactly, it shows that joint training can at least match the posterior mean for every observation y. This sharply contrasts with separately training f_θ and q_ϕ, which leads to bias almost surely, even at the level of the posterior mean.

Proposition 3.1. *Assume that $p(z) = \mathcal{N}(z|0, I)$, $p(x) = \mathcal{N}(x|m, C)$ for some $m \in \mathbb{R}^d$ and $C \in \mathbb{R}^{d\times d}$ symmetric positive definite, and $q_\phi(z|y) = \mathcal{N}(z|\mu_\phi(y), \mathtt{diag}(\sigma_\phi^2(y)))$. Then, for any linear observation $y = Ax + \varepsilon$, we have that*

1. *Joint optimization of f_θ and q_ϕ via loss (12) recovers the true posterior mean $\mathbb{E}_{p(x|y)}[x]$ exactly via the procedure $\mathbb{E}_{z\sim q_\phi(z|y)}[f_\theta(z)]$.*
2. *Training f_θ first to match $p(x)$ and then training q_ϕ via loss (12) with θ fixed almost surely fails to match the posterior mean, i.e., $\mathbb{E}_{z\sim q_\phi(z|y)}[f_\theta(z)] \neq \mathbb{E}_{p(x|y)}[x]$.*

Proof. See Proposition A.18 in Appendix A.2. □

Next, we relate the new term $\mathcal{L}_{\text{data}}(\theta, \phi)$ in (12) to the *mean flow loss* (Geng et al., 2025), which imposes structural constraints on the flow map.

Connection to mean flows. We briefly recall the mean flow objective from (Geng et al., 2025). Denoting

$$\begin{aligned} &\mathcal{E}_\theta(x, z, r, t) := (t - r)\left[u_\theta(\psi_t(x, z), r, t) - \dot{\psi}_t(x, z)\right], \\ &\text{where} \quad \psi_t(x, z) := (1 - t)x + tz, 0 \le r \le t \le 1, \end{aligned} \tag{16}$$

is the linear interpolant between data x and noise z, the mean flow loss is given by

$$\begin{aligned} &\mathbb{E}_{x,z,r,t}\left[\|\partial_t \mathcal{E}_\theta(x, z, r, t)\|^2\right] \approx \mathcal{L}_{\text{MF}}(\theta) \\ &:= \mathbb{E}_{x,z,r,t}\left[\|u_\theta(\psi_t(x, z), r, t) - \mathtt{stopgrad}(u_{\text{tgt}})\|^2\right], \end{aligned} \tag{17}$$

where $u_{\text{tgt}} := \dot{\psi}_t(x, z) - (t - r)\frac{d}{dt}u_\theta(\psi_t(x, z), r, t)$ is the effective regression target. Below, we establish a direct link between this objective and the term $\mathcal{L}_{\text{data}}(\theta, \phi)$ in (12).

Proposition 3.2. *Let the noise-to-data map f_θ be defined by $f_\theta(z) := z - u_\theta(z, 0, 1)$. Then we have*

$$\|x - f_\theta(z)\|^2 \le \int_0^1 \|\partial_t \mathcal{E}_\theta(x, z, 0, t)\|^2 dt. \tag{18}$$

Proof. See Appendix A.3. □

This result shows that the mean flow loss in the anchored case $r = 0$ and $t \sim U([0, 1])$ acts as an upper bound proxy to the reconstruction error $\|x - f_\theta(z)\|^2$ in (27). This specialized setting targets direct one-step transport to $r = 0$. Motivated by this connection, we opt to use the general mean flow loss (17), which distributes learning over (r, t) to additionally learn intermediate flow maps $f_\theta(x_t, r, t)$. While this does not ensure optimality for the one-step transport $x = f_\theta(z, 0, 1)$, in practice, it yields strong empirical performance and furthermore provides functionality for multi-step sampling (Section 3.3). Summarizing, we propose to train (θ, ϕ) using the following objective:

$$\mathcal{L}_{\theta,\phi} := \frac{1}{2\tau^2}\mathcal{L}_{\text{MF}}(\theta; \phi) + \frac{1}{2\sigma^2}\mathcal{L}_{\text{obs}}(\theta, \phi) + \mathcal{L}_{\text{KL}}(\phi), \tag{19}$$

where the mean flow term is evaluated using (x, z)-pairs sampled from the joint distribution $\pi_\phi(x, z) := \int q_\phi(z|y)p(y|x)p(x)dy$, in accordance with (27). This dependence induces an implicit coupling between θ and ϕ. To promote stable optimization, we further limit the interaction to this term by replacing θ in the observation loss $\mathcal{L}_{\text{obs}}$ with its exponential moving average (EMA), yielding $\mathcal{L}_{\text{obs}}(\theta^-, \phi)$, where θ^- denotes the EMA of θ.

Remark 3.3. Our framework can also be related to consistency model training by Proposition 6.1 in (Silvestri et al., 2025). In this case, the mean flow loss in (19) is replaced by an appropriate consistency loss.

Algorithm 1 Multi-Step Conditional Sampling with VFM

Input: Observation y, inverse problem class c, time partition $1 = t_0 > \cdots > t_K = 0$, adapter mean and standard deviation μ_ϕ, σ_ϕ, mean flow model u_θ
$\epsilon \sim \mathcal{N}(0, I)$
$z \leftarrow \mu_\phi(y, c) + \sigma_\phi(y, c) \odot \epsilon$
$x \leftarrow z$
for $k = 1$ to K **do**
  $x \leftarrow x + (t_k - t_{k-1}) u_\theta(x, t_k, t_{k-1})$
end for
Output: x

3.2. Amortizing Over Multiple Inverse Problems

In many applications, one is interested not in a single inverse problem defined by a fixed forward operator A, but rather a *family of inverse problems*. To accommodate this setting, we extend our framework by amortizing inference over multiple forward operators $A_1, \ldots, A_C$. This allows for a single model to handle multiple tasks, such as denoising, inpainting, and deblurring.

To achieve this, we consider a class-conditional noise adapter $q_\phi(z|y, c) = \mathcal{N}(z|\mu_\phi(y, c), \texttt{diag}(\sigma^2_\phi(y, c)))$, where $c \in \{1, \ldots, C\}$ is a categorical variable indicating which forward operator A_c was used to generate the observation y. Conditioning the adapter on c enables the model to adapt its posterior approximation to the specific structure of each inverse problem. We may further extend this by amortizing over *inverse problem classes*, where c now defines a collection of inverse problems $\mathcal{A}_c = \{A^\omega_c\}_{\omega \in \Omega}$. For example, these can define a family of random masks or a distribution of blurring kernels.

3.3. Single and Multi-Step Conditional Sampling

Given a trained noise adapter $q_\phi(z|y)$ and flow map $f_\theta(z)$, samples from the data-space posterior $p(x|y)$ can be approximately generated by first sampling $z \sim q_\phi(z|y)$ and then mapping $x = f_\theta(z)$. The validity of this procedure is justified by the following result.

Proposition 3.4. *Let the joint distribuion of (x, y, z) be given by $p(x, y, z) = p_\theta(x, y|z)p(z)$, for $p_\theta(x, y|z)$ in (11). Then, for any fixed observation y, the data-space posterior $p(x|y)$ converges weakly to the pushforward of the noise-space posterior $p(z|y)$ under the map f_θ, as $\tau \to 0$.*

Proof. See Appendix A.4. □

The proposition states that in the limiting case $\tau \to 0$, sampling from $p(x|y)$ is equivalent in distribution to first sampling $z \sim p(z|y)$ (approximated by $q_\phi(z|y)$) and then applying $x = f_\theta(z)$. While sound in theory, we find that when $\tau \ll \sigma$, joint optimization of (θ, ϕ) becomes difficult.

Algorithm 2 Joint training of the adapter and flow map

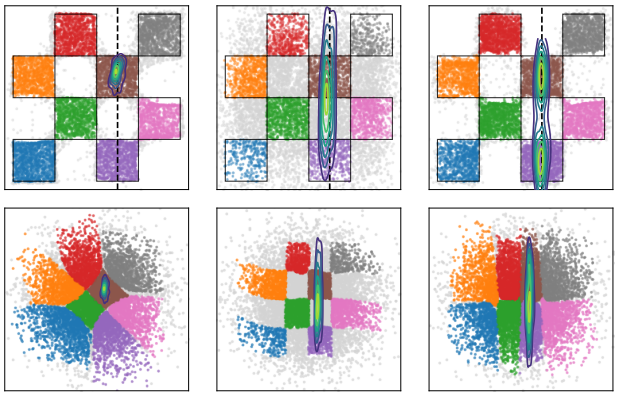
Input: Inverse problem classes $\mathcal{A}_1, \ldots, \mathcal{A}_C$, observation noise standard deviation σ, data misfit tolerance τ, conditional noise proportion α, learning rates η_1, η_2, EMA rate μ, adaptive loss constants γ, p
$\theta^- \leftarrow \texttt{stopgrad}(\theta)$
repeat
  Sample $c \sim p(c)$, $x \sim p(x)$
  Sample forward operator $A^\omega_c \in \mathcal{A}_c$
  $y \leftarrow A^\omega_c x + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$
  $z \leftarrow \mu_\phi(y, c) + \sigma_\phi(y, c) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$
  $\mathcal{L}_{\text{obs}}(\phi) \leftarrow \|y - A^\omega_c(f_{\theta^-}(z, 0, 1))\|^2$
  $\mathcal{L}_{\text{KL}}(\phi) \leftarrow \text{KL}\Big(\mathcal{N}(\mu_\phi(y, c), \sigma^2_\phi(y, c)I) \,||\, \mathcal{N}(0, I)\Big)$
  Sample $w \sim U([0, 1])$ and $(r, t) \sim p(r, t)$
  if $w > \alpha$ **then**
    $z \sim \mathcal{N}(0, I)$
  end if
  $\mathcal{L}_{\text{MF}}(\theta; \phi) \leftarrow \text{MeanFlowLoss}(x, z, r, t)$
  $\mathcal{L}(\theta, \phi) \leftarrow \frac{1}{2\tau^2}\mathcal{L}_{\text{MF}}(\theta; \phi) + \frac{1}{2\sigma^2}\mathcal{L}_{\text{obs}}(\theta) + \mathcal{L}_{\text{KL}}(\phi)$
  $\mathcal{L}(\theta, \phi) \leftarrow \mathcal{L}(\theta, \phi)/\texttt{stopgrad}(\|\mathcal{L}(\theta, \phi) + \gamma\|^p)$
  $\theta \leftarrow \theta - \eta_1 \nabla_\theta \mathcal{L}(\theta, \phi)$
  $\phi \leftarrow \phi - \eta_2 \nabla_\phi \mathcal{L}(\theta, \phi)$
  $\theta^- \leftarrow \texttt{stopgrad}(\mu\theta^- + (1 - \mu)\theta)$
until convergence

This is likely due to the RHS distribution in (10) concentrating sharply around the submanifold $\{(x, y, z) : x = f_\theta(z)\}$, making it nearly impossible to match using the LHS representation of (10), which remains a full distribution over (x, y, z). In practice, we find that using τ larger than σ yields stable optimization and the best empirical results.

Sample quality can also be improved by considering *multi-step sampling* instead of single-step sampling, as described in Algorithm 1. Empirically, high-quality samples can be obtained with only a small number of steps K, substantially fewer than the number of integration steps required for solving a full generative ODE or SDE.

3.4. Other Training Considerations

Mixing in the unconditional loss: We observe that training solely using the objective (19) can degrade the quality of unconditional samples $x = f_\theta(z)$, with $z \sim \mathcal{N}(0, I)$. This behaviour arises because latent samples drawn from $q_\phi(z|y)$ retain structural details of y, and therefore are not fully representative of pure noise drawn from $\mathcal{N}(0, I)$. Thus, during training, the mean flow loss is never evaluated on pure noise, imparining the model to generate unconditional samples. To address this, we modify the computation of the mean flow loss $\mathcal{L}_{\text{MF}}(\theta; \phi)$, by sampling $(x, z) \sim \pi_\phi(x, z)$ with probability α and with remaining probablility $1 - \alpha$, we sample $z \sim \mathcal{N}(0, I)$ independently of x.



(a) `frozen-`θ *(b)* `unconstrained-`θ *(c)* VFM (ours)

Figure 2. Prior 2D samples and posterior densities in data space (top row) and noise space (bottom row). We observe the x-component (black dashed lines) with $\sigma = 0.1$. The unconditional samples are color-coded by checkerboard cell; light grey for off-manifold samples. VFM successfully captures the bimodal nature of the posterior, while the baselines struggle to do so.

Adaptive loss: Similar to the mean flow training procedure of (Geng et al., 2025), we consider an adaptive loss scaling to stabilize optimization. Specifically, we use the rescaled loss $w \cdot \mathcal{L}_{\theta,\phi}$, where the weight w is given by $w = 1/\texttt{stopgrad}(\|\mathcal{L}_{\theta,\phi} + \gamma\|^p)$ for constants $\gamma, p > 0$.

We summarize the full training procedure in Algorithm 2.

4. Experiments

4.1. Illustration on a 2D Example

In this experiment, we illustrate the effects of jointly training (θ, ϕ) on a toy 2D example, and perform ablations on key design choices in VFM. Specifically, we take $p(x)$ to be a 4×4 checkerboard distribution supported on $[-2, 2] \times [-2, 2]$. For the forward problem, we observe only the first coordinate, i.e. $y = Ax + \varepsilon$ with $A = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 0.1$. We refer the readers to Appendix B.1 for details on the experimental setup.

Baselines and evaluation metrics. We consider two baselines: the first, `frozen-`θ trains only the noise adapter $q_\phi(z|y)$ via loss (9) (amortized over y), while keeping θ fixed to a pretrained flow map. The second, `unconstrained-`θ, optimizes the same objective but learns θ jointly with ϕ. These baselines are chosen to illustrate (i) the effect of joint optimization of θ and ϕ, and (ii) the failure mode that can occur when θ is trained without the structural constraints imposed by the mean flow loss.

For model evaluation, we use the following metrics: (1) The negative log predictive density (NLPD), evaluates how well generated samples are consistent with observations y; (2) the continuous ranked probability score (CRPS) measures uncertainty calibration around the ground truth x that generated y; (3) the maximum mean discrepancy (MMD) provides a sample-based distance between the true and approximate posteriors (Gretton et al., 2012); (4) the support accuracy (SACC) measures the proportion of samples $x = f_\theta(z)$ that lie on the checkerboard support. We compare MMD and SACC on both unconditional samples $\{f_\theta(z)\}_{z\sim\mathcal{N}(0,I)}$ and conditional samples $\{f_\theta(z)\}_{z\sim q_\phi(z|y)}$ to evaluate the quality of both prior and posterior approximations, respectively. For details, see Appendix B.1.3.

Ablation on the loss components. We compare VFM against `frozen-`θ and `unconstrained-`θ to isolate the effect of the mean flow term $\mathcal{L}_{\text{MF}}(\theta;\phi)$ in (19); results displayed in Figure 2. The `frozen-`θ baseline (Figure 2a) fails to capture the bimodality of the true posterior (support in the brown and purple cells), due to the limited flexibility of q_ϕ. On the other hand, `unconstrained-`θ (Figure 2b) is able to sample from both brown and purple cells, however, also produces many off-manifold samples. VFM (Figure 2c, $\tau = 100$, $\alpha = 1$) successfully captures both modes while preserving the checkerboard pattern; joint training improves the noise-to-data coupling, while $\mathcal{L}_{\text{MF}}$ pull samples towards the structured data manifold. This observation is supported by the improvements in CRPS and posterior MMD (see Figures 5 & 6, Appendix), and high support accuracy comparable to the pretrained flow map used in `frozen-`θ. Finally, removing $\mathcal{L}_{\text{KL}}(\phi)$ from (19) makes training unstable, owing to the ill-posedness of the inverse problem without prior regularization.

Ablation on τ and α. We sweep $\tau \in [10^{-2}, 10^2]$, and report metrics using a single-step and 4-step sampler (See Figures 5 & 6, Appendix). When $\tau \lesssim \sigma$, performance across metrics is generally worse than `frozen-`θ (Figures 8a & 8b, Appendix). For $\tau \geq 1$, results improve substantially, especially CRPS and posterior MMD, while SACC and prior MMD approach the strong values already achieved by the pretrained flow used in `frozen-`θ. We also ablate on α, fixing $\tau = 100$ (Figure 9, Appendix). Setting $\alpha = 0$ decouples the training of mean flow and the adapter, yielding behaviour close to `frozen-`θ. Increasing α strengthens the coupling, inducing a more pronounced warping of the latent space. In practice, $\alpha < 1$ is more stable and yields better prior fit (lower prior MMD, compare Figures 5d and 6d), whereas $\alpha = 1$ gives the best posterior fit (lower posterior MMD, see Figures 5c vs 6c, Appendix).

