TEST-TIME ANCHORING FOR DISCRETE DIFFUSION POSTERIOR SAMPLING

Litu Rout[1,2*] **Andreas Lugmayr**[1] **Yasamin Jafarian**[1] **Srivatsan Varadharajan**[1]
Constantine Caramanis[2] **Sanjay Shakkottai**[2] **Ira Kemelmacher-Shlizerman**[1]

[1] Google [2] UT Austin

{litu.rout,constantine,sanjay.shakkottai}@utexas.edu
{liturout,alugmayr,jafarian,srivatsanv,kemelmi}@google.com

ABSTRACT

We study the problem of posterior sampling using pretrained discrete diffusion foundation models, aiming to recover images from noisy measurements without retraining task-specific models. While diffusion models have achieved remarkable success in generative modeling, most advances rely on continuous Gaussian diffusion. In contrast, discrete diffusion offers a unified framework for jointly modeling categorical data such as text and images. Beyond unification, discrete diffusion provides faster inference, finer control, and principled training-free Bayesian inference, making it particularly well-suited for posterior sampling. However, existing approaches to discrete diffusion posterior sampling face severe challenges: derivative-free guidance yields sparse signals, continuous relaxations limit applicability, and split Gibbs samplers suffer from the curse of dimensionality. To overcome these limitations, we introduce Anchored Posterior Sampling (APS) for *masked diffusion* foundation models, built on two key innovations—*quantized expectation* for gradient-like guidance in discrete embedding space, and *anchored remasking* for adaptive decoding. Our approach achieves state-of-the-art performance among discrete diffusion samplers across linear and nonlinear inverse problems on the standard benchmarks. We further demonstrate the benefits of our approach in training-free stylization and text-guided editing.

1 INTRODUCTION

Diffusion models have become the state-of-the-art across a wide range of generative tasks, including images (Ramesh et al., 2021; Rombach et al., 2022; Baldridge et al., 2024; Esser et al., 2024; Black Forest Labs, 2024), audio (Huang et al., 2023; Veo, 2025), and video (Singer et al., 2023; OpenAI, 2024; Veo, 2025). Most of this progress has been driven by *continuous* diffusion models, where Gaussian noise is gradually added in pixel or latent space and then reversed by a learned denoiser (Sohl-Dickstein et al., 2015; Ho et al., 2020). Recently, however, *discrete* diffusion has emerged as a powerful alternative, showing superior performance in modeling categorical distributions such as text (Lou et al., 2024; Sahoo et al., 2024; Shi et al., 2024; Nie et al., 2025; Rout et al., 2025a) and images (Shi et al., 2024; Yang et al., 2025). Discrete diffusion further enables a unified framework for both image and text generation, supporting multimodal generation and editing.

Beyond unification, discrete diffusion offers several advantages over continuous diffusion that are particularly relevant for posterior sampling. First, it achieves *faster inference*, often generating high-quality samples in significantly fewer reverse steps (Shi et al., 2024; Schiff et al., 2025; Ma et al., 2025). Second, it provides *finer control*: the model predicts a normalized categorical distribution per token (e.g., a pixel or a patch), which decouples different parts of the image, unlike Gaussian diffusion where the entire image is coupled. Third, it enables *training-free posterior sampling*: since the model outputs full conditional distributions at each step, these can be reweighted by the likelihood to yield a better posterior estimate posterior (Murata et al., 2024; Chu et al., 2025). This property unlocks precise image editing and solving inverse problems without additional training (§4), motivating our approach to use discrete diffusion model (Yang et al., 2025) as a prior.

*This work was done during an internship at Google.

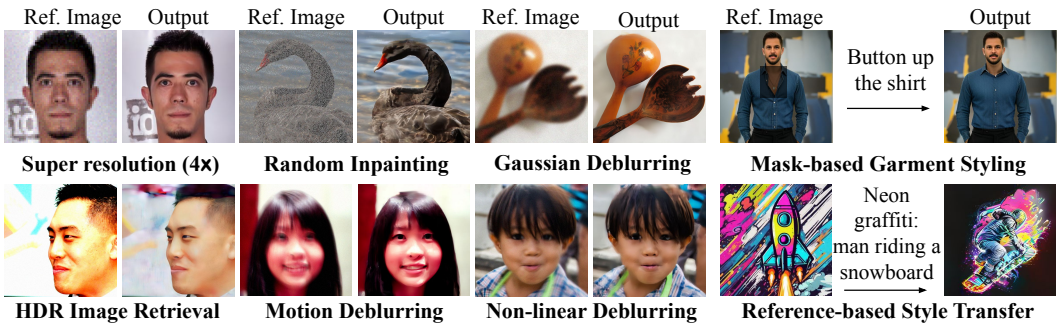

Figure 1: We introduce Anchored Posterior Sampling (APS) for *masked diffusion* foundation models, built on two key innovations: (i) *quantized expectation*, which provides gradient-like guidance in discrete embedding space, and (ii) *anchored remasking*, which enables adaptive decoding. Our method supports a variety of linear and nonlinear image restoration tasks (left three columns), as well as mask-based garment styling and reference-guided style transfer (last column).

State-of-the-art posterior samplers (Rout et al., 2023; 2024; Chung et al., 2024; Zhang et al., 2025) use continuous diffusion as a prior. These approaches rely on guiding the reverse diffusion process using likelihood gradients in continuous latent spaces (Rout et al., 2023; 2024; Chung et al., 2024; Zhang et al., 2025). This is infeasible for *discrete* diffusion due to non-differentiability in token space. Derivative-free discrete methods (Li et al., 2024) inspired by reinforcement learning provide weak signals. G2D2 (Murata et al., 2024) uses Gumbel-softmax relaxation but it is limited to discrete tokens with continuous embeddings. SGDD (Chu et al., 2025) introduces a split Gibbs sampler, but suffers from exponential complexity in sequence length (Chewi, 2023). Moreover, these methods unmask tokens in *random order*, which is suboptimal compared to adaptive decoding strategies in language modeling (Yang et al., 2025; Rout et al., 2025a). These limitations underscore the need for a discrete diffusion posterior sampler with adaptive decoding, leveraging next-generation multimodal models (Yang et al., 2025; Gemini Team, 2024).

Existing discrete diffusion methods explore posterior sampling under uniform (Chu et al., 2025) or mixed-noise processes (Murata et al., 2024), but these methods are tied to specific noising schemes and lack generalization (§4). Meanwhile, recent advances show that *masked diffusion* achieves state-of-the-art performance in image generation (Ma et al., 2025; Yang et al., 2025), yet its potential for posterior sampling remains underexplored. In this work, we take the first step towards leveraging multimodal masked diffusion models such as MMaDA (Yang et al., 2025) for posterior sampling. We introduce two key algorithmic innovations: (i) *quantized expectation*, which provides gradient-like guidance in purely discrete embedding space by updating the full conditional probability table (§3.2.1), and (ii) *anchored remasking*, an adaptive inference strategy that decodes important "anchor" tokens early in the reverse process (§3.2.2). Together, these techniques overcome the non-differentiability challenge in discrete diffusion and yield scalable posterior samplers. Extensive experiments (§4) show up to 31.36% LPIPS and 7.05% PSNR improvements on linear and nonlinear inverse problems, and illustrate training-free stylization results, as shown in Figure 1.

Our contributions are summarized below.

- **Theoretical results:** we derive (i) a *training* upper bound, $\mathcal{L}_{\text{DDPS}}$ (**Theorem 3.1**), that integrates measurements into the reverse diffusion process, and (ii) a *test-time* bound, $\mathcal{L}_{\text{APS}}$ (**Theorem 3.2**), that reuses a pretrained denoiser without expensive retraining per downstream task (§3).
- **Quantized expectation:** a novel strategy to update *all* entries of the conditional probability table, enabling strong *gradient-like* guidance in the discrete embedding space (§3.2.1).
- **Anchored remasking:** an adaptive decoding strategy that unmasks "anchor" tokens early in the reverse process, better utilizing model's capacity to decode remaining tokens (§3.2.2).
- **Extensive experiments:** comprehensive evaluation on linear (super resolution, Gaussian deblurring, inpainting, motion deblur) and nonlinear (HDR, nonlinear deblurring) inverse problems on FFHQ and ImageNet, where our method achieves up to 35.82% LPIPS and 10.94% PSNR gains on ImageNet super resolution (4×), and 31.36% LPIPS and 7.05% PSNR gains on FFHQ, over the prior state-of-the-art discrete sampler. We further demonstrate *training-free stylization* enabled by our discrete posterior sampler, highlighting flexibility beyond inverse problems (§4).

2 Related Works

Gaussian pixel-space diffusion methods (Chung et al., 2023; Kawar et al., 2022; Zhu et al., 2023) address noisy inverse problems by guiding the reverse process of diffusion models trained directly in pixel space. These priors are domain-specific—for instance, a model trained on ImageNet must be used for ImageNet tasks, and one trained on FFHQ for FFHQ tasks—yielding highly informative but narrow priors. Mixing domains during training dilutes this information, and at the extreme of internet-scale training, the prior remains valid for generation (Ramesh et al., 2021; Baldridge et al., 2024; Esser et al., 2024; Black Forest Labs, 2024) but becomes less informative for specific domains. This motivates the challenge of extracting domain-relevant information from general-purpose priors.

Gaussian latent diffusion. PSLD (Rout et al., 2023) introduced posterior sampling with latent diffusion, showing how domain-specific priors can be extracted from large-scale foundation models. This initiated a new line of work (Chung et al., 2024; Song et al., 2024; Rout et al., 2024; 2025b; Noroozi et al., 2024; Zhang et al., 2025; Chung et al., 2025) leveraging the advantages of latent diffusion: a single pretrained model can handle multiple domains, enabling inverse problems and semantic edits without retraining, while also being faster and more scalable to high-resolution synthesis. A drawback, however, is that posterior sampling often requires backpropagation through large denoisers (e.g., Flux (Black Forest Labs, 2024), SD3.5 (Esser et al., 2024)), which is prohibitively slow. RB-Modulation (Rout et al., 2025c) alleviates this by framing the problem as stochastic optimal control, directly optimizing the terminal latent state and reducing runtime from several minutes (PSLD: $\sim$12 min, P2L: $\sim$30 min, STSL: $\sim$3 min) to under 40 seconds. This efficiency relies on continuous, differentiable latent embeddings, a property that does not extend to discrete diffusion. Addressing this gap motivates the need for new posterior sampling approaches in discrete settings.

Uniform discrete diffusion. Recent works have explored posterior sampling with discrete diffusion. G2D2 (Murata et al., 2024) extends the proximal sampler of RB-Modulation to VQ-diffusion (Gu et al., 2022) using a star-shaped noising process and Gumbel-Softmax dequantization (Gumbel, 1954; Jang et al., 2017; Maddison et al., 2017). While it enables gradient guidance, G2D2 depends on continuous relaxations, requires storing log-probabilities from previous step, and struggles to generalize to purely discrete token embeddings (§4). SGDD (Chu et al., 2025) instead proposes a split Gibbs sampler with Hamming-distance reweighting and rejection sampling via Metropolis–Hastings, but its exponential rejection rate restricts results to low-resolution tasks.

Masked (absorbing) discrete diffusion. While G2D2 and SGDD can, in principle, be adapted to masked diffusion, they perform poorly with purely discrete token embeddings. In contrast, our method leverages a unified masked discrete diffusion model and introduces two key components: *quantized expectation* (§3.2.1) and *anchored remasking* (§3.2.2). Together, these yield an efficient and scalable posterior sampler for high-resolution inverse problems. To our knowledge, this is the first inverse problem solver tailored for masked discrete diffusion with purely discrete embeddings, outperforming prior discrete samplers and remaining competitive—often superior—to continuous diffusion methods at substantially lower inference cost (§4).

3 Method

3.1 Preliminaries

Visual Tokenizer. A cornerstone of modern image tokenization is the VQ-VAE (Van Den Oord et al., 2017), which maps images into discrete codebook indices. It consists of an encoder $\mathcal{E} : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{h\times w\times d}$ that projects an image into a latent embedding of size $h \times w \times d$. The embeddings are reshaped into a sequence $\mathbf{e} \in \mathbb{R}^{L\times d}$ of length $L = h \times w$. We denote images by $\hat{\mathbf{x}}$. The encoder produces $\mathbf{e} = \mathcal{E}(\hat{\mathbf{x}})$ where each embedding $\mathbf{e}^l$ for $l = 1, \ldots, L$ is quantized to the nearest codebook entry $\mathbf{c}_j \in \mathcal{C}$, where $\mathcal{C} \in \mathbb{R}^{K\times d}$ is a learned codebook:

$$\mathcal{Q}_{\text{vq}}(\mathbf{e}^l) \coloneqq \mathbf{c}_j, \quad j = \underset{k\in\{1,\ldots,K\}}{\arg\min} \left\|\mathbf{e}^l - \mathbf{c}_k\right\|_2. \tag{1}$$

Codebook entries may be continuous vectors $\mathbf{c}_j \in \mathbb{R}^d$ or purely discrete binary embeddings $\mathbf{c}_j \in \{-1, +1\}^d$ used in this paper. Binary embeddings are particularly appealing because prior studies (Yu et al., 2024) show that masked diffusion models degrade in generation quality as continuous vocabulary size grows, yet lookup-free quantization (LFQ) with binary embeddings achieves

both strong generation and reconstruction quality. In LFQ, the codebook is not learned but obtained by thresholding the encoder output:

$$\mathcal{Q}_{\text{lfq}}(\mathbf{e}^l) \coloneqq \text{sign}(\mathbf{e}^l), \text{ where } [\mathcal{Q}_{\text{lfq}}(\mathbf{e}^l)]_i = \begin{cases} +1 & \text{if } \mathbf{e}^l[i] > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{2}$$

Finally, each image is represented as a sequence of tokens $x = (x^1, \ldots, x^L)$ corresponding to the selected indices from (1) as $x^l = j$, or equivalently for LFQ the token index is obtained by $x^l = j = \sum_{i=1}^{d} 2^{i-1} \mathbf{1}_{\{\mathbf{e}_l[i]>0\}}$. Equivalently, we represent each sequence as a mixture of one-hot vectors $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L)$, where $\sum_{k=1}^{K} \mathbf{x}^l[k] = 1, \mathbf{x}^l[k] \geq 0$ and $\mathbf{x}^l[x^l] = 1$. This discrete token representation forms the basis of our masked diffusion posterior sampler.

Notation: We use '$\coloneqq$' to indicate architectural and parameterization choices in our model, to distinguish from '$=$' which are identities that follow from mathematical derivations.

Masked Diffusion. We now describe masked discrete diffusion, which defines a generative model over the discrete state space $\mathcal{S} = \mathcal{V}^L$, where $\mathcal{V} = \{1, \ldots, K, K+1\}$ consisting of K codebook indices and a special `[MASK]` token $\mathbf{m}$ corresponding to index $K+1$. Let $X = \{x^l\}_{l=1}^L \in \mathcal{S}$ denote a sequence of tokens (or equivalently its one-hot representation $\mathbf{x} = \{\mathbf{x}^l\}_{l=1}^L$). The data distribution is denoted by $q(\cdot)$ over $\mathcal{S}$. The goal of masked diffusion is to learn a generative model that samples from $q(\cdot)$. Masked diffusion models (MDMs) (Austin et al., 2021; Lou et al., 2024; Shi et al., 2024; Sahoo et al., 2024) construct a discrete-time Markov chain with T steps, parameterized by α_t with $t \in [0, 1]$. The forward process gradually replaces each token with the `[MASK]` token:

$$q(\mathbf{z}_t|\mathbf{x}) = \prod_{l=1}^{L} q(\mathbf{z}_t^l|\mathbf{x}), \quad q(\mathbf{z}_t^l|\mathbf{x}) = \text{Cat}\big(\mathbf{z}_t^l; \alpha_t \mathbf{x}^l + (1-\alpha_t)\mathbf{m}\big), \tag{3}$$

where $\mathbf{z}_t^l$ is either preserved from $\mathbf{x}^l$ with probability α_t or replaced with $\mathbf{m}$ otherwise. The corresponding reverse process is parameterized by a neural network p_θ, which predicts categorical distributions over tokens. For each token position l, the transition distribution is

$$p_\theta(\mathbf{z}_s^l|\mathbf{z}_t) \coloneqq q(\mathbf{z}_s^l|\mathbf{z}_t^l, \mathbf{x}_\theta(\mathbf{z}_t)) = \begin{cases} \text{Cat}(\mathbf{z}_s^l; \mathbf{z}_t^l), & \mathbf{z}_t^l \neq \mathbf{m}, \\ \text{Cat}\Big(\mathbf{z}_s^l; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\theta(\mathbf{z}_t) + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\Big), & \mathbf{z}_t^l = \mathbf{m}, \end{cases} \tag{4}$$

where $\mathbf{x}_\theta(\mathbf{z}_t)$ denotes the network prediction. Training minimizes the negative evidence lower bound (NELBO) by aligning the reverse transition $p_\theta(\mathbf{z}_s^l|\mathbf{z}_t)$ with the inference posterior $q(\mathbf{z}_s^l|\mathbf{z}_t^l, \mathbf{x})$ derived from the forward process (3). Concretely, the training objective $\mathcal{L}_{\text{NELBO}}(\mathbf{x}; \theta) \coloneqq$

$$\mathbb{E}_{Z_0 \sim q(\cdot|\mathbf{x})}\big[-\log p_\theta(\mathbf{x}|Z_0)\big] + \sum_{i=1}^{T} \mathbb{E}_{Z_t \sim q(\cdot|\mathbf{x})}\Big[\frac{\alpha_t - \alpha_s}{1-\alpha_t} \sum_{l=1}^{L} \log\langle \mathbf{x}_\theta^l(Z_t), \mathbf{x}^l\rangle \mathbf{1}_{\{Z_t^l = \mathbf{m}\}}\Big], \tag{5}$$

where, for brevity, we drop i from $t(i) = i/T$ and $s(i) = (i-1)/T$.

3.2 Test-Time Anchored Posterior Sampling

This section develops our theoretical framework for posterior sampling with discrete diffusion. We first derive a training-based objective (**Theorem 3.1**) that incorporates measurements into the reverse diffusion process, and then present a lightweight test-time training bound (**Theorem 3.2**) that reuses a pretrained denoiser without retraining. Building on this, we introduce two key mechanisms—*Quantized Expectation* (§3.2.1) for differentiable likelihood evaluation and *Anchored Remasking* (§3.2.2) for adaptive unmasking—that enable scalable and accurate posterior sampling.

