SUGAR: Spectral Uncertainty Guidance for Adaptive Reverse Diffusion Sampling

Tianqi Zhao[1,2] **Xixi Liu**[2] **Zhengrui Xiang**[2] **Liangrui Peng**[1]*
[1] Tsinghua University, Beijing, China
[2]Imperial College London, London, UK
ztq24@mails.tsinghua.edu.cn, penglr@tsinghua.edu.cn

Abstract

Although diffusion models achieve promising image generation performance, uncertainty during the denoising process can be associated with visual artifacts. Existing uncertainty estimation methods typically quantify the uncertainty at the pixel level. However, these methods assume independence among pixels, neglecting pixel correlations crucial for image structure. In contrast, we propose a spectral uncertainty estimation method as samples with artifacts exhibit higher frequency-domain uncertainty. We further derive a Louis-identity-based connection between the uncertainty and the optimal covariance of the Gaussian transition in the reverse process. To this end, we develop a structure-aware diffusion sampling guidance framework. For the mean of the Gaussian transition in the reverse process, the gradient of uncertainty is used to penalize specific frequency components; for the reverse process covariance, frequency-domain uncertainty is used as a proxy to modulate injected noise. Experiments on the ImageNet, LSUN-Churches, and DrawBench datasets across U-Net, U-ViT, and SD3 architectures validate our method.

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 and Belagiannis [2025] first proposed a pixel-wise uncertainty estimation capable of guiding the sampling process of diffusion models. 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, and such local associations are essential for image structure. Therefore, we propose measuring uncertainty in the frequency domain, which preserves the structural information. By em-

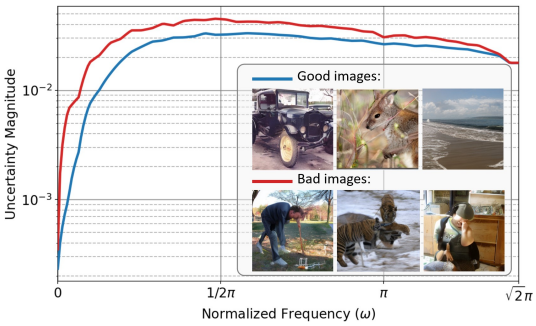

Figure 1: *Artifact samples exhibit higher frequency-domain uncertainty*. We manually divide 256 images generated by U-ViT [Bao et al., 2023] into structurally correct samples (Good) and samples with artifacts (Bad).

*Corresponding author: Liangrui Peng

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

ploying Monte Carlo sampling and applying a discrete cosine transform (DCT) to the score function, we obtain the frequency uncertainty map.

As shown in Figure 1, "bad" images exhibit higher uncertainty than "good" images across frequencies, especially in the low-frequency region. An image-level discrimination analysis is provided in Appendix B.1. This observation suggests a statistical association between frequency-domain uncertainty and visual artifacts. This finding aligns with the FreeU [Si et al., 2024], which suggests that low-frequency components carry richer semantic information. We therefore introduce score guidance via gradient descent to reduce uncertainty in selected frequency components.

Furthermore, we reinterpret the proposed uncertainty as a proxy for the reverse-process covariance. While Zhang et al. [2023], Ou et al. [2024] derive an analytical form of the optimal covariance, their method relies on auxiliary training to estimate its diagonal. In contrast, we use Louis' identity and a Fisher-information interpretation to show that the uncertainty can serve as a frequency-wise proxy for the optimal reverse-process covariance, without additional training.

We develop a structure-aware diffusion sampling guidance framework: (1) For the mean of the reverse process, we use the gradient of frequency-domain uncertainty to selectively penalize high-uncertainty frequency components. A frequency-dependent decay weighting function assigns stronger guidance to low-frequency components. (2) For the reverse-process covariance, we use the proposed uncertainty as a frequency-wise approximation to the diagonal structure of the optimal covariance. This leads to a training-free noise modulation strategy: we leverage the proposed uncertainty to modulate noise in the frequency domain. Intuitively, components with high uncertainty require greater noise injection to refine the structure.



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

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

- We propose a frequency-domain uncertainty estimation method for diffusion models using Monte Carlo sampling and the DCT to capture structural correlations.
- We develop a structure-aware sampling guidance framework that modulates the mean and covariance of the Gaussian transition in the reverse process. For the mean, we use the gradient of uncertainty to penalize frequency components with higher uncertainty, with stronger weights at lower frequencies. For the covariance, we derive a Louis-identity-based

connection between the optimal covariance and posterior noise variance, and use frequency-domain uncertainty as a local proxy for the latter, enabling training-free noise modulation.

- Experiments on the ImageNet, LSUN-Churches, and DrawBench datasets across U-Net, U-ViT, and SD3 architectures demonstrate the effectiveness of our method.

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 $\alpha_t \in (0,1)$:

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

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 recovers the data distribution by reversing the forward process. DDPM trains 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 $\beta_t = 1-\alpha_t$ or $\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)$.

According to Denoising Score Matching (DSM) [Vincent, 2011, Song and 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 aleatoric uncertainty of a model 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 (MC) sampling to implicitly approximate the Jacobian matrix. For each timestep t, we generate M perturbed samples $\{X_t^{(m)}\}_{m=1}^M$ from 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 perturbation, 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 and 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 aleatoric uncertainty of the diffusion model.

2.3 Frequency-Domain Uncertainty from Structured Score Covariance

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 and 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 detailed in Appendix A.1, we assume local shift stationarity of the denoising network's perturbation responses: within a local spatial region, the response depends primarily on the relative displacement between input and output positions. Under a local linearization with isotropic perturbations, this assumption yields an approximately Toeplitz covariance structure.