To EMA or not to EMA. Finally, we examine the role of using an EMA of θ in the observation loss $\mathcal{L}_{\text{obs}}(\theta, \phi)$. Without EMA, i.e., allowing θ-gradients to propagate through $\mathcal{L}_{\text{obs}}$, both prior and posterior support accuracy deteriorate as τ increases (orange curves in Figures 5 and 6). This can be explained by the fact that in the limit $\tau \to \infty$, this pushes training toward the `unconstrained-`θ fail-

Figure 3. Qualitative comparison on ImageNet 256×256 box inpainting. Top row: ground truth, measurement, and reconstructions from guidance-based baselines. Bottom row: conditional samples produced by VFM, showing diversity in the inpainted region.

ure mode, leading to unstructured sample generation. This can be seen in Figure 7c, where the no-EMA variant when $\tau = 100$ yields results similar to `unconstrained-`θ.

4.2. Image Inverse Problems

We evaluate VFM on standard image inverse problems using ImageNet 256×256, comparing against established guidance-based solvers, as well as the `frozen-`θ baseline considered in our earlier 2D experiment. For VFM, we amortize over the problems, as described in Section 3.2. All methods operate in the latent space of SD-VAE (Rombach et al., 2022). We provide further details of our experimental settings in Appendix B.2.

Comparison with guidance-based methods. Table 1 reports quantitative results on box inpainting and Gaussian deblurring tasks (additional tasks are in Table 2, Appendix). For VFM, we report results for both single posterior samples and averaged estimates over 10 posterior samples, shown as {`sample`}/{`average`}. For all guidance-based baselines, we use the same flow-matching backbone (SiT-B/2) used to initialize our mean-flow model.

Across both tasks, we observe that VFM is consistently better than the baselines on distributional metrics (FID, MMD & CRPS), e.g., on box inpainting, the FIDs on the baselines range between 63–76, while we achieve an FID of 33.3. These improvements align with the qualitative results in Figure 3, where we observe that VFM exhibits notable diversity in the inpainted region, while maintaining visual sharpness. Guidance-based methods generally struggle with box-inpainting, especially when operating in latent space.

On pixel-space fidelity metrics (PSNR, SSIM), guidance methods consistently scores higher than a single VFM draw. However, both PSNR and SSIM typically reward mean behavior and thus prefer smoother results (Zhang et al., 2018). To confirm this, we observe that averaging multiple VFM samples narrows this gap and even exceeds the baselines in some instances, e.g., on Gaussian deblurring. On LPIPS, which is a feature-space perceptual similarity metric, we find that VFM is competitive even without averaging; this is consistent with LPIPS being more aligned with the perceptual quality than PSNR or SSIM (Zhang et al., 2018).

We also note the significant speed advantage of VFM at inference time: we used 250 sampling steps for the guidance methods with an additional ×2 cost for classifier-free guidance (Ho & Salimans, 2022), while VFM requires only one step to achieve competitive results, as displayed. This results in around two orders of magnitude lower wall-clock time, e.g., DAPS (Zhang et al., 2025) has an inference cost close to a minute; in comparison, the ~ 0.03s cost of VFM is instantaneous.

Benefits of joint training. The `frozen-`θ baseline, while fastest at inference time, performs poorly across all metrics, exhibiting visible artifacts and blurriness. This highlights the importance of jointly training the flow map f_θ and the adapter q_ϕ, consistent with our observations from the 2D experiment that the flow map itself needs to adjust for the adapter to approximate the conditional distributions well. By training jointly, we observe surprisingly strong perceptual quality, despite the simple Gaussian structural assumption used in the variational posterior.

Unconditional generation. To assess the robustness of VFM, we also evaluate unconditional generation from the trained flow map. In Figure 4, we compare the FID on 50,000 unconditional samples generated from the flow map in VFM, against various baselines with similar architecture sizes (Song & Dhariwal, 2023; Frans et al., 2025; Lee et al., 2025; Zhou et al., 2025). We fine-tune the SiT-B/2 model (trained for 80 epochs) for an additional 100 epochs. We note, however, that the baselines are trained for longer ($\sim$240 epochs). Unconditional generation of VFM remains competitive, with 2-step sampling results achieving FID below 10 (see Figure 4 for visual results). To achieve this result, we emphasize the important role of the α parameter; we observe that the adapter's noise outputs retain some structure from the observations (see Figure 15, Appendix)

Task	Method	NFE	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	MMD (↓)	$\text{CRPS}_{\text{DINO}}$ (↓)	CRPS_{Inc} (↓)	Time (s) (↓)
Inpaint (box)	Latent DPS	250×2	22.80	0.704	0.349	62.89	0.132	0.511	0.389	7.223
	Latent DAPS	250×2	**23.98**	**0.707**	0.348	–	–	0.468	0.365	43.93
	PSLD	250×2	22.61	0.699	0.346	67.22	0.153	0.536	0.435	10.07
	MPGD	250×2	22.76	0.705	0.350	62.35	0.132	0.510	0.388	7.487
	FlowChef	250×2	22.80	0.704	0.349	63.20	0.133	0.512	0.389	7.612
	FlowDPS	250×2	23.21	0.706	0.364	75.62	0.166	0.606	0.482	14.47
	`frozen-`θ	1	19.41	0.531	0.528	136.12	0.255	0.814	0.601	**0.015**
	VFM (ours)	1 / 10	21.98 / 22.71	0.609 / 0.632	0.281 / **0.280**	**33.34**	**0.074**	**0.387**	**0.362**	0.025 / 0.252
Gaussian deblur	Latent DPS	250×2	23.21	0.592	0.434	83.11	0.180	0.613	0.498	7.724
	Latent DAPS	250×2	21.46	0.500	0.432	–	–	0.529	0.422	46.86
	PSLD	250×2	23.01	0.591	0.459	101.23	0.223	0.675	0.559	10.28
	MPGD	250×2	23.22	0.593	0.435	83.86	0.183	0.612	0.498	7.695
	FlowChef	250×2	23.21	0.592	0.434	83.19	0.180	0.613	0.499	7.525
	FlowDPS	250×2	23.41	0.615	0.449	92.13	0.209	0.699	0.569	14.91
	`frozen-`θ	1	20.02	0.419	0.597	172.74	0.306	0.927	0.657	**0.015**
	VFM (ours)	1 / 10	21.74 / **23.92**	0.510 / **0.619**	0.417 / **0.388**	**51.05**	**0.096**	**0.525**	**0.399**	0.027 / 0.268

Table 1. Quantitative comparison on ImageNet for box inpainting and Gaussian deblurring. Best results are in **bold**, second best are underlined. ↑: higher is better, ↓: lower is better. For VFM, we display results for single samples and average over 10 samples, displayed as {`sample`} / {`average`}. VFM achieves the best results on LPIPS, FID, MMD, CRPS, with a significantly reduced wall-clock time.

	NFE	FID (↓)
iCT	1	34.24
Shortcut-B/2	1	40.30
IMM-B/2	1×2	9.60
MF-B/2	1	6.17
DMF-B/2	1	5.63
VFM-B/2	1	10.77
	2	9.22

Figure 4. Unconditional generation on ImageNet 256×256. **Left:** unconditional samples from VFM-B/2. **Right:** unconditional FID comparison versus mean-flow baselines. VFM retains competitive performance despite it being trained for posterior sampling.

and are therefore not representative of pure standard Gaussian noise. Thus, using $\alpha < 1$ is necessary to achieve good unconditional performance. In our experiments, we used $\alpha = 0.5$.

5. Related Works

Variational/amortized inference with diffusion-based priors has been explored in previous works: (Feng et al., 2023) explores usage of score-based prior in variational inference to approximate posteriors $p(x|y)$ in data space and (Mammadov et al., 2024) extends this to the amortized inference setting. However, these approaches rely on normalizing flows to ensure sufficient flexibility for the variational posterior, making scaling to high-resolution settings difficult. The work (Mardani et al., 2023) on the other hand, uses a Gaussian variational posterior similar to ours, but still performs variational inference in data space.

Noise space posterior inference for arbitrary generative models has been considered in (Venkatraman et al., 2025). However, their method considers a frozen generator and compensates with a more flexible noise adapter based on neural SDEs, making training significantly more complex. In comparison, VFM uses a simpler adapter and instead unfreeze the generative flow map, so the model itself can adapt to the conditional task while keeping the objective simple.

We also note the work (Silvestri et al., 2025), which introduces Variational Consistency Training (VCT) to address instability issues in consistency model training by learning data-dependent noise couplings through a variational encoder that maps data into a better-behaved latent representation. While conceptually related to our work, the goal is different: VCT is aimed at improving stability of unconditional consistency training, whereas VFM is designed to amortize posterior sampling for conditional generation.

Finally, Noise Consistency Training (NCT) (Luo et al., 2025) also targets one-step conditional sampling, but via a differnt construction: they consider a diffusion process in (z, y)-space and learns a consistency map from intermediate states to (x, y). This is strongly tied to consistency models and therefore do not generalize naturally to flow maps, considered state-of-the-art in one-step generative modelling.

6. Conclusion

We proposed Variational Flow Maps (VFMs) to enable conditional generation using flow maps with just a single (or few) sampling steps. This leverages a variational objective to jointly train a flow map and the noise adapter, which infers noise from observations. A natural next step is to relax our current Gaussian adapter assumption by using more expressive noise adapters, e.g., energy-transformers (Hoover et al., 2023) that can capture richer, non-Gaussian conditional structure. In addition, extending VFMs to more general reward alignment problems (e.g., text-to-image) or video inverse problems would be of interest.

Impact Statement

The overarching goal of reducing the computational cost for conditional generation and posterior sampling has the potential not only to drive practical applications in scientific and engineering workflows that rely on fast generation of posterior samples, but also to help reduce the high energy cost for inference. This is especially valuable as generative models see widespread use in today's society; thus, the problem of lowering inference costs is an increasingly important challenge for machine learning. Variational flow maps take a step in this direction by enabling low-cost conditional sampling without sacrificing performance.

References

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Billingsley, P. *Convergence of probability measures*. John Wiley & Sons, 2013.

Boffi, N. M., Albergo, M. S., and Vanden-Eijnden, E. Flow Map Matching: A unifying framework for consistency models. *arXiv:2406.07507*, June 2024.

Boffi, N. M., Albergo, M. S., and Vanden-Eijnden, E. How to build a consistency model: Learning flow maps via self-distillation, 2025. URL https://arxiv.org/abs/2505.18825.

Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 35:25683–25696, 2022.

Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye, J. C. Diffusion posterior sampling for general noisy inverse problems, 2024. URL https://arxiv.org/abs/2209.14687.

Feng, B. T., Smith, J., Rubinstein, M., Chang, H., Bouman, K. L., and Freeman, W. T. Score-based diffusion models as principled priors for inverse imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10520–10531, 2023.

Frans, K., Hafner, D., Levine, S., and Abbeel, P. One step diffusion via shortcut models, 2025. URL https://arxiv.org/abs/2410.12557.

Geng, Z., Pokle, A., Luo, W., Lin, J., and Kolter, J. Z. Consistency Models Made Easy. *arXiv:2406.14548*, 2024.

Geng, Z., Deng, M., Bai, X., Kolter, J. Z., and He, K. Mean flows for one-step generative modeling, 2025. URL https://arxiv.org/abs/2505.13447.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The journal of machine learning research*, 13(1):723–773, 2012.

He, Y., Murata, N., Lai, C.-H., Takida, Y., Uesaka, T., Kim, D., Liao, W.-H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov, R., and Ermon, S. Manifold preserving guided diffusion, 2023. URL https://arxiv.org/abs/2311.16424.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, volume 33, pp. 6840–6851, 2020.

Hoover, B., Liang, Y., Pham, B., Panda, R., Strobelt, H., Chau, D. H., Zaki, M., and Krotov, D. Energy transformer. *Advances in neural information processing systems*, 36: 27532–27559, 2023.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *arXiv:2206.00364*, 2022.

Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Kim, J., Kim, B. S., and Ye, J. C. FlowDPS: Flow-driven posterior sampling for inverse problems. *arXiv preprint arXiv:2503.08136*, 2025.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Lee, K., Yu, S., and Shin, J. Decoupled meanflow: Turning flow models into flow maps for accelerated sampling. *arXiv preprint arXiv:2510.24474*, 2025.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2022.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022.

Luo, Y., Xue, S., Hu, T., and Tang, J. Noise consistency training: A native approach for one-step generator in learning additional controls. *arXiv preprint arXiv:2506.19741*, 2025.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers. *arXiv:2401.08740*, 2024.

Mammadov, A., Chung, H., and Ye, J. C. Amortized posterior sampling with diffusion prior distillation. *arXiv preprint arXiv:2407.17907*, 2024.

Mardani, M., Song, J., Kautz, J., and Vahdat, A. A variational perspective on solving inverse problems with diffusion models. *arXiv preprint arXiv:2305.04391*, 2023.

Patel, M., Wen, S., Metaxas, D. N., and Yang, Y. Flowchef: Steering of rectified flow models for controlled generations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15308–15318, 2025.

Perez, E., Strub, F., De Vries, H., Dumoulin, V., and Courville, A. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A., and Shakkottai, S. Solving linear inverse problems provably via posterior sampling with latent diffusion models. *Advances in Neural Information Processing Systems*, 36: 49960–49990, 2023.

Silvestri, G., Ambrogioni, L., Lai, C.-H., Takida, Y., and Mitsufuji, Y. Training consistency models with variational noise coupling. *arXiv preprint arXiv:2502.18197*, 2025.

Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv:1503.03585*, 2015.

Song, J., Vahdat, A., Mardani, M., and Kautz, J. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=9_gsMA8MRKQ.

Song, Y. and Dhariwal, P. Improved Techniques for Training Consistency Models. *arXiv:2310.14189*, 2023.

Song, Y. and Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. *arXiv:1907.05600*, 2020.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency Models. *arXiv:2303.01469*, 2023b.

Tang, Z., Bao, J., Chen, D., and Guo, B. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025.

Venkatraman, S., Hasan, M., Kim, M., Scimeca, L., Sendera, M., Bengio, Y., Berseth, G., and Malkin, N. Outsourced diffusion sampling: Efficient posterior inference in latent spaces of generative models. *arXiv preprint arXiv:2502.06999*, 2025.

Wang, Y., Yu, J., and Zhang, J. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

Zhang, B., Chu, W., Berner, J., Meng, C., Anandkumar, A., and Song, Y. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20895–20905, 2025.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. 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.

Zhou, L., Ermon, S., and Song, J. Inductive moment matching. *arXiv:2503.07565*, 2025.