In posterior sampling, our goal is to construct a Markov chain whose stationary distribution coincides with the Bayesian posterior: $q(\mathbf{x}|\mathbf{y}) \propto q(\mathbf{y}|\mathbf{x})\, q(\mathbf{x})$, where $\mathbf{y} = \mathcal{A}(\mathcal{D}(\mathbf{x})) + \sigma\varepsilon$ with measurement operator $\mathcal{A}(\cdot)$, image decoder $\mathcal{D}(\cdot)$, Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$, and standard deviation σ. When $\mathcal{A}$ is linear the task reduces to a *linear inverse problem*; otherwise a *nonlinear inverse problem*. We approximate $q(\mathbf{x}|\mathbf{y})$ with a tractable sampler $p_\varphi(\mathbf{x}|\mathbf{y})$ using only a masked diffusion model $p_\theta(\mathbf{x})$ previously trained to approximate $q(\mathbf{x})$.

To sample from the posterior $q(\cdot|\mathbf{y})$, we construct a Markov chain with the joint distribution defined as: $p_\varphi(\mathbf{x}, \mathbf{z}_{0:1}|\mathbf{y}) = p_\varphi(\mathbf{z}_1|\mathbf{y})\, p_\varphi(\mathbf{x}|\mathbf{z}_0, \mathbf{y}) \prod_{i=1}^{T} p_\varphi\big(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{y}\big)$, where $\mathbf{z}_{0:1} = \mathbf{z}_0, \mathbf{z}_{1/T}, \ldots, \mathbf{z}_1$. We parameterize measurement conditional transitions by tilting the $\mathbf{y}$-unconditional transition (4) with the likelihood of measurements given the current estimate:

$$p_\varphi(\mathbf{z}_s|\mathbf{z}_t,\mathbf{y}) \coloneqq \prod_{l=1}^{L} p_\varphi(\mathbf{z}_s^l|\mathbf{z}_t,\mathbf{y}),\ \text{where}\ p_\varphi(\mathbf{z}_s^l|\mathbf{z}_t,\mathbf{y}) \ \coloneqq\ q\big(\mathbf{z}_s^l|\mathbf{z}_t^l,\mathbf{x}_\varphi(\mathbf{z}_t)\big)\, q\big(\mathbf{y}|\mathbf{x}_\varphi(\mathbf{z}_t)\big). \tag{6}$$

Theorem 3.1 (Discrete Diffusion Posterior Sampling (DDPS))**.** *Given a sample* $\mathbf{x}\sim q$*, let* $q(Z_{0:1}|\mathbf{x})$ *denote the forward noising law of (3). Then, for any measurement* $\mathbf{y}\sim q(\cdot|\mathbf{x})$*, the negative log-posterior is bounded by* $-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathcal{L}_{\mathrm{DDPS}}(\mathbf{x},\mathbf{y};\varphi) \coloneqq \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0)]+$

$$\sum_{i=1}^{T}\mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\Big[\frac{\alpha_{t(i)}-\alpha_{s(i)}}{1-\alpha_{t(i)}}\sum_{l=1}^{L}\log\langle\mathbf{x}_\varphi^l(Z_{t(i)}),\mathbf{x}^l\rangle\mathbf{1}_{\{Z^l_{t(i)}=\mathbf{m}\}}\Big]-\sum_{i=1}^{T}\mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big].$$

Implications. Theorem 3.1 shows that $\mathcal{L}_{\mathrm{DDPS}}(\mathbf{x},\mathbf{y};\varphi)$ is a principled training criterion for discrete posterior samplers. The likelihood-based tilt $\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))$ enforces measurement consistency When $\mathbf{y}$ is absent, the tilting terms vanish and the objective reduces to the standard masked diffusion NELBO (5). The negative cross-entropy term $\log\langle\mathbf{x}_\varphi^l(Z_{t(i)}),\mathbf{x}^l\rangle$ is zero for revealed tokens and gets supervision for masked tokens, with weights determined by the noise schedule.

For retraining, one can minimize $\mathcal{L}_{\mathrm{DDPS}}(\mathbf{x},\mathbf{y};\varphi)$ with respect to φ to obtain a discrete posterior sampler. In practice, however, retraining a large-scale foundation model per task is often infeasible due to excessive compute and lack of training data. We therefore focus on the training-free case.

Theorem 3.2 (Test-time Anchored Posterior Sampling (APS))**.** *Given a sample* $\mathbf{x}\sim q$*, let* $q(Z_{0:1}|\mathbf{x})$ *denote the forward noising law of (3). Suppose the pretrained network* $p_\theta(\mathbf{x})$ *closely approximates the unconditional prior* $q(\mathbf{x})$*. Then, for any measurement* $\mathbf{y}\sim q(\cdot|\mathbf{x})$*, the negative log-posterior is bounded by* $-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathcal{L}_{\mathrm{APS}}(\mathbf{x},\mathbf{y};\varphi) \coloneqq \mathcal{L}_{\mathrm{NELBO}}(\mathbf{x};\theta)+$

$$\sum_{i=1}^{T}\mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\Big[\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\sum_{l=1}^{L}\log\frac{\langle\mathbf{x}_\theta^l(Z_{t(i)}),\mathbf{x}^l\rangle}{\langle\mathbf{x}_\varphi^l(Z_{t(i)}),\mathbf{x}^l\rangle}\,\mathbf{1}_{\{Z^l_{t(i)}=\mathbf{m}\}}\Big]-\sum_{i=1}^{T}\mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big].$$

Implications. Theorem 3.2 shows that posterior sampling can be performed *without retraining* by reusing a pretrained masked diffusion model.

- *Efficient test-time training.* The bound $\mathcal{L}_{\mathrm{APS}}(\mathbf{x},\mathbf{y};\varphi)$ is expressed in terms of the pretrained NELBO from (5). Since this term is constant with respect to the new parameters φ, it can be ignored during optimization. As a result, test-time training only needs to update the lightweight adaptation and measurement-consistency terms, while reusing the fixed pretrained network $\mathbf{x}_\theta(\cdot)$. This avoids backpropagation through the large-scale denoiser (e.g., billions of parameters), making posterior sampling feasible and efficient at test time.
- *Training-free inference.* Although efficient test-time training requires paired $(\mathbf{x},\mathbf{y})$ data, in posterior sampling we only observe $\mathbf{y}$. Interestingly, the bound $\mathcal{L}_{\mathrm{APS}}$ points to training-free inference by substituting the pretrained model prediction $\mathbf{x}_\theta^l(Z_{t(i)})$ in place of $\mathbf{x}^l$ (see §3.2.1). This makes posterior sampling feasible directly from measurements, without retraining or labeled pairs.
- *Adaptation gap.* The log-ratio terms capture the mismatch between unconditional predictions $\mathbf{x}_\theta(Z_{t(i)})$ and adapted posterior predictions $\mathbf{x}_\varphi(Z_{t(i)})$, active only at masked positions.
- *Measurement consistency.* The final summation enforces agreement with the measurement likelihood, ensuring the sampler produces samples consistent with observed measurements $\mathbf{y}$.

Discussion. In summary, retraining a new network φ for every task would require backpropagation through massive denoisers (e.g., 8B parameters in MMaDA), which is computationally prohibitive. Our theoretical results show that posterior sampling can be done efficiently by reusing the pretrained model and optimizing only lightweight parameters at test time. Next, we introduce two key ingredients—*Quantized Expectation* (§3.2.1) and *Anchored Remasking* (§3.2.2)—that make this training-free posterior sampling practically implementable. These two ideas together form our **Algorithm 1: Anchored Posterior Sampling (APS)**; please see Appendix B.2 for a detailed discussion.

3.2.1 Quantized Expectation

We propose *quantized expectation* to compute $\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))$ in a differentiable manner. Building on the *training-free inference* implication of **Theorem 3.2**, we note that in posterior sampling

Table 1: **Quantitative results on Super Resolution (4×) and Gaussian Deblurring.** APS consistently outperforms prior discrete samplers (G2D2, SGDD, and SVDD-PM) and remains competitive with strong continuous diffusion baselines (shaded gray).

Type	Method	(a) FFHQ				(b) ImageNet			
		SR (4×)		Deblur		SR (4×)		Deblur	
		LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑
Pixel	DPS	0.269	25.86	0.219	25.87	0.367	22.61	0.443	19.04
	DDRM	0.282	26.58	0.239	24.93	0.352	24.00	0.246	27.30
	DiffPIR	0.260	26.64	0.236	27.36	0.371	23.18	0.355	22.80
	DAPS	0.177	29.07	0.165	29.19	0.276	25.89	0.253	26.15
Latent	PSLD	0.276	27.62	0.304	27.37	0.332	24.43	0.365	24.04
	ReSample	0.507	22.98	0.329	25.69	0.382	22.63	0.438	22.32
	LatentDAPS	0.182	27.48	0.234	27.93	0.276	25.06	0.345	25.05
Uniform (Mask)	SVDD-PM	0.594	12.08	–	–	–	–	–	–
	G2D2	0.271	26.93	0.287	26.35	0.349	23.20	0.375	22.71
	SGDD	0.288	25.85	–	–	–	–	–	–
Mask	**APS**	**0.234**	**27.50**	**0.276**	**27.90**	**0.324**	**24.30**	**0.375**	**24.71**
	APS-L	**0.186**	**28.83**	**0.241**	**29.50**	**0.224**	**25.74**	**0.282**	**26.35**

only the measurements $\mathbf{y}$ are observed. Hence, we replace the ground-truth token $\mathbf{x}^l$ with the token mixture predicted by the pretrained model $\mathbf{x}^l_\theta(Z_{t(i)})$, which simplifies the upper bound to

$$\hat{\mathcal{L}}_{\text{APS}}(\mathbf{x},\mathbf{y};\varphi) = \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \left[\frac{\alpha_{t(i)} - \alpha_{s(i)}}{1-\alpha_{t(i)}} \sum_{l=1}^{L} \log \langle \mathbf{x}^l_\varphi(Z_{t(i)}), \mathbf{x}^l_\theta(Z_{t(i)}) \rangle \, \mathbf{1}_{\{Z^l_{t(i)}=\mathbf{m}\}} \right] - \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \left[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)})) \right] + \text{constants}. \tag{7}$$

The measurement likelihood takes the form $q(\mathbf{y}|\mathbf{x}) \propto \exp(-\frac{\|\mathbf{y}-\mathcal{A}(\mathcal{D}(\mathbf{x}))\|_2^2}{2\sigma^2})$, where $\mathcal{D}(\cdot)$ denotes the image decoder. The denoiser outputs token probabilities $\mathbf{x}_\varphi(\mathbf{z}_{t(i)}) = \{\mathbf{x}^l_\varphi(\mathbf{z}_{t(i)})\}_{l=1}^L$. Since $\mathcal{D}$ takes a sequence of discrete tokens (or their one-hot encodings) as input, a naive approach would sample $\mathbf{x} = (\mathbf{x}^1, \ldots, \mathbf{x}^L) \sim \mathbf{x}_\varphi(\mathbf{z}_{t(i)})$ and compute $q(\mathbf{y}|\mathbf{x}_\varphi(\mathbf{z}_{t(i)})) \propto \exp(-\frac{\|\mathbf{y}-\mathcal{A}(\mathcal{D}(\mathbf{x}))\|_2^2}{2\sigma^2})$. This introduces sampling noise and also makes $\hat{\mathcal{L}}_{\text{APS}}(\mathbf{x},\mathbf{y};\varphi)$ non-differentiable. Alternatively, one could update φ by employing *policy gradient* rule for non-differentiable rewards as in SVDD-PM (Li et al., 2024). However, this leads to sparse updates and inferior sample quality as shown in Table 1.

Differentiability has propelled posterior sampling to achieve state-of-the-art results using continuous diffusion (Chung et al., 2023; Rout et al., 2023; 2024; Zhang et al., 2025). To restore differentiability in discrete diffusion, we parameterize $\mathbf{x}_\varphi(\mathbf{z}_{t(i)}) = \text{Softmax}(\varphi_{t(i)})$ with $\varphi_{t(i)} = \{\varphi^l_{t(i)}\}_{l=1}^L \in \mathbb{R}^{K\times L}$ containing logits $\varphi^l_{t(i)}$ over the codebook $\{\mathbf{c}_k \in \mathcal{C}\}$ for each position l. We compute the expected embedding $\bar{\mathbf{x}}^l = \sum_{k=1}^K \mathbf{c}_k \, \mathbf{x}^l_\varphi(\mathbf{z}_{t(i)}) \in \mathbb{R}^d$, and then *quantize* it using LFQ (2) (Yu et al., 2024): $\mathbf{x}^l = \mathcal{Q}_{\text{lfq}}(\bar{\mathbf{x}}^l) \in \{-1,+1\}^d$. We then apply the straight-through estimator (Van Den Oord et al., 2017) to obtain an image $\hat{\mathbf{x}} = \mathcal{D}(\tilde{\mathbf{x}})$, $\tilde{\mathbf{x}} = \bar{\mathbf{x}} + \left[\mathbf{x} - \bar{\mathbf{x}}\right]_{\text{sg}}$, where sg denotes the stop-gradient operator. Finally, we compute the differentiable likelihood as $q(\mathbf{y}|\mathbf{x}_\varphi(\mathbf{z}_{t(i)})) \propto \exp(-\frac{\|\mathbf{y}-\mathcal{A}(\hat{\mathbf{x}})\|_2^2}{2\sigma^2})$.

Discussion. We minimize the objective (7) at every step t to obtain the optimal logits $\varphi^*_{t(i)}$. This ensures that the optimized probabilities $\mathbf{x}_{\varphi^*}(\mathbf{z}_{t(i)})$ remain close to the prior predictions $\mathbf{x}_\theta(\mathbf{z}_{t(i)})$ while being adapted to the measurements. To the best of our knowledge, this paper takes the first step towards *quantizing the expected codebook entry* for posterior sampling in discrete diffusion. This allows gradients from the measurement loss to propagate through $\bar{\mathbf{x}}$ and update the entire conditional probability table $\mathbf{x}_\varphi(\mathbf{z}_{t(i)})$, strengthening gradient-based guidance and accurate posterior sampling.

3.2.2 Anchored Remasking

In masked diffusion, the reverse process progressively unmasks tokens. Most discrete samplers (Austin et al., 2021; Chang et al., 2022; Lou et al., 2024; Sahoo et al., 2024; Shi et al., 2024) choose to unmask tokens in random order. ADLM (Rout et al., 2025a) shows that prioritizing "anchor" tokens (e.g., nouns or verbs in language, rather than articles or conjunctions) early in the

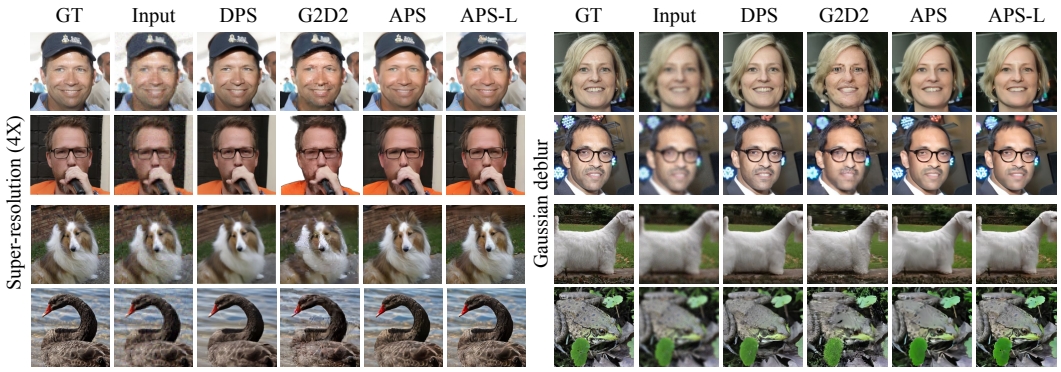

Figure 2: **Qualitative results on FFHQ and ImageNet for SR** ($4\times$) **and Gaussian deblur.** Compared to DPS and G2D2, APS yields better results with sharper texture and refined facial features. For instance, in the third row, APS reconstructs fine strands of the white and brown dog's fur.

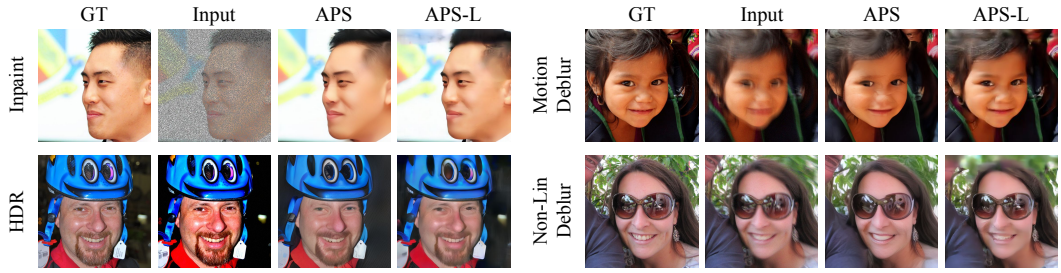

Figure 3: **Qualitative results on FFHQ** for linear (top row) and nonlinear (bottom row) inverse problems. APS and APS-L recover high-fidelity images from severely degraded inputs.

reverse process reduces conditional entropy of the remaining sequence and improves generation quality. To decode anchor tokens, ADLM jointly trains an anchor network in addition to the standard denoising network using an anchored NELBO objective. Distinct from ADLM, we propose a *training-free* variant of anchored denoising, enabling posterior sampling with a pretrained model.

Let φ be a minimizer of (7), resulting in $\mathbf{x}^*_\varphi(\mathbf{z}_t) = \{(\mathbf{x}^*_\varphi(\mathbf{z}_t))^l\}_{l=1}^L$ that represents categorical distributions over tokens at each position $l \in \{1, \dots, L\}$ at time step t. Anchored remasking selects a subset of positions $\mathcal{P}_t \subseteq \{1, \dots, L\}$ to decode early. The selection is based on the confidence of quantized tokens $\mathbf{x}$ (as defined in §3.2.1) under the posterior estimate $\mathbf{x}^*_\varphi(\mathbf{z}_t)$. Importantly, the posterior estimate is a function of the L-length sequence $\mathbf{z}_t$, and hence encodes the *joint* relation across all tokens; thus, anchored remasking is a function of *all* tokens, as compared to standard remasking (Chang et al., 2022; Yang et al., 2025) that depend only on per-token confidence. Formally, we compute the confidence of $\mathbf{x}^l$ as $\kappa^l_t = \langle (\mathbf{x}^*_\varphi(\mathbf{z}_t))^l, \mathbf{x}^l \rangle$, and choose anchor positions as $\mathcal{P}_t = \{\, l : \kappa^l_t \geq \tau_t \,\}$, where τ_t is an adaptive threshold based on the cosine schedule from MMaDA. We then update the state by fixing anchor tokens and remasking the rest: $\mathbf{z}^l_s = \mathbf{x}$ with probability $\frac{\alpha_s - \alpha_t}{1 - \alpha_t}$ if $l \in \mathcal{P}_t$ else $\mathbf{m}$. Once a token is unmasked, it remains fixed in subsequent steps.