Based on this analysis, 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. Under an approximately wide-sense stationary model of the perturbation responses, the frequency-domain diagonal approximation is expected to discard less off-diagonal energy, as discussed in Appendix A.1:

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

Applying 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:

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

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

2.4 Structure-Aware Score Guidance via Frequency-Domain Uncertainty

Empirical observations in Fig. 1 indicate that images with better structure exhibit lower frequency-domain uncertainty during the generation process. In another view, the empirical variance reflects the sensitivity of denoising predictions to input perturbations, and minimizing it encourages more consistent responses in the selected frequency components. Therefore, we minimize the uncertainty during sampling via one-step gradient descent.

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**
    $X_t^{(m)} = \sqrt{\bar{\alpha}_t} X_0^* + \sqrt{1-\bar{\alpha}_t} z^{(m)}$ $\triangleright$ $z^{(m)} \sim \mathcal{N}(0, \mathbf{I})$
    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 $U_t, \hat{\epsilon}_0$

Inspired by De Vita and 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, different frequency components within an image exhibit asymmetric behavior. Si et al. [2024] argue that low-frequency components evolve slowly and are more important for image semantics. Building on this, we introduce a negative exponential weighting function to apply stronger guidance to low frequencies, aggressively reducing uncertainty in large-scale structures. Let $\omega \in [0, \sqrt{2}\pi]$ 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 $\hat{\epsilon}_\theta(X_t, t)$, while the network parameters θ are frozen. The perturbed samples $X_t^{(m)}$ are treated as deterministic affine functions of $\hat{\epsilon}_\theta(X_t, t)$, thereby avoiding second-order derivatives of the network.

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 Frequency-Domain Uncertainty as a Reverse-Covariance Proxy

In the reverse process of DDPMs, learning the variance instead of adopting a fixed schedule can improve generation quality [Nichol and Dhariwal, 2021]. Recently, Zhang et al. [2023] and Ou et al. [2024] provide 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 avoid explicit Hessian estimation by connecting the optimal covariance to posterior noise variability. We introduce Louis' Identity [Louis, 1982] into DDPM to expand the marginal Hessian, stated as Theorem 1 in Appendix A.2:

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

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:

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

Note that ϵ denotes the true standard Gaussian noise injected during the forward process. Following the Fisher-information interpretation of Louis' identity in Appendix A.3, we use the empirical variance $\text{Var}\left(\{\hat{\epsilon}_\theta(X_t^{(m)}, t)\}_{m=1}^M\right)$ defined in Eq. 8 as a stochastic local proxy for the frequency-domain diagonal of the intractable posterior variance $\text{Var}_{q(X_0|X_t)}(\epsilon)$ required in Eq. 13. Since the absolute magnitude of this empirical variance depends on the local perturbation scale and score sensitivity, we introduce a scalar calibration factor η to control the overall strength of the covariance.

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

$$\text{diag}(\hat{\Sigma}_t^*) \approx \eta^2 \frac{1-\bar{\alpha}_t}{\bar{\alpha}_t} U_t \tag{14}$$

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 yields a training-free frequency-domain noise injection strategy. Intuitively, this modulation assigns relatively more noise to uncertain frequency components to explore plausible denoising outcomes, and less noise to confident components. As established in Eq. 7, the DCT is expected to reduce the diagonal truncation error compared with the spatial-domain approximation. 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 method. Our final sampling guidance framework is formulated as:

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

where $\odot$ represents the Hadamard product.

3 Experiment Results

3.1 Toy Demonstration

To verify whether the local MC based uncertainty estimation proposed in Section 2.5 can approximate the optimal covariance, we closely follow Ou et al. [2024] to conduct a toy experiment on a two-dimensional mixture of Gaussians (MoG). This distribution can be regarded as an image with 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.

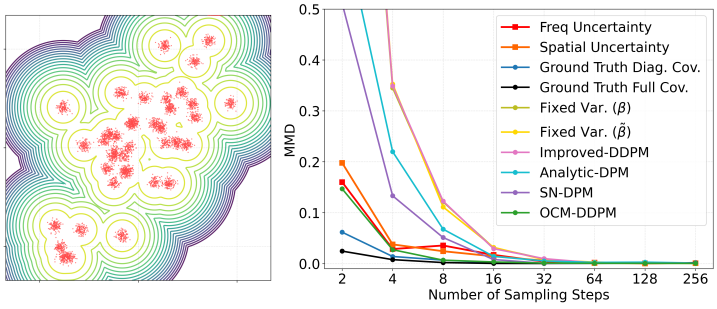

Figure 3: Toy demonstration on a correlated two-dimensional MoG distribution. The left shows the training data and the ground-truth density. The right reports the MMD of covariance estimation methods under different DDPM sampling steps.

We compare our method with several covariance estimation methods, including Analytic-DPM [Bao et al., 2022a], Improved-DDPM [Nichol and Dhariwal, 2021], SN-DPM [Bao et al., 2022b], and OCM-DDPM [Ou et al., 2024]. We also include two fixed-variance DDPM baselines, denoted by β

Table 1: Comparison results of the baseline, spatial-domain uncertainty guidance [De Vita and Belagiannis, 2025], and our method. U-Net and U-ViT models are evaluated on ImageNet with 60K samples, while SD3 is evaluated on 200 DrawBench prompts.