A. Theory

A.1. Derivation of the loss

We recall that the VFM objective is obtained by matching the following two representations of $p(x, y, z)$ using the KL divergence:

$$q_\phi(z|y)p(y|x)p(x) \approx p_\theta(x, y|z)p(z), \tag{20}$$

where we assumed that

$$p_\theta(x, y|z) = \mathcal{N}(x|f_\theta(z), \sigma^2 I)\,\mathcal{N}(y|Af_\theta(z), \tau^2 I). \tag{21}$$

By direct computation, this yields

$$\begin{aligned}
&\mathrm{KL}(q_\phi(z|y)p(y|x)p(x)\,||\,p_\theta(x, y|z)p(z)) && (22)\\
&= -\int \log \frac{p_\theta(x, y|z)p(z)}{q_\phi(z|y)p(y|x)p(x)} q_\phi(z|y)p(y|x)p(x)dxdydz && (23)\\
&= -\mathbb{E}_{q_\phi(z|y)p(y|x)p(x)}\left[\log p_\theta(x, y|z)\right] + \mathbb{E}_{p(y|x)p(x)}\left[\mathrm{KL}\left(q_\phi(z|y)\,||\,p(z)\right)\right] + \underbrace{\mathbb{E}_{p(y|x)p(x)}[\log(p(y|x)p(x))]}_{\leq 0} && (24)\\
&\leq -\mathbb{E}_{q_\phi(z|y)p(y|x)p(x)}\left[\log p_\theta(x, y|z)\right] + \mathbb{E}_{p(y)}\left[\mathrm{KL}\left(q_\phi(z|y)\,||\,p(z)\right)\right] && (25)\\
&\overset{(21)}{=} -\mathbb{E}_{q_\phi(z|y)p(y)}\left[\log\mathcal{N}(x|f_\theta(z), \sigma^2 I) + \log\mathcal{N}(y|\mathcal{A}f_\theta(z), \tau^2 I)\right] + \mathbb{E}_{p(y|x)p(x)}\left[\mathcal{KL}\left(q_\phi(z|y)\,||\,p(z)\right)\right], && (26)
\end{aligned}$$

where we used that $\mathbb{E}_{p(y|x)p(x)}[\log(p(y|x)p(x))] \leq 0$ since this is the negative Shannon entropy of the joint distribution $H(p(x, y)) := -\mathbb{E}_{p(x,y)}[\log(p(x, y))] \geq 0$. This yields

$$\mathrm{KL}(q_\phi(z|y)p(y|x)p(x)\,||\,p_\theta(x, y|z)p(z)) \leq \frac{1}{2\tau^2}\mathcal{L}_{\text{data}}(\theta, \phi) + \frac{1}{2\sigma^2}\mathcal{L}_{\text{obs}}(\theta, \phi) + \mathcal{L}_{\text{KL}}(\phi),$$

where

$$\begin{aligned}
\mathcal{L}_{\text{data}}(\theta, \phi) &= \mathbb{E}_{q_\phi(z|y)p(y|x)p(x)}\left[\|x - f_\theta(z)\|^2\right], && (27)\\
\mathcal{L}_{\text{obs}}(\theta, \phi) &= \mathbb{E}_{q_\phi(z|y)p(y)}\left[\|y - Af_\theta(z)\|^2\right], && (28)\\
\mathcal{L}_{\text{KL}}(\phi) &= \mathbb{E}_{p(y)}\left[\mathrm{KL}\left(q_\phi(z|y)\,||\,p(z)\right)\right]. && (29)
\end{aligned}$$

A.2. Proof of Proposition 3.1

This section provides the formal proofs for Proposition 3.1 within a Linear-Gaussian framework. We analyze the interaction between the generative map f_θ and the variational posterior q_ϕ to demonstrate that joint optimization is necessary for exact posterior mean recovery under diagonal constraints. The derivation proceeds from characterizing the optimal parameters to proving the almost sure failure of separate training in Proposition A.18. We conclude with Remark A.19, which discusses the extension of these results to non-linear cases through the lens of Jacobian alignment and symmetry restoration.

Data and Observation Model. We assume the ground truth data $x \in \mathbb{R}^d$ follows a Gaussian distribution:

$$x \sim p_{data}(x) = \mathcal{N}(m, C), \tag{30}$$

where $m \in \mathbb{R}^d$ is the data mean and $C \in \mathbb{R}^{d\times d}$ is the symmetric positive definite (SPD) covariance matrix. The observation $y \in \mathbb{R}^{d_y}$ is obtained via a linear operator $A \in \mathbb{R}^{d_y\times d}$ with additive Gaussian noise:

$$y = Ax + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I), \tag{31}$$

where $\sigma > 0$ is the noise level. Consequently, the marginal distribution of observations is given by

$$p(y) = \mathcal{N}(\mu_y, \Sigma_y), \quad \text{where } \mu_y = Am, \quad \Sigma_y = ACA^\top + \sigma^2 I. \tag{32}$$

Generative Model. We define the generative model $f_\theta : \mathbb{R}^d \to \mathbb{R}^d$ as a linear map acting on a standard Gaussian latent variable z:

$$z \sim p(z) = \mathcal{N}(0, I), \tag{33}$$
$$x = f_\theta(z) = K_\theta z + b_\theta, \tag{34}$$

where $\theta = \{K_\theta, b_\theta\}$ are the learnable parameters with $K_\theta \in \mathbb{R}^{d\times d}$ and $b_\theta \in \mathbb{R}^d$. The induced model distribution is $p_\theta(x) = \mathcal{N}(b_\theta, K_\theta K_\theta^\top)$.

Amortized Inference (Adapter). We parameterize the variational posterior (noise adapter) $q_\phi(z|y)$ as a multivariate Gaussian distribution:

$$q_\phi(z|y) = \mathcal{N}(\mu_\phi(y), \Sigma_\phi(y)), \tag{35}$$

where $\mu_\phi : \mathbb{R}^{d_y} \to \mathbb{R}^d$ and $\Sigma_\phi : \mathbb{R}^{d_y} \to \mathbb{R}^{d\times d}$ are generally parameterized by neural networks. While one may optionally restrict $\Sigma_\phi(y)$ to be a diagonal matrix for computational efficiency.

In the following sections, we will derive the optimal solutions for $\theta = \{K_\theta, b_\theta\}$ and ϕ under the separate training and joint training paradigms, respectively.

Training Objective. Recall that in the general framework, we minimized a joint objective consisting of a consistency regularization term, a data fitting term, and a KL divergence term. In this linear-Gaussian theoretical analysis, the generative map $f_\theta(z) = K_\theta z + b_\theta$ is explicitly parameterized as a single-step affine transformation.

$$\begin{aligned}\mathcal{L}(\theta,\phi) = &\underbrace{\mathbb{E}_{y\sim p_{\text{data}}(y)}\mathbb{E}_{z\sim q_\phi(z|y)}\left[\frac{1}{2\sigma^2}\|y - Af_\theta(z)\|^2\right]}_{\mathcal{L}_{\text{obs}}\text{: Reconstruction Loss}} \\ &+ \underbrace{\mathbb{E}_{x\sim p_{\text{data}}(x)}\mathbb{E}_{y\sim p(y|x)}\mathbb{E}_{z\sim q_\phi(z|y)}\left[\frac{1}{2\tau^2}\|x - f_\theta(z)\|^2\right]}_{\mathcal{L}_{\text{data}}\text{: Data Fitting Loss}} \\ &+ \underbrace{\mathbb{E}_{y\sim p_{\text{data}}(y)}\left[\mathrm{KL}(q_\phi(z|y)\,||\,p(z))\right]}_{\mathcal{L}_{\text{KL}}\text{: KL Loss}}.\end{aligned} \tag{36}$$

Note that $\mathcal{L}_{\text{data}}$ corresponds to the negative expected log-likelihood term $-\mathbb{E}\left[\log \mathcal{N}(x|f_\theta(z), \tau^2 I)\right]$.

Definition A.1 (Matrix Sets and Measure)**.** We denote the set of $d \times d$ orthogonal matrices as the orthogonal group $\mathcal{O}(d) = \{Q \in \mathbb{R}^{d\times d} \mid Q^\top Q = I\}$. The space $\mathcal{O}(d)$ is equipped with the unique normalized Haar measure $\nu_{\mathcal{O}(d)}$, representing the uniform distribution over the group. We denote the set of $d \times d$ real diagonal matrices as $\mathcal{D}(d) = \{\mathrm{diag}(d_1, \ldots, d_d) \mid d_i \in \mathbb{R}\}$.

Definition A.2 (Optimal Loss Value)**.** We define the optimal loss value for any $\theta \in \Theta^*$ and any ϕ as:

$$\mathcal{L}_{\text{opt}} = \min_{\theta\in\Theta^*,\phi} \mathcal{L}(\theta, \phi). \tag{37}$$

Proposition A.3 (Optimal Generative Parameters via KL Minimization)**.** *Consider the data distribution $p_{\text{data}}(x) = \mathcal{N}(m, C)$ and the induced model distribution $p_\theta(x) = \mathcal{N}(b_\theta, \Sigma_\theta)$ with $\Sigma_\theta = K_\theta K_\theta^\top$. Let $C = U\Lambda^2 U^\top$ be the eigen-decomposition where $U \in \mathcal{O}(d)$ and $\Lambda \in \mathcal{D}(d)$ has positive entries. The set of optimal parameters $\Theta^* = \arg\min_\theta \mathrm{KL}(p_{\text{data}}(x)\,||\,p_\theta(x))$ is given by:*

$$\Theta^* = \{\{K_\theta, b_\theta\} \mid b_\theta = m,\ K_\theta = U\Lambda Q,\ \forall Q \in \mathcal{O}(d)\}. \tag{38}$$

Proof. The KL divergence between two multivariate Gaussians is minimized if and only if their first and second moments match, i.e., $b_\theta = m$ and $\Sigma_\theta = C$. Substituting the parameterization $\Sigma_\theta = K_\theta K_\theta^\top$ and the eigen-decomposition of C, the second condition becomes $K_\theta K_\theta^\top = U\Lambda^2 U^\top = (U\Lambda)(U\Lambda)^\top$. This equality holds if and only if $K_\theta = U\Lambda Q$ for some $Q \in \mathbb{R}^{d\times d}$ such that $QQ^\top = I$, which implies $Q \in \mathcal{O}(d)$. □

Proposition A.4. *Consider the joint training objective $\mathcal{L}(\theta, \phi)$ in the Linear-Gaussian setting. For fixed generative parameters $\theta = \{K_\theta, b_\theta\}$, the optimal variational posterior $q_{\phi^*}(z|y) = \mathcal{N}(\mu^*(y), \Sigma^*(y))$ minimizes the loss. The optimal covariance Σ^* is a constant matrix independent of y, and the optimal mean $\mu^*(y)$ is an affine function of y:*

$$\mu^*(y) = K_\phi y + b_\phi, \tag{39}$$

$$\Sigma^*(y) = \Sigma_\phi, \tag{40}$$

where the optimal coefficients are given by

$$\Sigma_\phi = \left(I_d + \frac{1}{\tau^2} K_\theta^\top K_\theta + \frac{1}{\sigma^2} K_\theta^\top A^\top A K_\theta \right)^{-1}, \tag{41}$$

$$K_\phi = \Sigma_\phi K_\theta^\top \left(\frac{1}{\sigma^2} A^\top + \frac{1}{\tau^2} K \right), \tag{42}$$

$$b_\phi = \Sigma_\phi K_\theta^\top \left[-\frac{1}{\sigma^2} A^\top A b_\theta + \frac{1}{\tau^2} (I_d - KA) m - \frac{1}{\tau^2} b_\theta \right], \tag{43}$$

and $K = CA^\top (ACA^\top + \sigma^2 I_{d_y})^{-1}$ denotes the Kalman gain matrix associated with the data distribution.

Proof. The total loss is expressed as the expectation $\mathcal{L} = \mathbb{E}_{y \sim p(y)}[J(y; \mu, \Sigma)]$, where $\mu = \mu_\phi(y)$ and $\Sigma = \Sigma_\phi(y)$. The pointwise objective $J(y; \mu, \Sigma)$ is

$$\begin{aligned} J(y; \mu, \Sigma) = &\frac{1}{2\sigma^2} \left(\|y - Ab_\theta - AK_\theta \mu\|^2 + \mathrm{Tr}(K_\theta^\top A^\top A K_\theta \Sigma) \right) \\ &+ \frac{1}{2\tau^2} \left(\mathbb{E}_{x|y}[\|x - b_\theta - K_\theta \mu\|^2] + \mathrm{Tr}(K_\theta^\top K_\theta \Sigma) \right) \\ &+ \frac{1}{2} \left(\mathrm{Tr}(\Sigma) + \|\mu\|^2 - \ln |\Sigma| \right). \end{aligned} \tag{44}$$

Differentiating J with respect to Σ yields

$$\frac{\partial J}{\partial \Sigma} = \frac{1}{2} \left(\frac{1}{\sigma^2} K_\theta^\top A^\top A K_\theta + \frac{1}{\tau^2} K_\theta^\top K_\theta + I_d \right) - \frac{1}{2} \Sigma^{-1}. \tag{45}$$

The stationary point of this gradient corresponds to the constant optimal covariance matrix Σ_ϕ defined in (41). Similarly, the gradient with respect to the variational mean μ is given by

$$\nabla_\mu J = -\frac{1}{\sigma^2} K_\theta^\top A^\top (y - Ab_\theta - AK_\theta \mu) - \frac{1}{\tau^2} K_\theta^\top (\mathbb{E}[x|y] - b_\theta - K_\theta \mu) + \mu. \tag{46}$$

Rearranging the terms for the condition $\nabla_\mu J = 0$, it follows that

$$\left(I_d + \frac{1}{\sigma^2} K_\theta^\top A^\top A K_\theta + \frac{1}{\tau^2} K_\theta^\top K_\theta \right) \mu = \frac{1}{\sigma^2} K_\theta^\top A^\top (y - Ab_\theta) + \frac{1}{\tau^2} K_\theta^\top (\mathbb{E}[x|y] - b_\theta). \tag{47}$$

Observing that the coefficient matrix on the left-hand side is Σ_ϕ^{-1}, we obtain the expression for the optimal mean

$$\mu^*(y) = \Sigma_\phi K_\theta^\top \left[\frac{1}{\sigma^2} A^\top y - \frac{1}{\sigma^2} A^\top A b_\theta + \frac{1}{\tau^2} \mathbb{E}[x|y] - \frac{1}{\tau^2} b_\theta \right]. \tag{48}$$

Substituting the conditional expectation of the data distribution $\mathbb{E}[x|y] = Ky + (I_d - KA)m$ into (48) results in

$$\mu^*(y) = \Sigma_\phi K_\theta^\top \left(\frac{1}{\sigma^2} A^\top + \frac{1}{\tau^2} K \right) y + \Sigma_\phi K_\theta^\top \left[-\frac{1}{\sigma^2} A^\top A b_\theta + \frac{1}{\tau^2} (I_d - KA) m - \frac{1}{\tau^2} b_\theta \right]. \tag{49}$$

This affine structure identifies K_ϕ and b_ϕ as defined in (42) and (43). □

Corollary A.5. *The optimal variational covariance matrix Σ_ϕ given in (41) is SPD. Consequently, in the theoretical optimization landscape, the global minimum naturally satisfies the validity constraint for a covariance matrix, implying that explicitly enforcing positive semi-definite constraints is not required to define the solution at the global optimum.*

Proof. This is derived by the analytical form of the inverse optimal covariance derived in the proof of Proposition A.4:

$$(\Sigma_\phi)^{-1} = I + \frac{1}{\tau^2} K_\theta^\top K_\theta + \frac{1}{\sigma^2} (AK_\theta)^\top (AK_\theta). \tag{50}$$

□

Corollary A.6. *By Proposition A.4, for any fixed generative parameters $\theta \in \Theta^*$, the optimal variational posterior $q_{\phi^*}(z|y)$ is uniquely characterized by an affine mean $\mu^*(y) = K_\phi y + b_\phi$ and a constant covariance Σ_ϕ. Consequently, the optimization over the functional space Φ is equivalent to optimizing over the finite-dimensional parameter space:*

$$\Phi = \{(K_\phi, b_\phi, \Sigma_\phi) \mid K_\phi \in \mathbb{R}^{d \times d_y}, b_\phi \in \mathbb{R}^d, \Sigma_\phi \in \mathbb{S}^d_{++}\}. \tag{51}$$

Proof. The functional forms derived in Proposition A.4 show that any q_ϕ not belonging to this parametric family is strictly sub-optimal for the joint loss $\mathcal{L}(\theta, \phi)$, thus reducing the search space to the coefficients $\{K_\phi, b_\phi, \Sigma_\phi\}$. □

Definition A.7 (Separate Training)**.** The separate training paradigm consists of a two-stage sequential optimization. First, the generative parameters $\theta = \{K_\theta, b_\theta\}$ are obtained by minimizing the unconditional KL divergence

$$\theta^* = \operatorname*{argmin}_\theta \mathrm{KL}\left((f_\theta)_\sharp \mathcal{N}(0, I) \,||\, p_{\mathrm{data}}(x)\right), \tag{52}$$

which, in the linear-Gaussian case, implies $b_{\theta^*} = m$ and $K_{\theta^*} K_{\theta^*}^\top = C$. Subsequently, the variational parameters are determined by fixing θ^* and minimizing the joint objective

$$\phi^* = \operatorname*{argmin}_\phi \mathcal{L}(\theta^*, \phi). \tag{53}$$

Definition A.8 (Joint Training)**.** The joint training paradigm optimizes θ and ϕ simultaneously subject to a consistency constraint. The optimization problem is formulated as

$$\min_{\theta, \phi} \quad \mathcal{L}(\theta, \phi) \tag{54}$$

$$\text{subject to} \quad \mathrm{KL}\left((f_\theta)_\sharp \mathcal{N}(0, I) \,||\, p_{\mathrm{data}}(x)\right) = 0. \tag{55}$$

For the linear-Gaussian framework, this constraint restricts the search space of θ to the manifold $\Theta^* = \{\{K_\theta, b_\theta\} \mid b_\theta = m, K_\theta K_\theta^\top = C\}$.

