Beyond Pixel Space: Frequency-Domain Uncertainty for Structure-Aware Diffusion Guidance

Tianqi Zhao [1 2] **Xixi Liu** [2] **Yingzhen Li** [2] **Zhengrui Xiang** [2] **Liangrui Peng** [1]

Abstract

Although diffusion models achieve impressive generation quality, they inevitably produce samples with artifacts. Most existing approaches focus on quantifying pixel-level uncertainty to guide sample refinement. However, these methods erroneously assume independence among pixels. In this work, we turn to a frequency-based uncertainty metric that preserves structural information. We empirically show that samples with artifacts exhibit higher uncertainty, and further theoretically prove that the proposed uncertainty can be reinterpreted as an approximation of the optimal posterior covariance. To this end, we develop a structure-aware diffusion sampling guidance framework. For the mean of the reverse process, the gradient of the uncertainty is utilized to penalize specific frequency components; for the posterior covariance, we replace it with the proposed uncertainty. Experiments across U-Net, U-ViT, and SD3 architectures validate our approach.

1. Introduction

Although diffusion models have achieved remarkable breakthroughs in image (Rombach et al., 2021), video (Ho et al., 2022), and audio (Lemercier et al., 2024) generation, they still inevitably produce samples with artifacts or structural collapses during inference. Uncertainty estimation has been utilized to evaluate single-sample generation quality (Kou et al., 2023; Jazbec et al., 2025), yet these methods are mostly restricted to sample filtering. De Vita & Belagiannis (2025) first proposed a pixel-level uncertainty metric capable of guiding sample optimization during sampling. Nevertheless, such pixel-level approaches implicitly assume independence among pixels, thereby neglecting spatial correlations. In real images, adjacent pixels typically belong to the same semantic components, an such local associations represents the structural information.

Therefore, we propose measuring uncertainty in the frequency domain, which preserves the structural information. By employing Monte Carlo sampling and applying a Discrete Cosine transform (DCT) to the score function, we obtain the frequency uncertainty map.

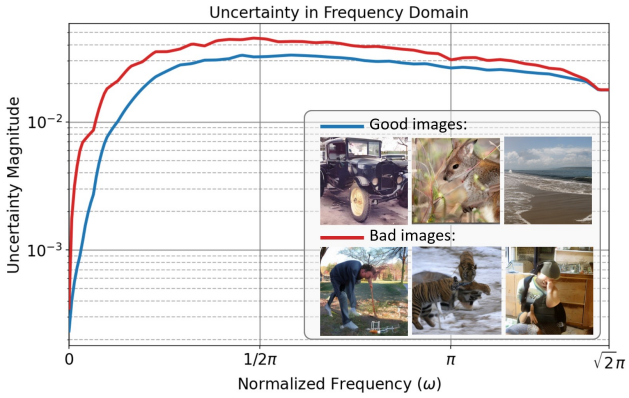


Figure 1. 256 images generated by U-ViT (Bao et al., 2023) are manually divided into two groups: structurally correct (Good) and with artifacts (Bad). We then compute their frequency-domain uncertainty during the denoising process.

We categorize the generated images into two groups based on visual quality. As shown in Figure 1, "bad" images exhibit an uncertainty gap across all frequencies compared to "good" images, with this gap being particularly significant in the low-frequency region. This finding aligns with the FreeU (Si et al., 2024), which suggest that low-frequency components encapsulate richer semantic information. Furthermore, we reinterpret our uncertainty as an approximation of noise covariance in the reverse process. Despite Ou et al. (2024) derived an analytical solution for the optimal noise variance, their approach requires auxiliary training. We theoretically proved that our proposed uncertainty metric approximates this analytical solution without the need for any additional training.

Building on this, we develop a structure-aware diffusion sampling guidance framework. (1) For the mean of the reverse process, we propose utilizing the gradient of the un-

[1]Department of Electronic Engineering, Tsinghua University, Beijing, China [2]Imperial College London, London, UK. Correspondence to: Liangrui Peng < penglr@tsinghua.edu.cn>, Yingzhen Li <yingzhen.li@imperial.ac.uk>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026.

Figure 2. Visual Comparisons on Stable Diffusion 3 (SD3) across the baseline (top row), spatial-domain uncertainty guidance (De Vita & Belagiannis, 2025) (middle row), and our frequency-domain uncertainty guidance (bottom row). The improved regions are highlighted with green boxes.

certainty to selectively penalize frequency components with the highest uncertainty. We also introduce a decay function to assign larger guidance weights to the low-frequency regions. (2) For the posterior covariance, we use the proposed uncertainty to approximate the diagonal of the optimal covariance matrix in the frequency domain. This facilitates a training-free noise injection strategy: we leverage the proposed uncertainty metric to modulate noise in the frequency domain. Intuitively, components with high uncertainty require greater noise injection to refine the structure.

Some results from our frequency-domain guidance framework and those from the spatial-domain guidance are shown in Figure 2. Our main contributions are summarized as follows:

- We shift the uncertainty estimation of diffusion models to the frequency-domain. We further prove that the proposed uncertainty can approximate the optimal covariance in the reverse process.
- We propose a structure-aware sampling guidance framework. For the mean of the reverse process, we use uncertainty gradients to selectively penalize frequency components. For the posterior covariance, we replace it with the frequency uncertainty.
- Experiments across U-Net, U-ViT, and SD3 demonstrate that our method reduces artifacts and improves generation quality.

2. Method

2.1. Preliminaries: Diffusion Models

Denoising Diffusion Probabilistic Models (DDPM) (Ho et al., 2020) model the data distribution by defining a forward noising process and learning its corresponding reverse denoising process.

Forward Process. Given a data sample $X_0 \sim q(X_0)$, the forward process is a Markov chain $\{X_t\}_{t=1}^T$ that gradually adds Gaussian noise according to a predefined noise schedule $\alpha_t \in (0, 1)$:

$$\begin{aligned} q(X_{0:T}) &= q(X_0) \prod_{t=1}^{T} q(X_t|X_{t-1}) \\ q(X_t|X_{t-1}) &= \mathcal{N}(X_t; \sqrt{\alpha_t} X_{t-1}, (1-\alpha_t)\mathbf{I}) \end{aligned} \tag{1}$$

The skip-time distribution $q(X_t|X_0)$ can be expressed in a closed form:

$$q(X_t|X_0) = \mathcal{N}(X_t; \sqrt{\bar{\alpha}_t} X_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. When the total number of timesteps T is sufficiently large, $q(X_T)$ converges to a standard isotropic Gaussian distribution, *i.e.*, $q(X_T) \approx p(X_T) = \mathcal{N}(0, \mathbf{I})$.

