Spectral Prior for Reducing Exposure Bias in Diffusion Models

Yuya Kobayashi[1], Masato Ishii[1], Yuhta Takida[1], Shibuya Takashi[1], and Yuki Mitsufuji[1,2]

[1] Sony AI
[2] Sony Group Corporation
https://ai.sony/

Abstract. Diffusion models typically suffer from error accumulation during iterative sampling, commonly referred to as exposure bias. We reveal systematic frequency-dependent discrepancies between training and inference, which can be interpreted as frequency-dependent SNR error. Crucially, the direction of this mismatch varies across models and timesteps, indicating that fixed correction rules do not generalize. We propose Spectral Alignment (SPA), a lightweight, guidance-based method that calibrates the power spectrum of intermediate predictions to a pre-computed prior. Our approach consists of two stages: (1) offline fitting of a parametric spectrum model from training data, and (2) inference-time guidance via efficient FFT-based gradient computation. SPA introduces minimal computational overhead (3-4%) and is complementary to Classifier-Free Guidance (CFG). We demonstrate consistent improvements across diverse architectures, from pixel-space models (DDPM, ADM) to latent diffusion models (SD2.0, SDXL) and flow-matching models (SD3.5, FLUX).

1 Introduction

Diffusion models have become a dominant paradigm for image generation [6, 10, 17, 26, 29]. Despite this rapid progress, a core challenge remains: during iterative sampling, small prediction errors can accumulate and lead to a distribution shift away from the target, a phenomenon often referred to as *exposure bias*. Prior studies have argued that exposure bias in diffusion sampling manifests primarily as a mismatch between the noise levels seen during training and those encountered at test time. Accordingly, these methods aim to correct the effective noise level to match the predefined noise schedule, e.g., by predicted-noise rescaling [21], time-shifted sampling [15], and wavelet-based frequency reweighting [37]. However, these approaches either assume frequency-uniform errors or rely on hand-designed correction rules and carefully tuned schedules.

In this work, we analyze the power spectrum of intermediate predictions $\hat{x}_{0|t}$ during diffusion inference and reveal a systematic spectral mismatch between the forward process and inference trajectories (Figure 2). We interpret this discrepancy as frequency-dependent effective SNR errors that vary across models

and timesteps, and show that correcting it leads to improved image quality. We discuss the details of this observation and its implications for method design in Section 2.

Based on this observation, we propose **Spectral Alignment (SPA)**, a lightweight guidance-based calibration method applicable to a wide range of models. This method shifts the power spectra of intermediate predictions toward a precomputed timestep-dependent *spectral prior*. The overview of the method is shown in Figure 1. Our approach consists of two stages: (1) pre-calculation of the spectral prior and (2) sampling with guidance using spectral prior. In the first stage, a parametric spectrum model are fitted using training data. In the second stage, samples are steered toward the spectral prior via efficient FFT-based gradient computation.

The method can be applied to various models adaptively. We conducted experiments across a wide range of text-to-image diffusion models, from early architectures (DDPM [10], ADM [6]) to LDMs (SD2.0 [31], SDXL [24]) and state-of-the-art flow-matching models (SD3.5 [32], FLUX [2]). We show that SPA outperforms other baseline methods, and consistently improves generation quality as measured by FID [9] and reward models such as HPSv3 [18]. In addition, SPA introduces minimal computational overhead (3–4%) and can be easily integrated into existing generation pipelines.

Our main contributions are as follows:

- We reveal that even modern flow-matching models exhibit spectral mismatch, and propose a calibration method using a precomputed reference power spectrum.
- We demonstrate that correcting spectral mismatch with our approach improves generation quality with minimal computational overhead (3–4%).
- We show that SPA is applicable to a wide variety of models without modification, from DDPM to state-of-the-art flow-matching models.

2 Motivation

In this section, we describe the key observations motivating the proposed method. We first introduce the notion of spectral mismatch from a frequency-dependent effective SNR error, and then present empirical observations that guided the design of our approach.