Definition A.9 (Solution Sets)**.** Let Θ^* be the set of optimal generative parameters from Proposition A.3, and let Φ and $\Phi_{\mathcal{D}}$ denote the unconstrained and diagonal variational parameter spaces, respectively. The solution sets for the training paradigms are defined as

$$\mathcal{S}^{\mathrm{sep}} = \{(\theta, \phi) \mid \theta \in \Theta^*, \phi = \operatorname*{argmin}_{\phi' \in \Phi} \mathcal{L}(\theta, \phi')\}, \tag{56}$$

$$\mathcal{S}^{\mathrm{sep}}_{\mathrm{diag}} = \{(\theta, \phi) \mid \theta \in \Theta^*, \phi = \operatorname*{argmin}_{\phi' \in \Phi_{\mathcal{D}}} \mathcal{L}(\theta, \phi')\}, \tag{57}$$

$$\mathcal{S}^{\mathrm{joint}} = \{(\theta, \phi) \mid (\theta, \phi) = \operatorname*{argmin}_{\theta' \in \Theta^*, \phi' \in \Phi} \mathcal{L}(\theta', \phi')\}, \tag{58}$$

$$\mathcal{S}^{\mathrm{joint}}_{\mathrm{diag}} = \{(\theta, \phi) \mid (\theta, \phi) = \operatorname*{argmin}_{\theta' \in \Theta^*, \phi' \in \Phi_{\mathcal{D}}} \mathcal{L}(\theta', \phi')\}. \tag{59}$$

Proposition A.10. *For every $Q \in \mathcal{O}(d)$, let $\theta(Q) = \{U\Lambda Q, m\} \in \Theta^*$. There exists a corresponding optimal variational parameter $\phi(Q)$ such that the joint loss is invariant to the choice of Q, satisfying*

$$\mathcal{L}(\theta(Q), \phi(Q)) = \mathcal{L}_{\mathrm{opt}}, \quad \forall Q \in \mathcal{O}(d). \tag{60}$$

Proof. For a fixed Q, let $K_\theta = U\Lambda Q$. From Proposition A.4, the optimal precision matrix $P_Q = (\Sigma^*_Q)^{-1}$ satisfies

$$\begin{aligned} P_Q &= I_d + \frac{1}{\tau^2} Q^\top \Lambda U^\top U \Lambda Q + \frac{1}{\sigma^2} Q^\top \Lambda U^\top A^\top A U \Lambda Q \\ &= Q^\top \left(I_d + \frac{1}{\tau^2} \Lambda^2 + \frac{1}{\sigma^2} \Lambda U^\top A^\top A U \Lambda \right) Q = Q^\top H Q, \end{aligned} \tag{61}$$

where H is a symmetric positive definite matrix independent of Q. The optimal covariance is thus $\Sigma_Q^* = Q^\top H^{-1} Q$. The invariance of the regularization loss $\mathcal{L}_{\mathrm{KL}}$ follows from the cyclic property of the trace, $\mathrm{Tr}(\Sigma_Q^*) = \mathrm{Tr}(H^{-1} Q Q^\top) = \mathrm{Tr}(H^{-1})$, and the determinant identity $|Q^\top H^{-1} Q| = |H^{-1}|$. For the reconstruction and fitting losses, let $\mu_Q^*(y) = \Sigma_Q^* K_\theta^\top v(y)$ as in (48), where $v(y)$ is independent of Q. Substitution yields

$$K_\theta \mu_Q^*(y) = (U\Lambda Q)(Q^\top H^{-1} Q)(Q^\top \Lambda U^\top) v(y) = U\Lambda H^{-1} \Lambda U^\top v(y). \tag{62}$$

The trace terms $\mathrm{Tr}(K_\theta^\top M K_\theta \Sigma_Q^*)$ and the norm $\|\mu_Q^*(y)\|^2$ similarly reduce to expressions involving only U, Λ, H, and M, all of which are independent of Q. Thus, the joint objective is invariant to the rotation Q, and every (θ, ϕ) with $\theta \in \Theta^*$ achieves the global minimum $\mathcal{L}_{\mathrm{opt}}$. □

Corollary A.11. *Under the Linear-Gaussian setting with full-rank variational covariance Σ_ϕ, the solution sets for separate and joint training are*

$$\mathcal{S}^{\mathrm{sep}} = \{(\theta(Q_{\mathrm{sep}}), \phi(Q_{\mathrm{sep}}))\}, \quad \textit{for a fixed } Q_{\mathrm{sep}} \in \mathcal{O}(d), \tag{63}$$

$$\mathcal{S}^{\mathrm{joint}} = \{(\theta(Q), \phi(Q)) \mid \forall Q \in \mathcal{O}(d)\}, \tag{64}$$

where $\theta(Q) = \{U\Lambda Q, m\}$, and the components of $\phi(Q)$ are given by

$$\Sigma_\phi(Q) = Q^\top \left(I_d + \frac{1}{\tau^2}\Lambda^2 + \frac{1}{\sigma^2} \Lambda U^\top A^\top A U \Lambda \right)^{-1} Q, \tag{65}$$

$$K_\phi(Q) = \Sigma_\phi(Q) Q^\top \Lambda U^\top \left(\frac{1}{\sigma^2} A^\top + \frac{1}{\tau^2} K \right), \tag{66}$$

$$b_\phi(Q) = \Sigma_\phi(Q) Q^\top \Lambda U^\top \left[-\frac{1}{\sigma^2} A^\top A m + \frac{1}{\tau^2}(I_d - KA)m - \frac{1}{\tau^2} m \right]. \tag{67}$$

Proof. Substituting $K_\theta = U\Lambda Q$ and $b_\theta = m$ into the optimality conditions (41)–(43) from Proposition A.4 directly yields the parameterized forms of $\Sigma_\phi(Q)$, $K_\phi(Q)$, and $b_\phi(Q)$. The rotational invariance established in Proposition A.10 ensures that for every $Q \in \mathcal{O}(d)$, the pair $(\theta(Q), \phi(Q))$ achieves the global minimum $\mathcal{L}_{\mathrm{opt}}$, forming the manifold $\mathcal{S}^{\mathrm{joint}}$. In contrast, the separate training paradigm uniquely determines Q_{sep} during the pre-training of the generative map, thereby restricting the solution to a singleton. □

Proposition A.12. *For any $(\theta, \phi) \in \mathcal{S}^{\mathrm{joint}}$, and the Kalman gain $K = CA^\top (ACA^\top + \sigma^2 I_{d_y})^{-1}$, we have*

$$K_\theta K_\phi = K. \tag{68}$$

Proof. Substituting Σ_ϕ from (41) into the expression for K_ϕ in (42), and applying the push-through identity $K_\theta (I_d + K_\theta^\top \mathcal{M} K_\theta)^{-1} = (I_d + K_\theta K_\theta^\top \mathcal{M})^{-1} K_\theta$ with $\mathcal{M} = \sigma^{-2} A^\top A + \tau^{-2} I_d$, we have:

$$\begin{aligned} K_\theta K_\phi &= (I_d + C(\sigma^{-2} A^\top A + \tau^{-2} I_d))^{-1} C \left(\sigma^{-2} A^\top + \tau^{-2} K \right) \\ &= (C^{-1} + \sigma^{-2} A^\top A + \tau^{-2} I_d)^{-1} \left(\sigma^{-2} A^\top + \tau^{-2} K \right). \end{aligned} \tag{69}$$

Using the Kalman identity $(C^{-1} + \sigma^{-2} A^\top A) K = \sigma^{-2} A^\top$ and adding $\tau^{-2} K$ to both sides yields:

$$(C^{-1} + \sigma^{-2} A^\top A + \tau^{-2} I_d) K = \sigma^{-2} A^\top + \tau^{-2} K. \tag{70}$$

Left-multiplying by $(C^{-1} + \sigma^{-2} A^\top A + \tau^{-2} I_d)^{-1}$ and comparing with (69), we obtain $K_\theta K_\phi = K$. □

Proposition A.13. *For any $(\theta, \phi) \in \mathcal{S}^{\mathrm{joint}}$, the expected output of the generative inference process recovers the exact Bayesian posterior mean:*

$$\mathbb{E}_{z \sim q_\phi(z|y)}[f_\theta(z)] = \mathbb{E}_{p_{\mathrm{data}}(x|y)}[x]. \tag{71}$$

Proof. The expected reconstruction is $\hat{x} = K_\theta(K_\phi y + b_\phi) + b_\theta$. Given $(\theta, \phi) \in \mathcal{S}^{\text{joint}}$, we have $b_\theta = m$ and $K_\theta K_\phi = K$. Expanding $K_\theta b_\phi$ via (43) with $b_\theta = m$:

$$\begin{aligned} K_\theta b_\phi &= K_\theta \Sigma_\phi K_\theta^\top \left[-\sigma^{-2} A^\top A m + \tau^{-2}(I_d - KA)m - \tau^{-2} m\right] \\ &= -(C^{-1} + \sigma^{-2} A^\top A + \tau^{-2} I_d)^{-1}(\sigma^{-2} A^\top A + \tau^{-2} KA)m \\ &= -(C^{-1} + \sigma^{-2} A^\top A + \tau^{-2} I_d)^{-1}(\sigma^{-2} A^\top + \tau^{-2} K)Am. \end{aligned} \tag{72}$$

From the proof of Proposition A.12, the term $(C^{-1} + \sigma^{-2} A^\top A + \tau^{-2} I_d)^{-1}(\sigma^{-2} A^\top + \tau^{-2} K)$ is exactly K. Thus, $K_\theta b_\phi = -KAm$. Substituting these into the expression for $\hat{x}$:

$$\hat{x} = Ky - KAm + m = m + K(y - Am), \tag{73}$$

which is the analytical conditional expectation $\mathbb{E}_{p_{\text{data}}(x|y)}[x]$. □

Corollary A.14. *For any $(\theta, \phi) \in \mathcal{S}^{\text{sep}}$, the generative inference process also recovers the exact Bayesian posterior mean, i.e., $\mathbb{E}_{z \sim q_\phi(z|y)}[f_\theta(z)] = \mathbb{E}_{p_{\text{data}}(x|y)}[x]$.*

Proof. This follows immediately from Proposition A.13 and the fact that $\mathcal{S}^{\text{sep}} \subset \mathcal{S}^{\text{joint}}$ according to Corollary A.11. □

Proposition A.15. *Assuming the observation operator A and data covariance C are in general position such that they are not simultaneously diagonalizable in the canonical basis, the following properties hold:*

1. ***Sub-optimality of Separate Training:*** $\mathcal{S}^{\text{sep}} \cap \mathcal{S}^{\text{sep}}_{\text{diag}} = \emptyset$ *a.s. w.r.t.* $\nu_{\mathcal{O}(d)}$ *(Definition A.1).*
2. ***Optimality of Joint Training:*** $\mathcal{S}^{\text{joint}} \cap \mathcal{S}^{\text{joint}}_{\text{diag}} \neq \emptyset$.

Proof. For any $\theta(Q) \in \Theta^*$, define

$$P(Q) = Q^\top H Q. \tag{74}$$

Let $\Sigma_\phi = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d) \in \mathcal{D}(d)$. The covariance-dependent objective $J(\Sigma_\phi)$ and its minimizer $\Sigma^*_{\text{diag},\phi}$ are:

$$J(\Sigma_\phi) = \frac{1}{2} \sum_{i=1}^{d} \left(P_{ii}(Q)\sigma_i^2 - \ln \sigma_i^2\right) + \text{const}, \tag{75}$$

$$\Sigma^*_{\text{diag},\phi}(Q) = [\text{diag}(P(Q))]^{-1} = \text{diag}(P(Q)^{-1}) = \text{diag}(\Sigma^*_\phi(Q)). \tag{76}$$

The optimality gap $\Delta J(Q)$ between $\Sigma^*_{\text{diag},\phi}(Q)$ and the unconstrained $\Sigma^*_\phi(Q) = P(Q)^{-1}$ is:

$$\Delta J(Q) = J(\Sigma^*_{\text{diag},\phi}(Q)) - J(\Sigma^*_\phi(Q)) = \frac{1}{2} \ln \left(\frac{\prod_{i=1}^d P_{ii}(Q)}{\det(P(Q))} \right). \tag{77}$$

By Hadamard's inequality, $\Delta J(Q) \geq 0$ and the equality holds if and only if $P(Q) \in \mathcal{D}(d)$.

In separate training, Q_{sep} is fixed during pre-training. Global optimality requires $P(Q_{\text{sep}}) = Q_{\text{sep}}^\top H Q_{\text{sep}} \in \mathcal{D}(d)$. By the general position assumption of A and C, $H = I_d + \tau^{-2}\Lambda^2 + \sigma^{-2}\Lambda U^\top A^\top A U \Lambda$ is not a diagonal matrix. Then the set $\{Q \in \mathcal{O}(d) \mid Q^\top H Q \in \mathcal{D}(d)\}$ consists of the eigenvectors of H, which has measure zero in $\mathcal{O}(d)$ with respect to the normalized Haar measure (Definition A.1). Thus, $P(Q_{\text{sep}}) \notin \mathcal{D}(d)$ a.s., leading to $\Delta J(Q) > 0$.

In joint training, Q is a learnable parameter. By the Spectral Theorem, there exists $Q^* \in \mathcal{O}(d)$ such that:

$$(Q^*)^\top H Q^* = \Lambda_H \in \mathcal{D}(d). \tag{78}$$

Setting $\Sigma^*_\phi = \Lambda_H^{-1} \in \mathcal{D}(d)$ and $\theta^* = \theta(Q^*)$ yields $\Delta J(Q) = 0$. This configuration achieves the unconstrained global minimum $\mathcal{L}_{\text{opt}}$, thus $\{\theta(Q^*), \phi(Q^*)\} \subset \mathcal{S}^{\text{joint}} \cap \mathcal{S}^{\text{joint}}_{\text{diag}}$. □

Corollary A.16. *We have $\mathcal{S}^{\text{joint}}_{\text{diag}} \subset \mathcal{S}^{\text{joint}}$. Specifically,*

$$\mathcal{S}^{\text{joint}}_{\text{diag}} = \{(\theta(Q), \phi(Q)) \mid Q \in \mathcal{V}(H)\}, \tag{79}$$

where $\mathcal{V}(H) = \{Q \in \mathcal{O}(d) \mid Q^\top H Q \in \mathcal{D}(d)\}$ is the set of eigen-bases of H.

Proof. Proposition A.15 establishes that the global minimum $\mathcal{L}_{\text{opt}}$ is achieved if and only if the optimality gap $\Delta J(Q)$ vanishes. By Hadamard's inequality, the condition $\Delta J(Q) = 0$ is equivalent to $P(Q) = Q^\top H Q$ being a diagonal matrix, which implies $Q \in \mathcal{V}(H)$. For any such rotation, the optimal variational covariance $\Sigma^*_\phi = P(Q)^{-1}$ is elements of $\mathcal{D}(d)$, ensuring that the pair $(\theta(Q), \phi(Q))$ belongs to $\mathcal{S}^{\text{joint}}_{\text{diag}}$. Since these solutions simultaneously minimize the unconstrained objective, the inclusion $\mathcal{S}^{\text{joint}}_{\text{diag}} \subset \mathcal{S}^{\text{joint}}$ holds. □

Lemma A.17. *Let $\mathcal{O}(d)$ be the orthogonal group equipped with the normalized Haar measure $\nu_{\mathcal{O}(d)}$. Let $\text{Sym}_0(d) = \{M \in \mathbb{R}^{d\times d} \mid M = M^\top, \text{diag}(M) = 0\}$. Define the map $\Phi : \mathcal{O}(d) \to \text{Sym}_0(d)$ by $\Phi(Q) = QHQ^\top - \text{diag}(QHQ^\top)$, where $H \in \mathbb{R}^{d\times d}$ is a fixed symmetric matrix with distinct eigenvalues. Let $V \subset \mathbb{R}^{d\times d}$ be a proper subspace such that $\text{Sym}_0(d) \not\subseteq V$. Then the set*

$$S = \{Q \in \mathcal{O}(d) \mid \Phi(Q) \in V\}$$

has measure $\nu_{\mathcal{O}(d)}(S) = 0$.