Reverse Process. The generation process recover the data distribution by reversing the forward process. DDPM train a neural network $\epsilon_\theta(X_t, t)$ to approximate the reverse transitions $p_\theta(X_{t-1}|X_t) = \mathcal{N}(X_{t-1}; \mu_\theta(X_t, t), \Sigma_t)$, where the mean μ_θ is parameterized as:

$$\mu_\theta(X_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta(X_t, t) \right) \tag{3}$$

For the covariance Σ_t, the original DDPM adopts a fixed, isotropic variance schedule $\Sigma_t = \sigma_t^2 \mathbf{I}$, where σ_t^2 is typically set to $1-\alpha_t$ or $\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)$.

According to Denoising Score Matching (DSM) (Vincent, 2011; Song & Ermon, 2019), training a neural network ϵ_θ to predict the Gaussian noise is equivalent to estimating the score function:

$$\epsilon_\theta(X_t, t) \approx -\sqrt{1-\bar{\alpha}_t} \nabla_{X_t} \log q(X_t) \tag{4}$$

2.2. Uncertainty Estimation by Monte Carlo sampling

Mi et al. (2022) demonstrated that a model's uncertainty can be quantified by the sensitivity to input perturbations, defined as Jacobian Frobenius norm $\|J(x)\|_F$ (Novak et al., 2018). However, computing the full partial derivative matrix of diffusion models is infeasible.

Inspired by Wang et al. (2019), we propose using Monte Carlo sampling to implicitly approximate the Jacobian matrix. For each timestep t, we generate M perturbed samples $\{X_t^{(m)}\}_{m=1}^M$ from input X_t and compute the empirical covariance of the predicted score function:

$$\mathbf{C}_t = \frac{1}{M} \sum_{m=1}^{M} \left(\epsilon_\theta(X_t^{(m)}, t) - \bar{\epsilon}_\theta \right) \left(\epsilon_\theta(X_t^{(m)}, t) - \bar{\epsilon}_\theta \right)^T \tag{5}$$

Based on a first-order Taylor expansion, the empirical covariance satisfies $\mathbf{C}_t \propto JJ^T$. However, the perturbation strategy in diffusion models is non-trivial. Isotropic Gaussian perturbations, i.e., $\tilde{X}_t = X_t + \sigma z$, fails because diffusion models are trained on marginal distributions. Samples at timestep t must satisfy the form given by Eq. 2, but adding additional noise σz inflates the variance.

To address this, we adopt the resampling strategy of De Vita & Belagiannis (2025). Using Tweedie's formula (Chung et al., 2022), we compute a pseudo-ground-truth anchor X_0^* and generate M Monte Carlo samples $\{X_t^{(m)}\}_{m=1}^M$ via forward noise injection.

In this way, through M network forward passes, we achieve an implicit approximation of the computationally expensive Jacobian matrix, estimating the uncertainty of the diffusion model.

2.3. Structure-Aware Uncertainty Estimation in Frequency-Domain

Although the empirical covariance matrix $\mathbf{C}_t$ obtained in Eq. 5 encapsulates the whole uncertainty information, explicitly computing and storing the full matrix remains infeasible. For high-dimensional images with a flattened dimensionality of D, $\mathbf{C}_t$ is a large $D \times D$ dense matrix.

De Vita & Belagiannis (2025) proposed using pixel-wise empirical variance to compute the uncertainty map, which is mathematically equivalent to substituting the covariance matrix with its diagonal. While this diagonal approximation renders the computation tractable, it completely discards the off-diagonal terms, which encode the relation among the pixels. In natural images, there exists strong spatial correlation among pixels (especially in adjacent ones). Performing noise injection or sampling guidance based on a pixel-independent uncertainty map leads to the loss of structural information.

To address this issue, we propose to shift the uncertainty estimation to the frequency domain. We introduce the Discrete Cosine Transform (DCT) (Ahmed et al., 1974) to approximately diagonalize the covariance matrix $\mathbf{C}_t$.

For any vectorized spatial signal x, we denote its Discrete Cosine Transform as

$$\hat{x} = \mathrm{DCT}(x) = \mathbf{Q}x, \tag{6}$$

where $\mathbf{Q}$ denotes the DCT basis matrix.

Studies in signal processing demonstrate that natural images can be modeled as first-order Markov processes with strong adjacent pixel correlations, and their spatial covariance matrix is a Toeplitz matrix (Jain, 1981). For such matrices, the orthogonal basis provided by the DCT is asymptotically equivalent to the eigenvectors of the optimal decoupling transform for this Toeplitz matrix, the Karhunen-Loève Transform (KLT) (Clarke, 1981).

As discussed in Appendix A, Tweedie's formula gives an affine relation between the predicted score $\epsilon_\theta(X_t, t)$ and the reconstructed clean image. Consequently, the local responses of the score function are expected to retain similar spatial correlations, suggesting its empirical covariance matrix $\mathbf{C}_t$ to exhibit an approximately Toeplitz-like structure.

Motivated by this observation, we apply the 2D-DCT to each perturbed score function to obtain its frequency-domain representation $\hat{\epsilon}_\theta(X_t^{(m)}, t)$. Substituting this for $\epsilon_\theta(X_t^{(m)}, t)$ in Eq. 5 yields the frequency-domain covariance matrix $\hat{\mathbf{C}}_t$. Equivalently, the corresponding frequency-domain covariance can be written as $\hat{\mathbf{C}}_t = \mathbf{Q}\mathbf{C}_t\mathbf{Q}^\top$. Because the DCT is expected to decorrelate the Toeplitz-like structure, $\hat{\mathbf{C}}_t$ becomes more diagonal-dominant than its spatial-domain counterpart.