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**

and $\tilde{\beta}$ variances introduced in Section 2.1. In addition, we report the results using the ground-truth diagonal covariance (GT Diag. Cov.) and the ground-truth full covariance (GT Full Cov.). We use Maximum Mean Discrepancy (MMD) [Gretton et al., 2012] as the evaluation metric. For all methods, the reverse-process mean is computed using the analytical 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 in Figure 3 show that our training-free uncertainty approximation achieves performance close to OCM-DDPM and outperforms several methods that require additional training. This supports the connection between local uncertainty and the optimal reverse covariance. Moreover, frequency-domain uncertainty outperforms spatial-domain uncertainty under most sampling steps.

3.2 Experiment Setup

Datasets We evaluate our method on both conditional and unconditional generation settings. For conditional generation, we use ImageNet [Deng et al., 2009] at 128^2, 256^2, and 512^2 resolutions for class-conditioned image generation, and DrawBench [Saharia et al., 2022] for text-conditioned generation. For unconditional generation, we evaluate on LSUN-Churches at 256^2 [Yu et al., 2015].

Model Architectures We evaluate our method across four representative settings: (1) a conditional ADM U-Net [Dhariwal and Nichol, 2021] on ImageNet 128^2; (2) latent-space U-ViT models [Bao et al., 2023] on ImageNet 256^2 and 512^2; (3) the large-scale text-to-image model SD3 [Esser et al., 2024]; and (4) a pretrained unconditional DDPM U-Net [Ho et al., 2020] on LSUN-Churches256.

Evaluation Metrics We evaluate visual fidelity using FID [Heusel et al., 2017], which computes the Fréchet distance between Inception-v3 [Szegedy et al., 2016] feature distributions of real and generated images. For conditional generation, we additionally report CLIP score [Radford et al., 2021] to measure semantic alignment between generated images and their conditions. For class-conditional ImageNet experiments, class labels are converted into standard text prompts, e.g., "A photo of a [CLASS]". For text-to-image SD3 experiments, we further report HPSv2 [Wu et al., 2023], a human-preference score from a fine-tuned vision-language model.

Implementation Details To reduce computational overhead, we apply our guided sampling only to steps 3–5 for diffusion models with 50 sampling steps. An evaluation of guidance across all 50 steps is provided in Appendix B.5. For Sec. 2.2, we set $M = 5$. For Sec. 2.4, we use $\lambda = 0.1$, $p = 90$, $T = 5$ for U-ViT and SD3; and $\lambda = 0.1$, $p = 95$, $T = 1$ for U-Net-based models. For Sec. 2.5, $\eta = 0.1$. All experiments are conducted on four 32GB NVIDIA V100 GPUs.

3.3 Comparison with Other Methods

Conditional Generation. To validate the effectiveness of our frequency-domain uncertainty guidance in conditional generation, we compare our method with the baseline and the spatial-domain uncertainty guidance method [De Vita and Belagiannis, 2025]. For class-conditional generation, we evaluate U-Net and U-ViT models by generating 60,000 images at each resolution and reporting FID and CLIP scores. For text-conditioned generation, we evaluate SD3 on the DrawBench [Saharia et al., 2022] benchmark, which contains 200 challenging prompts across 11 categories, and report CLIP and HPSv2 scores. The quantitative results are presented in Table 1.

Table 2: Unconditional generation results on LSUN-Churches256 with 60K samples, comparing the baseline, spatial-domain uncertainty guidance [De Vita and Belagiannis, 2025], and our method.

Method	FID ↓
Baseline	12.25
Spatial Guidance	12.34
Frequency Guidance (Ours)	**11.44**

The results demonstrate that the proposed structure-aware uncertainty estimation and guidance in the frequency domain outperform spatial-domain methods that rely on the assumption of pixel-wise independence. Additional qualitative comparisons on ImageNet are provided in Appendix B.2.

Figure 4 shows qualitative examples of SD3 on DrawBench prompts. Spatial-domain guidance mainly adjusts local image regions, whereas our method modulates uncertain components in the frequency domain. This allows the guidance to affect long-range structures and better resolve semantic mismatches. Figure 2 further presents comparisons on realistic natural scenes. The spatial-domain guidance method [De Vita and Belagiannis, 2025] often introduces structural discontinuities, especially in complex regions such as human hands, due to its independent pixel-level refinement. In contrast, our method better preserves global coherence while improving local semantics. Notably, since frequency-domain modulation has a global effect, correcting artifacts may also induce changes in other plausible regions, such as the woman's glasses in the second example.



(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 and Belagiannis, 2025], Right column: Ours.

Unconditional Generation. We further evaluate whether our method generalizes to unconditional generation. Specifically, we use the pretrained DDPM U-Net [Ho et al., 2020] on LSUN-Churches256 [Yu et al., 2015]. As shown in Table 2, our method improves unconditional diffusion sampling, suggesting that frequency-domain uncertainty guidance does not rely on class labels or text prompts. Qualitative comparisons are provided in Appendix B.4.

Inference Overhead. We compare the inference overhead of baseline sampling, spatial-domain uncertainty guidance, and our frequency-domain uncertainty guidance using U-ViT on ImageNet-256. All experiments are conducted with a batch size of 16, generating 256 images per method. For 50-step sampling, both guidance methods are applied at steps 3–5 with $M = 5$ Monte Carlo samples. Table 3 reports generation time, throughput, peak allocated memory, and the number of denoising network function evaluations per image (NFE/image).

Table 3: Inference overhead of baseline sampling and uncertainty-guided sampling using U-ViT on ImageNet-256.