Discussion. Diffusion language models tend to be overconfident on low-information tokens such as articles ("a", "an" or "the") or conjunctions (Rout et al., 2025a); similarly, discrete image samplers using *independent per-token* confidence often unmask background pixels first. In contrast, our method leverages the *joint posterior* $\mathbf{x}_\varphi(\mathbf{z}_t)$ to identify anchor tokens consistent with the measurements. This leads to earlier decoding of informative tokens (e.g., a bird against a flat background), enabling faster reconstruction and improved likelihood of generated samples; refer §B.3 for details.

4 Experiments

Baselines. Since our focus is on discrete diffusion, we first compare against existing discrete posterior samplers: G2D2 (Murata et al., 2024) and SGDD (Chu et al., 2025) (§2). To provide a comprehensive evaluation, we also include established continuous baselines, both in pixel space (DPS (Chung et al., 2023), DDRM (Kawar et al., 2022), and DiffPIR (Zhu et al., 2023)) and in la-

Table 2: **Quantitative results on general inverse problems.** We report results on two additional linear (random inpainting, motion deblurring) and nonlinear (HDR, nonlinear blur) tasks. Since G2D2 and SGDD do not evaluate on these tasks, we compare our *discrete* sampler against representative *continuous* baselines: DPS (pixel-space) and PSLD (latent-space).

Type	Method	Random Inpainting		Motion Deblur		HDR		Nonlinear Blur	
		LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑
FFHQ									
Pixel	DPS	**0.203**	25.46	**0.246**	24.52	**0.264**	22.73	0.278	23.39
Latent	PSLD	0.221	**30.31**	0.336	22.31	–	–	–	–
Mask	**APS (ours)**	0.304	27.38	0.317	26.58	0.282	**23.89**	0.263	27.19
Discrete	**APS-L (ours)**	0.291	28.11	0.298	**27.98**	0.323	23.56	**0.262**	**28.46**
ImageNet									
Pixel	DPS	**0.297**	23.52	0.423	18.96	0.503	19.23	**0.306**	22.49
Latent	PSLD	0.337	**31.30**	0.511	20.85	–	–	–	–
Mask	**APS (ours)**	0.378	24.59	0.410	23.37	0.345	21.92	0.330	24.18
Discrete	**APS-L (ours)**	0.338	25.39	**0.318**	**25.19**	**0.346**	**22.68**	0.309	**25.35**

Figure 4: **Qualitative results on stylization.** We present four style–content combinations. For each case, our APS algorithm conditions on a single reference style image together with a text prompt to generate the stylized output images. The prompt follows the template: "Generate an image in `[style]` style. A `[class]`, high detail, photorealistic." Here, `[style]` denotes the reference style (e.g., *Celestial Artwork*), and `[class]` corresponds to the label shown below (e.g., *Carousel*).

tent space (PSLD (Rout et al., 2023) and ReSample (Song et al., 2024)). This ensures our evaluation spans both discrete and continuous posterior sampling paradigms.

Benchmarks. We evaluate on the standard inverse problem benchmarks used in prior works. For high-resolution faces we use FFHQ at 256×256 (Karras et al., 2019), and for diverse natural images we use ImageNet at 256×256 (Deng et al., 2009). Performance is measured using three standard metrics: Learned Perceptual Image Patch Similarity (Zhang et al., 2018) (LPIPS), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) (Wang et al., 2004). All methods are evaluated at the same resolution and on the same images following G2D2 (Murata et al., 2024).

Tasks. We consider both linear and nonlinear inverse problems. Following prior works (e.g., G2D2 (Murata et al., 2024) and SGDD (Chu et al., 2025)), we evaluate on *Super Resolution (SR) (4×)* and *Gaussian deblurring* on both FFHQ and ImageNet. Additionally, we evaluate on more linear (*random inpainting* and *motion blur*) and nonlinear (*high dynamic range (HDR) recovery* and *nonlinear deblurring*) tasks. Beyond inverse problems, we conduct experiments on *training-free stylization*, an emerging area largely dominated by continuous diffusion (Hertz et al., 2023; Wang et al., 2024; Rout et al., 2025c).

To demonstrate scalability, we upsample the benchmark images to 1024×1024 and apply our posterior sampler to these scaled images (termed **APS-L** in discussion below). We defer implementation details to §B.1 and computational complexity to §B.5.

4.1 Results on Linear Inverse Problems

Evaluation on FFHQ: Table 1 (a) shows that our approach outperforms prior discrete diffusion samplers across both super resolution and Gaussian deblurring tasks. For *super resolution*, APS reduces LPIPS by 13.65% and improves PSNR by 2.11%. On *Gaussian deblurring*, APS lowers LPIPS by 3.83% compared to G2D2 and raises PSNR by 5.88%. The large variant, APS-L, pushes performance even further, achieving up to a 31.36% gain over G2D2 in terms of LPIPS. These results show that our quantized expectation (§3.2.1) and anchored remasking (§3.2.2) strategies not

only outperform discrete diffusion baselines but often surpass strong continuous-diffusion methods, such as DiffPIR (which uses pixel-space diffusion) and PSLD (which uses latent-space diffusion), resulting in higher perceptual and reconstruction quality.

Figure 2 (top two rows) presents qualitative comparisons for super resolution ($4\times$) and Gaussian deblurring on FFHQ. DPS produces over-smoothed results with blurry facial details, while G2D2 often introduces artifacts and fails to restore a natural facial structure. Our APS sampler yields sharper textures and more faithful reconstructions, recovering details such as hair strands, facial contours, and eyeglass edges with higher perceptual quality. The large variant, APS-L, further enhances structure and realism, generating photo-realistic outputs with finer details and fewer artifacts. These results confirm that APS and APS-L deliver superior qualitative performance on face datasets, where perceptual fidelity is especially critical.

Evaluation on ImageNet: Table 1 (b) quantifies that our method achieves consistent improvements across both SR and Gaussian deblurring tasks on the ImageNet benchmark. For *super resolution*, APS reduces LPIPS by 7.16% compared to G2D2 and improves PSNR by 4.74%. On *Gaussian deblurring*, APS improves PSNR by 8.81% while maintaining comparable LPIPS. APS-L improves LPIPS by up to 35.82% compared to G2D2. These results confirm that APS outperforms prior discrete posterior samplers and often surpasses continuous baselines such as DiffPIR and PSLD, delivering superior perceptual quality and reconstruction fidelity.

Figure 2 (bottom two rows) shows SR ($4\times$) and Gaussian deblurring results on ImageNet. Notably, DPS produces overly smooth outputs with a loss of fine details, while G2D2 introduces struggles to recover sharp edges. In contrast, APS reconstructs sharper textures (e.g., the fur of the dog and the feathers of the swan) and yields more natural color. The large variant, APS-L, further enhances structural fidelity, recovering finer details in challenging regions such as a frog's skin texture and a goat's fur. These examples highlight that our approach achieves superior perceptual quality and faithful structure reconstruction compared to both pixel-based and discrete diffusion baselines.

4.2 Results on General Inverse Problems

Table 2 shows that APS generalizes effectively to more challenging linear (random inpainting, motion deblurring) and nonlinear (HDR, nonlinear blur) inverse problems on FFHQ and ImageNet. Unlike prior discrete samplers such as G2D2 and SGDD, which were demonstrated only on limited tasks, APS consistently achieves strong perceptual quality (lower LPIPS) and reconstruction fidelity (higher PSNR). For instance, on ImageNet motion deblurring APS-L attains 0.318 LPIPS and 25.19 PSNR, substantially outperforming continuous baselines DPS and PSLD. Similarly, in nonlinear tasks such as HDR and nonlinear blur, APS delivers sharper, more consistent reconstructions, closing the gap with continuous diffusion while operating within a purely discrete framework. These results highlight the broader applicability and robustness of our approach compared to existing discrete diffusion samplers. Figure 3 shows the qualitative results of general inverse problems.

4.3 Results on Reference-based Stylization

We compare APS with the discrete diffusion baseline MMaDA (Yang et al., 2025) and continuous methods (shaded gray). Table 3 reports quantitative results, while Figure 4 shows qualitative examples. APS improves over MMaDA on ImageReward (Xu et al., 2024), CLIP-T (Radford et al., 2021), and DINO (Caron et al., 2021) scores, demonstrating that our posterior sampler unlocks stylization capabilities absent in the base model. Notably, APS also surpasses continuous baselines such as IP-Adapter (Ye et al., 2023) and StyleAligned (Hertz et al., 2023) on ImageReward and CLIP-T, despite relying on a weaker generative prior (Yang et al., 2025).

	ImageReward ↑	CLIP-T ↑	DINO ↑
IP-Adapter	-1.51	0.26	0.89
StyleAligned	0.01	0.31	0.85
InstantStyle	0.72	0.33	0.72
RB-Modulation	1.18	0.34	0.73
MMaDA	0.48	0.33	0.32
APS (ours)	0.63	0.34	0.41

Table 3: **Quantitative results on stylization.**

4.3.1 Results on Text-Guided Block Inpainting

We compare our approach against large-scale continuous diffusion baselines: Imagen3 (Baldridge et al., 2024), Flux (Black Forest Labs, 2024), and HDPainter (Manukyan et al., 2025). As illustrated in Figure 5, our method generates realistic clothing textures, with stronger alignment to the reference prompts and fewer artifacts (red boxes highlight failure regions of competing methods). In contrast,

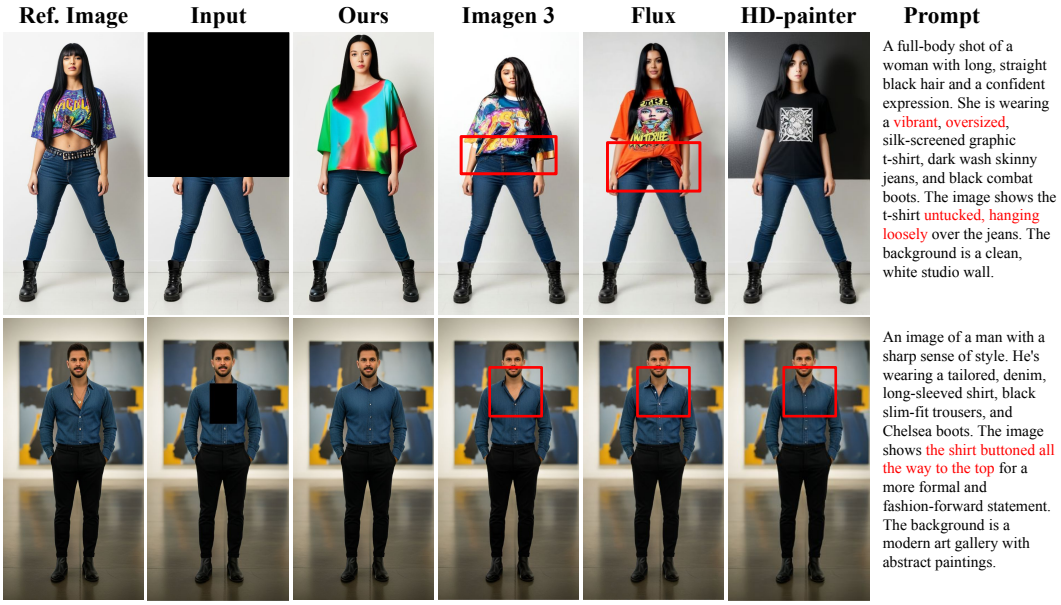

Figure 5: **Text-guided block inpainting on high-resolution (1024×512) images.** Our approach (APS) generates prompt-aligned garment completions compared to prior methods. Red boxes highlight prompt-misalignment in competing approaches. In the first row, the prompt specifies an untucked, oversized T-shirt with vibrant colors—details missed by Imagen 3, Flux, and HD-Painter. In the second row, our approach correctly buttons the shirt all the way to the top.

Imagen3 and Flux often introduce distorted or inconsistent garment regions, while HD-painter produces less faithful completions with mismatched styles.

The appendix provides extensive details on our Anchored Posterior Sampling method. We present the theoretical derivation of variational bounds (§A), all implementation specifics and hyperparameters (Algorithm 1, Table 4), and a detailed ablation study (§B.3, Figure 6) demonstrating the impact of our innovations. We also showcase the superior computational efficiency (Table 6) and numerous additional qualitative results (§B) across complex inverse problems, details of text-guided block inpainting (§B.8.2), and extended stylization results (Figure 12 and 13).

5 Conclusion

We introduce **Anchored Posterior Sampling (APS)**, a training-free posterior sampler with discrete diffusion foundation models. Grounded in our theoretical bounds, APS enables the reuse of a pretrained denoiser *without* task-specific retraining. APS is built on two algorithmic innovations: *Quantized Expectation*, which provides fine-grained, gradient-like guidance in discrete spaces, and *Anchored Remasking*, which adaptively decodes important tokens early in the reverse process. Together, these yield an efficient, scalable, and training-free posterior sampler.

Extensive experiments on linear and nonlinear inverse problems demonstrate that APS achieves state-of-the-art results among discrete samplers and is competitive with continuous baselines—while operating at significantly lower inference cost. Beyond inverse problems, APS also unlocks new capabilities such as training-free stylization, underscoring the flexibility of discrete diffusion models when paired with effective posterior sampling. We believe this work establishes discrete diffusion as a practical and scalable alternative for posterior sampling, with promising extensions to video, multimodal generation, and other structured domains.

Acknowledgments

The authors thank the **Google's ARML Commerce team** for their support and for providing a stimulating environment for this research, which was conducted while the first author was an intern at Google. We are also grateful to **Akash Sengupta** and **Yingwei Li** for their insightful discussions during the early stages of this project. This research has been partially supported by NSF Grants 2112471, 2505865 and the UT Austin Machine Learning Lab.

Reproducibility Statement

Our experiments are built upon the publicly available MMaDA codebase (Yang et al., 2025). All modifications and implementation details are described in Appendix B.1, which includes the pseudocode in Algorithm 1 and the specific parameters used for every experiment. Furthermore, Appendix B.2 provides comprehensive ablation studies and hyperparameter sweeps (Table 4). The experiments use the widely-used public datasets FFHQ (Karras et al., 2019) and ImageNet (Deng et al., 2009). The combination of a public codebase and datasets, along with our detailed Algorithm 1 and description (§B.2), should ensure that our results are readily reproducible.

Ethics Statement

The proposed method for controlled image editing contributes to the democratization of advanced image editing. While this has positive societal benefits, we acknowledge the dual-use nature of generative models and the potential for misuse or misinterpretation of their outputs.

A primary concern arises in the context of inverse problems such as super resolution, inpainting, and deblurring. Our method generates a plausible, high-quality image, but this output is a sample from a posterior distribution and does not represent a unique or guaranteed reconstruction of the original ground truth. A misunderstanding of this core limitation could lead to dangerous assumptions, particularly in sensitive applications like forensic analysis, where a generated image might be misinterpreted as factual evidence. We therefore emphasize that the method's intended purpose is to enhance perceptual realism for creative or aesthetic applications, and it is not suitable for tasks requiring high-fidelity reconstruction or person identification.

In stylization tasks, there is a risk of unintentional information leakage, where details from the source content image could persist in the stylized output. This could lead to the inadvertent disclosure of sensitive information that a user did not intend to share.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=h7-XixPCAL.

Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluis Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, Hongliang Fei, Nando de Freitas, Yilin Gao, Evgeny Gladchenko, Sergio Gómez Colmenarejo, Mandy Guo, Alex Haig, Will Hawkins, Hexiang Hu, Huilian Huang, Tobenna Peter Igwe, Christos Kaplanis, Siavash Khodadadeh, Yelin Kim, Ksenia Konyushkova, Karol Langner, Eric Lau, Rory Lawton, Shixin Luo, Soňa Mokrá, Henna Nandwani, Yasumasa Onoe, Aäron van den Oord, Zarana Parekh, Jordi Pont-Tuset, Hang Qi, Rui Qian, Deepak Ramachandran, Poorva Rane, Abdullah Rashwan, Ali Razavi, Robert Riachi, Hansa Srinivasan, Srivatsan Srinivasan, Robin Strudel, Benigno Uria, Oliver Wang, Su Wang, Austin Waters, Chris Wolff, Auriel Wright, Zhisheng Xiao, Hao Xiong, Keyang Xu, Marc van Zee, Junlin Zhang, Katie Zhang, Wenlei Zhou, Konrad Zolna, Ola Aboubakar, Canfer Akbulut, Oscar Akerlund, Isabela Albuquerque, Nina Anderson, Marco Andreetto, Lora Aroyo, Ben Bariach, David Barker, Sherry Ben, Dana Berman, Courtney Biles, Irina Blok, Pankil Botadra, Jenny Brennan, Karla Brown, John Buckley, Rudy Bunel, Elie Bursztein, Christina Butterfield, Ben Caine, Viral Carpenter, Norman Casagrande, Ming-Wei Chang, Solomon Chang, Shamik Chaudhuri, Tony Chen, John Choi, Dmitry Churbanau, Nathan Clement, Matan Cohen, Forrester Cole, Mikhail Dektiarev, Vincent Du, Praneet Dutta, Tom Eccles, Ndidi Elue, Ashley Feden, Shlomi Fruchter, Frankie Garcia, Roopal Garg, Weina Ge, Ahmed Ghazy, Bryant Gipson, Andrew Goodman, Dawid Górny, Sven Gowal, Khyatti Gupta, Yoni Halpern, Yena Han, Susan Hao, Jamie Hayes, Jonathan Heek, Amir Hertz, Ed Hirst, Emiel Hoogeboom, Tingbo Hou, Heidi Howard, Mohamed Ibrahim, Dirichi Ike-Njoku, Joana Iljazi, Vlad Ionescu, William Isaac, Reena Jana, Gemma Jennings, Donovon Jenson, Xuhui Jia, Kerry Jones, Xiaoen Ju, Ivana Kajic, Christos Kaplanis, Burcu Karagol Ayan, Jacob Kelly, Suraj Kothawade, Christina Kouridi, Ira Ktena, Jolanda Kumakaw, Dana Kurniawan, Dmitry Lagun,

Lily Lavitas, Jason Lee, Tao Li, Marco Liang, Maggie Li-Calis, Yuchi Liu, Javier Lopez Alberca, Matthieu Kim Lorrain, Peggy Lu, Kristian Lum, Yukun Ma, Chase Malik, John Mellor, Thomas Mensink, Inbar Mosseri, Tom Murray, Aida Nematzadeh, Paul Nicholas, Signe Nørly, João Gabriel Oliveira, Guillermo Ortiz-Jimenez, Michela Paganini, Tom Le Paine, Roni Paiss, Alicia Parrish, Anne Peckham, Vikas Peswani, Igor Petrovski, Tobias Pfaff, Alex Pirozhenko, Ryan Poplin, Utsav Prabhu, Yuan Qi, Matthew Rahtz, Cyrus Rashtchian, Charvi Rastogi, Amit Raul, Ali Razavi, Sylvestre-Alvise Rebuffi, Susanna Ricco, Felix Riedel, Dirk Robinson, Pankaj Rohatgi, Bill Rosgen, Sarah Rumbley, Moonkyung Ryu, Anthony Salgado, Tim Salimans, Sahil Singla, Florian Schroff, Candice Schumann, Tanmay Shah, Eleni Shaw, Gregory Shaw, Brendan Shillingford, Kaushik Shivakumar, Dennis Shtatnov, Zach Singer, Evgeny Sluzhaev, Valerii Sokolov, Thibault Sottiaux, Florian Stimberg, Brad Stone, David Stutz, Yu-Chuan Su, Eric Tabellion, Shuai Tang, David Tao, Kurt Thomas, Gregory Thornton, Andeep Toor, Cristian Udrescu, Aayush Upadhyay, Cristina Vasconcelos, Alex Vasiloff, Andrey Voynov, Amanda Walker, Luyu Wang, Miaosen Wang, Simon Wang, Stanley Wang, Qifei Wang, Yuxiao Wang, Ágoston Weisz, Olivia Wiles, Chenxia Wu, Xingyu Federico Xu, Andrew Xue, Jianbo Yang, Luo Yu, Mete Yurtoglu, Ali Zand, Han Zhang, Jiageng Zhang, Catherine Zhao, Adilet Zhaxybay, Miao Zhou, Shengqi Zhu, Zhenkai Zhu, Dawn Bloxwich, Mahyar Bordbar, Luis C. Cobo, Eli Collins, Shengyang Dai, Tulsee Doshi, Anca Dragan, Douglas Eck, Demis Hassabis, Sissie Hsiao, Tom Hume, Koray Kavukcuoglu, Helen King, Jack Krawczyk, Yeqing Li, Kathy Meier-Hellstern, Andras Orban, Yury Pinsky, Amar Subramanya, Oriol Vinyals, Ting Yu, and Yori Zwols. Imagen 3, 2024. URL https://arxiv.org/abs/2408.07009.

Black Forest Labs. Black forest labs, 2024. URL https://blackforestlabs.ai/. Accessed: September 1, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. URL https://arxiv.org/abs/2104.14294.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022. URL https://arxiv.org/pdf/2202.04200.

Sinho Chewi. Log-concave sampling. *Book draft available at https://chewisinho. github. io*, 9: 17–18, 2023. URL https://chewisinho.github.io/main.pdf.

Wenda Chu, Zihui Wu, Yifan Chen, Yang Song, and Yisong Yue. Split gibbs discrete diffusion posterior sampling. *arXiv preprint arXiv:2503.01161*, 2025. URL https://arxiv.org/pdf/2503.01161.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8941–8967. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/chung24b.html.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024. URL https://dl.acm.org/doi/10.5555/3692070.3692573.

Google Gemini Team. Gemini 1.5: Unlocking multimodal understanding across long contexts. Technical report, Google, February 2024. URL https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf. Technical Report.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10696–10706, 2022. URL https://arxiv.org/pdf/2111.14822.

Emil Julius Gumbel. *Statistical theory of extreme values and some practical applications: a series of lectures*, volume 33. US Government Printing Office, 1954.

Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018. URL https://arxiv.org/pdf/1711.08447.

Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. *arXiv preprint arXiv:2312.02133*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. URL https://arxiv.org/pdf/2006.11239.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pp. 13916–13932. PMLR, 2023. URL https://proceedings.mlr.press/v202/huang23i/huang23i.pdf.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/pdf/2201.11793.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, et al. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *arXiv preprint arXiv:2408.08252*, 2024. URL https://arxiv.org/pdf/2408.08252.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=CNicRIVIPA.

Tianren Ma, Xiaosong Zhang, Boyu Yang, Junlan Feng, and Qixiang Ye. Reddit: Rehashing noise for discrete visual generation. *arXiv preprint arXiv:2505.19656*, 2025.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=S1jE5L5gl.

Hayk Manukyan, Andranik Sargsyan, Barsegh Atanyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. HD-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6lB5qtdYAg.

Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Bac Nguyen, Stefano Ermon, and Yuki Mitsufuji. G2d2: Gradient-guided discrete diffusion for image inverse problem solving. *arXiv preprint arXiv:2410.14710v1*, 2024. URL https://arxiv.org/abs/2410.14710v1.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. URL https://arxiv.org/pdf/2502.09992.

Mehdi Noroozi, Isma Hadji, Brais Martinez, Adrian Bulat, and Georgios Tzimiropoulos. You only need one step: Fast super-resolution with stable diffusion via scale distillation. In *European Conference on Computer Vision*, pp. 145–161. Springer, 2024. URL https://arxiv.org/abs/2401.17258.

OpenAI. Sora: Creating video from text, 2024. URL https://openai.com/index/sora/. Accessed: 2025-09-13.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021. URL https://proceedings.mlr.press/v139/ramesh21a/ramesh21a.pdf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. URL https://arxiv.org/abs/2112.10752.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G Dimakis, and Sanjay Shakkottai. Solving inverse problems provably via posterior sampling with latent diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=XKBFdYwfRo.

Litu Rout, Yujia Chen, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Beyond first-order tweedie: Solving inverse problems using latent diffusion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/pdf/2312.00852.

Litu Rout, Constantine Caramanis, and Sanjay Shakkottai. Anchored diffusion language model. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=E8adS5srds.

Litu Rout, Yujia Chen, Nataniel Ruiz, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. Semantic image inversion and editing using rectified stochastic differential equations. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=Hu0FSOSEyS.

Litu Rout, Yujia Chen, Nataniel Ruiz, Abhishek Kumar, Constantine Caramanis, Sanjay Shakkottai, and Wen-Sheng Chu. RB-modulation: Training-free stylization using reference-based modulation. In *The Thirteenth International Conference on Learning Representations*, 2025c. URL https://openreview.net/forum?id=bnINPG5A32.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Yair Schiff, Subham Sekhar Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-torre, Bernardo P de Almeida, Alexander M Rush, Thomas PIERROT, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=i5MrJ6g5G1.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nJfylDvgzlq.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, Proceedings of Machine Learning Research, pp. 2256–2265. PMLR, 07–09 Jul 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. Styledrop: Text-to-image generation in any style. In *37th Conference on Neural Information Processing Systems (NeurIPS)*. Neural Information Processing Systems Foundation, 2023. URL https://arxiv.org/pdf/2306.00983.

Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. 2024. URL https://arxiv.org/pdf/2404.01292v1.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=j8hdRqOUhN.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. URL https://arxiv.org/pdf/1711.00937.

Team Veo. Veo: A Generalist Video Generation Model with State-of-the-Art Spatiotemporal Consistency, Resolution and Quality. Technical report, Google, May 2025. URL https://storage.googleapis.com/deepmind-media/veo/Veo-3-Tech-Report.pdf. Accessed: 2025-09-13.

Haofan Wang, Qixun Wang, Xu Bai, Zekui Qin, and Anthony Chen. Instantstyle: Free lunch towards style-preserving in text-to-image generation. *arXiv preprint arXiv:2404.02733*, 2024.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1284395.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. URL https://arxiv.org/pdf/2505.15809.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Lijun Yu, Jose Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gzqrANCF4g.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20895–20905, 2025. URL https://arxiv.org/pdf/2407.01521.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018. URL https://arxiv.org/abs/1801.03924.

Luyang Zhu, Yingwei Li, Nan Liu, Hao Peng, Dawei Yang, and Ira Kemelmacher-Shlizerman. M&m vto: Multi-garment virtual try-on and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1346–1356, 2024. URL https://arxiv.org/abs/2406.04542.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1219–1229, 2023. URL https://arxiv.org/pdf/2305.08995.

A Additional Theoretical Results

This appendix develops complementary theory for masked discrete diffusion posterior sampling. We first derive a pathwise variational bound for *training* a likelihood-tilted reverse process (**Theorem A.1**), showing that $-\log p_\varphi(\mathbf{x}|\mathbf{y})$ is upper bounded by a reconstruction term, a sum of token-wise KL-divergence terms, and a sequence of measurement likelihood-based tilting terms. We then specialize this analysis to the *training-free* setting where the token-to-image decoder is shared between unconditional generation and posterior sampling, yielding a bound for test-time anchored posterior sampling (**Theorem A.2**). We provide theoretical insights drawn from each theorem in **Implication** subsections after the corresponding proofs.

Theorem A.1 (Discrete Diffusion Posterior Sampling(DDPS))**.** *Let $Z_{0:1} = \{Z_{t(i)}\}_{i=0}^{T}$ with $t(i) = i/T$ and $s(i) = (i-1)/T$ be the latent path of a masked discrete diffusion model, and let $q(Z_{0:1}|\mathbf{x})$ be the forward noising law from (3). Consider reverse kernels and a terminal decoder that factorize as follows:*

$$p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y}) = p_\varphi(Z_1|\mathbf{y})\, p_\varphi(\mathbf{x}|Z_0, \mathbf{y}) \prod_{i=1}^{T} p_\varphi\big(Z_{s(i)}|Z_{t(i)}, \mathbf{y}\big)\,,$$

with token-wise reverse transitions given by the inference posterior in (4) tilted by the likelihood,

$$p_\varphi(Z_s^l|Z_t, \mathbf{y}) \;:=\; q\big(Z_s^l|Z_t^l, \mathbf{x}_\varphi(Z_t)\big)\, q\big(\mathbf{y}|\mathbf{x}_\varphi(Z_t)\big).$$

Then, for any $(\mathbf{x}, \mathbf{y})$,

$$\begin{aligned} -\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathcal{L}_{\mathrm{DDPS}}(\mathbf{x}, \mathbf{y}; \varphi) \coloneqq\; & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0)] \\ & + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \sum_{l=1}^{L} \mathrm{CE}\big(\mathbf{x}^l, \mathbf{x}_\varphi^l(Z_{t(i)})\big)\mathbf{1}_{\{Z_{t(i)}^l = \mathbf{m}\}}\right] \\ & - \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big]. \end{aligned} \tag{8}$$

Proof. We present the derivation for sequence length $L = 1$; the extension to $L > 1$ follows by token-wise factorization as in (Sohl-Dickstein et al., 2015). Recall our parameterized reverse kernel (identical in form to (4) up to a likelihood tilt):

$$q(Z_s^l|Z_t^l, \mathbf{x}_\varphi(Z_t)) = \begin{cases} \mathrm{Cat}(Z_s^l; Z_t^l), & Z_t^l \neq \mathbf{m}, \\ \mathrm{Cat}\Big(Z_s^l; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\varphi(Z_t) + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\Big), & Z_t^l = \mathbf{m}. \end{cases} \tag{9}$$

Starting from the conditional likelihood,

$$\begin{aligned} -\log p_\varphi(\mathbf{x}|\mathbf{y}) &= -\log \int p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y})\, \mathrm{d}Z_{0:1} \\ &= -\log \int p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y})\, \frac{q(Z_{0:1}|\mathbf{x}, \mathbf{y})}{q(Z_{0:1}|\mathbf{x}, \mathbf{y})}\, \mathrm{d}Z_{0:1}. \end{aligned}$$

By the conditional independence of the forward process (the noising path is conditionally independent of $\mathbf{y}$ given $\mathbf{x}$), we have $q(Z_{0:1}|\mathbf{x}, \mathbf{y}) = q(Z_{0:1}|\mathbf{x})$, hence

$$-\log p_\varphi(\mathbf{x}|\mathbf{y}) = -\log \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\left[\frac{p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y})}{q(Z_{0:1}|\mathbf{x})}\right].$$

Applying Jensen's inequality,

$$-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\left[-\log \frac{p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y})}{q(Z_{0:1}|\mathbf{x})}\right].$$

Using the factorization $p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y}) = p_\varphi(Z_1|\mathbf{y})\, p_\varphi(\mathbf{x}|Z_0, \mathbf{y}) \prod_{i=1}^{T} p_\varphi(Z_{s(i)}|Z_{t(i)}, \mathbf{y})$ and the forward process factorization $q(Z_{0:1}|\mathbf{x}) = q(Z_1|\mathbf{x}) \prod_{i=1}^{T} q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})$, we obtain

$$
\begin{aligned}
&-\log \frac{p_\varphi(\mathbf{x}, Z_{0:1}|\mathbf{y})}{q(Z_{0:1}|\mathbf{x})} \\
&= -\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y}) + \log \frac{q(Z_1|\mathbf{x})}{p_\varphi(Z_1|\mathbf{y})} + \sum_{i=1}^{T} \log \frac{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}{p_\varphi(Z_{s(i)}|Z_{t(i)}, \mathbf{y})}.
\end{aligned}
$$

Taking expectation under $q(Z_{0:1}|\mathbf{x})$ yields

$$
\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le\ & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\big[-\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y})\big] + \mathbb{E}_{q(Z_1|\mathbf{x})}\Big[\log \frac{q(Z_1|\mathbf{x})}{p_\varphi(Z_1|\mathbf{y})}\Big] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}\Big[\log \frac{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}{p_\varphi(Z_{s(i)}|Z_{t(i)}, \mathbf{y})}\Big]. \qquad (10)
\end{aligned}
$$

Expanding the tilted reverse kernel using the parameterization $p_\varphi(Z_{s(i)}|Z_{t(i)}, \mathbf{y}) := q(Z_{s(i)}|Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))\, q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))$, we can rewrite the last term as

$$
\sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}\Big[\log \frac{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))} - \log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\Big].
$$

This yields the final decomposition:

$$
\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le\ & \underbrace{\mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y})] + \mathrm{KL}\big(q(Z_1|\mathbf{x}) \,\|\, p_\varphi(Z_1|\mathbf{y})\big)}_{\text{reconstruction + boundary KL}} \\
&+ \underbrace{\sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\Big[\mathrm{KL}\big(q(Z_{s(i)}|Z_{t(i)}, \mathbf{x}) \,\|\, q(Z_{s(i)}|Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))\big)\Big]}_{\text{per-step KLs}} \\
&- \underbrace{\sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big]}_{\text{likelihood tilt terms}} := \mathcal{L}_{\mathrm{DDPS}}(\mathbf{x}, \mathbf{y}; \varphi).
\end{aligned}
$$

The first two groups of terms correspond to the standard masked diffusion posterior NELBO objective (Sohl-Dickstein et al., 2015; Austin et al., 2021) denoted by:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{PNELBO}}(\mathbf{x}, \mathbf{y}; \varphi) :=\ & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\big[-\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y})\big] \\
&+ \mathrm{KL}\big(q(Z_1|\mathbf{x}) \,\|\, p_\varphi(Z_1|\mathbf{y})\big) \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\Big[\mathrm{KL}\big(q(Z_{s(i)}|Z_{t(i)}, \mathbf{x}) \,\|\, q(Z_{s(i)}|Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))\big)\Big].
\end{aligned}
$$

The remaining likelihood terms,

$$
-\sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big],
$$

capture the effect of incorporating observations $\mathbf{y}$ into posterior sampling. Thus, the overall objective is the PNELBO plus a sequence of likelihood corrections at every step.

We now bound the per-step KL terms. Having connected the decomposition to the NELBO, it suffices to compute, for each step i, the divergence

$$
\mathrm{KL}\big(q(Z_{s(i)} \mid Z_{t(i)}, \mathbf{x}) \,\|\, q(Z_{s(i)} \mid Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))\big).
$$

Since masked diffusion induces a two–state posterior at each step (either remain masked or reveal the data token), we treat the two cases for $Z_{t(i)}$ separately.