Therefore, we adopt a diagonal approximation to $\hat{\mathbf{C}}_t$ in the frequency domain. A diagonal covariance in the DCT domain corresponds to a generally non-diagonal covariance after inverse transformation, thereby inducing spatially correlated modulation of the pixels. Moreover, under the assumption that the DCT asymptotically approximates the KLT, the diagonal approximation in the frequency domain is expected to incur a smaller truncation error, i.e., smaller discarded off-diagonal energy:

$$||\hat{\mathbf{C}}_t - \text{diag}(\hat{\mathbf{C}}_t)||_F^2 \lesssim ||\mathbf{C}_t - \text{diag}(\mathbf{C}_t)||_F^2 \tag{7}$$

Apply the diagonal approximation to $\hat{\mathbf{C}}_t$, which is equivalent to computing the frequency-wise empirical variance of the score function, yielding our frequency-domain uncertainty map:

$$\begin{aligned} U_t = \text{diag}(\hat{\mathbf{C}}_t) &= \text{Var}\left(\left\{\hat{\epsilon}_\theta(X_t^{(m)}, t)\right\}_{m=1}^{M}\right) \\ &= \frac{1}{M}\sum_{m=1}^{M}\left(\hat{\epsilon}_\theta(X_t^{(m)}, t) - \bar{\hat{\epsilon}}_\theta\right)^2 \end{aligned} \tag{8}$$

where the squaring operation is applied element-wise across the frequency.

Algorithm 1 Frequency-Domain Uncertainty Estimation

Input: Noisy image X_t, timestep t, noise schedule $\bar{\alpha}_t$, number of samples M
Output: Frequency uncertainty map U_t, base score in frequency domain $\hat{\epsilon}_0$
Compute base score in spatial and frequency domain: $\epsilon_0 = \epsilon_\theta(X_t, t)$, $\hat{\epsilon}_0 = \mathcal{DCT}(\epsilon_0)$
Anchor of perturbation: $X_0^* = (X_t - \sqrt{1-\bar{\alpha}_t}\epsilon_0)/\sqrt{\bar{\alpha}_t}$
for $m = 1, \ldots, M$ **do**
  $X_t^{(m)} = \sqrt{\bar{\alpha}_t}X_0^* + \sqrt{1-\bar{\alpha}_t}z^{(m)}$ ▷ $z^{(m)} \sim \mathcal{N}(0, \mathbf{I})$
  Compute perturbed score in frequency domain: $\hat{\epsilon}^{(m)} = \mathcal{DCT}(\epsilon_\theta(X_t^{(m)}, t))$
end for
Compute frequency-wise empirical variance: $U_t = \text{Var}(\{\hat{\epsilon}_\theta^{(m)}\}_{m=1}^M)$
return $U_t, \hat{\epsilon}_0$

2.4. Structure-Aware Score Guidance via Frequency-Domain Uncertainty

Empirical observations indicate that images with better structural exhibit pronounced lower uncertainty in frequency domain during the generation process. Therefore, we minimize the uncertainty during sampling via one-step gradient descent.

Inspired by De Vita & Belagiannis (2025), we apply uncertainty guidance only to high-uncertainty frequencies. These frequencies indicate the scales and structures where the model is currently most uncertain. We define a frequency mask: $M_t = \mathbb{I}(U_t > \tau_p)$ where $\mathbb{I}$ is the indicator function, and τ_p is a threshold determined by the p-th percentile of the uncertainty among the dataset.

Unlike spatial pixels, the behaviors of different frequency components within an image is asymmetrical. Si et al. (2024) argues low-frequency components evolve slowly and are more important to image semantic. Building on this, we introduce a negative exponential weighting function to apply stronger guidance to low-frequency, aggressively reducing uncertainty in large-scale structures. Let $\omega \in [0, \sqrt{2}]$ denote the normalized frequency radius. The weight matrix $W(\omega)$ is defined as: $W(\omega) = \exp(-\omega/T)$ where T is the temperature coefficient.

To minimize the targeted structural uncertainty, we define the guidance objective as the weighted sum of the masked uncertainties. For the original frequency-domain score prediction $\hat{\epsilon}_\theta(X_t, t)$, the uncertainty-guided score is updated via gradient descent:

$$\hat{\epsilon}_\theta^{\text{guided}}(X_t, t) = \hat{\epsilon}_\theta(X_t, t) - \lambda\left(M_t \odot W(\omega)\right)\frac{\partial U_t}{\partial \hat{\epsilon}_\theta(X_t, t)} \tag{9}$$

where λ controls the guidance strength, and $\odot$ denotes the Hadamard product. The gradient is computed only with respect to $\hat{\epsilon}_\theta(X_t, t)$. The MC noises are fixed, and $X_t^{(m)}$ are treated as deterministic affine functions of $\hat{\epsilon}_\theta(X_t, t)$.

Finally, we apply the Inverse Discrete Cosine Transform (IDCT) to obtain the refined spatial score function, $\epsilon_\theta^{\text{guided}}(X_t, t)$.

2.5. Reinterpreted Frequency-Domain Uncertainty as Covariance of the Reverse Process

In the reverse process of DDPMs, learning the variance instead of adopting a fixed schedule can improve generation quality (Nichol & Dhariwal, 2021). Recently, Ou et al. (2024) provided a closed-form solution for the optimal covariance during the diffusion denoising process:

$$\Sigma_t^* = (\sigma^4 \nabla_{X_t}^2 \log q(X_t) + \sigma^2 \mathbf{I})/\alpha^2 \tag{10}$$

where $\alpha = \sqrt{\bar{\alpha}_t}$ and $\sigma^2 = 1 - \bar{\alpha}_t$.

However, computing the marginal Hessian matrix $\nabla_{X_t}^2 \log q(X_t)$ is intractable. Ou et al. (2024) proposed training an auxiliary network to approximate its diagonal, which not only introduces additional training costs but also discards pixel correlations due to spatial diagonal truncation.

In contrast, we demonstrate that this complex Hessian estimation can be bypassed. We introduce Louis' Identity (Louis, 1982) into the DDPM framework to expand the

Table 1. Quantitative comparison of different sampling guidance methods, including the baseline, spatial-domain uncertainty guidance (De Vita & Belagiannis, 2025), and our method. We evaluate U-Net and U-ViT using 60K samples, and SD3 using 200 prompts from DrawBench (Saharia et al., 2022).