Method	Time (s) ↓	Throughput (img/s) ↑	Peak memory (GiB) ↓	NFE/image ↓
Baseline	78.69	3.253	6.703	50
Spatial Guidance	107.87	2.373	6.703	65
Frequency Guidance (Ours)	108.37	2.362	6.703	65

The two guidance methods incur similar computational overhead, primarily from the additional MC evaluations at guided steps. The small runtime difference suggests that DCT/IDCT operations and frequency-domain noise modulation introduce little additional cost.

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 4 demonstrate the effectiveness of our framework, which simultaneously adjusts the mean and variance during the diffusion sampling. Additional ablations on the frequency weighting function and frequency mask are provided in Appendix B.5, further validating the design choices of our frequency-domain guidance.

Few-Step Generation. Covariance matching in the reverse diffusion process can improve generation quality with fewer sampling steps [Ou et al., 2024]. We evaluate score guidance, reverse covariance estimation, and their combination under 5-step sampling. Table 4 shows that reverse covariance estimation provides the main improvement in this setting, and combining both components still achieves the best performance. Qualitative comparisons using only posterior variance modulation are provided in Appendix B.6.

Table 4: Component ablation and few-step generation results. Experiments are conducted using the U-ViT model at 256^2 resolution with 4,096 images.

Method	50 Steps		5 Steps	
	CLIP ↑	FID ↓	CLIP ↑	FID ↓
Baseline	28.80	16.46	27.92	25.89
Score Guidance	28.98	16.02	28.01	25.36
Reverse Covariance Estimation	28.81	16.26	28.11	24.67
Score Guidance + Reverse Covariance Estimation	**29.07**	**15.89**	**28.13**	**24.59**

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 and 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 and Wilson, 2017] applies Bayesian inference to approximate the posterior distributions of the generator and discriminator, while UGAC [Upadhyay et al., 2021] estimates pixel-wise uncertainty. VAEs [Kingma and Welling, 2013] can naturally provide uncertainty through the predicted mean and variance of the latent distribution [An and 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 and Belagiannis [2025] estimate 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 may ignore adjacent-pixel correlations and introduce artifacts. We instead estimate uncertainty in the frequency domain, enabling frequency-component guidance while 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. Nichol and Dhariwal [2021] show that learning this variance can improve generation quality with fewer sampling steps.

Existing reverse-variance estimation methods mainly adopt either isotropic variance or pixel-wise diagonal covariance. Analytic-DPM [Bao et al., 2022a] derives a globally optimal reverse variance and estimates it via Monte Carlo approximation over score functions. For pixel-wise diagonal covariance, Improved-DDPM [Nichol and Dhariwal, 2021] learns the variance with a variational lower-bound objective, while Bao et al. [2022b] train an auxiliary network to predict the second-order moment of the noise.

Zhang et al. [2023] derive a Gaussian denoising posterior covariance identity involving the log-density Hessian and estimate its diagonal using Rademacher vectors. OCM-DDPM [Ou et al., 2024] studies the same optimal diagonal covariance matching in diffusion models. Due to the computational intractability of the full Hessian, the authors utilize a network to approximate its diagonal elements.

However, pixel-wise diagonal covariance neglects spatial correlations, and most existing approaches require additional training. In contrast, our method estimates uncertainty in the frequency domain, enabling structure-aware noise injection without extra training.

Concurrent work by Schioppa and Salimans [2026] also estimates reverse covariance using JVPs and DCT-based representations. Our method additionally uses uncertainty gradients to guide the reverse-process mean.

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 and Discussion

In this paper, we present a frequency-domain uncertainty estimation method and a structure-aware guidance framework for diffusion sampling. Building upon this estimator, we introduce a diffusion guidance framework: (1) score guidance that selectively penalizes frequency components; (2) training-free reverse covariance approximation for frequency-domain noise injection. Experiments on both conditional and unconditional generation across architectures including U-Net, U-ViT, and SD3 validate our approach.

The global nature of the DCT limits our method's ability to explicitly localize artifact regions, since individual frequency components do not correspond to specific image locations. Consequently, guidance aimed at correcting localized artifacts may also alter other image regions, as observed in the SD3 experiments. Moreover, our method requires access to intermediate diffusion states, limiting its direct applicability to closed-source models that expose only generated images. In future work, we plan to explore patch-level or multi-scale DCT transforms and wavelet transforms for more spatially localized uncertainty guidance.

Acknowledgments and Disclosure of Funding

This work was partially supported by Huawei Noah's Ark Lab. Tianqi Zhao acknowledges support from Tsinghua University's short-term overseas research visit funding for doctoral students. Xixi Liu acknowledges support from the WASP International Postdoctoral Scholarship, funded by the Knut and Alice Wallenberg Foundation.

The authors declare no competing interests.

References

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution image synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*, 2021.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. URL `https://arxiv.org/abs/2204.03458`.

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. Diffusion models for audio restoration, 2024. URL `https://arxiv.org/abs/2402.09821`.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A ViT backbone for diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22669–22679, 2023.

Siqi Kou, Lei Gan, Dequan Wang, Chongxuan Li, and Zhijie Deng. BayesDiff: Estimating pixel-wise uncertainty in diffusion via Bayesian inference. *arXiv preprint arXiv:2310.11142*, 2023.

Metod Jazbec, Eliot Wong-Toi, Guoxuan Xia, Dan Zhang, Eric Nalisnick, and Stephan Mandt. Generative uncertainty in diffusion models. *arXiv preprint arXiv:2502.20946*, 2025.

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

Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free lunch in diffusion U-Net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4733–4743, 2024.