Case I: $Z_{t(i)} = \mathbf{m}$. From the masked diffusion forward process (3), when $Z_{t(i)} = \mathbf{m}$ the true posterior over $Z_{s(i)}$ has mass

$$q(Z_{s(i)} = \mathbf{m} \mid Z_{t(i)} = \mathbf{m}, \mathbf{x}) = \frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}, \qquad q(Z_{s(i)} = \mathbf{x} \mid Z_{t(i)} = \mathbf{m}, \mathbf{x}) = \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}}.$$

Under the model with prediction $\mathbf{x}_\varphi(Z_{t(i)})$ (a categorical distribution) as in (4), the "unmask" branch is weighted by the model's probability of the true token, $\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x} \rangle$, while the mask probability is unchanged. Hence,

$$\begin{aligned}
&\mathrm{KL}\big(q(Z_{s(i)} \mid Z_{t(i)} = \mathbf{m}, \mathbf{x}) \,\|\, q(Z_{s(i)} \mid Z_{t(i)} = \mathbf{m}, \mathbf{x}_\varphi(Z_{t(i)}))\big) \\
&= \frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}} \log \frac{\frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}}{\frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}} + \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \log \frac{\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}}{\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}} \langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x} \rangle} \\
&= \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \log \frac{1}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x} \rangle} = \frac{\alpha_{t(i)} - \alpha_{s(i)}}{1-\alpha_{t(i)}} \log \langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x} \rangle.
\end{aligned}$$

Equivalently, writing the cross-entropy with the one-hot target $\mathbf{x}$ as $\mathrm{CE}(\mathbf{x}, \mathbf{x}_\varphi(Z_{t(i)})) = -\log \langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x} \rangle$,

$$\mathrm{KL}\big(q(Z_{s(i)} \mid Z_{t(i)} = \mathbf{m}, \mathbf{x}) \,\|\, q(Z_{s(i)} \mid Z_{t(i)} = \mathbf{m}, \mathbf{x}_\varphi(Z_{t(i)}))\big) = \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \mathrm{CE}(\mathbf{x}, \mathbf{x}_\varphi(Z_{t(i)})).$$

Case II: $Z_{t(i)} \neq \mathbf{m}$. When the current token is already unmasked, the posterior is deterministic: $q(Z_{s(i)} = Z_{t(i)} \mid Z_{t(i)} \neq \mathbf{m}, \mathbf{x}) = 1$. Thus,

$$\mathrm{KL}\big(q(Z_{s(i)} \mid Z_{t(i)} \neq \mathbf{m}, \mathbf{x}) \,\|\, q(Z_{s(i)} \mid Z_{t(i)} \neq \mathbf{m}, \mathbf{x}_\varphi(Z_{t(i)}))\big) = 0.$$

Only masked coordinates contribute to the per-step KL, yielding for each step i,

$$\mathrm{KL}\big(q(Z_{s(i)} | Z_{t(i)}, \mathbf{x}) \,\|\, q(Z_{s(i)} | Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))\big) = \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \mathrm{CE}\big(\mathbf{x}, \mathbf{x}_\varphi(Z_{t(i)})\big) \mathbf{1}_{\{Z_{t(i)} = \mathbf{m}\}}.$$

This recovers the standard masked-diffusion NELBO weighting (cf. (5)): per-step contributions are cross-entropies at masked positions, scaled by $(\alpha_{s(i)} - \alpha_{t(i)})/(1 - \alpha_{t(i)})$. Generalizing this to sequences with length $L > 1$ yields:

$$\mathrm{KL}\big(q(Z_{s(i)} \mid Z_{t(i)}, \mathbf{x}) \,\|\, q(Z_{s(i)} \mid Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))\big) = \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \sum_{l=1}^{L} \mathrm{CE}\big(\mathbf{x}^l, \mathbf{x}^l_\varphi(Z_{t(i)})\big) \mathbf{1}_{\{Z^l_{t(i)} = \mathbf{m}\}}.$$

Combining this with the likelihood tilt terms, we get

$$\begin{aligned}
\mathcal{L}_{\mathrm{DDPS}}(\mathbf{x}, \mathbf{y}; \varphi) = \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y})] + \mathrm{KL}\big(q(Z_1|\mathbf{x}) \,\|\, p_\varphi(Z_1|\mathbf{y})\big) \\
+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \sum_{l=1}^{L} \mathrm{CE}\big(\mathbf{x}^l, \mathbf{x}^l_\varphi(Z_{t(i)})\big) \mathbf{1}_{\{Z^l_{t(i)} = \mathbf{m}\}} \right] \\
- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)})) \big].
\end{aligned}$$

In masked diffusion Z_1 is typically the fully masked state or absorbing state, so $q(Z_1|\mathbf{x})$ is degenerate and, with $p_\varphi(Z_1|\mathbf{y}) = q(Z_1|\mathbf{x})$, the boundary KL vanishes. Thus,

$$\begin{aligned}
\mathcal{L}_{\mathrm{DDPS}}(\mathbf{x}, \mathbf{y}; \varphi) = \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y})] \\
+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \sum_{l=1}^{L} \mathrm{CE}\big(\mathbf{x}^l, \mathbf{x}^l_\varphi(Z_{t(i)})\big) \mathbf{1}_{\{Z^l_{t(i)} = \mathbf{m}\}} \right] \\
- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})} \big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)})) \big].
\end{aligned}$$

Furthermore, $\mathbf{y}$ imposes a distribution over $\mathbf{x}$ from which we wish to sample. However, when Z_0 is given then $\mathbf{x}$ is uniquely determined by the decoder as $\mathbf{x} = \mathrm{Dec}(Z_0)$. Therefore,

$$\begin{aligned}\mathcal{L}_{\mathrm{DDPS}}(\mathbf{x}, \mathbf{y}; \varphi) = \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0)] \\ + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1 - \alpha_{t(i)}} \sum_{l=1}^{L} \mathrm{CE}\big(\mathbf{x}^l, \mathbf{x}^l_\varphi(Z_{t(i)})\big)\mathbf{1}_{\{Z^l_{t(i)}=\mathbf{m}\}}\right] \\ - \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big],\end{aligned}$$

which completes the proof of the statement. □

Implications. Theorem A.1 provides a principled upper bound on the negative log-posterior likelihood in discrete masked diffusion models.

- *Trainable upper bound.* The pathwise upper bound $\mathcal{L}_{\mathrm{DDPS}}(\mathbf{x}, \mathbf{y}; \varphi)$ in (8) provides a principled training criterion for discrete diffusion posterior sampling.
- *Reduction to standard training when* $\mathbf{y}$ *is absent.* Setting the likelihood to a constant (no measurements) removes the tilt terms and recovers the masked-diffusion training objective: only the reconstruction term and per-step KL-divergence terms remain.
- *Masked-token supervision.* The per-step KL-divergence terms vanish on already-revealed tokens and reduce to weighted cross-entropies on masked tokens, focusing learning signal exactly where denoising must occur. The weights $(\alpha_{s(i)} - \alpha_{t(i)})/(1 - \alpha_{t(i)})$ expose how the noise schedule shapes gradient magnitude.
- *Data-consistency via tilt.* The additive terms $-\sum_i \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))]$ encourage reverse transitions that produce intermediate predictions consistent with the measurement model, integrating task specific information at every denoising step.
- *Boundary conditions.* With an absorbing mask state, the boundary KL-divergence is constant (often zero), so optimization concentrates on reconstruction, token-level KL-divergence terms, and measurement consistency.
- *Compatibility with efficient parameterizations.* Because (8) is written in terms of token-wise categoricals, it directly supports time-independent or lightweight parameterizations (e.g., shared denoisers), helping scalability to long sequences and high resolution.

Theorem A.2 (Test-time Anchored Posterior Sampling)**.** *Let $Z_{0:1} = \{Z_{t(i)}\}_{i=0}^{T}$ with $t(i) = i/T$ and $s(i) = (i-1)/T$ denote the latent path of a masked discrete diffusion model, and let $q(Z_{0:1}|\mathbf{x})$ be the forward noising law from (3). Assume the decoder is shared between unconditional generation and posterior sampling ($p_\varphi(\mathbf{x}|Z_0) = p_\theta(\mathbf{x}|Z_0)$), and the unconditional reverse transitions are parameterized as in (4). Define $\mathcal{L}_{\mathrm{NELBO}}(\mathbf{x}; \theta)$ as in (5). Then, for any $(\mathbf{x}, \mathbf{y})$,*

$$-\log p_\varphi(\mathbf{x}|\mathbf{y}) \;\le\; \mathcal{L}_{\mathrm{APS}}(\mathbf{x}, \mathbf{y}; \varphi),$$

where

$$\begin{aligned}\mathcal{L}_{\mathrm{APS}}(\mathbf{x}, \mathbf{y}; \varphi) \coloneqq \mathcal{L}_{\mathrm{NELBO}}(\mathbf{x}; \theta) + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1 - \alpha_{t(i)}} \log \frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle} \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\ - \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big].\end{aligned}$$

Proof. From the proof of **Theorem A.1**, the conditional likelihood satisfies

$$\begin{aligned}-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\big[-\log p_\varphi(\mathbf{x}|Z_0, \mathbf{y})\big] + \mathbb{E}_{q(Z_1|\mathbf{x})}\Big[\log \frac{q(Z_1|\mathbf{x})}{p_\varphi(Z_1|\mathbf{y})}\Big] \\ + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}\Big[\log \frac{q(Z_{s(i)}|Z_{t(i)}, \mathbf{x})}{p_\varphi(Z_{s(i)}|Z_{t(i)}, \mathbf{y})}\Big]. \qquad (11)\end{aligned}$$

Compute each term inside summation. We now focus on the i^{th} term inside the summation. Using the pretrained network $p_\theta(Z_{s(i)}|Z_{t(i)})$ as given in (4), we decompose the term as:

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right]$$
$$= \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}{p_\theta(Z_{s(i)}|Z_{t(i)})} + \log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right].$$

Isolating the first-term as a KL-divergence term, this yields

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right]$$
$$= \underbrace{\mathrm{KL}\Big(q(Z_{s(i)}|Z_{t(i)},\mathbf{x}) \,\|\, p_\theta(Z_{s(i)}|Z_{t(i)})\Big)}_{\text{divergence w.r.t. pretrained network}} + \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right].$$

Since $p_\theta(Z_{s(i)}|Z_{t(i)})$ is parameterized via the pretrained network prediction $\mathbf{x}_\theta(Z_{t(i)})$, we get

$$\mathrm{KL}\Big(q(Z_{s(i)}|Z_{t(i)},\mathbf{x}) \,\|\, p_\theta(Z_{s(i)}|Z_{t(i)},\mathbf{x})\Big) = \mathrm{KL}\Big(q(Z_{s(i)}|Z_{t(i)},\mathbf{x}) \,\|\, q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\theta(Z_{t(i)}))\Big).$$

Following the argument of **Theorem A.1**, we distinguish two cases for $Z_{t(i)}$. For $Z_{t(i)} = \mathbf{m}$, the KL-divergence evaluates to

$$\mathrm{KL}\big(q(Z_{s(i)}|Z_{t(i)},\mathbf{x}) \,\|\, q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\theta(Z_{t(i)}))\big) = \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}}\,\mathrm{CE}\big(\mathbf{x},\mathbf{x}_\theta(Z_{t(i)})\big),$$

while for $Z_{t(i)} \neq \mathbf{m}$ the KL-divergence vanishes. Thus we obtain

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right]$$
$$= \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}}\,\mathrm{CE}\big(\mathbf{x},\mathbf{x}_\theta(Z_{t(i)})\big)\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}} + \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right].$$

Substituting the above expression in the conditional likelihood (11), we get

$$\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \leq\; & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\big[-\log p_\varphi(\mathbf{x}|Z_0,\mathbf{y})\big] + \mathbb{E}_{q(Z_1|\mathbf{x})}\Big[\log \frac{q(Z_1|\mathbf{x})}{p_\varphi(Z_1|\mathbf{y})}\Big] \\
& + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}}\,\mathrm{CE}\big(\mathbf{x},\mathbf{x}_\theta(Z_{t(i)})\big)\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
& + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right]. \qquad (12)
\end{aligned}$$

Expected log-likelihood ratio under q. We now examine the expected difference in log-likelihoods between the prior and posterior transition distributions under the conditional law of the forward process. Recall that the reverse transition under our parameterization is

$$p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y}) = q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))\,q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)})).$$

Consider the expectation of the log-likelihood ratio under the forward posterior $q(Z_{s(i)}|Z_{t(i)},\mathbf{x})$:

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right].$$

Substituting the definition of p_φ, we obtain

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))\,q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))}\right]$$
$$= \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))} - \log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\right]$$
$$= \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))}\right] - \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\right].$$

Since $p_\theta(Z_{s(i)}|Z_{t(i)})$ is represented by the network prediction $\mathbf{x}_\theta(Z_{t(i)})$, we can rewrite the first expectation as:

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))}\right] = \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\theta(Z_{t(i)}))}{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))}\right].$$

Putting everything together, the expected log-likelihood ratio under q becomes

$$\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right] \tag{13}$$
$$= \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\theta(Z_{t(i)}))}{q(Z_{s(i)}|Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))}\right] - \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\right]. \tag{14}$$

Similarly to the proof of **Theorem A.1**, we consider two cases to compute the first expectation. **Case I:** $Z_{t(i)} = \mathbf{m}$**.** Using the masked diffusion inference posterior derived from (3), when $Z_{t(i)} = \mathbf{m}$ we have

$$q(Z_{s(i)} = \mathbf{m}\,|\,Z_{t(i)} = \mathbf{m},\mathbf{x}) = \frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}, \quad q(Z_{s(i)} = \mathbf{x}\,|\,Z_{t(i)} = \mathbf{m},\mathbf{x}) = \frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}.$$

Under the pretrained model parameterization (4), the corresponding terms are

$$q(Z_{s(i)} = \mathbf{m}\,|\,Z_{t(i)} = \mathbf{m},\mathbf{x}_\theta(Z_{t(i)})) = \frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}},$$
$$q(Z_{s(i)} = \mathbf{x}\,|\,Z_{t(i)} = \mathbf{m},\mathbf{x}_\theta(Z_{t(i)})) = \frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\langle \mathbf{x}_\theta(Z_{t(i)}),\mathbf{x}\rangle,$$

and likewise with $\mathbf{x}_\varphi(Z_{t(i)})$ replacing $\mathbf{x}_\theta(Z_{t(i)})$ for our parameterization. Hence,

$$\mathbb{E}_{q(Z_{s(i)}\,|\,Z_{t(i)}=\mathbf{m},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}\,|\,Z_{t(i)}=\mathbf{m},\mathbf{x}_\theta(Z_{t(i)}))}{q(Z_{s(i)}\,|\,Z_{t(i)}=\mathbf{m},\mathbf{x}_\varphi(Z_{t(i)}))}\right]$$
$$= \frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}\log\frac{\frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}}{\frac{1-\alpha_{s(i)}}{1-\alpha_{t(i)}}} + \frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\log\frac{\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\langle \mathbf{x}_\theta(Z_{t(i)}),\mathbf{x}\rangle}{\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\langle \mathbf{x}_\varphi(Z_{t(i)}),\mathbf{x}\rangle}$$
$$= \frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\log\frac{\langle \mathbf{x}_\theta(Z_{t(i)}),\mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}),\mathbf{x}\rangle}.$$

Case II: $Z_{t(i)} \neq \mathbf{m}$**.** When the token is already revealed, the posterior is deterministic:

$$q(Z_{s(i)} = Z_{t(i)}\,|\,Z_{t(i)} \neq \mathbf{m},\mathbf{x}) = 1,$$

and this form is identical for the $\mathbf{x}_\theta$- and $\mathbf{x}_\varphi$-parameterized posteriors as well:

$$q(Z_{s(i)} = Z_{t(i)}\,|\,Z_{t(i)} \neq \mathbf{m},\mathbf{x}_\theta(Z_{t(i)})) = 1, \quad q(Z_{s(i)} = Z_{t(i)}\,|\,Z_{t(i)} \neq \mathbf{m},\mathbf{x}_\varphi(Z_{t(i)})) = 1.$$

Therefore,

$$\mathbb{E}_{q(Z_{s(i)}\,|\,Z_{t(i)}\neq\mathbf{m},\mathbf{x})}\left[\log \frac{q(Z_{s(i)}\,|\,Z_{t(i)},\mathbf{x}_\theta(Z_{t(i)}))}{q(Z_{s(i)}\,|\,Z_{t(i)},\mathbf{x}_\varphi(Z_{t(i)}))}\right] = 0$$

Combining both cases,

$$\mathbb{E}_{q(Z_{s(i)} \mid Z_{t(i)},\mathbf{x})}\left[\log \frac{q(Z_{s(i)} \mid Z_{t(i)}, \mathbf{x}_\theta(Z_{t(i)}))}{q(Z_{s(i)} \mid Z_{t(i)}, \mathbf{x}_\varphi(Z_{t(i)}))}\right] = \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \log \frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle} \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}},$$

which simplifies (13) as follows:

$$\begin{aligned}
&\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right] \\
&= \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \log \frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle} \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}} - \mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\Big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\Big]. \qquad (15)
\end{aligned}$$

Substituting (15) into (12), the conditional likelihood can be bounded as

$$\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le\ & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0,\mathbf{y})] + \mathbb{E}_{q(Z_1|\mathbf{x})}\Big[\log \frac{q(Z_1|\mathbf{x})}{p_\varphi(Z_1|\mathbf{y})}\Big] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \mathrm{CE}(\mathbf{x}, \mathbf{x}_\theta(Z_{t(i)}))\, \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\mathbb{E}_{q(Z_{s(i)}|Z_{t(i)},\mathbf{x})}\left[\log \frac{p_\theta(Z_{s(i)}|Z_{t(i)})}{p_\varphi(Z_{s(i)}|Z_{t(i)},\mathbf{y})}\right] \\
=\ & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}[-\log p_\varphi(\mathbf{x}|Z_0,\mathbf{y})] + \mathbb{E}_{q(Z_1|\mathbf{x})}\Big[\log \frac{q(Z_1|\mathbf{x})}{p_\varphi(Z_1|\mathbf{y})}\Big] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \mathrm{CE}(\mathbf{x}, \mathbf{x}_\theta(Z_{t(i)}))\, \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \log \frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle} \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\Big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\Big], \qquad (16)
\end{aligned}$$

where in the last step we use the fact that $q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))$ is independent of $Z_{s(i)}$ given $Z_{t(i)}$.

Treatment of boundary conditions. In masked diffusion the terminal state Z_1 is absorbing (all-mask), hence $q(Z_1|\mathbf{x}) = p_\varphi(Z_1|\mathbf{y})$ and the boundary KL-divergence vanishes. Thus (16) becomes

$$\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le\ & \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\big[-\log p_\varphi(\mathbf{x}|Z_0,\mathbf{y})\big] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \mathrm{CE}\big(\mathbf{x}, \mathbf{x}_\theta(Z_{t(i)})\big)\, \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)} - \alpha_{t(i)}}{1-\alpha_{t(i)}} \log \frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle} \mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big]. \qquad (17)
\end{aligned}$$

Since $\mathbf{x}$ is a deterministic function of Z_0, the reconstruction term in (17) simplifies to:

$$
\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathbb{E}_{q(Z_0|\mathbf{x})}\big[-\log p_\varphi(\mathbf{x}|Z_0)\big] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\,\mathrm{CE}\big(\mathbf{x}, \mathbf{x}_\theta(Z_{t(i)})\big)\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\,\log\frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle}\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big]. \qquad (18)
\end{aligned}
$$

Finally, the decoder is same for unconditional generation and posterior sampling, i.e., $p_\varphi(\mathbf{x}|Z_0) = p_\theta(\mathbf{x}|Z_0)$. Substituting this property in (18) yields

$$
\begin{aligned}
-\log p_\varphi(\mathbf{x}|\mathbf{y}) \le \mathbb{E}_{q(Z_{0:1}|\mathbf{x})}\big[-\log p_\theta(\mathbf{x}|Z_0)\big] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\,\mathrm{CE}\big(\mathbf{x}, \mathbf{x}_\theta(Z_{t(i)})\big)\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&+ \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\,\log\frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle}\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
&- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big] \coloneqq \mathcal{L}_{\mathrm{APS}}(\mathbf{x}, \mathbf{y}; \varphi) \qquad (19)
\end{aligned}
$$

Since $\mathrm{CE}\big(\mathbf{x}, \mathbf{x}_\theta(Z_{t(i)})\big) = -\log\langle\rangle$, the first two terms in (19) equals to the standard NELBO (5) used to train the masked diffusion model. Therefore, we have

$$
\begin{aligned}
\mathcal{L}_{\mathrm{APS}}(\mathbf{x}, \mathbf{y}; \varphi) = \mathcal{L}_{\mathrm{NELBO}}(\mathbf{x}; \theta) + \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\left[\frac{\alpha_{s(i)}-\alpha_{t(i)}}{1-\alpha_{t(i)}}\,\log\frac{\langle \mathbf{x}_\theta(Z_{t(i)}), \mathbf{x}\rangle}{\langle \mathbf{x}_\varphi(Z_{t(i)}), \mathbf{x}\rangle}\,\mathbf{1}_{\{Z_{t(i)}=\mathbf{m}\}}\right] \\
- \sum_{i=1}^{T} \mathbb{E}_{q(Z_{t(i)}|\mathbf{x})}\big[\log q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))\big],
\end{aligned}
$$

completing the statement of the theorem. □

Implications. **Theorem A.2** establishes a principled upper bound on the negative log-posterior likelihood when posterior sampling is performed without additional training.