Method	U-Net (128^2)		U-ViT (256^2)		U-ViT (512^2)		SD3 (512^2)	
	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	FID ↓	CLIP ↑	HPSv2 ↑
Baseline	27.61	15.76	28.71	5.88	27.73	13.85	31.95	28.35
Spatial Guidance	27.42	15.68	28.91	5.81	27.88	14.06	32.05	28.39
Frequency Guidance (Ours)	**27.93**	**13.21**	**29.07**	**5.42**	**27.92**	**13.46**	**32.27**	**28.52**

marginal Hessian:

$$\nabla^2_{X_t} \log q(X_t) = \mathbb{E}_{q(X_0|X_t)}[\nabla^2_{X_t} \log q(X_t|X_0)] + \mathrm{Var}_{q(X_0|X_t)} \left(\nabla_{X_t} \log q(X_t|X_0)\right) \tag{11}$$

Equation 2 provides the explicit form of the forward noising distribution. By expressing the injected standard Gaussian noise as $\epsilon = (X_t - \alpha X_0)/\sigma$, taking the first and second derivatives of the log-density with respect to X_t yields:

$$\nabla_{X_t} \log q(X_t|X_0) = -\frac{\epsilon}{\sigma}, \quad \nabla^2_{X_t} \log q(X_t|X_0) = -\frac{1}{\sigma^2}\mathbf{I} \tag{12}$$

Substituting this result back into Eq. 11 and Eq. 10 gives:

$$\Sigma_t^* = \frac{\sigma^2}{\alpha^2} \mathrm{Var}_{q(X_0|X_t)}(\epsilon) \tag{13}$$

Note that ϵ denotes the true standard Gaussian noise injected during the forward process. Following the Fisher-information interpretation of Louis' identity in Appendix C, we use the empirical variance $\mathrm{Var}\left(\{\hat{\epsilon}_\theta(X_t^{(m)}, t)\}_{m=1}^M\right)$ defined in Eq. 8 as a stochastic approximation to the intractable posterior variance $\mathrm{diag}\left(\mathrm{Var}_{q(X_0|X_t)}(\hat{\epsilon})\right)$ required in Eq. 13, where $\hat{\epsilon} = \mathrm{DCT}(\epsilon)$.

Consequently, scaling U_t in Eq. 8 gives the diagonal of the approximation for the optimal posterior covariance in the frequency domain, denoted as $\hat{\Sigma}_t^*$:

$$\mathrm{diag}(\hat{\Sigma}_t^*) \approx \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} U_t \tag{14}$$

We sample a standard white noise vector $\hat{\mathbf{z}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in the frequency domain and modulate its amplitude using our uncertainty. The modulated noise is subsequently projected back to the spatial domain via the Inverse Discrete Cosine Transform (IDCT) and injected into the spatial mean. This design provides a training-free, diagonal estimation of the optimal noise covariance in the frequency domain. As established in Eq. 7, the DCT ensures the diagonal approximation has a lower truncation error than its spatial counterpart. Furthermore, modulating noise in the frequency domain accounts for spatial dependencies. Upon applying the IDCT, pixel correlations are restored, resulting in a structure-aware noise injection mechanism.

Our final sampling guidance framework is formulated as:

$$X_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(X_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta^{guided}(X_t, t) \right) + \mathcal{IDCT} \left(\sqrt{\frac{1-\bar{\alpha}_t}{\bar{\alpha}_t}} U_t \odot \hat{\mathbf{z}} \right) \tag{15}$$

where $\odot$ represents the Hadamard product.

3. Experiment Results

3.1. Toy Demonstration

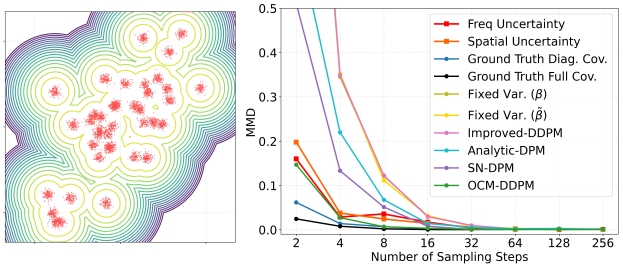


Figure 3. Toy demonstration on a correlated two-dimensional MoG distribution. The left shows the training data and the ground-truth density. The right reports the MMD of different covariance estimation methods under different DDPM sampling steps.

To verify whether the local Monte Carlo based uncertainty estimation proposed in Section 2.5 can approximate the optimal covariance in the diffusion reverse process, we conduct a toy experiment on a two-dimensional mixture of Gaussians (MoG). This distribution can be regarded as an image with only 2 pixels. To mimic the statistical correlation between adjacent pixels in natural images, we place the centers of the Gaussian components roughly along the $45°$ direction.

We compare our method with several covariance estimation methods, including Analytic-DPM (Bao et al., 2022b), Improved-DDPM (Nichol & Dhariwal, 2021), SN-DPM (Bao et al., 2022a), and OCM-DDPM (Ou et al., 2024). We also include two fixed-variance DDPM baselines, denoted by β and $\tilde{\beta}$ variances introduced in Section 2.1. In addition, we report the results using the ground-truth diago-

nal covariance (GT Diag. Cov.) and the ground-truth full covariance (GT Full Cov.) as references.

We use Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) as the evaluation metric. For all methods, the reverse-process mean is estimated using the analytically computed ground-truth score. We sample $5{,}000$ points from the MoG distribution as the test set. For methods requiring additional training, we use another $10{,}000$ points as the training set.

The results show that our training-free uncertainty approximation achieves performance close to OCM-DDPM and outperforms several methods that require additional training, such as SN-DPM. This empirically supports the connection between local uncertainty and the optimal reverse covariance. Moreover, frequency-domain uncertainty outperforms spatial-domain uncertainty under most sampling steps. This is because the two-point DCT decomposes the coordinates into the 45° diagonal direction and its orthogonal direction, where the diagonal direction aligns with the dominant variation of this MoG distribution.

3.2. Experiment Setup

Model Architectures To validate the applicability of our frequency-domain uncertainty guidance, we evaluate across three representative diffusion architectures: (1) U-Net model ADM (Dhariwal & Nichol, 2021) , trained on ImageNet (Deng et al., 2009) 128^2, which performs denoising in pixel space, (2) U-ViT model (Bao et al., 2023) trained on ImageNet 256^2 and 512^2, which demonstrates that our method is applicable to latent space, and (3) The large-scale SD3 (Esser et al., 2024), to validate its scalability to state-of-the-art foundational models.

Evaluation Metrics We evaluate our method using FID (Heusel et al., 2017), CLIP score (Radford et al., 2021), and HPSv2 (Wu et al., 2023). Specifically, FID assesses overall visual quality by computing the Fréchet distance between the Inception-v3 (Szegedy et al., 2016) feature distributions of real and generated images. For text-conditioned evaluation, CLIP score calculates the cosine similarity between cross-modal text and image embeddings, while HPSv2 extracts a human preference score using a fine-tuned vision-language model. Specifically, for the class-conditional generation on ImageNet, we convert the labels into standard text templates (e.g., "A photo of a [CLASS]").