Proof. The orthogonal group $\mathcal{O}(d)$ is a compact real analytic manifold. Let $\mathfrak{so}(d) = \{B \in \mathbb{R}^{d\times d} \mid B = -B^\top\}$ denote its Lie algebra. The map Φ is real analytic since its entries are polynomial functions of the elements of Q. Let $P_{V^\perp}$ be the projection operator onto the orthogonal complement of V in $\mathbb{R}^{d\times d}$. The condition $\Phi(Q) \in V$ is equivalent to $f(Q) \coloneqq P_{V^\perp}\Phi(Q) = 0$. Since f is real analytic, $\nu_{\mathcal{O}(d)}(S) = 0$ follows if f is not identically zero on the connected components of $\mathcal{O}(d)$.

Since H is symmetric with distinct eigenvalues, there exists $Q_0 \in \mathcal{O}(d)$ such that $Q_0 H Q_0^\top = \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$, where $\lambda_i \neq \lambda_j$ for $i \neq j$. We evaluate the differential $\mathrm{D}\Phi(Q_0)$ by considering the variation $Q(\epsilon) = e^{\epsilon B} Q_0$ for $B \in \mathfrak{so}(d)$. The directional derivative at Q_0 is given by

$$\mathrm{D}\Phi(Q_0)[B] = [B, \Lambda] - \text{diag}([B, \Lambda]).$$

For the off-diagonal entries $i \neq j$, the commutator yields $[B, \Lambda]_{ij} = \sum_k (B_{ik}\Lambda_{kj} - \Lambda_{ik}B_{kj}) = B_{ij}\lambda_j - \lambda_i B_{ij} = (\lambda_j - \lambda_i)B_{ij}$. For the diagonal entries, $[B, \Lambda]_{ii} = B_{ii}\lambda_i - \lambda_i B_{ii} = 0$, which implies $\text{diag}([B, \Lambda]) = 0$. Thus, for any $i \neq j$, we have

$$(\mathrm{D}\Phi(Q_0)[B])_{ij} = (\lambda_j - \lambda_i)B_{ij}.$$

Given that λ_i are pairwise distinct, for any target matrix $M \in \text{Sym}_0(d)$, we can uniquely determine $B \in \mathfrak{so}(d)$ by setting $B_{ij} = M_{ij}/(\lambda_j - \lambda_i)$ for $i < j$. This proves that the differential $\mathrm{D}\Phi(Q_0) : \mathfrak{so}(d) \to \text{Sym}_0(d)$ is a linear isomorphism.

By the assumption $\text{Sym}_0(d) \not\subseteq V$, there exists some $B \in \mathfrak{so}(d)$ such that $\mathrm{D}\Phi(Q_0)[B] \in \text{Sym}_0(d) \setminus V$. It follows that $P_{V^\perp}\mathrm{D}\Phi(Q_0)[B] \neq 0$, implying that f is not identically zero in a neighborhood of Q_0. By the identity theorem for real analytic functions, the zero set $S \cap \mathcal{O}(d)^\circ$ has Haar measure zero, where $\mathcal{O}(d)^\circ$ denotes the connected component containing Q_0. A similar argument holds for the remaining connected component of $\mathcal{O}(d)$. □

Proposition A.18 (Mean Recovery Gap under Diagonal Constraint)**.** *Assuming A and C are in general position, the following properties hold under the diagonal constraint $\Sigma_\phi \in \mathcal{D}(d)$:*

1. ***Separate Training:*** *For any $(\theta, \phi) \in \mathcal{S}^{\text{sep}}_{\text{diag}}$, the inference process fails to recover the posterior mean almost surely:*

$$\mathbb{E}_{z\sim q_\phi(z|y)}[f_\theta(z)] \neq \mathbb{E}_{p_{\text{data}}(x|y)}[x] \quad \textit{a.s. w.r.t. } p(y) = \mathcal{N}(y|Am, \Sigma_y), \nu_{\mathcal{O}(d)}(Q). \tag{80}$$

2. ***Joint Training:*** *For any $(\theta, \phi) \in \mathcal{S}^{\text{joint}}_{\text{diag}}$, the inference process recovers the exact posterior mean:*

$$\mathbb{E}_{z\sim q_\phi(z|y)}[f_\theta(z)] = \mathbb{E}_{p_{\text{data}}(x|y)}[x]. \tag{81}$$

Proof. The expected reconstruction is $\hat{x} = K_\theta K_\phi y + K_\theta b_\phi + b_\theta$, while the analytical posterior mean is $\mathbb{E}[x|y] = m + K(y - Am)$. In the separate training paradigm, where $(\theta, \phi) \in \mathcal{S}^{\text{sep}}_{\text{diag}}$, the rotation Q_{sep} is fixed. Let $P = Q_{\text{sep}} H Q_{\text{sep}}^\top$ and denote $E = [\text{diag}(P)]^{-1} P - I$. Under the diagonal constraint, the gain matrix becomes $K_{\text{diag}} = K_\theta [\text{diag}(P)]^{-1} K_\theta^\top (\sigma^{-2} A^\top + \tau^{-2} K)$. The recovery error simplifies to

$$\hat{x} - \mathbb{E}[x|y] = (K_{\text{diag}} - K)(y - Am) = K_\theta E K_\theta^{-1} K (y - Am). \tag{82}$$

Since A and C are in general position, Q_{sep} does not diagonalize H almost surely, implying that P is non-diagonal and thus E is a non-zero matrix with a vanishing diagonal. In addition, this general position assumption implies that H has distinct eigenvalues and that $K = CA^\top (ACA^\top + \sigma^2 I)^{-1}$ has rank d_y.

Define $\text{Sym}_0(d) = \{M \in \mathbb{R}^{d \times d} \mid M = M^\top, \text{diag}(M) = 0\}$, and $V = \{E \in \text{Sym}_0(d) | K_\theta E K_\theta^{-1} K = 0\}$ as a subspace of $\mathbb{R}^{d \times d}$. Since K_θ is invertible due to $K_\theta K_\theta^\top = C$, we know $V \neq \text{Sym}_0(d)$. Therefore we can use the Lemma A.17 for H and V to show that

$$\nu_{\mathcal{O}(d)}(\{E | K_\theta E K_\theta^{-1} K = 0\}) = \nu_{\mathcal{O}(d)}(\{Q \in \mathcal{O}(d) | QHQ^\top - \text{diag}(QHQ^\top) \in V\}) = 0, \tag{83}$$

where $\nu_{\mathcal{O}(d)}$ is the normalized Haar measure on $\mathcal{O}(d)$ (Definition A.1). For non-zero $K_\theta E K_\theta^{-1} K$, $\{y \in \mathbb{R}^{d_y} \mid K_\theta E K_\theta^{-1} K(y - Am) = 0\}$ constitutes a proper affine subspace of $\mathbb{R}^{d_y}$ with dimension at most $d_y - 1$. Since $p(y) = \mathcal{N}(p | Am, \Sigma_y)$, it follows that $\hat{x} - \mathbb{E}[x|y] \neq 0$ almost surely with respect to $p(y)$.

For joint training, Corollary A.16 establishes that $\mathcal{S}^{\text{joint}}_{\text{diag}} \subset \mathcal{S}^{\text{joint}}$. Since every pair in $\mathcal{S}^{\text{joint}}$ satisfies the unconstrained optimality condition $\hat{x} = \mathbb{E}[x|y]$ by Proposition A.13, the identity holds for all $(\theta, \phi) \in \mathcal{S}^{\text{joint}}_{\text{diag}}$ for all $y \in \mathbb{R}^{d_y}$. □

Remark A.19 (Implications for Coordinate Alignment and Non-linear Extensions). The theoretical findings in Propositions A.15 and A.18 elucidate a fundamental synergy between the generative map and the amortized inference network.

- **Coordinate Alignment in Joint Training:** In the separate training paradigm, the generative model fixes a rigid coordinate system within the noise space. When the variational posterior is restricted to a diagonal covariance Σ_ϕ for computational efficiency, it must approximate a potentially dense precision matrix $P(Q_{\text{sep}})$, leading to a systematic recovery gap. In contrast, joint training facilitates the optimization of the orthogonal matrix $Q \in \mathcal{O}(d)$, effectively aligning the principal axes of the required posterior precision with the canonical basis of the prior distribution. This alignment ensures that a diagonal parameterization can achieve the global minimum $\mathcal{L}_{\text{opt}}$ without loss of expressivity.
- **Extension to Non-linear Manifolds:** While this analysis assumes a linear-Gaussian setting, the core principle of symmetry restoration suggests broader implications for non-linear generative models. In non-linear cases, the joint optimization allows the generator to learn a representation where the local geometry of the data distribution is consistent with the inductive bias of the adapter. For instance, the generator can adjust its Jacobian $\nabla_z f_\theta$ such that the pull-back metric is approximately diagonalized in the noise space, thereby validating the use of simplified posterior approximations in complex inference tasks.

A.3. Proof of Proposition 3.2

First, noting that $f_\theta(z) = z - u_\theta(z, 0, 1)$, we have

$$\|x - f_\theta(z)\|^2 = \|x - (z - u_\theta(z, 0, 1))\|^2 = \|u_\theta(z, 0, 1) - (z - x)\|^2. \tag{84}$$

Then, by Jensen's inequality, and recalling that $\psi_t(x, z) := tz + (1 - t)x$, we get

$$\int_0^1 \|\partial_t \mathcal{E}_\theta(x, z, 0, t)\|^2 dt \tag{85}$$

$$= \int_0^1 \left\| \frac{d}{dt} \left[t u_\theta(\psi_t(x, z), 0, t) - \int_0^t \dot{\psi}_t(x, z) ds \right] \right\|^2 dt \tag{86}$$

$$\overset{\text{Jensen}}{\geq} \left\| \int_0^1 \frac{d}{dt} \left[t u_\theta(\psi_t(x, z), 0, t) - t(z - x) \right] dt \right\|^2 \tag{87}$$

$$= \|u_\theta(z, 0, 1) - (z - x)\|^2. \tag{88}$$

Putting these together, we establish our desired bound. □

A.4. Proof of Proposition 3.4

From our assumptions, we can compute

$$p_\tau(x|y) := \int_{\mathbb{R}^d} \mathcal{N}(x|f_\theta(z), \tau^2 I) p(z|y) dz, \tag{89}$$

where

$$p(z|y) := \frac{\mathcal{N}(y|Af_\theta(z), \sigma^2 I)p(z)}{\int_{\mathbb{R}^d} \mathcal{N}(y|Af_\theta(z), \sigma^2 I)p(z)dz}. \tag{90}$$

Denoting by $\mu_\tau^y(dx) := p_\tau(x|y)dx$ and $\nu^y(dz) := p(z|y)dz$ the posterior measures in x and z spaces, respectively, for any $g \in C_b(\mathbb{R}^d)$, we have

$$\int_{\mathbb{R}^d} g(x)\mu_\tau^y(dx) \overset{(89)}{=} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} g(x)\mathcal{N}(x|f_\theta(z), \tau^2 I)\nu^y(dz)dx \tag{91}$$

$$\overset{\tau\to 0}{\longrightarrow} \int_{\mathbb{R}^d} g\big(f_\theta(z)\big)\nu^y(dz) \tag{92}$$

$$= \int_{\mathbb{R}^d} g\big(x\big)(f_\theta)_\sharp \nu^y(dx), \tag{93}$$

where we used the standard result that $\mathcal{N}(x|f_\theta(z), \tau^2 I)$ converges weakly to the delta measure around $f_\theta(z)$ as $\tau \to 0$ (Billingsley, 2013), and we used the dominated convergence theorem and Fubini's theorem, both justified by the bound

$$\left|\int_{\mathbb{R}^d} g(x)\mathcal{N}(x|f_\theta(z), \tau^2 I)dx\right| \le \|g\|_\infty. \tag{94}$$

This proves the weak convergence of measures $\mu_\tau^y \Rightarrow (f_\theta)_\sharp \nu^y$ as $\tau \to 0$. □

B. Experimental Details

B.1. 2D Checkerboard Data

We use a 2D checkerboard distribution supported on alternating squares in $[-2, 2]^2$. To sample, we first draw $u \sim \mathrm{Unif}([0, 1]^2)$ and partition the unit square into a 4×4 uniform grid. Then, we accept samples that lie on one of the checkerboard cells. Finally, we center and scale via $x = 4(u - (0.5, 0.5))$, so the support lies in $[-2, 2]^2$ and each retained square has side length 1. We used $20,000$ samples from this distribution to train our models.

B.1.1. Model architectures

For the mean-flow network u_θ, we use a SiLU MLP with six layers and width 512. We initialize this model from a flow-matching velocity network pretrained on the checkerboard samples. The noise adapter is a smaller SiLU MLP with four layers and width 256, trained from scratch. Each model is trained for $50{,}000$ iterations with batch size 2048 using the `AdamW` optimizer with learning rate 2×10^{-4} and weight decay 1×10^{-4}.

B.1.2. Problem formulation

The task in this experiment is to solve the Bayesian inverse problem

$$p(x|y) \propto \exp\left(-\frac{|y - Ax|^2}{2\sigma^2}\right) p(x), \tag{95}$$

where $p(x)$ is the 2D checkerboard distribution and the forward operator is given by $A = \begin{pmatrix}1 & 0\end{pmatrix}$, that is, observing only the first component. For the observation noise, we take $\sigma = 0.1$.

B.1.3. Metrics

To evaluate our results, we use the following metrics.

Negative log predictive density (NLPD). Given an observation $y \in \mathbb{R}$ and posterior samples $\{x^{(j)}\}_{j=1}^{J}$ with $x^{(j)} \sim p(x|y)$, the predictive density is approximated by Monte Carlo:

$$p(y'|y) \;=\; \int p(y' \mid x)\, p(x|y)\, dx \;\approx\; \frac{1}{J}\sum_{j=1}^{J} \mathcal{N}\big(y'|Ax^{(j)},\, \sigma^2\big), \tag{96}$$

where y' is a fresh observation independent of y. We report the *negative log predictive density (NLPD)*,

$$\mathrm{NLPD}(y'; y) \;=\; -\log p(y'|y) \;\approx\; -\log\left(\frac{1}{J}\sum_{j=1}^{J} \mathcal{N}\big(y'|Ax^{(j)},\, \sigma^2\big)\right), \tag{97}$$

which is a proper scoring rule. To sample from $p(x|y)$ approximately using VFM, we first sample $z^{(j)} \sim q_\phi(z|y)$ and then set $x^{(j)} = f_\theta(z^{(j)})$. We report the averaged NLPD over a batch $\{y'_b, y_b, \{x_b^{(j)}\}_{j=1}^{J}\}_{b=1}^{B}$. We take $B = 10,000$ and $J = 100$.

Continuous ranked probability score (CRPS). Given ground-truth targets $x^\dagger \in \mathbb{R}^2$ and J predictive samples $\{x^{(j)}\}_{j=1}^{J}$ corresponding to an observation $y^\dagger = Ax^\dagger + \varepsilon^\dagger$ for some noise realisation $\varepsilon^\dagger$ (i.e. we take $x^{(j)} = f_\theta(z^{(j)})$ for $z^{(j)} \sim q_\phi(z|y^\dagger)$), we estimate the CRPS as:

$$\mathrm{CRPS}(x^\dagger; y^\dagger) \;\approx\; \frac{1}{J}\sum_{j=1}^{J}\big\|x^{(j)} - x^\dagger\big\| \;-\; \frac{1}{2J^2}\sum_{j=1}^{J}\sum_{k=1}^{J}\big\|x^{(j)} - x^{(k)}\big\|. \tag{98}$$

The first term measures the average distance of samples to the truth, while the second term rewards diversity. We report the averaged CRPS over a batch $\{x_b^\dagger, y_b^\dagger, \{x_b^{(j)}\}_{j=1}^{J}\}_{b=1}^{B}$. We take $B = 10,000$ and $J = 100$.