- *Reuse of pretrained objective.* The bound $\mathcal{L}_{\mathrm{APS}}(\mathbf{x}, \mathbf{y}; \varphi)$ is expressed in terms of the standard masked diffusion $\mathcal{L}_{\mathrm{NELBO}}(\mathbf{x}; \theta)$, meaning that posterior sampling can be performed using pretrained masked diffusion models.
- *Adaptation gap.* The log-ratio correction term quantifies the mismatch between the pretrained transitions $\mathbf{x}_\theta$ and the proposed posterior transitions $\mathbf{x}_\varphi$, effective only at masked positions. This isolates the additional cost of posterior sampling.
- *Measurement consistency.* The final summation enforces alignment with the measurement likelihood $q(\mathbf{y}|\mathbf{x}_\varphi(Z_{t(i)}))$, ensuring that the sampler accounts for observations at each diffusion step.
- *Boundary conditions.* As in the training bound, the absorbing mask state renders the boundary KL-divergence constant (often zero), so the effective objective simplifies to the pretrained NELBO plus adaptation and measurement terms.
- *Test-time posterior sampling.* Together, the decomposition clarifies how posterior sampling can be performed without retraining: start from the pretrained NELBO and add corrections for model adaptation at masked positions while incorporating measurements via tilting.

B Additional Experiments

This section provides supplementary details and evaluations of our APS method. We first describe implementation details for both inverse problems and stylization (§B.1), followed by algorithmic analysis including pseudocode and hyperparameter studies (§B.2). We then examine the impact of design choices in our ablations (§B.3), evaluate using the same prior (§B.4), and discuss computational complexity (§B.5). Next, we outline the compared baselines (§B.6) and summarize benchmarks and metrics (§B.7). Finally, we present additional results (§B.8) and limitations (§B.9).

B.1 Implementation Details

B.1.1 Inverse Problems

For inverse problems, we implement our test-time optimization using two main configurations corresponding to the APS (512×512) and APS-L (1024×1024) results reported in §4. The full reverse sampling process is discretized into 15 time steps following a cosine mask schedule, using a classifier-free guidance scale of 3.5.

At each of the 15 reverse steps, we perform an inner optimization loop to ensure measurement consistency. This loop consists of 100 optimization steps using the Adam optimizer with a learning rate of $\eta = 1.0$. The total loss function is a weighted sum of three components: (1) a reconstruction loss, which could be L1 or L2 norm; we choose L1 norm $\|\mathbf{y} - \mathcal{A}(\hat{\mathbf{x}})\|_1$ because it is known to generate sharper quality (Van Den Oord et al., 2017; Yu et al., 2024), (2) a VGG perceptual loss with a coefficient of $\lambda_{\text{p}} = 10^{-3}$, and (3) a prior preservation loss $\mathcal{L}_{\text{prior}}(\mathbf{x}_\varphi(\mathbf{z}_{t(i)}), \mathbf{x}_\theta(\mathbf{z}_{t(i)})) = -\frac{1}{L}\sum_{l=1}^{L} \log\langle \mathbf{x}_\varphi(\mathbf{z}_{t(i)}), \mathbf{x}_\theta(\mathbf{z}_{t(i)})\rangle \mathbf{1}_{\{\mathbf{z}^l_{t(i)}=\mathbf{m}\}}$ with coefficient $\lambda_{\text{pp}} = 10^{-4}$. To reduce the number of hyper-parameters introduced by our **Algorithm 1** and ensure fast convergence, we make the following practical implementation choices:

- initialize $\mathbf{x}_\varphi(\mathbf{z}_{t(i)})$ with $\mathbf{x}_\theta(\mathbf{z}_{t(i)})$: Lines 6 and 9,
- choose a small decaying learning rate $\eta = 1.0$ in Adam: Line 15,
- optimize for a fewer iterations $T = 15$ and $M = 100$: Lines 5 and 8.

This has an equivalent effect of prior preservation. We note that the optimization loop (Lines 8-16) computes gradients with respect to the logits φ. This is significantly cheaper than the expensive backpropagation steps through denoising network (here, 8B parameters for MMaDA) used in prior continuous diffusion approaches (Rout et al., 2023; 2024; Chung et al., 2024). Thus, we set $\lambda_{\text{pp}} = 0$ by default. Alternatively, one could initialize φ with all zeros or randomly and set M very large until convergence while enforcing the prior preservation loss ($\lambda_{\text{pp}} \neq 0$) in Line 14.

The specific parameters for each degradation operator $\mathcal{A}(\cdot)$ vary by task, with all measurements simulated by adding Gaussian noise of $\sigma = 0.05$. For super resolution, we use a $4\times$ downsampling factor. For Gaussian Deblurring, the operator is a Gaussian kernel of size 61×61 with a standard deviation of 3.0. Motion Deblurring uses a kernel of the same size with an intensity parameter of 0.5 (on a scale of 0 for linear to 1 for highly nonlinear), corresponding to a moderately nonlinear motion path. For Inpainting, we randomly remove 70% of pixels. Nonlinear Deblurring kernels are generated using the `KernelWizard`[1] model from the `bkse`[2] library. Finally, High Dynamic Range (HDR) reconstruction is modeled by scaling the image data and clipping the result, following the operation $\text{clip}(\text{data} \times 2, -1, 1)$.

B.1.2 Reference-based Stylization

Our approach to training-free, reference-based stylization leverages the core APS framework by framing the task as a *highly nonlinear* inverse problem. Let $\hat{\mathbf{x}}_{\text{ref}}$ denote the conditional reference image providing the style. The target measurement, a style vector $\mathbf{y}$, is obtained by applying a pretrained Contrastive Style Descriptor (CSD) (Somepalli et al., 2024) model as our measurement

[1] https://github.com/LeviBorodenko/motionblur
[2] https://github.com/VinAIResearch/blur-kernel-space-exploring

operator $\mathcal{A}(\cdot)$ to this reference image, i.e., $\mathbf{y} = \mathcal{A}(\hat{\mathbf{x}}_{\text{ref}})$. At each reverse diffusion step t, APS aims to generate a sample whose style matches this target.

As discussed in §3.2, there are two main stages of APS. In the first stage (§3.2.1), we perform a differentiable forward pass by computing the expected codebook embedding $\bar{\mathbf{x}}^l = \sum_{k=1}^{K} \mathbf{c}_k \mathbf{x}_\varphi^l(\mathbf{z}_{t(i)})$ for $l = 1, \ldots, L$ using the model's output probabilities $\mathbf{x}_\theta(\mathbf{z}_t)^l$ as initial condition for $\varphi_{t(i)}^l$. The straight-through estimator is then used to obtain a differentiable image representation $\hat{\mathbf{x}} = \mathcal{D}(\tilde{\mathbf{x}})$. To guide the optimization, the measurement consistency loss is calculated as the cosine distance between the style vector of the generated image and the target style vector $\mathbf{y}$ as follows:

$$\mathcal{L}_{\text{style}}(\varphi_{t(i)}) = 1 - \frac{\langle \mathcal{A}(\hat{\mathbf{x}}), \mathbf{y} \rangle}{\|\mathcal{A}(\hat{\mathbf{x}})\| \|\mathbf{y}\|}.$$

For each reverse step, we perform 100 optimization steps using an Adam optimizer with a learning rate of 0.1. Gradients from this style loss are backpropagated through the frozen VQ-VAE decoder $\mathcal{D}(\cdot)$ and the straight-through estimator $\tilde{\mathbf{x}}$ to update all the entries of the conditional probability table by directly updating the logits $\varphi_{t(i)} = \{\varphi_{t(i)}^l\}_{l=1}^L$.

In the second stage (§3.2.2), the posterior estimate $\mathbf{x}_\varphi(\mathbf{z}_{t(i)})$ obtained from the first stage is used adaptively unmask anchor tokens in the sequence. Both the processes continue over 15 total steps following a cosine mask schedule. The fully unmasked sequence satisfies the stylistic constraints without requiring any task-specific retraining of the foundation model.

B.1.3 High-Resolution Inference

To demonstrate the scalability of our method, we experiment with a higher-resolution setting. For a fair comparison on our 256×256 benchmark, we follow the upsampling protocol described in G2D2 (Murata et al., 2024). Specifically, for both our standard (512×512) and large-scale (APS-L, 1024×1024) models, we first upsample the benchmark images to the model's native resolution before applying the forward operator. For the APS-L configuration, this increases the number of visual tokens by a factor of four (from 1024 to 4096)[3]. The Transformer-based MMaDA (Yang et al., 2025) model accommodates this by processing a longer sequence without architectural changes. After the high-resolution reconstruction is complete, we downsample the output back to the benchmark's native 256×256 resolution for evaluation. As demonstrated in our experiments (§4.1, §4.2), this approach further improves performance, achieving substantial gains in both PSNR and LPIPS.

B.2 Algorithm Details

Algorithm 1 details our APS procedure. The process begins with a fully masked latent space, $\mathbf{z}_1$, and iteratively refines the image over T reverse diffusion steps. Each step features an inner optimization phase designed to align the model's predictions with the given measurement $\mathbf{y}$. To enable gradient-based optimization through the discrete quantization step which assigns each dimension of an embedding to its nearest value in $\{-1, 1\}$ we employ the straight-through estimator (STE) via the stop-gradient operator $\lfloor \cdot \rceil_{\text{sg}}$.

This optimization is guided by a composite loss function. For reconstruction, we primarily use the Mean Absolute Error (MAE, or L1 loss), which we find produces perceptually superior results compared to the Mean Squared Error (MSE). While MSE corresponds to maximizing the Gaussian log-likelihood, MAE is typically more robust. This is supplemented with a VGG perceptual loss to enforce similarity in the features extracted from the measurements, further improving visual quality in the image space[4]. Once the optimization at a given step is complete, the Anchored Remasking strategy uses the tilted measure to estimate confidence assigned to the token chosen via quantization. This is followed for each position to selectively unmask tokens for the next iteration.

To determine the optimal weight for the perceptual loss, we conducted an ablation study on its coefficient, λ_{p}. As shown in Table 4, a value of 10^{-3} provides the optimal trade-off between reconstruction fidelity (PSNR) and perceptual quality (LPIPS) across our benchmarks.

[3]The visual tokenizer (Yu et al., 2024) used in MMaDA (Yang et al., 2025) uses 16× downscaling, generating $1024 = 32 \times 32$ tokens for an image of size 512×512.

[4]The LPIPS metric measures visual quality in the image space.

Algorithm 1: Test-Time Anchored Posterior Sampling (APS)

Input: Diffusion steps T, measurement $\mathbf{y}$, denoiser $\mathbf{x}_\theta^{\text{logits}}(\cdot)$, operator $\mathcal{A}(\cdot)$, decoder $\mathcal{D}(\cdot)$, Codebook $\mathcal{C} = \{\mathbf{c}_k\}_{k=1}^K$, Lookup-Free Quantizer $\mathcal{Q}_{\text{lfq}}$.
Tunable parameters: Optimization steps M, learning rate η, loss coefficients $\lambda_{\text{p}}, \lambda_e$.
Output: Reconstructed image $\hat{\mathbf{x}}_{\text{final}}$.
Initialize latent state $\mathbf{z}_1 \leftarrow \{\mathbf{m}\}^L$ (all masked)
for $i = 1$ **to** T **do**
  $\varphi \leftarrow \mathbf{x}_\theta^{\text{logits}}(\mathbf{z}_{t(i)})$ ▷ Quantized Expectation §3.2.1
  $\mathbf{x}_\theta(\mathbf{z}_{t(i)}) = \text{Softmax}(\mathbf{x}_\theta^{\text{logits}}(\mathbf{z}_{t(i)})$
  for $m = 1$ **to** M **do**
    $\mathbf{x}_\varphi(\mathbf{z}_{t(i)}) = \text{Softmax}(\varphi)$
    $\bar{\mathbf{x}} \leftarrow \sum_{k=1}^K \mathbf{c}_k \mathbf{x}_\varphi(\mathbf{z}_{t(i)})$
    $\mathbf{x} = \mathcal{Q}_{\text{lfq}}(\bar{\mathbf{x}})$
    $\tilde{\mathbf{x}} \leftarrow \bar{\mathbf{x}} + \left[\mathbf{x} - \bar{\mathbf{x}}\right]_{\text{sg}}$
    $\hat{\mathbf{x}} \leftarrow \mathcal{D}(\tilde{\mathbf{x}})$
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{measurement}}(\mathcal{A}(\hat{\mathbf{x}}), \mathbf{y}) + \lambda_{\text{p}} \mathcal{L}_{\text{perceptual}}(\mathcal{A}(\hat{\mathbf{x}}), \mathbf{y}) + \lambda_{\text{pp}} \mathcal{L}_{\text{prior}}(\mathbf{x}_\varphi(\mathbf{z}_{t(i)}), \mathbf{x}_\theta(\mathbf{z}_{t(i)}))$
    $\varphi \leftarrow \text{Adam}(\varphi, \nabla_\varphi \mathcal{L}, \eta)$
  end for
  $\mathbf{x}_\varphi^*(\mathbf{z}_{t(i)}) = \text{Softmax}(\varphi)$ ▷ Anchored Remasking §3.2.2
  $\bar{\mathbf{x}}^* \leftarrow \sum_{k=1}^K \mathbf{c}_k \mathbf{x}_\varphi^*(\mathbf{z}_{t(i)})$
  $\mathbf{x} = \mathcal{Q}_{\text{lfq}}(\bar{\mathbf{x}}^*)$
  $\kappa^l \leftarrow \langle (\mathbf{x}_\varphi^*(\mathbf{z}_{t(i)})^l, \mathbf{x}^l \rangle$ for $l = 1, \ldots, L$
  $\mathcal{P}_{t(i)} \leftarrow \{l : \kappa^l \geq \tau_{t(i)}\}$
  $\mathbf{z}_{s(i)} \leftarrow \text{UpdateState}(\mathbf{z}_{t(i)}, \mathbf{x}, \mathcal{P}_{t(i)})$
end for
$\hat{\mathbf{x}}_{\text{final}} \leftarrow \mathcal{D}(\mathbf{z}_0)$
return $\hat{\mathbf{x}}_{\text{final}}$

Table 4: **Hyperparameter sweeps on perceptual loss coefficient (λ_{p}) on ImageNet and FFHQ.** Highlighted rows denote the chosen setting (1e-3), which offers a trade-off among perceptual quality (LPIPS), distortion (PSNR), and structure (SSIM).

(a) ImageNet

λ_{p}	LPIPS ↓	PSNR ↑	SSIM ↑
0.0	0.416	23.98	0.652
1e-5	0.402	24.05	**0.658**
1e-4	0.380	**24.06**	0.655
1e-3	0.334	23.61	0.639
1e-2	**0.325**	22.06	0.566
5e-2	0.328	21.44	0.543
1e-1	0.327	21.29	0.539

(b) FFHQ

λ_{p}	LPIPS ↓	PSNR ↑	SSIM ↑
0.0	0.311	27.32	0.801
1e-5	0.301	27.40	0.803
1e-4	0.268	**27.67**	**0.812**
1e-3	**0.247**	26.61	0.781
5e-3	0.252	24.90	0.729
1e-2	0.256	24.49	0.715
1e-1	0.260	23.44	0.692

Effect of perceptual loss Coefficient. Table 4 reports the effect of varying the perceptual loss coefficient on ImageNet and FFHQ. Small weights (10^{-5}, 10^{-4}) only marginally improve perceptual quality over the baseline, while larger weights (10^{-2}, 10^{-1}) overly emphasize perceptual similarity at the cost of distortion and structure. A coefficient of 10^{-3} provides the best balance: on ImageNet, it reduces LPIPS from 0.416 to 0.334 while maintaining PSNR 23.61 and SSIM 0.639, and on FFHQ, it achieves the lowest LPIPS (0.247) with a slight drop in PSNR (26.61) and SSIM (0.781). We therefore adopt 10^{-3} as the default across datasets. Notably, our approach introduces only this single hyperparameter, whereas G2D2 relies on multiple carefully tuned schedules (e.g., four different coefficients) that must be re-optimized for each task. In contrast, we use the same setting for all inverse problems, underscoring the robustness and simplicity of our posterior sampler relative to prior discrete diffusion methods.