Mingtian Zhang, Alex Hawkins-Hooker, Brooks Paige, and David Barber. Moment matching denoising Gibbs sampling. *Advances in Neural Information Processing Systems*, 36:23590–23606, 2023.

Zijing Ou, Mingtian Zhang, Andi Zhang, Tim Z Xiao, Yingzhen Li, and David Barber. Improving probabilistic diffusion models with optimal diagonal covariance matching. *arXiv preprint arXiv:2406.10808*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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

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

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

Roman Novak, Yasaman Bahri, Daniel A Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. *arXiv preprint arXiv:1802.08760*, 2018.

Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neuro computing*, 338:34–45, 2019.

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

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

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

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

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171, 2021.

Thomas A Louis. 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.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-DPM: An analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022a.

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022b.

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

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*, pages 248–255. Ieee, 2009.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. 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`.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

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*, pages 8748–8763, 2021.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

David JC MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E Hopcroft, and Kilian Q Weinberger. Snapshot ensembles: Train 1, get m for free. *arXiv preprint arXiv:1704.00109*, 2017.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Yunus Saatci and Andrew G Wilson. Bayesian GAN. *Advances in neural information processing systems*, 30, 2017.

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

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

Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015. URL `https://api.semanticscholar.org/CorpusID:36663713`.

Andrea Schioppa and Tim Salimans. Covariance-aware sampling for diffusion models, 2026. URL `https://arxiv.org/abs/2605.13910`.

Jonathan Crabbé, Nicolas Huynh, Jan Stanczuk, and Mihaela Van Der Schaar. Time series diffusion in the frequency domain. *arXiv preprint arXiv:2402.05933*, 2024.

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

Yiyang Wang, Xi Chen, Xiaogang Xu, Sihui Ji, Yu Liu, Yujun Shen, and Hengshuang Zhao. Diffdoctor: Diagnosing image diffusion models before treating. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18917–18926. IEEE, 2025.

A Analytical Derivations and Approximation Justifications

A.1 Toeplitz-like Structure of Score Covariance

We provide a local modeling argument for the spatial structure of the perturbation-response covariance and explain why the DCT is used as an approximate decorrelation basis.

Local perturbation-response covariance. For simplicity, consider a one-dimensional spatial signal $X_t \in \mathbb{R}^D$ and a denoising network $\epsilon_\theta(\cdot, t) : \mathbb{R}^D \to \mathbb{R}^D$. For this local analysis, let $\boldsymbol{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$ be a small isotropic input perturbation, where σ_p denotes the perturbation scale. A first-order Taylor expansion gives

$$\Delta\boldsymbol{\epsilon} = \epsilon_\theta(X_t + \boldsymbol{\delta}, t) - \epsilon_\theta(X_t, t) \approx \mathbf{J}_t \boldsymbol{\delta}, \tag{16}$$

where $\mathbf{J}_t = \partial \epsilon_\theta(X_t, t)/\partial X_t \in \mathbb{R}^{D\times D}$ is the Jacobian of the denoising network with respect to its input. The corresponding response covariance is approximated by

$$\mathbf{C}_t = \mathrm{Cov}(\Delta\boldsymbol{\epsilon}) \approx \mathbf{J}_t \,\mathrm{Cov}(\boldsymbol{\delta}) \mathbf{J}_t^\top = \sigma_p^2 \mathbf{J}_t \mathbf{J}_t^\top. \qquad \mathbf{C}_t \in \mathbb{R}^{D\times D}. \tag{17}$$

Consequently, let J_{ik} denote the influence of a small perturbation at input position k on output position i. Then,

$$[\mathbf{C}_t]_{ij} \approx \sigma_p^2 \sum_{k=1}^{D} J_{t,ik} J_{t,jk}. \tag{18}$$

Assumption 1 (Local shift stationarity of perturbation responses)**.** *Within a local spatial region, the response to an input perturbation depends approximately on the relative displacement between the input and output positions:*

$$J_{t,ik} \approx R_t(k - i), \tag{19}$$

where R_t is a local response kernel.

Weight sharing in convolutional operations motivates this assumption, but does not guarantee shift-stationary responses at a fixed input. Content-dependent interactions, nonlinearities, positional encodings, and finite boundaries can violate the assumption. We therefore use it as a local approximation rather than a global architectural property.

Substituting Eq. 19 into Eq. 18 gives

$$[\mathbf{C}_t]_{ij} \approx \sigma_p^2 \sum_{k} R_t(k - i) R_t(k - j). \tag{20}$$

Let $m = k - i$ and $r = j - i$. For positions away from boundaries, neglecting finite-boundary effects yields

$$[\mathbf{C}_t]_{ij} \approx \sigma_p^2 \sum_{m} R_t(m) R_t(m - r) = \kappa_t(r) = \kappa_t(j - i). \tag{21}$$

Thus, the covariance depends primarily on relative displacement, giving an approximately Toeplitz structure within the region where the assumption holds. For a two-dimensional spatial field, the analogous argument gives covariance depending on two-dimensional displacement, corresponding to a block Toeplitz with Toeplitz blocks (BTTB) structure.

Frequency-domain approximation and diagonal energy. The spatial covariance structure derived above motivates using the DCT as an approximate decorrelation basis. We further assume that the DCT provides an approximate KLT representation of the local perturbation-response covariance.