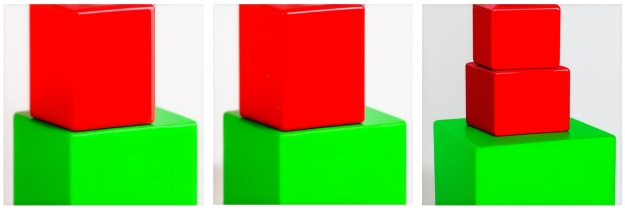
Implementation Details To mitigate additional computational overhead, we apply our guided sampling method exclusively during steps 3 to 5 for diffusion models with 50 sampling steps. In Section 2.4, the threshold percentage p is set to 90, the temperature coefficient T to 5, and the guidance scale λ to 0.1.

3.3. Comparison with Other Methods

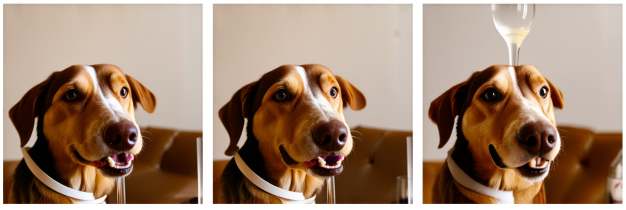
To validate the effectiveness of our frequency-domain uncertainty guidance method in improving visual fidelity and enhancing semantic consistency, we conduct comparisons with the spatial-domain uncertainty-based generative guidance method (De Vita & Belagiannis, 2025). For the U-Net and U-ViT architectures evaluated on ImageNet, we randomly generate 60,000 images at each resolution to compute the FID and CLIP scores. For the SD3 experiments, we evaluate the CLIP and HPSv2 scores using the DrawBench (Saharia et al., 2022) benchmark. This benchmark comprises 200 challenging prompts of 11 evaluation categories. The quantitative results are presented in Table 1, and visualization results in Appendix D.1.

The experimental results demonstrate that the proposed structural-aware uncertainty estimation and guidance in the frequency domain outperform spatial-domain methods that rely on the assumption of pixel-wise independence.

Figure 4 illustrates qualitative examples of SD3 on the DrawBench prompts. These results demonstrate that our method possesses a superior capacity for complex compositional generation. While spatial-domain guidance is largely limited to optimizing local structures, our approach modulates the most uncertain components in the frequency domain. This global optimization strategy can resolve long-range structural dependencies and semantic mismatches.



(a) Prompt: *A stack of 3 cubes. A red cube is on the top, sitting on a red cube. The red cube is in the middle, sitting on a green cube. The green cube is on the bottom.*



(b) Prompt: *A wine glass on top of a dog.*

Figure 4. Qualitative comparison of SD3 on DrawBench. Left column: baseline, Middle column: spatial guidance (De Vita & Belagiannis, 2025), Right column: Ours.

Figure 2 visualizes more qualitative comparisons on SD3 with realistic natural scenes. The results indicate that images generated by the spatial-domain guidance method (De Vita

Table 2. Ablation study and few-step generation analysis. Experiments are conducted using the U-ViT model at 256^2 resolution with 4,092 images. Uncertainty guidance is applied in steps 0-2 for the 5-step setting.

Method	50 Steps		5 Steps	
	CLIP ↑	FID ↓	CLIP ↑	FID ↓
Baseline	28.80	16.46	27.92	25.89
Score Guidance	28.98	16.02	-	-
Variance Estimation	28.81	16.26	**28.11**	**24.67**
Score Guidance + Variance Estimation	**29.07**	**15.89**	-	-

& Belagiannis, 2025) frequently suffers from structural discontinuities, particularly in complex regions such as human hands. This degradation is largely attributable to its independent pixel-level refinement. In contrast, our method preserves structural coherence and enhances local semantics. Notably, while our method effectively repairs the artifacts, it may simultaneously induce variations in other areas of the image (e.g., the woman's glasses in the second example). This is an inherent characteristic of our method, as modulating frequency components exerts a global influence across the entire spatial domain.

3.4. Ablation Study

Component Ablation. To validate our proposed sampling guidance framework, we conduct an ablation study. We independently apply the score function guidance detailed in Section 2.4 and the noise variance estimation detailed in Section 2.5. The results in Table 2 demonstrate the effectiveness of our framework, which simultaneously adjusts the mean and variance during the reverse diffusion process.

Effectiveness in Few-Step Generation. Variance matching during the reverse diffusion process can enhance generation quality under few sampling steps (Ou et al., 2024). To validate the effectiveness of the uncertainty-based variance estimation proposed in Section 2.5, we conduct experiments with only 5 sampling steps incorporating only the noise variance estimation. As shown in Table 2, our noise variance estimation effectively improves generation quality. The visualization results are presented Appendix D.2.

4. Related Works

4.1. Uncertainty Estimation for Generative Models

Bayesian Neural Networks (BNNs) (MacKay, 1992) are a longstanding approach for uncertainty estimation, which quantify model uncertainty by estimating the posterior distribution of the model weights. In practice, this posterior is often approximated by Monte Carlo Dropout (Gal & Ghahramani, 2016) or model ensembles (Lakshminarayanan et al., 2017; Huang et al., 2017).

Recent studies have extended uncertainty estimation to generative models. For GANs (Goodfellow et al., 2014), Bayesian GAN (Saatci & Wilson, 2017) applies Bayesian inference to approximate the posterior distributions of the generator and discriminator, while UGAC (Upadhyay et al., 2021) proposes the UGAC framework estimates pixel-wise uncertainty. VAEs (Kingma & Welling, 2013) can naturally provide uncertainty through the predicted mean and variance of the latent distribution, which has been used for anomaly detection (An & Cho, 2015).

Diffusion models (Ho et al., 2020) have recently achieved strong performance in high-quality image generation, motivating uncertainty estimation for diffusion sampling. BayesDiff (Kou et al., 2023) and Jazbec et al. (2025) introduce Bayesian inference into diffusion models via Last-Layer Laplace Approximation (LLLA), focusing on pixel-wise uncertainty estimation and image-level quality filtering, respectively. More recently, De Vita & Belagiannis (2025) estimates uncertainty during sampling by perturbing the current state and measuring the variance of denoising scores, then uses its gradient to guide the sampling trajectory.

However, existing methods mainly operate in the spatial domain, where pixel-wise estimation can neglect correlations among adjacent pixels and introduce disconnected results and artifacts. We instead propose estimating uncertainty in the frequency domain, enabling guidance on frequency components while better preserving structural information.