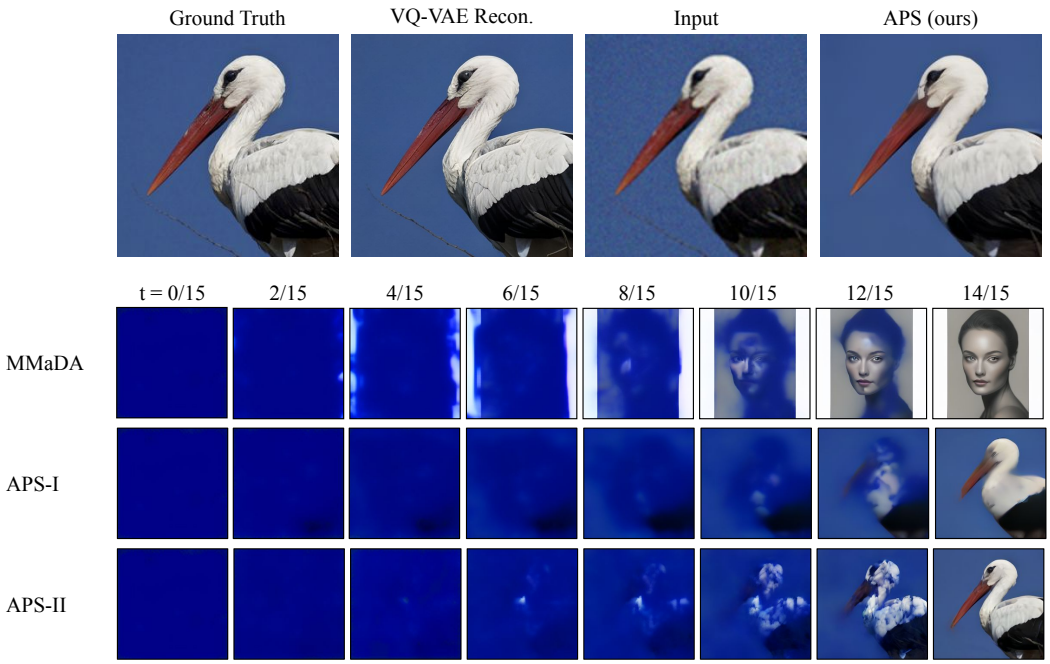

Figure 6: **Ablation study of design choices in APS.** We study the posterior sampling problem with the input of the noisy image of a bird (top row, third image from left). In the base case of using MMada (second row), since it does not take an image as an input (but only text), the final image is inconsistent with the noisy input image (image of person instead of bird). Thus, this shows the effect of *standard remasking*, where high-confidence background tokens are unmasked early. The third row (APS-I) adds *quantized expectation*, which mitigates sampling bias and improves consistency with the measurement, but still relies on prior-based confidence remasking. This results in blurry reconstructions. Finally, the fourth row (APS-II) combines *quantized expectation* with *anchored remasking*, preserving optimized anchor tokens while remasking uninformative background tokens. This combination yields the most stable and measurement-consistent generations, as discussed in §B.3. Indeed, we note that best case one could hope for is the direct VQ-VAE reconstruction (top row, second from left) of the Ground Truth image (top row, leftmost). We observe that APS (top row, rightmost) is comparable in quality to the VQ-VAE reconstruction.

B.3 Ablation Study

To better understand the impact of our algorithmic innovations, we conduct a systematic ablation study on ImageNet SR (4×), as shown in Figure 6 (top row). Our analysis focuses on three core design choices: (i) *standard remasking*, which highlights the limitations of confidence-based token selection under the prior; (ii) *quantized expectation*, which addresses sampling bias and improves measurement consistency; and (iii) *anchored remasking*, which preserves informative tokens identified by optimization while suppressing spurious high-confidence background tokens. Together, these ablations disentangle the contributions of each component, providing both theoretical insight and qualitative evidence (Figure 6) into how APS achieves stable and measurement-consistent posterior sampling.

B.3.1 Effect of Standard Remasking

Standard remasking in MaskGIT (Chang et al., 2022) and MMaDA (Yang et al., 2025) proceeds as follows. At each iteration, the base denoiser produces a distribution over all tokens $\mathbf{x}_\theta(\mathbf{z}_{t(i)}) = \{\mathbf{x}_\theta^l(\mathbf{z}_{t(i)})\}_{l=1}^L$ given the current partially unmasked sequence $\mathbf{z}_t$. Sampled tokens $\mathbf{x} = \{\mathbf{x}^l\}_{l=1}^L$ are drawn independently at each position, and previously unmasked tokens are carried over to the next

state $\mathbf{z}_s$. For masked positions, the confidence of each sampled token $\mathbf{x}^l$ is defined as

$$\kappa_t^l = \langle \mathbf{x}_\theta^l(\mathbf{z}_{t(i)}), \mathbf{x}^l \rangle.$$

Tokens to unmask are then selected based on an adaptive threshold schedule (e.g., cosine),

$$\mathcal{P}_t = \{l : \kappa_t^l \geq \tau_t\},$$

which favors unmasking the most confident tokens under the prior distribution $\mathbf{x}_\theta(\mathbf{z}_t)$.

In language generation, this corresponds to unmasking frequent, low-information tokens (e.g., articles or conjunctions) (Rout et al., 2025a). Analogously in images, the model tends to unmask background regions first, since they dominate training statistics and are easier to predict. As a result, informative foreground tokens remain masked until late in the process, limiting semantic guidance and increasing conditional entropy.

Qualitatively, Figure 6 (second row) illustrates this phenomenon. While the model confidently unmasks background tokens early, the salient object is revealed only much later, highlighting the limitation of standard remasking for posterior sampling.

B.3.2 Effect of Quantized Expectation

Standard remasking suffers from two key issues: (i) sampling bias, and (ii) independent token-wise confidence. When sampling tokens directly from the unconditional distribution $\mathbf{x}_\theta(\mathbf{z}_t)$, the model may pick unrelated or spurious tokens, which—once unmasked—remain fixed in all future steps. This introduces inconsistency and often locks the model into poor generations.

To address this, we propose *quantized expectation* (§3.2.1). Instead of sampling, we tilt the unconditional distribution $\mathbf{x}_\theta(\mathbf{z}_t)$ towards the measurement likelihood, obtaining an approximate posterior $\mathbf{x}_\varphi(\mathbf{z}_t)$. We then optimize in the span of codebook embeddings by treating the tilted probabilities as coefficients in a linear combination and passing their expectation through the decoder using a straight-through estimator. The resulting embedding is then quantized back to the nearest valid token. This procedure implicitly maximizes the measurement likelihood, avoids sampling noise, and enables the discovery of tokens with zero prior probability mass but strong measurement consistency–leading to a better posterior sample.

We treat such tokens as "anchor tokens," since they minimize reconstruction error and provide critical guidance under the measurement operator. As shown in Figure 6 (third row, APS-I), quantized expectation corrects sampling bias and yields reconstructions that remain consistent with observations throughout the denoising trajectory.

B.3.3 Effect of Anchored Remasking

A limitation of confidence-based remasking under the prior is that anchor tokens, obtained through our optimization procedure, may receive near-zero probability mass under $\mathbf{x}_\theta(\mathbf{z}_t)$. As a result, the model would discard these informative tokens in favor of background tokens, which the prior predicts with high confidence. This reintroduces the very bias we aim to avoid.

To address this, we compute token confidence using the posterior estimate $\mathbf{x}_\varphi(\mathbf{z}_t)$ rather than the unconditional prior. In this way, anchor tokens identified via quantized expectation are preserved, while low-likelihood background tokens are remasked. Qualitatively, this effect is evident at $t = 6/15$ in Figure 6 (fourth row, APS-II): unlike the standard prior-based strategy (second row), which prematurely unmasks background pixels, our approach commits to anchor tokens aligned with the bird's body. This ensures that the background (blue sky in the measurement) is correctly downweighted, as it is inconsistent with the prior white background generated by $\mathbf{x}_\theta$. Subsequent steps therefore refine the image conditioned on these anchor tokens, reducing conditional entropy and producing reconstructions that remain faithful to the measurements.

B.4 Evaluation Under Identical Discrete Diffusion Prior

We compare APS to G2D2 using the *official* G2D2 codebase (§B.6) on 100 images from the FFHQ validation set. Both methods use an identical compute budget: 100 reverse diffusion steps, each with 30 inner optimization steps. As shown in Table 5, APS consistently improves over G2D2

Table 5: **Quantitative results for super resolution (4×) on FFHQ.** APS consistently outperforms G2D2 across standard evaluation metrics: PSNR, SSIM and LPIPS while the same base generative model VQ-Diffusion. This also shows compatibility of our approach with uniform (mask) discrete diffusion.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
G2D2	25.46	0.717	0.350
APS (Ours)	**26.80**	**0.759**	**0.310**

Table 6: **Quantitative results on sampling efficiency of continuous and discrete samplers.** PDM/LDM denote pixel-/latent-space continuous diffusion models, respectively; VQ-Diffusion and MMaDA are discrete diffusion models. Rows shaded gray report runtimes on a single NVIDIA A6000 GPU copied from G2D2 (Murata et al., 2024); rows shaded orange are measured by us on a single NVIDIA A100 GPU.

Method	Model	Resolution	GPU (GiB)	Time (s)	#Steps
DPS (Chung et al., 2023)	PDM	256	10.7	277	1000
DDRM (Kawar et al., 2022)	PDM	256	5.8	4	20
PSLD (Rout et al., 2023)	LDM	512	20.9	738	1000
ReSample (Song et al., 2024)	LDM	256	7.1	555	500
G2D2 (Murata et al., 2024)	VQ-Diffusion	256	4.7	194	100
DPS (Chung et al., 2023)	PDM	256	10.7	180	1000
LDPS (Rout et al., 2023)	LDM	256	15.4	190	1000
PSLD (Rout et al., 2023)	LDM	256	15.5	194	1000
G2D2 (Murata et al., 2024)	VQ-Diffusion	256	4.7	107	100
APS (ours)	VQ-Diffusion	256	4.6	**106**	100
APS (ours)	MMaDA	256	19.2	**55**	15
PSLD (Rout et al., 2023)	LDM	512	20.9	720	1000
APS (ours)	MMaDA	512	26.2	**121**	15
APS-L (ours)	MMaDA	1024	51.8	**484**	15

across all metrics: PSNR improves from 25.46 to 26.80, SSIM from 0.717 to 0.759, and LPIPS decreases from 0.350 to 0.310. Since the generative prior and compute budget are identical, these performance gains arise purely from our algorithmic innovations: quantized expectation (§3.2.1), anchored remasking (§3.2.2), and the use of a perceptual loss. This experiment demonstrates that APS is not only effective but also compatible with both *mask-based* and *uniform* discrete diffusion frameworks.

B.5 Computational Complexity

Analysis. Table 6 highlights the *sampling efficiency* of our discrete posterior sampler (APS) relative to both continuous-diffusion baselines and the prior discrete sampler G2D2.

Against continuous (pixel/latent) diffusion. Pixel-space samplers (DPS, DDRM) either require very long Markov chains (e.g., DPS: 1000 steps, 277 s) or sacrifice quality when shortened; latent-space samplers (PSLD, ReSample) still need 500–1000 steps and hundreds of seconds per image at 256×256–512×512 resolutions (e.g., PSLD: 738 s at 512×512). In contrast, APS runs with only 15 reverse steps on the MMaDA backbone: 55 s at 256×256 and 121 s at 512×512 on a single A100—*66× shorter chain* for comparable or better perceptual quality. Crucially, at 512×512 APS matches the quality of PSLD as shown in Table 1 while being $\sim 6\times$ faster (PSLD: ∼720–740 s vs. APS: 121 s). When scaling the sequence length to $L{=}4096$ tokens (corresponding resolution 1024×1024), APS-L completes in 484 s and becomes significantly more accurate: ImageNet Gaussian deblur improves by 22.7% LPIPS and 9.6% PSNR over PSLD—with $\sim 1.5\times$ less time.

Against prior discrete diffusion. Under the same budget, APS matches or improves G2D2's runtime while yielding better reconstructions: at 256×256 with VQ-Diffusion both methods use 100 steps, but APS yields higher quality (see Table 5); with the stronger MMaDA prior at 512×512, APS achieves consistently better LPIPS/PSNR as discussed in §4. Moreover, our 1024×1024 configuration (APS-L) demonstrates better test-time scaling behavior compared continuous diffusion: we keep 15 steps and still obtain substantial quality gains at reasonable cost (484 s).

Importantly, continuous methods struggle to match this performance without prohibitive runtimes. PSLD (Rout et al., 2023) already takes nearly 12 minutes to process a single 512×512 image and more complex methods such as P2L (Chung et al., 2024) take around 30 minutes for the same resolution. Therefore, training-free posterior sampling using continuous diffusion at very high resolutions such as 1024×1024 becomes computationally prohibitive.

B.6 Compared Baselines

We compare our method against state-of-the-art posterior samplers using pixel-/latent-space continuous and discrete diffusion models. Each baseline is evaluated under the same data as ours. We follow the experimental setup from G2D2 and reuse the baseline implementations to ensure a fair comparison. To address the resolution mismatch between our model and the benchmark datasets, we adopt the protocol from G2D2 (Murata et al., 2024). The benchmark images are first upsampled to match the input resolution of our base model MMaDA (Yang et al., 2025). The forward corruption operator and our posterior sampling method are then applied in this high-resolution space. Finally, the resulting output is downsampled to the original 256×256 resolution for a fair evaluation. A brief description of each baseline and links to available source code are provided below:

- **DPS** (Chung et al., 2023): A *continuous* diffusion-based method operating in pixel space that solves noisy inverse problems by employing a one-step gradient update in the pixel domain. Source: https://github.com/DPS2022/diffusion-posterior-sampling
- **DDRM** (Kawar et al., 2022): A *continuous* diffusion-based method in pixel space, evaluated using the same base models as DPS. Source: https://github.com/bahjat-kawar/ddrm
- **DiffPIR** (Zhu et al., 2023): A pixel-space *continuous* diffusion-based method for plug-and-play image restoration. Source: https://github.com/yuanzhi-zhu/DiffPIR
- **DAPS** (Zhang et al., 2025): A *continuous* diffusion-based method that employs a decoupled noise annealing strategy to solve inverse problems. Source: https://github.com/zhangbingliang2019/DAPS
- **PSLD** (Rout et al., 2023): A latent-space *continuous* diffusion method that solves inverse problems by performing a one-step gradient update in the latent space and optimizing towards the fixed point of a VAE. Source: https://github.com/LituRout/PSLD
- **ReSample** (Song et al., 2024): A *continuous* latent-space diffusion method that enforces a hard data-consistency constraint during sampling. Source: https://github.com/soominkwon/resample
- **G2D2** (Murata et al., 2024): A *discrete* diffusion posterior sampler that uses a star-shaped noising process and Gumbel-Softmax continuous relaxation to enable gradient guidance in discrete space. We use the source code with the exact implementation and hyperparameters provided in the original paper. Specifically, we use 100 reverse diffusion steps and 30 optimization steps per reverse step. The learning rate and the coefficient for the KL-divergence loss are scheduled logarithmically, as proposed. Source: https://github.com/sony/g2d2
- **SGDD** (Chu et al., 2025): A *discrete* diffusion posterior sampler that uses a split Gibbs sampler, reweights probabilities by Hamming distance, and employs rejection sampling via Metropolis–Hastings. Source: https://github.com/chuwd19/Split-Gibbs-Discrete-Diffusion-Posterior-Sampling

B.7 Benchmarks & Metrics

The APS method is evaluated on standard inverse problem benchmarks and is also shown to generalize to more complex tasks, including non-linear inverse problems and training-free stylization. The evaluation uses two main datasets to cover diverse image types and resolutions, with performance measured using Learned Perceptual Image Patch Similarity (LPIPS) (↓: lower the better), Peak Signal-to-Noise Ratio (PSNR) (↑: higher the better), and Structural Similarity Index (SSIM) (↑: higher the better).

FFHQ (Flickr-Faces-HQ) (Karras et al., 2019):

- **Dataset Focus**: High-resolution face images.
- **Evaluation Set**: To maintain a fair comparison with prior work, specifically SGDD (Chu et al., 2025), G2D2 (Murata et al., 2024) and DAPS (Zhang et al., 2025), our APS algorithm is evaluated on **100 images** (indices $0, 1, \ldots, 99$) from the FFHQ validation set.
- **Tasks**: The evaluation includes (1) linear inverse problems: SR ($4\times$), Gaussian Deblurring, random inpainting, and motion deblur and (2) nonlinear inverse problems: high dynamic range (HDR) and nonlinear blur.

ImageNet (Deng et al., 2009):

- **Dataset Focus**: Diverse natural images.
- **Evaluation Set**: Following the experimental setup by G2D2, a subset of **100 images** is selected from the validation set, ensuring diverse class representation by sampling from classes with indices $0, 10, \ldots, 990$. The specific image list is publicly available in the following text file: `imagenet_val_1k.txt` → https://github.com/XingangPan/deep-generative-prior/.
- **Tasks**: The evaluation includes the same linear and nonlinear inverse problems as in FFHQ.

Licenses and Usage. Both FFHQ and ImageNet datasets used in this work are publicly available and licensed for research use. FFHQ dataset, including its documentation and metadata, is distributed by NVIDIA Corporation under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. ImageNet data is provided free of charge to researchers, for non-commercial research and educational purposes.

B.8 Additional Results

B.8.1 General Inverse Problems

Figure 7 illustrates super resolution ($4\times$) results on ImageNet (Deng et al., 2009). Competing methods—DPS (Chung et al., 2023), DDRM (Kawar et al., 2022), PSLD (Rout et al., 2023), and ReSample (Song et al., 2024)—recover coarse structures but often yield blurry textures or color shifts, while G2D2 (Murata et al., 2024) sharpens details at the cost of noticeable artifacts. In contrast, APS produces sharper and more natural reconstructions across both object and animal categories, closely adhering to the ground truth.

Similarly, Figure 8 shows results for Gaussian deblurring. Continuous methods again capture overall structure but leave residual blur or noise, and G2D2 (Murata et al., 2024) partially enhances details yet struggles with fine textures. APS delivers cleaner and more faithful reconstructions, effectively balancing sharpness and natural appearance across diverse scenes.

Figure 9 compares APS against continuous (DPS (Chung et al., 2023), DDRM (Kawar et al., 2022), PSLD (Rout et al., 2023), ReSample (Song et al., 2024)) and discrete (G2D2 (Murata et al., 2024)) approaches on FFHQ super resolution. Continuous methods capture overall facial structure but tend to oversmooth, leaving blurred or distorted skin textures, while ReSample introduces strong artifacts. G2D2 sharpens details but produces unnatural appearances. APS, by contrast, reconstructs sharper features with natural skin tones and clean edges, yielding perceptually faithful faces across diverse examples and demonstrating clear advantages for high-resolution face restoration.