To explain this approximation, let $\mathbf{C}_t = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ be its eigendecomposition. The KLT basis $\mathbf{U}^\top$ diagonalizes the covariance:

$$\mathbf{U}^\top \mathbf{C}_t \mathbf{U} = \boldsymbol{\Lambda}.$$

Since orthogonal transformations preserve Frobenius energy, the KLT concentrates all covariance energy on the diagonal. The DCT approximation to this basis therefore motivates an expected increase in diagonal energy.

Let $\mathbf{Q}$ denote the orthogonal DCT matrix and $\hat{\mathbf{C}}_t = \mathbf{Q}\mathbf{C}_t\mathbf{Q}^\top$. Then

$$\|\hat{\mathbf{C}}_t\|_F^2 = \|\mathbf{C}_t\|_F^2.$$

For each representation, the discarded off-diagonal energy equals the total energy minus the diagonal energy:

$$\|\mathbf{C}_t - \operatorname{diag}(\mathbf{C}_t)\|_F^2 = \|\mathbf{C}_t\|_F^2 - \|\operatorname{diag}(\mathbf{C}_t)\|_F^2,$$
$$\|\hat{\mathbf{C}}_t - \operatorname{diag}(\hat{\mathbf{C}}_t)\|_F^2 = \|\mathbf{C}_t\|_F^2 - \|\operatorname{diag}(\hat{\mathbf{C}}_t)\|_F^2.$$

Here, $\operatorname{diag}(\mathbf{A})$ denotes the diagonal matrix retaining the diagonal entries of $\mathbf{A}$. Under the approximate KLT representation, we expect

$$\|\operatorname{diag}(\hat{\mathbf{C}}_t)\|_F^2 \gtrsim \|\operatorname{diag}(\mathbf{C}_t)\|_F^2,$$

corresponding to the expected reduction in diagonal truncation error expressed in Eq. 7.

A.2 Proof of Louis' Identity in DDPM

Theorem 1 (Louis' identity for DDPM marginals)**.** *Let the DDPM marginal distribution be*

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

Assume that differentiation and integration can be interchanged and that the involved densities are twice differentiable with appropriate boundary conditions. Then,

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

Proof 1. We aim to compute the Hessian matrix of the log marginal density, $\nabla_{X_t}^2 \log q(X_t)$. 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:

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

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 $\frac{\nabla_{X_t}^2 q(X_t)}{q(X_t)}$, expanding the marginal distribution into its integral form and moving the Hessian operator inside the integral yields:

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

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:

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

□

A.3 Local MC Score Variance as a Posterior Noise Covariance Proxy

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.

Assumption 2 (Score covariance as a Hessian proxy)**.** *For the diffusion marginal $q(X_t)$, the variability of the marginal score $\nabla_{X_t} \log q(X_t)$ reflects the Hessian structure $\nabla^2_{X_t} \log q(X_t)$ that appears in the optimal reverse-process covariance.*

Explanation. 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:

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

This supports Assumption 2: in expectation, the variability of the marginal score reflects the Hessian structure involved in the optimal reverse-process covariance.

Assumption 3 (Local MC approximation to posterior noise variance)**.** *The Tweedie anchor X_0^* approximates the posterior center $\mathbb{E}[X_0|X_t]$. Monte Carlo samples $\{X_t^{(m)}\}_{m=1}^M$ generated by forward noising around X_0^* therefore provide a local approximation to the posterior noise variability induced by $q(X_0|X_t)$.*

Explanation. 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.

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

Taking the local covariance over the perturbation distribution, we obtain

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

This indicates that the local covariance of the predicted score reflects the local variability of the marginal score, and thus provides a stochastic proxy for the local second-order structure discussed above.

However, the absolute magnitude of this empirical covariance depends on the local perturbation scale and the sensitivity of the score network. Therefore, we do not directly treat it as a calibrated posterior covariance value. Instead, we use it to estimate the relative covariance structure and introduce a scalar coefficient η in the main text to calibrate the overall noise magnitude.

In our method, this covariance is estimated after applying DCT. Let $\hat{\epsilon}_\theta(X_t^{(m)}, t) = \mathrm{DCT}(\epsilon_\theta(X_t^{(m)}, t))$. Our frequency uncertainty can be written as the frequency-wise variance:

$$U_t = \mathrm{diag}(\mathrm{Cov}_{\mathrm{local}}\left(\hat{\epsilon}_\theta(X_t^{(m)}, t)\right)) = \mathrm{Var}_{\mathrm{local}}\left(\hat{\epsilon}_\theta(X_t^{(m)}, t)\right).$$

Where $\mathrm{Var}_{\mathrm{local}}(\cdot)$ denotes the element-wise variance.

Under above assumptions, the frequency-domain uncertainty U_t serves as a frequency-wise diagonal proxy for the relative structure of the posterior noise covariance $\mathrm{Var}_{q(X_0|X_t)}(\hat{\epsilon})$, where $\hat{\epsilon} = \mathrm{DCT}(\epsilon)$. After the scaling in Eq. 14, it provides a calibrated frequency-domain approximation to the optimal reverse-process covariance.

B Additional Experiment Results

B.1 Association Between Uncertainty and Artifacts

We quantitatively examine whether frequency-domain uncertainty is associated with visual artifacts, complementing the qualitative observations in Fig. 1.

We use the 256 images generated by U-ViT [Bao et al., 2023] shown in Fig. 1. Through visual inspection, we label images with visible artifacts as "bad" and structurally correct images as "good". The resulting set contains 166 bad images and 90 good images, with bad images treated as the positive class.