2.1 Preliminaries

Diffusion Models The forward process of diffusion models gradually adds Gaussian noise to data x_0 according to a predefined noise schedule. The latent variable at an intermediate timestep, denoted by x_t, can be obtained as:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_t$ determines the noise level at timestep t. The reverse process learns to denoise x_t by predicting the noise $\epsilon_\theta(x_t, c, t)$ using a neural network, given

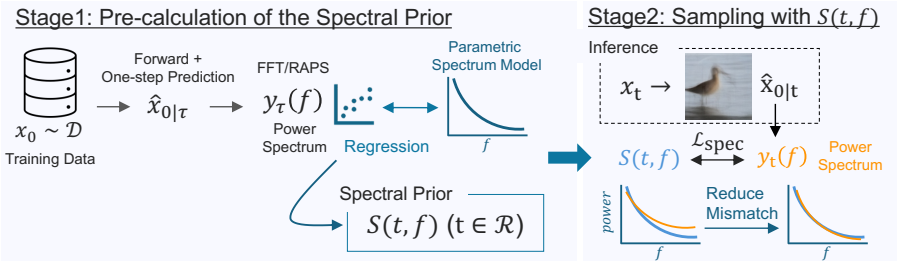

Fig. 1: Overview of the proposed method. We pre-calculate the spectral prior in the first stage, and spectral mismatch is corrected using the prior in the second stage during inference. For latent diffusion models, apply the same procedure on latent z.

condition c. The denoising step from x_t to x_{t-1} using DDIM [29] is given by a deterministic update:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\hat{x}_{0|t} + \sqrt{1-\bar{\alpha}_{t-1}} \cdot \epsilon_\theta(x_t, c, t), \tag{2}$$

where $\hat{x}_{0|t}$ is the predicted clean data estimated from x_t using Tweedie's formula:

$$\hat{x}_{0|t} = \frac{x_t - \sqrt{1-\bar{\alpha}_t} \cdot \epsilon_\theta(x_t, c, t)}{\sqrt{\bar{\alpha}_t}}. \tag{3}$$

Classifier-Free Guidance (CFG) For conditional generation, CFG [11] computes a guided noise prediction by extrapolating between the conditional and unconditional estimates:

$$\epsilon_\theta^{\text{CFG}} = \epsilon_\theta(x_t, t, \emptyset) + w \cdot \big(\epsilon_\theta(x_t, t, c) - \epsilon_\theta(x_t, t, \emptyset)\big), \tag{4}$$

where $w > 1$ is the guidance scale and $\emptyset$ denotes null conditioning. While CFG improves prompt adherence, high guidance scales often lead to oversaturation artifacts [27, 35].

2.2 Spectral Mismatch

Frequency-Dependent SNR Exposure bias in diffusion models refers to the mismatch between the distribution of intermediate variables observed during training (*i.e.*, forward process) and those during inference (*i.e.*, reverse process). Prior work suggests quantifying this mismatch through deviations from the predefined noise schedule, which is often referred to as an SNR error. Ning *et al.* [21] proposed rescaling the predicted noise to correct such errors. Furthermore, Yu *et al.* [37] showed that the mismatch differs between low- and high-frequency bands, indicating a frequency-dependent structure. In this work, we go beyond these perspectives and examine how the mismatch varies across the entire frequency spectrum, timesteps, and model choices, and how it can be corrected.

To analyze this, we define the signal-to-noise ratio at frequency ω and timestep t as:

$$\mathrm{SNR}(\omega, t) = \frac{\bar{\alpha}_t \, |\tilde{x}_0(\omega)|^2}{(1 - \bar{\alpha}_t) \, |\tilde{\epsilon}(\omega)|^2}, \tag{5}$$

where $\tilde{x}_0(\omega) = \mathcal{F}[x_0]$ and $\tilde{\epsilon}(\omega) = \mathcal{F}[\epsilon]$ denote the Fourier coefficients at frequency ω. Since Gaussian noise has a flat power spectrum (*i.e.*, $\mathbb{E}[|\tilde{\epsilon}(\omega)|^2]$ is constant across frequencies), the frequency dependence of $\mathrm{SNR}(\omega, t)$ is primarily determined by the signal power $|\tilde{x}_0(\omega)|^2$. Accordingly, rather than directly estimating an SNR error, we focus on the mismatch in signal strength (power spectrum) between training and inference in this study.