4.2. Covariance Estimation in Diffusion Reverse Process

In the original DDPMs (Ho et al., 2020), the variance of the reverse process is set to a predefined schedule. This state-independent variance restricts the expressive capacity of the model and results in slow generation speeds. To address this, Nichol & Dhariwal (2021) demonstrated that learning the variance in the reverse process can improve generation quality with fewer sampling steps.

Existing methods for estimating the reverse variance in diffusion models can be broadly categorized into isotropic variance and pixel-wise diagonal covariance matrices. Analytic-DPM (Bao et al., 2022b) derives a global optimal reverse

variance and estimates it via Monte Carlo approximation over score functions. For state-dependent diagonal covariance, Improved-DDPM (Nichol & Dhariwal, 2021) learns the variance using a variational lower-bound objective, while Bao et al. (2022a) train an auxiliary network to predict the second-order moment of the noise. Recently, Optimal Covariance Matching (OCM) (Ou et al., 2024) provides the analytic form of the posterior optimal covariance matrix, which is determined by the Hessian of the log-likelihood. Due to the computational intractability of the full Hessian, the authors utilize a lightweight network to approximate its diagonal elements.

Although these methods improve sampling efficiency, pixel-wise diagonal covariance still treats pixels independently and neglects spatial correlations. Moreover, most existing approaches require additional training. In contrast, our method estimates uncertainty in the frequency domain, enabling structure-aware noise injection without extra training.

4.3. Frequency-Domain Analysis of Diffusion Models

Frequency-domain analysis is useful for capturing long-range dependencies and global spatial correlations. Frequency Diffusion Models (Crabbé et al., 2024) apply the Discrete Fourier Transform (DFT) to time-domain SDEs and perform score matching in the frequency space. FreeU (Si et al., 2024) analyzes the different evolution patterns of low- and high-frequency components, and uses this insight to modulate U-Net backbone and skip-connection features.

However, frequency-domain variance or uncertainty estimation in the reverse diffusion process remains underexplored. In this work, we shift uncertainty estimation to the frequency domain and use it for training-free sampling guidance.

5. Conclusion

In this paper, we present a novel frequency-domain uncertainty estimation and guidance framework for diffusion model sampling, achieving structure-aware uncertainty estimation and modulation. Building upon this, we introduce a dual-guidance mechanism: (1) a score function guidance that selectively penalizes frequency components; (2) and a training-free reverse covariance approximation method for noise injection. Evaluations across diverse architectures, including U-Net, U-ViT, and Stable Diffusion 3, validate the effectiveness of our approach. Notably, as observed in our SD3 experiments, since the DCT is a global transformation, our method may alter other image regions while correcting localized artifacts. In future work, we plan to incorporate Wavelet Transforms to achieve more controllable, localized frequency-domain uncertainty modulation.

References

Ahmed, N., Natarajan, T., and Rao, K. R. Discrete Cosine Transform. *IEEE transactions on Computers*, 100(1): 90–93, 1974.

An, J. and Cho, S. Variational autoencoder based anomaly detection using reconstruction probability. 2015. URL https://api.semanticscholar.org/CorpusID:36663713.

Bao, F., Li, C., Sun, J., Zhu, J., and Zhang, B. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022a.

Bao, F., Li, C., Zhu, J., and Zhang, B. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022b.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22669–22679, 2023.

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

Clarke, R. Relation between the karhunen loeve and cosine transforms. In *IEE Proceedings F (Communications, Radar and Signal Processing)*, volume 128, pp. 359–360. IET, 1981.

Crabbé, J., Huynh, N., Stanczuk, J., and Van Der Schaar, M. Time series diffusion in the frequency domain. *arXiv preprint arXiv:2402.05933*, 2024.

De Vita, M. and Belagiannis, V. Diffusion model guided sampling with pixel-wise aleatoric uncertainty estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3844–3854. IEEE, 2025.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL http://arxiv.org/abs/1706.08500.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models, 2022. URL https://arxiv.org/abs/2204.03458.

Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J. E., and Weinberger, K. Q. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Jain, A. K. Advances in mathematical models for image processing. *Proceedings of the IEEE*, 69(5):502–528, 1981.

Jazbec, M., Wong-Toi, E., Xia, G., Zhang, D., Nalisnick, E., and Mandt, S. Generative uncertainty in diffusion models. *arXiv preprint arXiv:2502.20946*, 2025.

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

Kou, S., Gan, L., Wang, D., Li, C., and Deng, Z. BayesDiff: Estimating pixel-wise uncertainty in diffusion via bayesian inference. *arXiv preprint arXiv:2310.11142*, 2023.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Lemercier, J.-M., Richter, J., Welker, S., Moliner, E., Välimäki, V., and Gerkmann, T. Diffusion models for audio restoration, 2024. URL https://arxiv.org/abs/2402.09821.

Louis, T. A. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 44(2):226–233, 1982.

MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Mi, L., Wang, H., Tian, Y., He, H., and Shavit, N. N. Training-free uncertainty estimation for dense regression: Sensitivity as a surrogate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10042–10050, 2022.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J., and Sohl-Dickstein, J. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Ou, Z., Zhang, M., Zhang, A., Xiao, T. Z., Li, Y., and Barber, D. Improving probabilistic diffusion models with optimal diagonal covariance matching. *arXiv preprint arXiv:2406.10808*, 2024.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution image synthesis with Latent Diffusion Models, 2021.

Saatci, Y. and Wilson, A. G. Bayesian gan. *Advances in neural information processing systems*, 30, 2017.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35: 36479–36494, 2022.

Si, C., Huang, Z., Jiang, Y., and Liu, Z. FreeU: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4733–4743, 2024.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Upadhyay, U., Chen, Y., and Akata, Z. Robustness via uncertainty-aware cycle consistency. *Advances in neural information processing systems*, 34:28261–28273, 2021.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., and Vercauteren, T. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019.

Wasserman, L. *All of statistics: a concise course in statistical inference*, volume 26. Springer, 2004.

Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., and Li, H. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

A. Proof of the Toeplitz Structure of the Score Function

According to Tweedie's formula, there exists a linear mapping between the network-predicted score function $\epsilon_\theta(X_t, t)$ and the reconstructed clean image $\hat{X}_0$:

$$\epsilon_\theta(X_t, t) = \frac{X_t - \sqrt{\bar{\alpha}_t}\hat{X}_0}{\sqrt{1-\bar{\alpha}_t}}$$

Natural images follow a first-order Markov process (AR(1)), and their spatial autocorrelation matrix $\mathbf{\Sigma}_{data}$ is a Toeplitz matrix. For data with an adjacent pixel correlation coefficient $\rho \to 1$, the covariance matrix takes the following form:

$$\mathbf{\Sigma}_{data} \propto \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{N-1} \\ \rho & 1 & \rho & \cdots & \rho^{N-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{N-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{N-1} & \rho^{N-2} & \rho^{N-3} & \cdots & 1 \end{bmatrix}$$

The prediction target of the network $\hat{X}_0(X_t)$ is the natural image. Assuming training convergence, its Jacobian matrix $\mathbf{J}_X$ naturally inherits the data structure, with its spatial autocorrelation coefficient satisfying $\rho_x \to 1$.

Injecting a local white noise perturbation $z \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ into the input X_t, let the Jacobian matrix of the score function be $\mathbf{J}_\epsilon = \nabla_{X_t}\epsilon_\theta$, and the Jacobian matrix of the denoised image be $\mathbf{J}_X \triangleq \nabla_{X_t}\hat{X}_0$. Differentiation yields the linear decomposition:

$$\mathbf{J}_\epsilon = c_1\mathbf{I} - c_2\mathbf{J}_X, \quad c_1 = \frac{1}{\sqrt{1-\bar{\alpha}_t}}, c_2 = \frac{\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}$$

Based on the first-order Taylor expansion, the empirical covariance matrix $\mathbf{C}_t$ of the perturbation response can be approximated as:

$$\mathbf{C}_t = \mathbb{E}[\delta\epsilon_\theta \delta\epsilon_\theta^T] \approx \sigma^2 \mathbf{J}_\epsilon \mathbf{J}_\epsilon^T = \sigma^2(c_1^2\mathbf{I} - 2c_1c_2\mathbf{J}_X + c_2^2\mathbf{J}_X\mathbf{J}_X^T)$$

This indicates that $\mathbf{C}_t$ is a linear combination of $\mathbf{I}$, $\mathbf{J}_X$, and $\mathbf{J}_X\mathbf{J}_X^T$. The identity matrix $\mathbf{I}$ is a trivial Toeplitz matrix. Given the algebraic closure of Toeplitz matrices under linear combinations and asymptotic multiplication, the empirical covariance matrix $\mathbf{C}_t$ rigorously preserves the Toeplitz structure. This provides the mathematical foundation for why applying our uncertainty estimation in the frequency domain significantly reduces the diagonal truncation error.

B. Proof of Louis' Identity in DDPM

In diffusion models, the prior distribution of real data is $q(X_0)$, the conditional probability distribution of the forward noising process is $q(X_t|X_0)$, and the marginal distribution at time step t is $q(X_t)$. The marginal distribution can be obtained by integrating over X_0:

$$q(X_t) = \int q(X_t|X_0)q(X_0)dX_0$$

We aim to compute the Hessian matrix of the log marginal density, $\nabla^2_{X_t}\log q(X_t)$. First, we compute its first derivative (i.e., the score function):

$$\nabla_{X_t}\log q(X_t) = \frac{\nabla_{X_t}q(X_t)}{q(X_t)}$$

Substituting the integral form and applying Bayes' theorem $q(X_0|X_t) = \frac{q(X_t|X_0)q(X_0)}{q(X_t)}$, we obtain the relationship between the forward and posterior distributions:

$$\nabla_{X_t}\log q(X_t) = \int \frac{\nabla_{X_t}q(X_t|X_0)}{q(X_t|X_0)}\frac{q(X_t|X_0)q(X_0)}{q(X_t)}dX_0 = \mathbb{E}_{q(X_0|X_t)}[\nabla_{X_t}\log q(X_t|X_0)]$$

Next, taking the derivative of the above first derivative yields the Hessian matrix:

$$\nabla^2_{X_t}\log q(X_t) = \nabla_{X_t}\left(\frac{\nabla_{X_t}q(X_t)}{q(X_t)}\right) = \frac{\nabla^2_{X_t}q(X_t)}{q(X_t)} - \frac{\nabla_{X_t}q(X_t)\left(\nabla_{X_t}q(X_t)\right)^T}{q(X_t)^2}$$

For the second term, note that $\frac{\nabla_{X_t} q(X_t)}{q(X_t)}$ equals the previously defined $\nabla_{X_t} \log q(X_t)$. Thus, the second term can be written as:

$$-\mathbb{E}_{q(X_0|X_t)}[\nabla_{X_t} \log q(X_t|X_0)]\mathbb{E}_{q(X_0|X_t)}[\nabla_{X_t} \log q(X_t|X_0)]^T$$

For the first term $\frac{\nabla^2_{X_t} q(X_t)}{q(X_t)}$, expanding the marginal distribution into its integral form and moving the Laplacian operator inside the integral yields:

$$\begin{aligned}
\frac{\nabla^2_{X_t} q(X_t)}{q(X_t)} &= \frac{\nabla^2_{X_t} \int q(X_t|X_0)q(X_0)dX_0}{\int q(X_t|X_0)q(X_0)dX_0} = \frac{\int \nabla^2_{X_t} q(X_t|X_0)q(X_0)dX_0}{q(X_t)} \\
&= \frac{\int \nabla_{X_t} \left(\nabla_{X_t} \log q(X_t|X_0) \cdot q(X_t|X_0)\right) q(X_0)dX_0}{q(X_t)} \\
&= \frac{\int \left[\nabla^2_{X_t} \log q(X_t|X_0) \cdot q(X_t|X_0) + \nabla_{X_t} \log q(X_t|X_0) \left(\nabla_{X_t} q(X_t|X_0)\right)^T\right] q(X_0)dX_0}{q(X_t)} \\
&= \frac{\int \left[\nabla^2_{X_t} \log q(X_t|X_0) + \nabla_{X_t} \log q(X_t|X_0) \left(\nabla_{X_t} \log q(X_t|X_0)\right)^T\right] q(X_t|X_0)q(X_0)dX_0}{q(X_t)} \\
&= \int \left[\nabla^2_{X_t} \log q(X_t|X_0) + \nabla_{X_t} \log q(X_t|X_0) \left(\nabla_{X_t} \log q(X_t|X_0)\right)^T\right] \frac{q(X_t|X_0)q(X_0)}{q(X_t)} dX_0 \\
&= \mathbb{E}_{q(X_0|X_t)} \left[\nabla^2_{X_t} \log q(X_t|X_0) + \nabla_{X_t} \log q(X_t|X_0) \left(\nabla_{X_t} \log q(X_t|X_0)\right)^T\right]
\end{aligned}$$