For each image, we use frequency-domain uncertainty at sampling step 3 as the image-level discrimination score, without training an additional classifier. We evaluate uncertainty scores computed from three frequency ranges: low frequency ($\omega < \pi/2$), high frequency ($\omega \geq \pi/2$), and the full frequency range, where ω follows the radial-frequency convention used in our frequency weighting analysis. We report the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), with higher values indicating better discrimination.

Table 5: Image-level artifact discrimination using frequency-domain uncertainty at sampling step 3. The evaluation uses 256 U-ViT images, including 166 images with artifacts as positive samples. No additional classifier is trained.

Frequency range	AUROC ↑	AUPRC ↑
Random reference	0.5000	0.6485
Low frequency	**0.6459**	0.8109
High frequency	0.6300	0.7981
Full frequency	0.6446	**0.8116**

As shown in Table 5, uncertainty scores from all three frequency ranges outperform the random reference. Low-frequency uncertainty achieves a higher AUROC than high-frequency uncertainty, while full-frequency uncertainty achieves the highest AUPRC. These observations support a statistical association between frequency-domain uncertainty and image-level artifacts. They are also consistent with our design choice of assigning stronger guidance weights to low-frequency components.

B.2 Qualitative Results on ImageNet

This section provides additional visual comparisons for the conditional generation experiments in Sec. 3.3. Figure 5 compares samples generated by the baseline, spatial-domain uncertainty guidance [De Vita and Belagiannis, 2025], and our frequency-domain uncertainty guidance on ImageNet using U-ViT [Bao et al., 2023] at 256^2 resolution.

As observed from the results, while the method in [De Vita and 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.

B.3 Artifact-Specific Evaluation on ImageNet and LSUN-Churches256

FID and CLIP evaluate distribution-level visual fidelity and image–condition alignment, respectively, but do not directly measure localized artifacts. We therefore complement these metrics with the artifact detector provided by DiffDoctor [Wang et al., 2025], which is trained on human-annotated artifact data and produces pixel-wise artifact confidence maps.

Experimental setup. We evaluate 4,096 generated images per method on each of two datasets: ImageNet-256 for conditional generation with U-ViT, and LSUN-Churches256 for unconditional generation with a U-Net. We compare baseline sampling, spatial-domain uncertainty guidance [De Vita and Belagiannis, 2025], and our frequency-domain uncertainty guidance. For each guidance method, generated images are paired with their corresponding baseline images to assess changes in artifact area.



(a) Baseline



(b) Spatial Uncertainty Guidance [De Vita and Belagiannis, 2025]



(c) Frequency Uncertainty Guidance (Ours)

Figure 5: Additional qualitative comparison on ImageNet with U-ViT at 256^2 resolution.

Evaluation metrics. We report aggregated artifact confidence and artifact area, with lower values indicating fewer detector-estimated artifacts. For each dataset, we set a threshold τ to the 95th percentile of the baseline pixel-wise artifact confidence distribution. This threshold is then kept fixed when evaluating all three methods.

Let $c_i(p)$ denote the artifact confidence at pixel p of image i, and let $\mathcal{P}_i$ be its pixel set. The artifact area of image i is defined as

$$A_i = \frac{100}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathbb{I}[c_i(p) > \tau] \,, \tag{22}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. We report the mean artifact area over the evaluated images. For each guidance method, we additionally report the percentages of paired images whose artifact area decreases, increases, or remains unchanged relative to the baseline. By construction, the baseline artifact area is always equal to 5%.

Results. Table 6 shows that our method achieves the lowest aggregated artifact confidence and mean artifact area on both datasets. These results provide artifact-specific evidence that complements the FID and CLIP evaluations and supports the effectiveness of frequency-domain uncertainty guidance in both conditional and unconditional generation.

B.4 Qualitative Results on LSUN-Churches256

This section provides additional visual comparisons for the unconditional generation experiments in Sec. 3.3. Figure 6 compares the baseline, spatial-domain uncertainty guid-

Table 6: Artifact-specific evaluation using DiffDoctor on 4,096 images per method and dataset. Confidence denotes aggregated artifact confidence; area denotes mean artifact area. Dec., Inc., and Same denote the percentages of paired images whose artifact area decreases, increases, or remains unchanged relative to baseline sampling.

Method	Confidence ↓	Area (%) ↓	Dec. (%)	Inc. (%)	Same (%)
ImageNet-256: conditional generation					
Baseline	357.17	5.00	–	–	–
Spatial guidance	353.24	4.53	61.33	38.57	0.10
Frequency guidance (Ours)	**344.72**	**4.26**	69.85	30.10	0.05
LSUN-Churches256: unconditional generation					
Baseline	406.32	5.00	–	–	–
Spatial guidance	408.29	4.83	52.91	47.09	0.00
Frequency guidance (Ours)	**388.63**	**3.54**	71.41	28.49	0.10

ance [De Vita and Belagiannis, 2025], and our method on LSUN-Churches256 [Yu et al., 2015] using a pretrained unconditional DDPM U-Net [Ho et al., 2020]. The results show that our method better preserves complex architectural structures, such as church spires, while producing fewer artifacts. This further supports the generalization of our frequency-domain uncertainty guidance to unconditional generation.



(a) Baseline



(b) Spatial Uncertainty Guidance [De Vita and Belagiannis, 2025]



(c) Frequency Uncertainty Guidance (Ours)

Figure 6: Additional qualitative comparison on LSUN-Churches256 using a pretrained unconditional U-Net DDPM

B.5 Additional Ablation Study

We ablate two design choices in the frequency-domain score guidance of Sec. 2.4: the frequency weighting function and the frequency mask. Experiments are conducted on ImageNet 256^2 with U-ViT.