During training, ground-truth images x_0 are available, and their frequency characteristics can be computed directly. During inference, however, x_0 is unknown, preventing a direct comparison of $\mathrm{SNR}(\omega, t)$ between the two settings. To address this, we use the estimated clean image $\hat{x}_{0|t}$ as a common proxy in both cases. Specifically, we compute $\hat{x}_{0|t}^{\mathrm{train}}$ by applying a trained denoiser to x_t sampled from the forward process, and $\hat{x}_{0|t}^{\mathrm{infer}}$ from the reverse process at the corresponding timestep. By comparing the power spectra of these two estimates, we isolate the mismatch caused by the distributional gap between training-time and inference-time inputs, excluding the intrinsic high-frequency attenuation inherent to the posterior mean estimator. This allows us to quantify how the spectral mismatch varies across timesteps and models.

Empirical Observation Figure 2(a) shows the averaged power spectra for each channel of $\hat{x}_{0|t}^{\mathrm{train}}$ and $\hat{x}_{0|t}^{\mathrm{infer}}$ obtained from ADM and SDXL. We observe systematic spectral mismatches across all models tested, but crucially, the direction and pattern of the mismatch vary across models and timesteps. For instance, ADM [6] exhibits high-frequency attenuation, while Stable Diffusion 2.0 [31] shows low-frequency attenuation. More recent models such as SDXL [24], SD3.5 [32], and FLUX [2] show complex, channel-wise, time-dependent patterns. More examples can be found in the Appendix.

2.3 Motivation for the Proposed Method

A natural question is whether reducing the spectral mismatch directly improves sample quality, and if so, how to reduce it efficiently. In this paper, we propose a guidance-based method to reduce the spectral mismatch for two reasons. First, we aim to establish that spectral mismatch is a meaningful source of quality degradation and that correcting it yields measurable improvements. Second, we seek a practical, lightweight add-on applicable to existing pretrained models without retraining.

The model-dependent diversity shown in the observations above has two important implications. First, fixed correction rules (e.g., "boost high frequencies") cannot generalize across architectures and denoising schedules. Second, the mismatch carries spatial structure that cannot be resolved by scalar variance cor-

rection alone. These observations motivate a data-driven approach that learns the appropriate target spectrum for each model.

3 Method

3.1 Overview

In this section, we describe our guidance-based spectral mismatch correction method. Our approach consists of two stages. In the first stage, we characterize the expected power spectrum of intermediate predictions $\hat{x}_{0|t}$ by fitting a parametric model using training data and a pretrained diffusion model. This is an offline process required only once per model. In the second stage, during inference, we steer each denoising step toward the target spectrum via an efficient FFT-based gradient correction. We describe the details of the first stage in Section 3.2 and those of the second stage in Section 3.3. Figure 1 shows an overview of our method.

Remark. This modeling process involves single-step model predictions (i.e., the forward process followed by one denoising step). While single-step prediction is not free of prediction error, it does not suffer from the error accumulation that arises during multi-step rollouts.

3.2 Target Power Spectrum Modeling

To define a correction target, we characterize the expected frequency properties of $\hat{x}_{0|t}$ by fitting a parametric spectrum model from training data.

Power Spectrum Extraction For each channel of $\hat{x}_{0|t}$, we compute the 2D power spectrum $\mathbf{Y}_t = |\mathbf{F}\hat{x}_{0|t}|^2$, where $\mathbf{F}$ denotes the 2D DFT matrix, and $|\cdot|^2$ denotes the element-wise squared magnitude. We then reduce this to a 1D representation via radial averaging, commonly referred to as the Radially Averaged Power Spectrum (RAPS) [30]:

$$y_t(f) = \text{RadialAvg}(\mathbf{Y}_t), \tag{6}$$

where $f \in (0, 0.5]$ is the normalized spatial frequency ($f = 0.5$ is the Nyquist frequency). All operations are applied independently to each channel.