Maximum mean discrepancy (MMD). To compare two measures μ_P and μ_Q, we can compute their maximum mean discrepancy, which is a distance on the space of measures, whose square is given by (Gretton et al., 2012)

$$\mathrm{MMD}^2(X, Y) = \mathbb{E}[k(x, x')] + \mathbb{E}[k(y, y')] - 2\,\mathbb{E}[k(x, y)], \tag{99}$$

with $x, x' \sim \mu_P$ and $y, y' \sim \mu_Q$ i.i.d., and $k(\cdot,\cdot)$ is a choice of kernel such as the squared exponential kernel

$$k(u, v) := \exp\left(-\frac{\|u - v\|_2^2}{2\ell^2}\right). \tag{100}$$

In practice, we use the unbiased estimator:

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{N(N-1)}\sum_{i\neq i'} k(x^{(i)}, x^{(i')}) + \frac{1}{M(M-1)}\sum_{j\neq j'} k(y^{(j)}, y^{(j')}) - \frac{2}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} k(x^{(i)}, y^{(j)}), \tag{101}$$

For the lengthscale hyperparameter ℓ, we choose the median heuristic computed from pairwise distances between samples. In our computations, we choose $N = M = 10,000$ samples to compare the prior distributions and the posterior distributions. Here, our true prior distribution is the checkerboard distribution, and our approximate prior is obtained by $\{f_\theta(z)\}_{z\sim\mathcal{N}(0,I)}$. For the true posterior, we compute it using rejection sampling (see Algorithm 3) and the approximate posterior is obtained by $\{f_\theta(z)\}_{z\sim q_\phi(z|y)}$.

Support accuracy (SACC). We measure *support accuracy* as the proportion (percentage) of generated samples that fall inside one of the filled checkerboard squares. Concretely, for samples $\{x^{(j)}\}_{j=1}^{J}$, we compute

$$\mathrm{Acc}(\{x^{(j)}\}_{j=1}^{J}) \;=\; \frac{1}{J}\sum_{j=1}^{J} \mathbf{1}\Big[x^{(j)} \text{ lies in a checkerboard cell}\Big]. \tag{102}$$

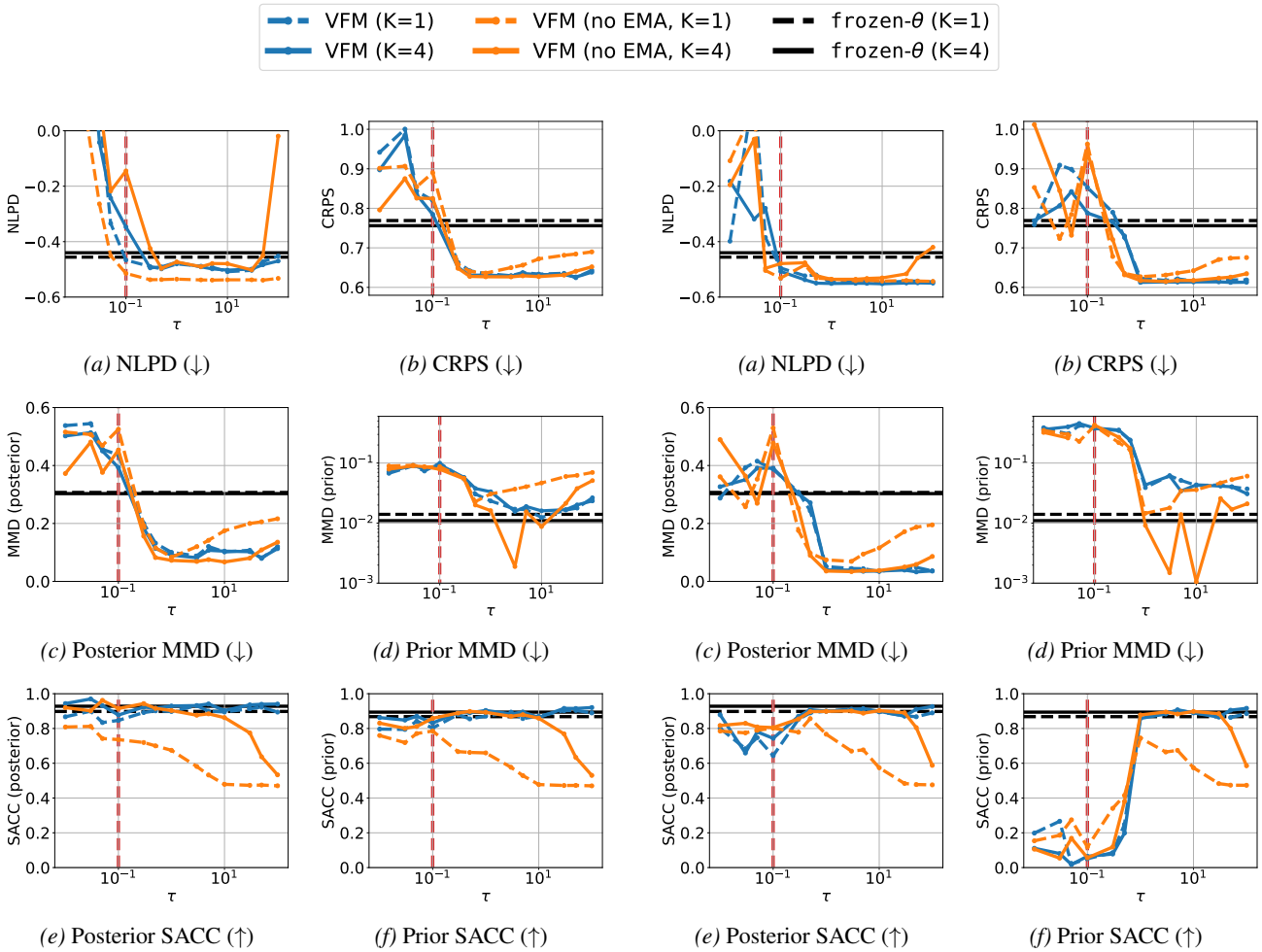
We compute the support accuracy for both prior samples $\{f_\theta(z)\}_{z\sim\mathcal{N}(0,I)}$ and posterior samples $\{f_\theta(z)\}_{z\sim q_\phi(z|y)}$.

Algorithm 3 Rejection sampling for $p(x \mid y)$

Input: observation $y \in \mathbb{R}$, noise $\sigma > 0$, number of samples J, prior $p(x)$
Initialize accepted set $\mathcal{S} \leftarrow \varnothing$
while $|\mathcal{S}| < J$ **do**
  Propose $x \sim p(x)$
  Compute $a \leftarrow \exp\left(-\frac{(y-Ax)^2}{2\sigma^2}\right)$
  Draw $u \sim \mathrm{Unif}(0, 1)$
  if $u < a$ **then**
    Append x to $\mathcal{S}$
  end if
end while
Output: $\{x^{(j)}\}_{j=1}^{J} \leftarrow \mathcal{S}$

B.1.4. Ablation plots

- Figure 5: Ablation of VFM for all metrics with respect to the parameter τ. The parameter α is set to 0.5. We also display the results of the `frozen-`θ baseline for reference.
- Figure 6: Ablation of VFM for all the metrics with respect to τ. The parameter α is set to 1.0. We also display the results of the `frozen-`θ baseline for reference.
- Figure 7: Plots displaying the noise-to-data alignment in VFM with or without various modeling choices in the loss to isolate their effects on the final results. In particular, we consider: (1) `frozen-`θ, (2) `unconstrained-`θ, (3) VFM with no EMA, (4) VFM without KL loss.
- Figure 8: Plots displaying how the noise-to-data alignment for VFM changes with respect to τ. Here, α is set to 1.0.
- Figure 9: Plots displaying how the noise-to-data alignment for VFM changes with respect to α. Here, τ is set to 100.0.

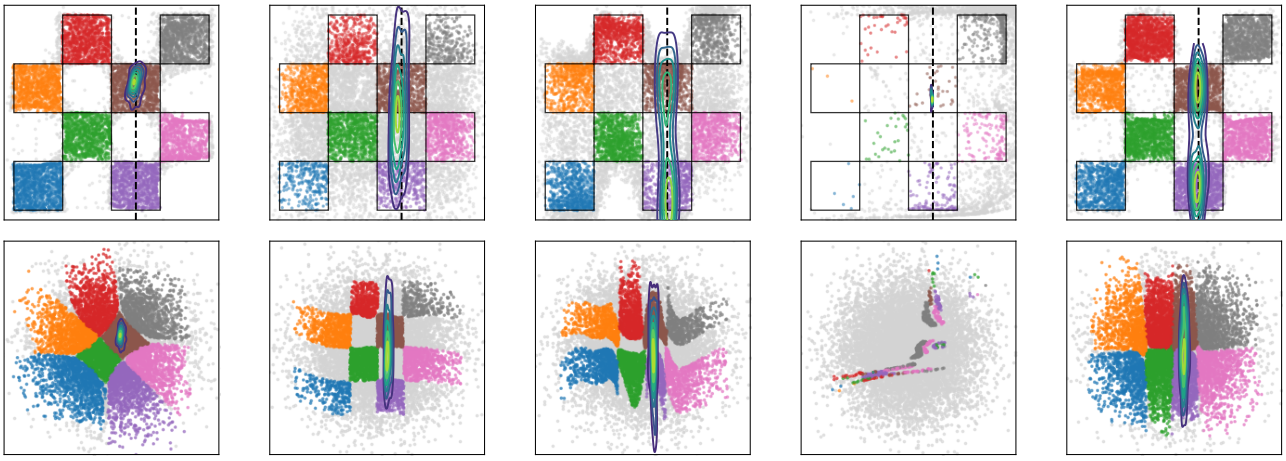


(a) NLPD (↓) *(b)* CRPS (↓) *(c)* Posterior MMD (↓) *(d)* Prior MMD (↓) *(e)* Posterior SACC (↑) *(f)* Prior SACC (↑)

Figure 5. Metrics for VFM with $\boldsymbol{\alpha} = \mathbf{0.5}$ and varying τ. Dashed vertical (red) line indicates the reference value $\sigma = 0.1$. The baseline model (black lines) is `frozen-`θ. We compare the results of VFM with EMA used in the observation loss term (blue lines) vs. without using EMA (orange line) for $K = 1, 4$.

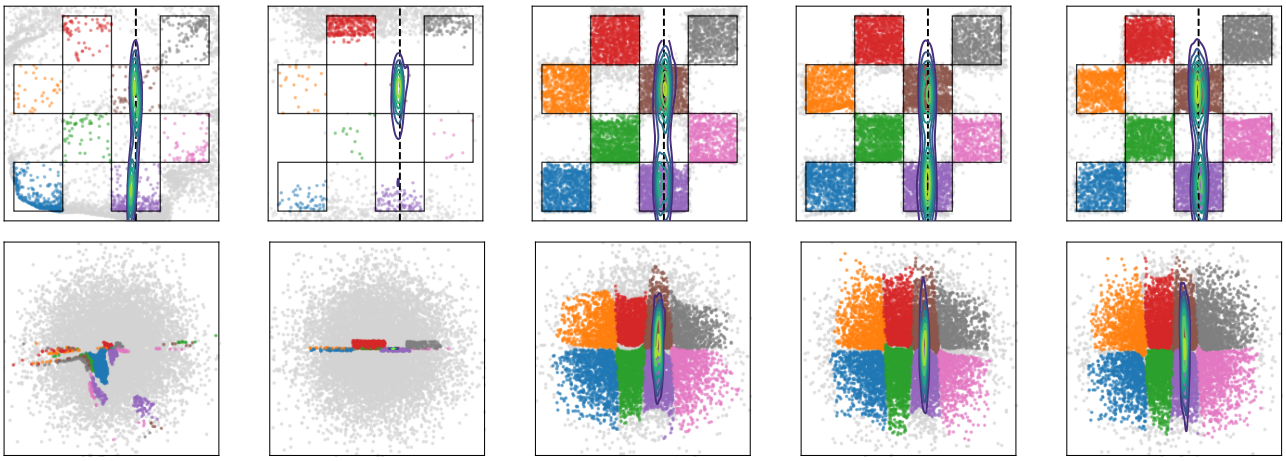
(a) NLPD (↓) *(b)* CRPS (↓) *(c)* Posterior MMD (↓) *(d)* Prior MMD (↓) *(e)* Posterior SACC (↑) *(f)* Prior SACC (↑)

Figure 6. Metrics for VFM with $\boldsymbol{\alpha} = \mathbf{1.0}$ and varying τ. Dashed vertical (red) line indicates the reference value $\sigma = 0.1$. The baseline model (black lines) is `frozen-`θ. We compare the results of VFM with EMA used in the observation loss term (blue lines) vs. without using EMA (orange line) for $K = 1, 4$.



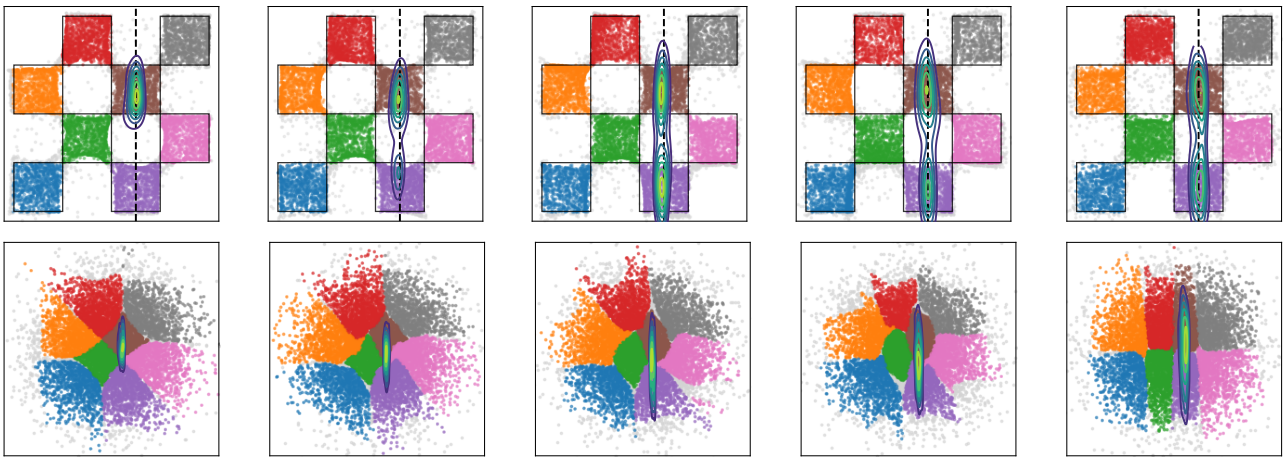
(a) `frozen-`θ *(b)* `unconstrained-`θ *(c)* VFM (no EMA) *(d)* VFM (no KL) *(e)* VFM

Figure 7. Ablation of VFM with respect to key modeling choices in the loss. Observation in black dots and $\sigma = 0.1$. For VFM (7c, 7d, 7e), we used $\tau = 100.0$, $\alpha = 1.0$ and $K = 4$. We observe that 7a: `frozen-`θ fails to capture the bimodal nature of the posterior; 7b: `unconstrained-`θ produces many off-manifold samples; 7c: removing EMA from the term $\mathcal{L}_{\text{obs}}$ in VFM also produces many off-manifold samples when τ is large; 7d: removing the KL term $\mathcal{L}_{\text{KL}}$ in the VFM loss leads to unstable optimization and results in poor approximations of both the prior and posterior.



(a) $\tau = 0.01$ *(b)* $\tau = 0.1$ *(c)* $\tau = 1.0$ *(d)* $\tau = 10.0$ *(e)* $\tau = 100.0$

Figure 8. Ablation of VFM with respect to the τ parameter. For each plot, we set $\alpha = 1.0$ and $K = 4$. We fix $y = 0.5$ and $\sigma = 0.1$. We observe that for $\tau \lesssim \sigma$, the quality of prior/posterior approximations are poor, yielding many off-manifold samples. This is likely due to the difficulty of optimization as we tighten the correspondence between x and z. For $\tau \geq 1$, we observe significant improvements in results and surprising robustness with respect to large values of τ.



(a) $\alpha = 0.0$ *(b)* $\alpha = 0.25$ *(c)* $\alpha = 0.5$ *(d)* $\alpha = 0.75$ *(e)* $\alpha = 1.0$

Figure 9. Ablation of VFM with respect to the α parameter. For each plot, we set $\tau = 100.0$ and $K = 4$. We observe that the warping of the latent space becomes stronger as $\alpha \to 1$, making it easier to sample from the bimodal posterior using the simple Gaussian variational posterior in latent space.

B.2. ImageNet experiment

In this section, we provide a detailed breakdown of the architectures, training objectives, and the extensive tuning process conducted for the baselines used in the ImageNet 256×256 experiments.

B.2.1. Model Architectures

Flow Map Backbone (f_θ). We employ a SiT-B/2 architecture (Ma et al., 2024) (130M parameters) initialized from a flow-matching model pre-trained for 80 epochs. Following the design of Decoupled Mean Flow (DMF) (Lee et al., 2025), we utilize decoupled encoder/decoder embeddings for the timesteps to better capture the flow dynamics. Our fine-tuning is performed for 100 epochs, making the total training process to be 180 epochs.