Figure 10 presents Gaussian deblurring on FFHQ (Karras et al., 2019). DPS (Chung et al., 2023) and DDRM (Kawar et al., 2022) again oversmooth, suppressing fine facial detail; PSLD (Rout et al.,

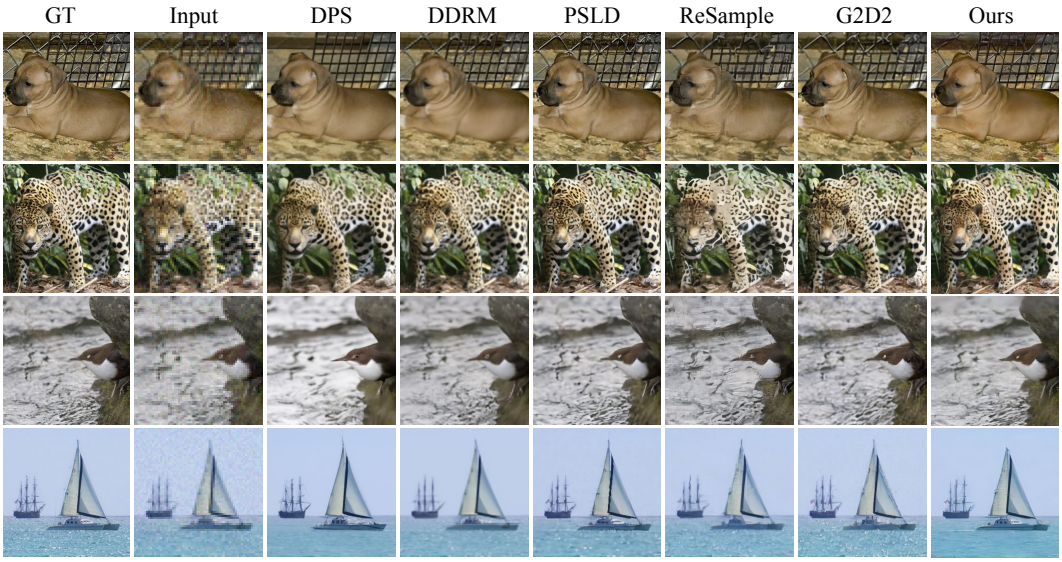

Figure 7: **Additional qualitative results for super resolution ($4\times$) on ImageNet.** Compared to continuous baselines (DPS, DDRM, PSLD, ReSample) and the discrete baseline G2D2, APS produces sharper details and more faithful reconstructions across diverse examples. For instance, in the second row, the leopard's eyes are reconstructed with finer details.

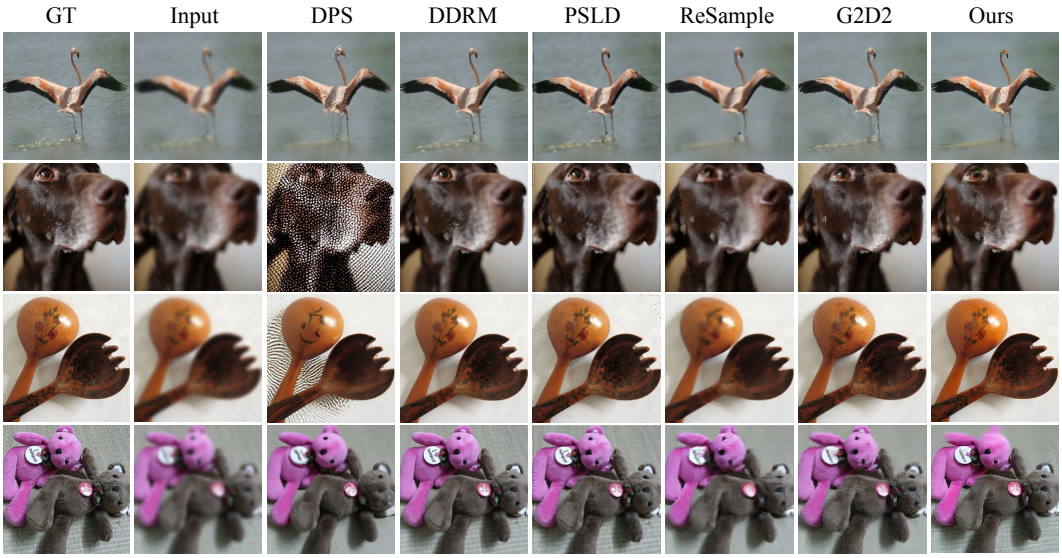

Figure 8: **Additional qualitative results for Gaussian deblurring on ImageNet.** Compared to continuous baselines (DPS, DDRM, PSLD, ReSample) and the discrete baseline G2D2, our APS sampler achieves sharper textures, less artifacts, and more faithful reconstructions across diverse examples. For instance, in the third row, both the floral pattern on the outside of the wooden spoon and the artistic pattern inside it are accurately preserved by our method, whereas most baselines either miss or misrepresent these details.

2023) and ReSample (Song et al., 2024) introduce ringing and plastic-like skin; and G2D2 (Murata et al., 2024) struggles to remove noise from the noisy measurements (Input), creating halos along edges. APS recovers crisp structures such as hair, eyeglass frames, and lip contours while preserving natural highlights and avoiding artifacts, producing reconstructions that are perceptually closer to the ground truth and aligned with our quantitative improvements.

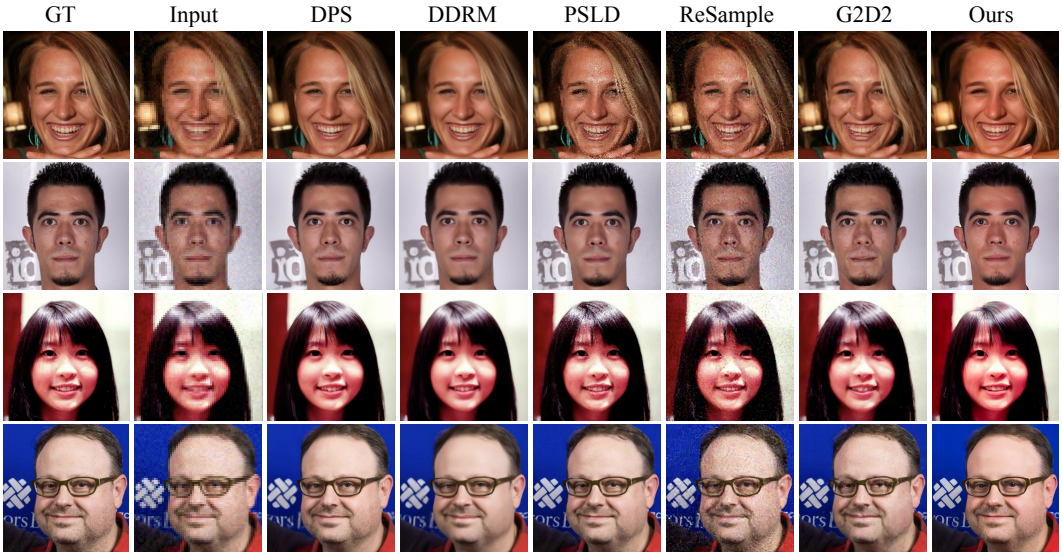

Figure 9: **Additional qualitative results for super resolution** ($4\times$) **on FFHQ.** Continuous baselines (DPS, DDRM, PSLD, ReSample) generate plausible but oversmoothed faces, while the discrete baseline G2D2 often introduces artifacts. In contrast, our APS algorithm reconstructs sharper, more natural faces that closely align with the ground truth. For example, in the second row, our method successfully recovers the small mole on the person's left cheek, a detail overlooked by the baselines.

Figure 10: **Additional qualitative results for Gaussian deblurring on FFHQ.** Our proposed approach recovers sharper facial details and edges with fewer artifacts (e.g., reduced ringing and texture distortions), leading to more natural reconstructions. For instance, in the first row, our method accurately reconstructs the hair strands and their shadow on the face, whereas the prior baselines fail to capture these details as precisely.

Figure 11 shows additional qualitative results on complex linear and nonlinear inverse problems on FHHQ dataset (Karras et al., 2019), showcasing the performance of both APS and APS-L.

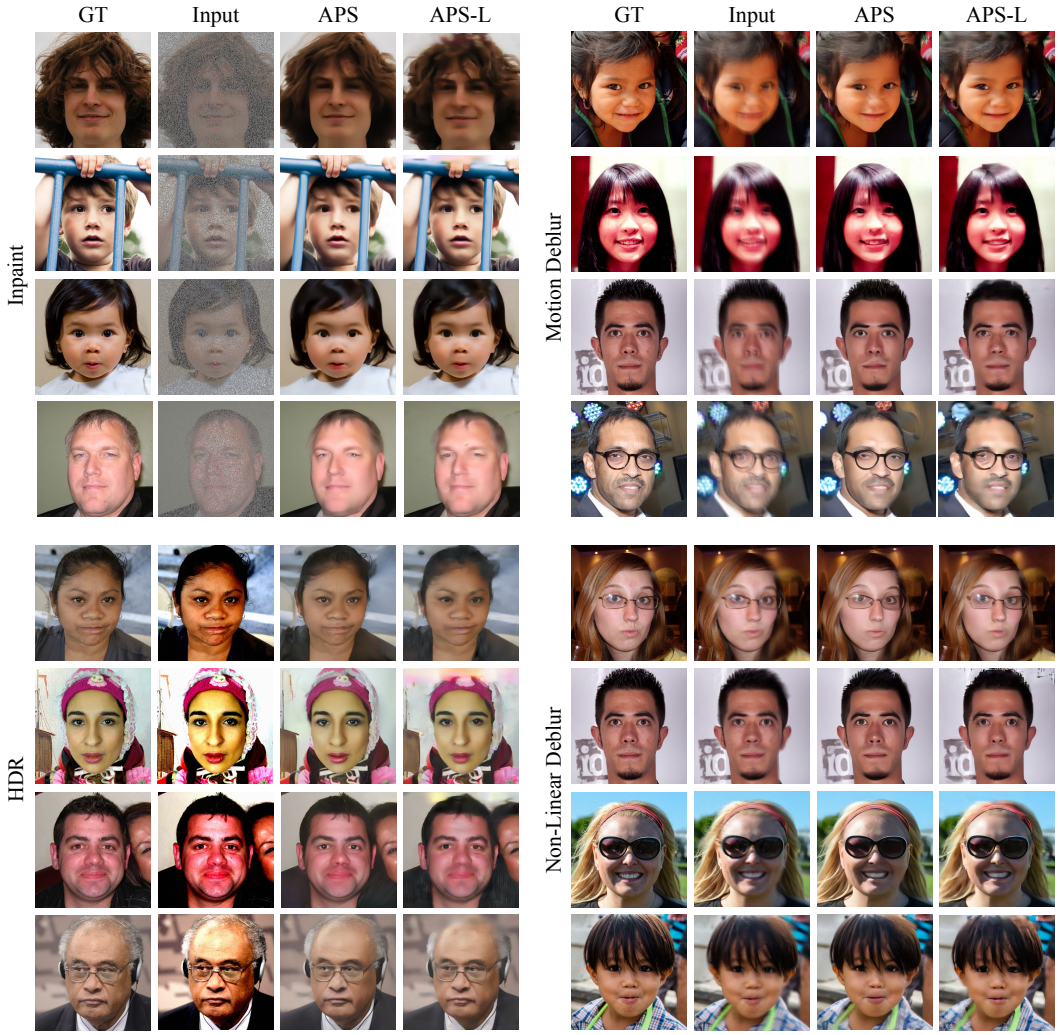

Figure 11: **Qualitative results on FFHQ** for linear (top 4 rows: inpaint and motion deblur) and nonlinear (bottom 4 rows: HDR and non-linear deblur) inverse problems .

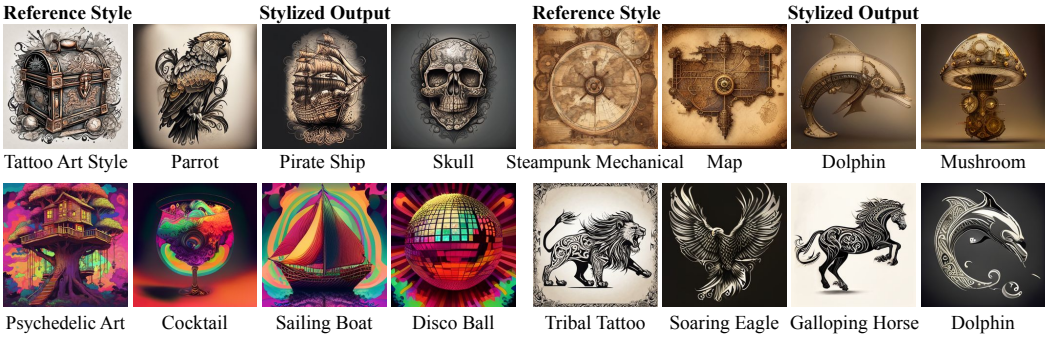

Figure 12: **Additional qualitative results on reference-based stylization.** We show four style-content combinations. For each, our APS optimizer conditions on a single Reference Style image and a text prompt describing the content to generate the Stylized Output images.

B.8.2 Text-Guided Large Block Inpainting

Large block inpainting is a particularly challenging setting for generative models, as it requires filling in large missing regions with semantically coherent and high-fidelity content guided by text descriptions. One interesting application of large block inpainting is virtual try-on (Han et al., 2018; Zhu et al., 2024), where models must realistically generate clothing or accessories consistent with both a reference garment and the overall body pose.

In a typical real-world fashion catalog, the full body images naturally have rectangular aspect ratios. Since most existing multimodal foundation models are trained on square images (e.g., 512×512 for MMaDA), we fine-tune MMaDA using our training objective $\mathcal{L}_{\text{DDPS}}$ derived in **Theorem 3.1** on a collection of 1024×512 full-body images. This dataset is curated from a fashion dataset (Zhu et al., 2024), following preprocessing with segmentation-based cropping and padding to standardize framing. This adaptation enables our base model to better handle rectangular image structures. For training, we have randomly selected 100K images from this dataset, providing a diverse and challenging testbed for inpainting at scale.

Figure 5 shows that our method generates realistic clothing textures, with better alignment to the reference prompts and fewer artifacts. APS leverages discrete diffusion's ability to directly reweight categorical distributions under posterior guidance, that helps generate visually appealing and semantically accurate completions.

B.8.3 Stylization

Figure 12 presents additional qualitative results on reference-based stylization. In each case, our APS optimizer conditions on a single reference style image and a text prompt describing the desired content. The outputs demonstrate that APS effectively transfers diverse artistic styles—including tattoo art, steampunk mechanical, psychedelic art, and tribal tattoo—while preserving semantic fidelity to the target content. These results highlight the robustness and versatility of APS in handling a wide range of style-content combinations.

Figure 13 provides a qualitative comparison of our full method (MMaDA + APS) against the base model (MMaDA) and state-of-the-art continuous diffusion approaches. This experiment is designed to demonstrate the novel capability of discrete diffusion models for challenging nonlinear style transfer, not solely to outperform continuous alternatives. We observe that competing methods struggle to balance style fidelity with content alignment. Training-free methods like StyleAligned (Hertz et al., 2023) and InstantStyle (Wang et al., 2024) often drift towards generic textures. Conversely, the training-based StyleDrop (Sohn et al., 2023) tends to overfit to superficial color patterns, which compromises semantic coherence with the text prompt. Our base model, MMaDA (Yang et al., 2025), maintains reasonable content fidelity but fails to transfer fine-grained style attributes, such as material textures or stroke-level details. In contrast, our full method (MMaDA + APS) consistently produces outputs that preserve the reference style while maintaining strong semantic alignment. For instance, our result for the "letter" prompt retains the intricate, flowing smoke design, while the "milkshake" example accurately captures the specified retro diner aesthetic. These results highlight the effectiveness of APS in discrete diffusion models for complex style transfer tasks.

Figure 13: **Additional qualitative comparison on reference-based stylization.** We compare our full method (MMaDA + APS) with the base model (MMaDA) and several state-of-the-art continuous diffusion methods across four style-prompt pairs. For each row, all methods use the same Style Reference image (left) and text prompt (shown below the images).

Table 7: **Super Resolution (4×) on FFHQ.** Performance comparison of APS against prior works. Continuous methods are shaded gray. Evaluation over 1000 samples from validation set.

Type	Method	LPIPS ↓	PSNR ↑	SSIM ↑
Pixel-domain	DPS	0.238	26.07	0.756
	DDRM	0.252	28.09	0.804
LDM	PSLD	0.282	27.12	0.757
	ReSample	0.508	23.07	0.445
Discrete	G2D2 w/ star-shaped noise process	<u>0.265</u>	<u>27.29</u>	<u>0.763</u>
	G2D2 w/ Markov noise process	0.369	25.15	0.699
Mask	**APS**	**0.232**	**27.81**	**0.808**

B.8.4 Additional Quantitative Results

We perform a larger-scale evaluation for 4× super resolution on FFHQ, extending our analysis to 1000 samples. Our results are compared against the numbers reported for the same task in G2D2 (Murata et al., 2024). Importantly, our observation in the main draft extends to the larger-scale setting and our APS algorithm consistently outperforms G2D2 in all metrics.

B.9 Limitations

Despite achieving state-of-the-art performance among discrete diffusion samplers on (1) complex (linear and nonlinear) inverse problems and (2) reference-based stylization tasks, APS exhibits the following limitations:

1. **Tokenizer quality.** MMaDA uses MagVIT-v2 (Yu et al., 2024) tokenizer which has limited reconstruction quality compared to modern visual tokenizers (Black Forest Labs, 2024). Therefore, future improvements in discrete visual tokenizers could directly benefit APS.

Figure 14: **Failure cases of APS.** Under extreme circumstances such as out-of-distribution styles or highly nonlinear measurement operators, our method can sometimes fail, producing over-smoothed reconstructions or noticeable artifacts.

2. **Base model performance.** Discrete diffusion backbones are still in an early stage of development and, at present, underperform large-scale continuous diffusion foundation models such as Flux (Black Forest Labs, 2024), SD3.5 (Esser et al., 2024), and Imagen (Baldridge et al., 2024) in unconditional generative quality. Nevertheless, our theoretical and empirical results indicate that discrete diffusion shows promising potential for posterior sampling and could, with further advances, become a viable alternative to the continuous models that dominate current practice.
3. **Stylization dependence.** The performance of APS in stylization tasks depends both on the pretrained discrete diffusion backbone and the quality of the style feature extractor (e.g., CSD). If the style extractor has not been trained on a particular style, our sampler struggles to transfer it faithfully, limiting its applicability to out-of-distribution styles.

Failure Cases. Figure 14 illustrates failure cases of our approach in stylization. We observe that APS sometimes produces over-smoothed outputs when the reference style is out-of-distribution, or introduces artifacts when the measurement operator is poorly aligned with the pretrained backbone. These examples highlight opportunities for improving robustness and generalization.