Frequency weighting function. We compare three weighting functions. Let the radial frequency be $\omega(u, v) = \pi\sqrt{u^2 + v^2}, \quad u, v \in [0, 1]$, where $\omega \in [0, \sqrt{2}\pi]$. The three weighting functions are defined as follow.

(1) Low-Frequency Decay function: $W_{\text{low}}(\omega) = \exp(-\omega/T)$, (2) High-Frequency Growth function: $W_{\text{high}}(\omega) = \exp(-(\sqrt{2}\pi - \omega)/T)$, (3) Uniform Weighting function: $W_{\text{uni}}(\omega) = c, \quad c = \frac{T}{\sqrt{2}\pi}\left(1 - \exp(-\sqrt{2}\pi/T)\right)$. The constant c makes the total weighting strength of uniform weighting comparable to W_{low} and W_{high}. In all settings, we set temperature coefficient $T = 5$, guidance strength $\lambda = 0.1$, and keep the frequency-mask percentile as $p = 90$.

Table 7: Ablation of frequency weighting functions on ImageNet 256^2 with U-ViT.

Weighting Function	CLIP ↑	FID ↓
High-Frequency Growth	29.00	15.97
Uniform Weighting	29.03	15.92
Low-Frequency Decay	**29.07**	**15.89**

As shown in Table 7, assigning stronger guidance to low frequencies yields the best performance. This suggests that aggressively reducing uncertainty in large-scale structures is beneficial, consistent with Fig. 1, where "bad" samples exhibit a larger uncertainty gap in low-frequency components than "good" samples.

Frequency mask. We further ablate the frequency mask by varying the percentile threshold p. We use three settings: (1) $p = 90$ with $\lambda = 0.1$, (2) $p = 50$ with $\lambda = 0.02$, and (3) $p = 0$ with $\lambda = 0.01$, where $p = 0$ means applying guidance to all frequency components. The guidance strength λ is reduced for lower thresholds to keep the overall guidance magnitude comparable.

Table 8: Ablation of frequency masks on ImageNet 256^2 with U-ViT.

Mask Percentile p	CLIP ↑	FID ↓
0	28.93	16.18
50	28.98	16.08
90	**29.07**	**15.89**

Table 8 validates the use of the frequency mask to guide only high-uncertainty frequency components. These components correspond to the scales and structures where the model is most uncertain, while leaving most frequency components unchanged helps preserve the overall image structure.

Guidance Steps. We examine the effect of applying uncertainty guidance at sampling steps 3–5 versus all 50 sampling steps. Experiments use U-ViT on ImageNet-256 with 4,096 generated images per setting and $M = 5$ Monte Carlo samples at each guided step.

The steps 3–5 setting follows De Vita and Belagiannis [2025] and uses guidance strength $\lambda = 0.1$. For guidance across all 50 steps, we set $\lambda = 0.006$ to match the nominal sum of the per-step guidance coefficients. We report CLIP, FID, throughput, and the number of denoising network function evaluations per image (NFE/image).

As shown in Table 9, applying guidance at all 50 steps achieves comparable CLIP and FID scores while substantially increasing computational cost. The corresponding throughput decreases from 2.313 to 0.520 images per second. These results support the use of guidance at steps 3–5 as an efficient sampling configuration.

Table 9: Comparison of guidance at sampling steps 3–5 and all 50 steps using U-ViT on ImageNet-256. Each setting is evaluated with 4,096 generated images and $M = 5$ MC samples per guided step. Throughput is measured in images per second.

Guidance steps	CLIP ↑	FID ↓	Throughput ↑	NFE/image ↓
3–5	**29.07**	**15.89**	**2.313**	**65**
All 50	29.04	16.08	0.520	300

B.6 Visualizations under Few-Step Sampling

This section provides qualitative comparisons for the few-step generation experiments discussed in Sec. 3.4. Here, we apply only posterior variance modulation, without score guidance. Figure 7 compares the baseline with our frequency-domain uncertainty-based variance modulation under 5-step sampling, using U-ViT [Bao et al., 2023] at a 256^2 resolution.



(a) Baseline



(b) Our Covariance Estimation

Figure 7: Qualitative comparison of extreme few-step generation (5 sampling steps) using the UViT architecture [Bao et al., 2023].

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 accelerating diffusion sampling.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: frequency-domain uncertainty estimation and structure-aware diffusion guidance. The stated scope is limited to diffusion sampling guidance.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 discusses a limitation of the proposed method.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the assumptions behind the frequency-domain diagonalization and reverse-covariance approximations, the derivations are provided in Sections 2.3–2.5 and Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the proposed algorithm in Section 2, and the evaluated datasets, model architectures, metrics, sample counts, and hyperparameters needed to reproduce the main experimental setup in Section 3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code is not publicly released at submission time.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.2 specifies the experimental datasets, model architectures, evaluation metrics, sampling settings, and hyperparameters used in the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper reports standard point-estimate metrics such as FID, CLIP score, and HPSv2, but does not report error bars, confidence intervals, or statistical significance tests.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 3.2 reports the compute workers used for the experiments, including four NVIDIA V100 GPUs with 32GB memory each.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses public datasets and pretrained diffusion models, does not involve human subjects, and is intended for improving the reliability and quality of generative model sampling.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Section 5 discusses the positive impact of improving diffusion sampling quality and the potential limitation that frequency-domain guidance may introduce non-local modifications, which requires caution in applications where localized changes should be strictly controlled.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release new high-risk pretrained generative models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources of the datasets and models used in the experiments.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.