Noise Adapter (q_ϕ). To map high-dimensional observations y and inverse problem classes c to the latent noise distribution $q_\phi(z|y, c)$, we design a lightweight U-Net style adapter (10M parameters). The adapter is conditioned on the inverse problem class c using Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). The class embedding modifies the features at multiple resolutions via affine transformations $\gamma \cdot x + \beta$. The network processes the 256×256 input observation through a series of residual blocks and downsampling layers (channel multipliers: 1, 2, 4, 4), which compresses the spatial resolution to 32×32. The final projection layer outputs the mean μ and log-variance $\log \sigma^2$ (clamped between -10.0 and 2.0) for the latent distribution, from which we sample z using the reparameterization trick.

Latent space encoding. Modern generative models often operate in a lower-dimensional latent space obtained via an autoencoder or similar compression mechanism (Rombach et al., 2022). We adopt this setting in this experiment by defining the flow map and adapter in the latent space of SD-VAE (Rombach et al., 2022) rather than pixel space, and applying the forward operator to decoded samples. Measurement encoding can be incorporated directly into the adapter architecture; specifically, our U-Net-based architecture for the adapter maps the high-dimensional input observations into a lower-dimensional latent representation, which allows us to estimate the mean (μ) and variance (σ) within the latent space.

Training. For VFM training, we employ standard model guidance techniques during the flow map training phase. Following previous works (Tang et al., 2025; Lee et al., 2025), we utilize a prefixed CFG probability to redefine the target velocity, which allows us to perform robust one-step generation during sampling.

B.2.2. Baselines and Tuning

We compare VFM against a comprehensive suite of guidance-based solvers. A major challenge in this comparison is the high sensitivity of these methods to hyperparameters. To ensure a fair comparison, we performed an exhaustive hyperparameter sweep for every baseline, task, and backbone. We found that most inference-time methods require significant per-task tuning, which makes them computationally burdensome compared to the one-step nature of VFM.

Unless otherwise stated, all baselines use 250 ODE steps. To maximize their performance, we also applied Classifier-Free Guidance (CFG) with a scale of 2.0, which we found empirically boosts results across methods, even those that do not originally prescribe it.

Latent DPS (Chung et al., 2024). We extend Diffusion Posterior Sampling (DPS) to the latent flow matching setting. Through extensive sweeping, we identified a novel gradient scaling technique that provided the best stability. We normalize the likelihood gradient update to have a magnitude of 1, i.e., using a step size of $1/||\nabla_z \log p(y|z)||$.

Latent DAPS (Zhang et al., 2025). We implemented DAPS in the latent flow matching space, strictly following the original paper's settings. This involves 5 ODE rollout steps followed by 50 annealing steps and 50 Langevin steps, which makes the optimization extremely slow.

PSLD (Rout et al., 2023) Our implementation follows the original paper. We tuned the coefficients and found the optimal values to match the original recommendations, where DPS and gluing coefficients are chosen to be 1.0 and 0.1, respectively.

MPGD (He et al., 2023). We extended Manifold Preserving Guidance (MPGD) to flow matching, which approximates the Jacobian as identity. We utilized DDIM-type deterministic velocity maps and, similar to DPS, found that a gradient scaling of $1/||\nabla||$ yielded the best performance.

FlowChef (Patel et al., 2025). We followed the exact implementation from the original paper. After heavy tuning, we found that it behaved similarly to MPGD and performed best with the $1/||\nabla||$ gradient scaling.

FlowDPS (Kim et al., 2025). We followed the official implementation. Tuning revealed that a larger step size coefficient of $10.0/||\nabla||$ was optimal. We adhered to the original protocol of repeating the update 3 times per ODE iteration. Importantly, we disabled the stochasticity parameter as it was found to degrade performance, instead we relied on deterministic velocity updates.

B.2.3. Metrics and Evaluation

We evaluate performance using two distinct categories of metrics:

Pixel-Space Fidelity (PSNR/SSIM). While we report these standard metrics, we note that inference-time optimization methods (like DPS) tend to produce smooth estimates that maximize these scores by converging toward the conditional mean. This often results in a loss of high-frequency texture and realistic detail (Zhang et al., 2018).

Semantic and Distributional Fidelity (LPIPS, FID, MMD, CRPS). To assess whether the model captures the true posterior distribution rather than just the mean, we prioritize metrics in embedding space. We evaluate methods by using standard LPIPS and FID by using 1024 reconstructions from the validation set of ImageNet. We further evaluate Maximum Mean Discrepancy (MMD) metric in the embedding space of Inception network (used also in FID). This measures the distance between the true and approximate posterior distributions in the semantic space. To evaluate the generation quality along with its diversity (which is very important in posterior sampling and uncertainty quantification), we also use the Continuous Ranked Probability Score (CRPS) scoring rule. It assesses the calibration and coverage of the posterior. We compute this in the embedding spaces of both Inception and DINO models to ensure semantic consistency. Refer to Appendix B.1.3 for further details on the computation of MMD and CRPS.

We evaluate PSNR, SSIM, LPIPS, FID, and MMD on the randomly selected 1024 samples from the validation set of the ImageNet dataset. As for CRPS metric, we generate 10 different reconstructions of 128 samples from validation set. We follow this recipe for all the baselines and our VFM experiments, except for Latent DAPS, where, due to the slower generation we only generated 128 samples instead of 1024 (all the rest of the settings are followed as stated above).

Our results show that while baselines may achieve high PSNR/SSIM due to mean-seeking behavior, VFM significantly outperforms them on distributional metrics (FID, MMD, CRPS), which indicates superior perceptual quality and a more accurate approximation of the complex posterior. We also observe that generating multiple samples through VFM in 1-step and then taking the average smoothes the reconstructions, which achieves competitive or better PSNR/SSIM values as well.

Projection trick Measurement space projection is very common to improve the pixel-wise metrics (PSNR/SSIM) in guidance world. In most of the methods (also gluing term in PSLD), it is common to use projection to guide the samples further towards measurement space (Rout et al., 2023; Chung et al., 2022; Wang et al., 2022). Specifically, given that we have a generation z_0 and observation y, we can project generated samples by applying $\hat{z}_0 = \mathcal{E}(A^T y + (I - A^T A)\mathcal{D}(z_0))$, where $\mathcal{E}$ and $\mathcal{D}$ denotes encoder and decoder, respectively. We found this useful in inpainting and gaussian debluring tasks, where the 1-step output of VFM is corrected by this formula.

B.2.4. Inverse Problems and Evaluation Setup

We evaluate VFM and all baselines on a diverse set of standard linear inverse problems frequently used in the literature. Our VFM model was trained jointly to handle denoising, random inpainting, box inpainting, super-resolution, Gaussian deblurring, and motion deblurring via the amortized conditioning mechanism described in Section 3.2.

For quantitative evaluation, we focus on the structurally challenging tasks (inpainting, super-resolution, and deblurring) and omit pure denoising. To ensure a rigorous and fair comparison, all baselines utilize the exact same pre-trained SiT-B/2 backbone that was used to initialize VFM. This strictly isolates the performance differences to the sampling method (iterative guidance vs. one-step VFM) rather than the generative prior quality. Consequently, the reported numbers for VFM can serve as a reliable reference for future benchmarking on the SiT-B/2 architecture. We also followed the best practices from SiT-B/2 unconditional sampling to get the best results.

The specific forward operators for the evaluated tasks are defined as follows. For random inpainting, we apply a random

Task	Method	NFE	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	MMD (↓)	$\text{CRPS}_{\text{DINO}}$ (↓)	CRPS_{Inc} (↓)	Time (s) (↓)
Inpaint (random)	Latent DPS	250×2	26.01	0.721	0.337	55.81	0.113	0.472	0.363	7.2164
	Latent DAPS	250×2	25.09	0.671	0.384	–	–	0.474	**0.356**	44.347
	PSLD	250×2	25.63	0.713	0.338	56.13	0.123	0.462	0.386	10.286
	MPGD	250×2	**26.03**	0.720	0.339	55.82	0.112	0.470	0.363	7.3512
	FlowChef	250×2	26.01	0.720	0.338	55.73	0.111	0.471	0.364	7.3885
	FlowDPS	250×2	25.80	**0.729**	0.344	62.62	0.139	0.557	0.453	14.054
	`frozen-`θ	1	21.07	0.534	0.530	126.45	0.236	0.787	0.580	**0.015**
	VFM (ours)	1 / 10	23.59 / 24.89	0.598 / 0.677	0.367 / **0.336**	**51.35**	**0.110**	**0.447**	0.444	0.025 / 0.252
Super-res. (×4)	Latent DPS	250×2	23.91	0.641	0.388	68.73	0.154	0.554	0.447	7.4195
	Latent DAPS	250×2	21.73	0.511	0.473	–	–	0.575	0.400	44.424
	PSLD	250×2	23.92	0.639	0.401	74.59	0.169	0.565	0.453	10.375
	MPGD	250×2	23.93	0.642	0.388	69.01	0.157	0.553	0.446	7.3801
	FlowChef	250×2	23.91	0.641	0.388	68.63	0.154	0.553	0.447	7.4914
	FlowDPS	250×2	24.13	0.655	0.413	81.47	0.193	0.633	0.547	14.303
	`frozen-`θ	1	20.61	0.469	0.557	148.50	0.270	0.837	0.637	**0.015**
	VFM (ours)	1 / 10	22.69 / **24.16**	0.600 / **0.658**	**0.382**	**47.61**	**0.068**	**0.539**	**0.392**	**0.015** / 0.148
Motion deblur	Latent DPS	250×2	22.17	0.555	0.478	103.35	0.203	0.716	0.519	7.5214
	Latent DAPS	250×2	21.26	0.499	0.480	–	–	**0.558**	**0.392**	46.691
	PSLD	250×2	21.62	0.537	0.516	136.63	0.260	0.819	0.588	10.129
	MPGD	250×2	22.20	0.557	0.478	102.97	0.203	0.715	0.519	7.5031
	FlowChef	250×2	22.18	0.556	0.477	103.35	0.203	0.715	0.519	7.4681
	FlowDPS	250×2	**22.31**	**0.579**	0.498	122.09	0.240	0.804	0.597	14.715
	`frozen-`θ	1	18.30	0.348	0.651	214.29	0.365	1.099	0.720	**0.015**
	VFM (ours)	1 / 10	18.72 / 20.22	0.400 / 0.506	0.480 / **0.471**	**60.28**	**0.098**	0.683	0.421	**0.015** / 0.148

Table 2. Quantitative comparison on ImageNet for various inverse problems. Best results are in **bold**, second best are underlined. ↑: higher is better, ↓: lower is better.

noise mask where the occlusion probability is sampled uniformly from the interval $(0.3, 0.7)$ for each image. In the case of box inpainting, we utilize a rectangular mask with a random location and aspect ratio, where the height and width are sampled independently from the interval $(32, 128)$. Super-resolution (x4) is implemented by downsampling the input image by a factor of 4 using bicubic interpolation. Finally, for the deblurring tasks, we employ a 61×61 kernel size, using a standard deviation of $\sigma = 3.0$ for Gaussian deblurring and an intensity value of 0.5 for motion deblurring. Additionally, for all inverse problems, the measurements are further corrupted by additive Gaussian noise with a standard deviation of $\sigma = 0.05$.

B.3. Additional Results

In this section, we present a comprehensive set of qualitative results on ImageNet 256×256 to further validate the effectiveness of Variational Flow Maps.

Qualitative Comparisons. Figures 10, 11, 12, 13, 14 provide side-by-side comparisons of VFM against seven state-of-the-art baselines across five distinct inverse problems. In all cases, VFM produces sharp, coherent, and consistent samples in a single forward pass, whereas baselines often exhibit artifacts or require hundreds of function evaluations to achieve comparable fidelity.

Uncertainty Quantification. A key advantage of VFM is its ability to learn a proper posterior distribution rather than collapsing to a single mode. In Figure 16, we visualize the pixel-wise mean and standard deviation computed from multiple posterior samples (10 samples). The uncertainty maps clearly highlight that VFM localizes variance in ambiguous regions (e.g., occluded areas or fine details lost to blur), which provides valuable information about the posterior that is typically infeasible to extract with slow or mode-collapsing baselines.

Structured Noise ("Make Some Noise"). Figure 15 visualizes the internal operation of the noise adapter $q_\phi(z|y)$. We display the predicted latent mean μ, the standard deviation σ, and the resulting reparameterized noise samples z. We observe strong structural patterns in the learned noise, which indicates that the adapter actively aligns the latent space to the data manifold. This validates our core premise: by "learning the proper noise" via optimization, we bridge the guidance gap without requiring iterative steering.

Diversity and Mode Coverage. In Figures 18 and 19, we examine diverse generation scenarios. While the baselines

frequently fail or collapse to a single (often incorrect) solution, VFM successfully generates diverse, high-quality samples that are all consistent with the measurements. We observe that greater ill-posedness naturally leads to higher diversity in our generations, confirming that the model captures the multimodal nature of the posterior.

Unconditional Generation. Finally, Figure 17 presents additional curated unconditional samples generated by the trained flow map. It further highlights the generative quality of our backbone model.

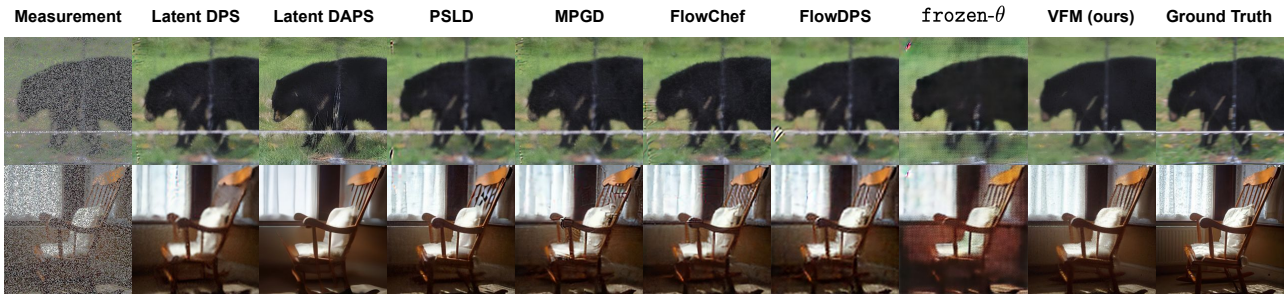


Figure 10. **Qualitative comparison on Random Inpainting.** We compare one-step VFM samples against seven baselines. VFM recovers fine details and texture consistent with the unmasked regions, while maintaining high perceptual quality.

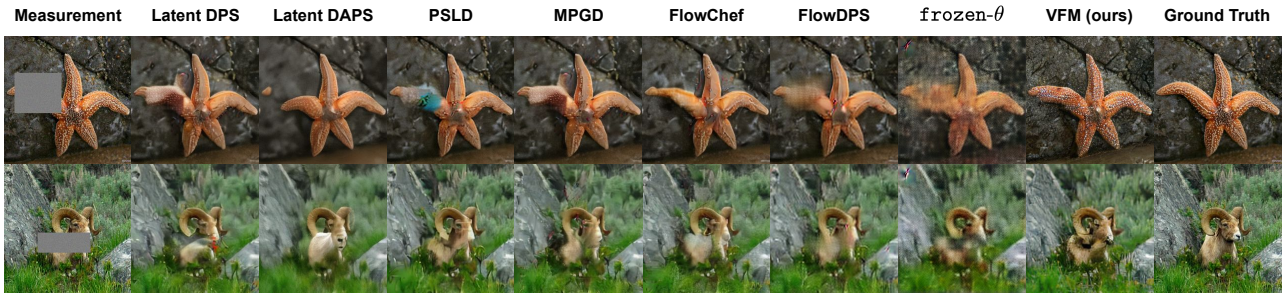


Figure 11. **Qualitative comparison on Box Inpainting.** Comparison of VFM against baselines for large occlusions. VFM generates plausible semantic content to fill the missing regions in a single step.

Figure 12. **Qualitative comparison on Super-Resolution** ($\times 4$)**.** VFM effectively upsamples the low-resolution inputs, leading to sharp edges and realistic textures compared to the often over-smoothed baseline results.