Recombining the two terms of $\nabla^2_{X_t} \log q(X_t)$, we extract the conditional expectation of the Hessian and group the remaining terms together:

$$\nabla^2_{X_t} \log q(X_t) = \mathbb{E}_{q(X_0|X_t)}[\nabla^2_{X_t} \log q(X_t|X_0)]$$

$$+ \underbrace{\mathbb{E}_{q(X_0|X_t)} \left[\nabla_{X_t} \log q(X_t|X_0) \left(\nabla_{X_t} \log q(X_t|X_0)\right)^T\right] - \mathbb{E}_{q(X_0|X_t)}[\nabla_{X_t} \log q]\mathbb{E}_{q(X_0|X_t)}[\nabla_{X_t} \log q]^T}_{\text{Variance of } \nabla_{X_t} \log q(X_t|X_0)}$$

The part inside the underbrace in the above equation matches the definition of the covariance matrix $\text{Var}(Y) = \mathbb{E}[YY^T] - \mathbb{E}[Y]\mathbb{E}[Y]^T$. Therefore, we finally obtain Louis' Identity applied to DDPM:

$$\nabla^2_{X_t} \log q(X_t) = \mathbb{E}_{q(X_0|X_t)}[\nabla^2_{X_t} \log q(X_t|X_0)] + \text{Var}_{q(X_0|X_t)} \left(\nabla_{X_t} \log q(X_t|X_0)\right)$$

C. Approximating Posterior Noise Covariance with Local Monte Carlo Sampling

The derivation in Sec. 2.5 identifies the posterior noise covariance as the target quantity. We now explain why the empirical MC variance can serve as a local approximation.

According to Louis' identity, the optimal covariance in the reverse process depends on the second-order structure of the marginal distribution $q(X_t)$ at the current sample, i.e., the Hessian $\nabla^2 \log q(X_t)$, which characterizes the local curvature of the log-density.

For the marginal distribution $q(X_t)$ in diffusion models, it follows from its construction as a Gaussian-smoothed data distribution that it satisfies standard regularity conditions, including (i) twice differentiability, (ii) interchangeability of differentiation and integration, and (iii) vanishing boundary conditions at infinity.

Under these conditions, classical Fisher information theory (Wasserman, 2004) relates the curvature of the log-density to the Fisher information matrix:

$$\mathcal{I} = -\mathbb{E}_{q(X_t)}[\nabla^2 \log q(X_t)],$$

and equivalently,

$$\mathcal{I} = \mathbb{E}_{q(X_t)} \left[\nabla \log q(X_t) \nabla \log q(X_t)^T\right].$$

Since the score function has zero mean under $q(X_t)$, i.e.,

$$\mathbb{E}_{q(X_t)}[\nabla \log q(X_t)] = 0,$$

the Fisher information can also be interpreted as the covariance of the score function:

$$\mathcal{I} = \mathrm{Cov}_{q(X_t)} \left(\nabla \log q(X_t) \right).$$

This suggests that, in expectation, the variability of the score function reflects the curvature of the underlying data distribution.

However, the above identities are defined over the global distribution $q(X_t)$, whereas our method estimates uncertainty locally around a given sample X_t. To bridge this gap, we introduce stochastic perturbations in a local neighborhood of X_t and measure the variation of the score function within this region. Specifically, this local neighborhood is constructed around the Tweedie anchor X_0^*, which can be interpreted as an approximation to the posterior mean $\mathbb{E}[X_0|X_t]$ when the score model is accurate. Therefore, the resulting Monte Carlo perturbations probe the score variability around the estimated posterior center rather than around an arbitrary noisy point.

According to denoising score matching (DSM), the trained network satisfies:

$$\epsilon_\theta(X_t, t) \approx -\sigma \nabla_{X_t} \log q(X_t),$$

where $\sigma^2 = 1 - \bar{\alpha}_t$.

Therefore, the empirical uncertainty computed in our method can be expressed as:

$$U_t = \mathrm{Var}_{\text{local}}(\epsilon_\theta(X_t^{(m)}, t)) \approx \sigma^2 \mathrm{Var}_{\text{local}} \left(\nabla_{X_t} \log q(X_t^{(m)}) \right),$$

where the variance is taken over the local perturbation distribution. In our method, this variance is computed after applying DCT, yielding the frequency-domain diagonal approximation.

Under the assumption that the score function varies smoothly within the neighborhood induced by the Tweedie posterior anchor, this empirical variance captures the local variability of the score function around the estimated posterior center. In this sense, it can be viewed as a stochastic local proxy that reflects the curvature properties of $\log q(X_t)$. This provides an intuitive justification for using U_t as a stochastic local surrogate for the posterior noise covariance in the reverse process.

D. Additional Experiment Results

D.1. Additional Qualitative Results for Sec. 3.3

This section provides additional visual samples for the comparative experiments discussed in Sec. 3.3. Figure 5 displays a comparison of image samples generated by the baseline model, the spatial-domain uncertainty guidance method (De Vita & Belagiannis, 2025), and our proposed method, evaluated on the UViT architecture (Bao et al., 2023) at a 256^2 resolution.

As observed from the results, while the method in (De Vita & Belagiannis, 2025) primarily introduces modifications to local image details, in contrast, our approach is capable of correcting structural and semantic errors at a much larger scale. Overall, our method produces fewer generative artifacts and achieves better semantic coherence.

D.2. Additional Qualitative Results for Sec. 3.4

This section provides sample visualizations for the few-step generation experiments discussed in Sec. 3.4. Figure 6 illustrates the qualitative effects of our proposed frequency domain uncertainty-based posterior noise covariance estimation when generating images in only 5 steps, using the UViT architecture (Bao et al., 2023) at a 256^2 resolution.

As shown in the results, although both methods suffer from noticeable blur due to the extreme low-step setting, our generated images exhibit perceptibly sharper and more coherent global structures (e.g., the dog structure in the bottom-left). This demonstrates that our method effectively enhances generation quality under limited steps, thus accelerate the sampling for diffusion models.

(a) Baseline

(b) Spatial Uncertainty Guidance (De Vita & Belagiannis, 2025)

(c) Frequency Uncertainty Guidance (Ours)

Figure 5. Comparison of generated images evaluated on the UViT architecture (Bao et al., 2023) at 256^2 resolution.

(a) Baseline

(b) Our Covariance Estimation

Figure 6. Qualitative comparison of extreme few-step generation (5 sampling steps) using the UViT architecture (Bao et al., 2023).