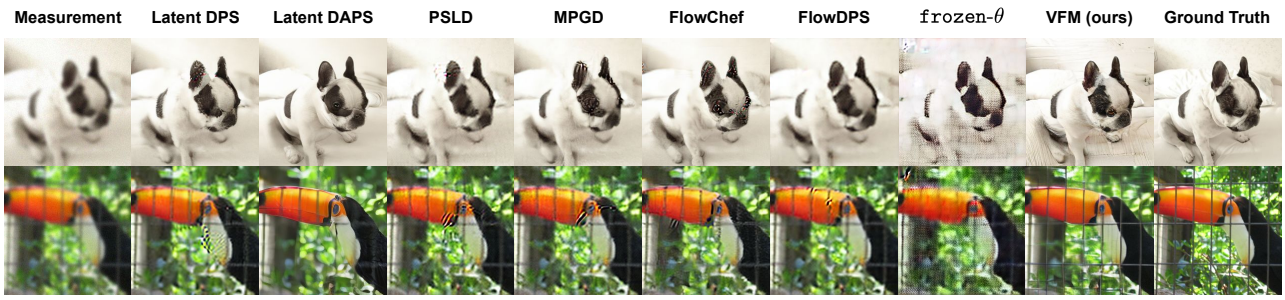


Figure 13. **Qualitative comparison on Gaussian Deblurring.** VFM successfully restores sharpness from heavily blurred observations ($\sigma = 3.0$), and it also avoids the artifacts common in guidance-based methods.

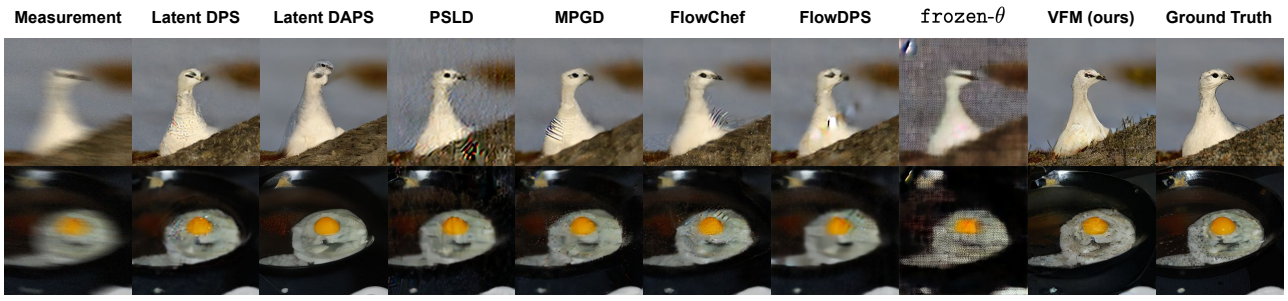


Figure 14. **Qualitative comparison on Motion Deblurring.** Comparison of deblurring performance on motion-blurred inputs. VFM resolves the motion streaks into coherent structures.

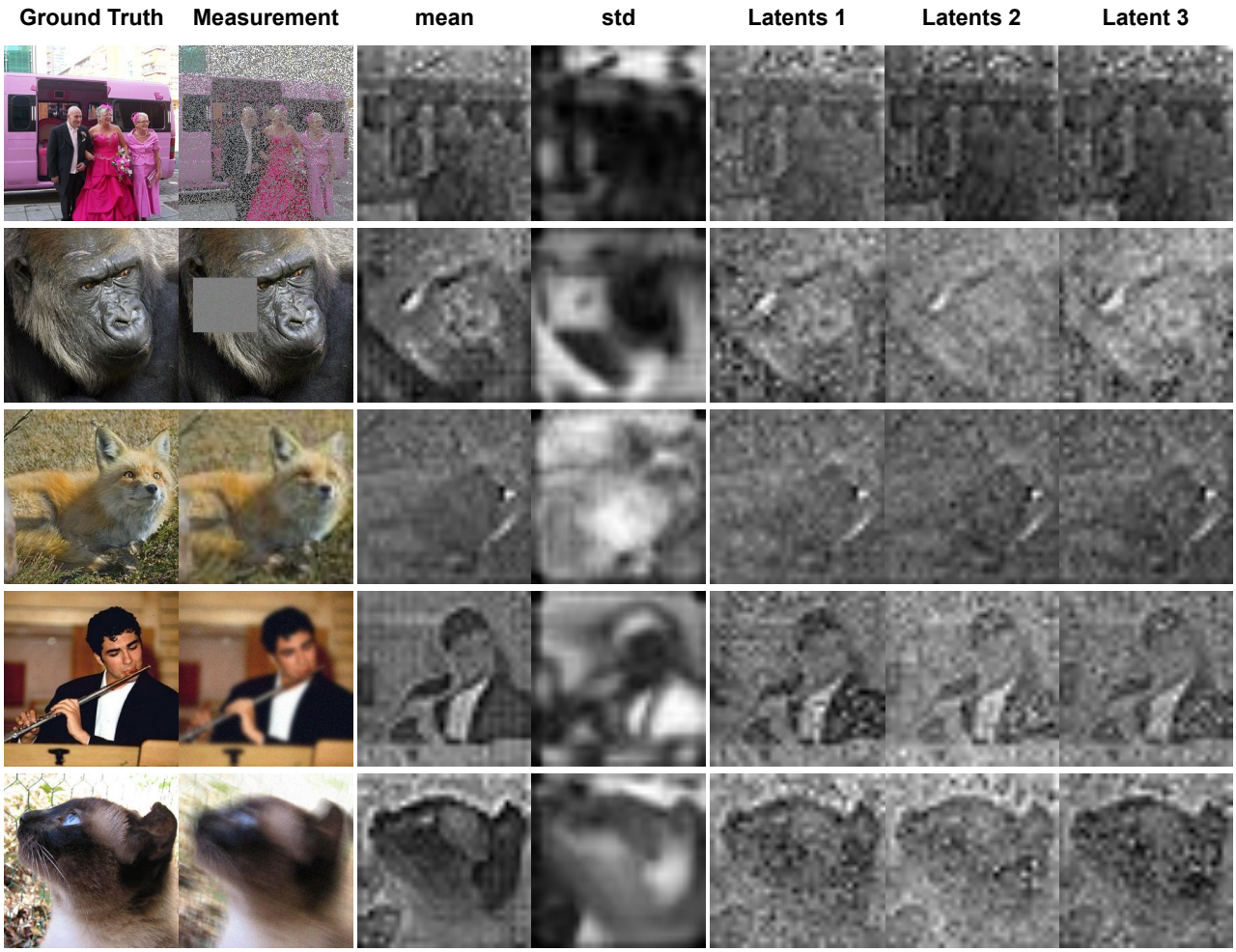


Figure 15. **Visualizing the Learned Noise Space.** We visualize the outputs of the noise adapter $q_\phi(z|y)$. From left to right: ground truth, measurement, the predicted latent mean μ, standard deviation σ, and three independent latent samples drawn from the distribution. The visible structure in the "noise" confirms that the adapter optimizes the latent initialization to align with the conditional data manifold. From top to bottom, rows correspond to: random inpainting, box inpainting, super-resolution, gaussian deblurring, and motion deblurring.

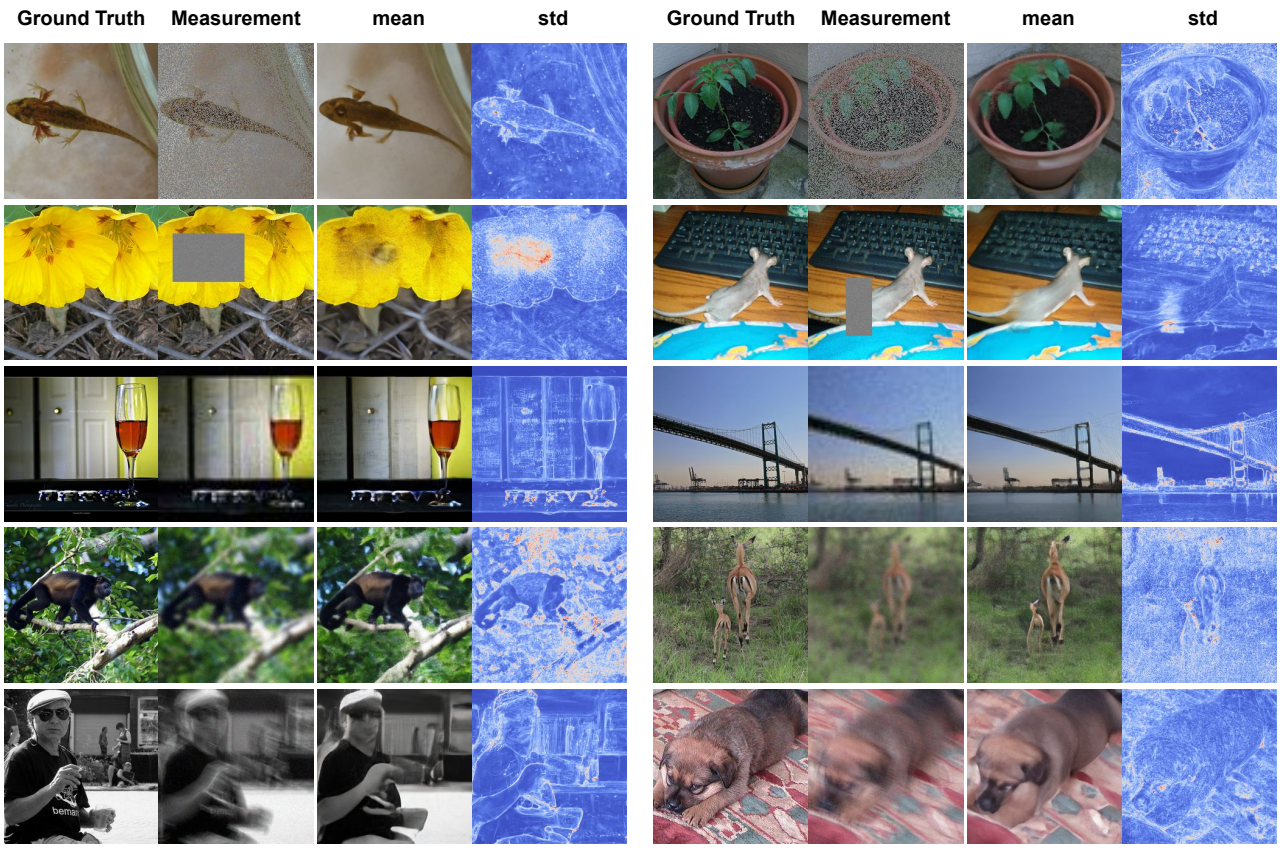


Figure 16. **Posterior Uncertainty Quantification.** We display the pixel-wise mean and standard deviation computed from 10 conditional samples generated by VFM. The standard deviation maps (right column) accurately capture the uncertainty inherent in the inverse problem, which highlights ambiguous regions where the model generates diverse solutions.

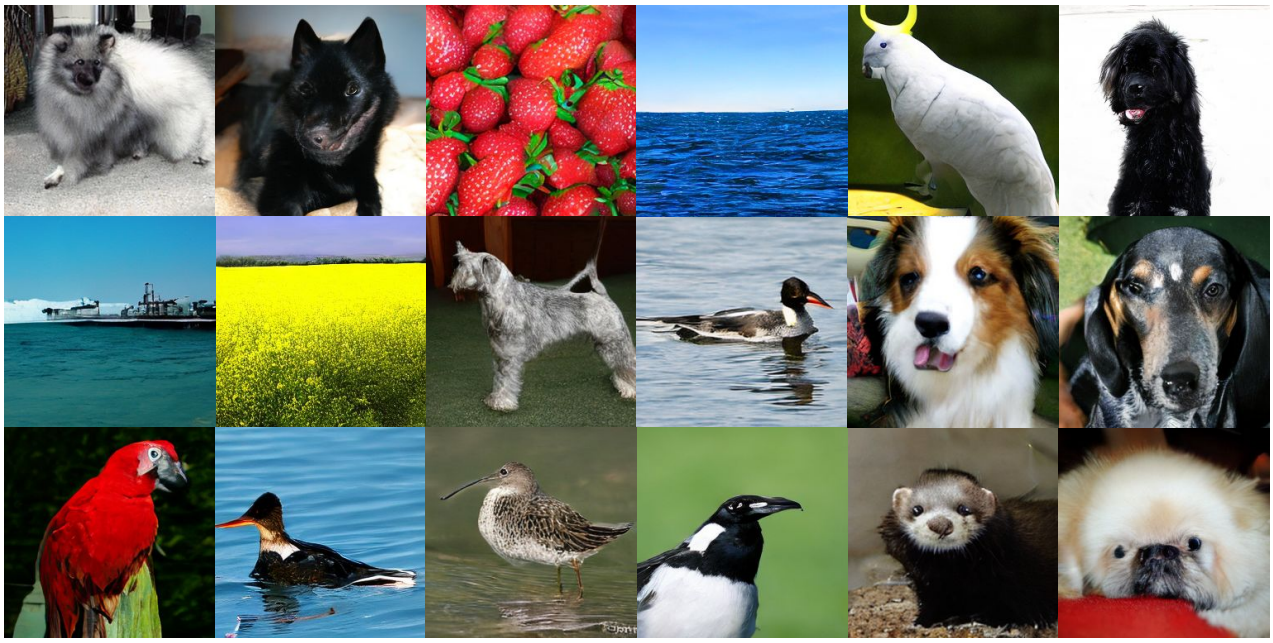


Figure 17. **Unconditional Samples.** Curated unconditional samples generated by the VFM.

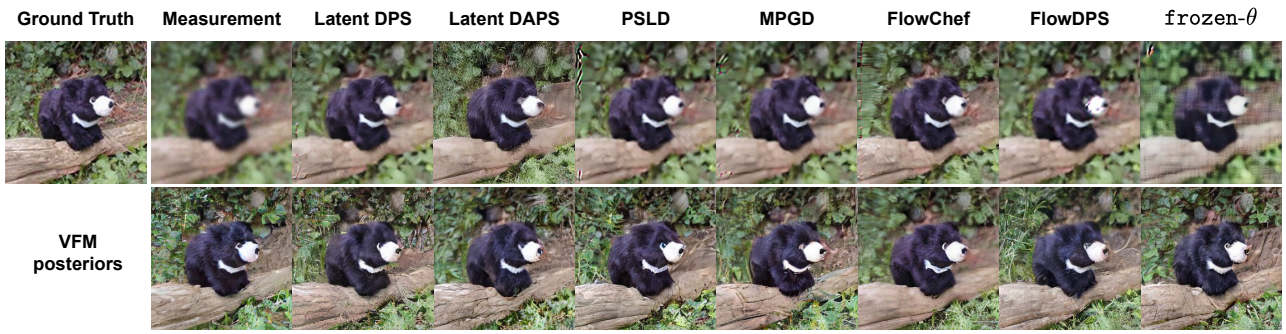


Figure 18. **Posterior Diversity (Sample Set 1).** Evaluation of sample diversity on gaussian deblurring. While baselines often collapse to a single mode or fail to produce valid results, VFM generates eight distinct, plausible, and measurement-consistent posterior samples.

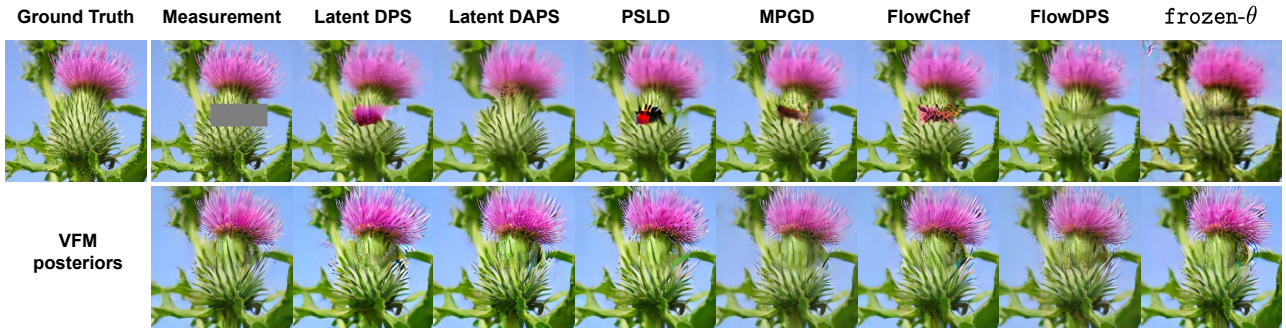


Figure 19. **Posterior Diversity (Sample Set 2).** Additional examples of diverse posterior sampling on box inpainting task. The high variance among the VFM samples reflects the multimodal nature of the posterior distribution for these ill-posed tasks.