Parametric Spectrum Model It is well known that the power spectrum of natural images follows a power law [7, 33]. We confirmed that $\hat{x}_{0|t}$, which is basically a blurry image, also follows the power law with slightly different parameters from natural images. We model the target spectrum as:

$$S(t, f) = p_t \cdot f^{-q_t} + r_t \cdot f + s_t, \tag{7}$$

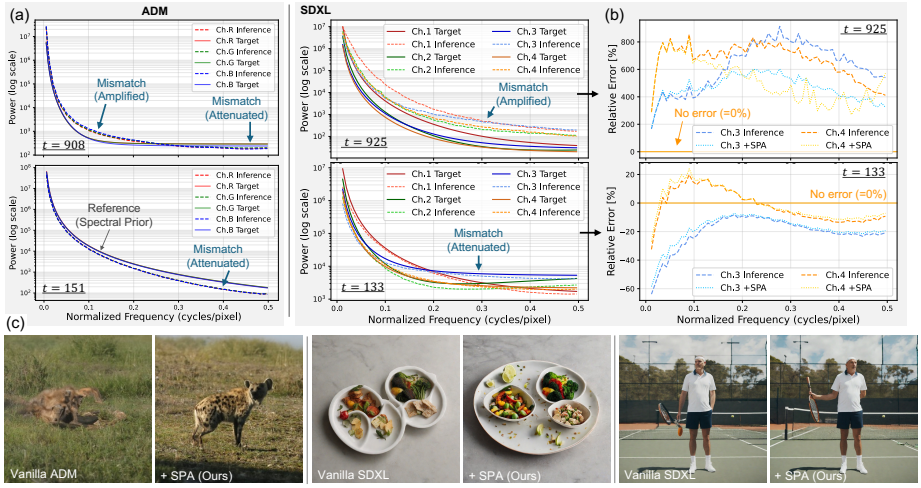

Fig. 2: Visualization of spectral mismatch in ADM and SDXL. The channel-wise spectra are obtained from $\hat{x}_{0|t}$ (or $\hat{z}_{0|t}$ with SDXL) by averaging multiple samples. (a) Spectra from the target spectral prior (real lines) and from inference (broken lines). (b) Relative error for SDXL between the target and inference corresponding to the left figure. Results with SPA are also shown (dotted lines). They have smaller errors across frequencies (broken lines). Only two channels are visualized for visibility. (c) Examples of generated images used for these spectra. Each pair of images is sampled with the same initial noise and a random seed.

where p_t, q_t, r_t, and s_t are timestep-dependent parameters. The first term captures the power-law decay characteristic of natural images. The bias term s_t and the linear term $r_t \cdot f$ represent deviations from the pure power law. For the linear term in particular, we set $r_t = 0$ by default and enable it only when it significantly improves the fit (e.g., SDXL, FLUX). All channels are modeled independently.

Model design. It is technically possible to use a 2D representation directly; however, we chose to use RAPS for the following reasons. (1) A 1D power spectrum is easier to visualize, which makes analysis, explanation, and hyperparameter tuning much simpler, and (2) the regression model becomes much simpler in the 1D case, which makes the fitting process more stable and more robust.

Parameter Estimation Steps

Step 1: Per-timestep fitting. For a discrete set of timesteps $\tau \in \{t_1, t_2, \ldots, t_K\}$, we collect reference power spectra via the following procedure:

1. Sample clean images $x_0^{(i)}$ from the training dataset and apply the forward process to obtain $x_\tau^{(i)} = \sqrt{\bar{\alpha}_\tau}\, x_0^{(i)} + \sigma_\tau\, \epsilon^{(i)}$.

2. Compute $\hat{x}_{0|\tau}^{(i)}$ using Eq. (3) with the pretrained model.
3. Extract the RAPS $y_\tau^{(i)}(f)$ via Eq. (6) and average across samples: $\bar{y}_\tau(f) = \frac{1}{N}\sum_{i=1}^{N} y_\tau^{(i)}(f)$.
4. Fit the parametric model (Eq. (7)) to $\bar{y}_\tau(f)$ via least-squares regression, obtaining $\{p_\tau, q_\tau, r_\tau, s_\tau\}$ for each channel and timestep independently.

Step 2: Temporal Interpolation. To obtain smooth parameter functions over all timesteps $t \in [0, T]$, we apply cubic spline interpolation to each parameter:

$$g(t) = \text{CubicSpline}(t; \{\tau_k, g_{\tau_k}\}), \tag{8}$$

where $g \in \{p, q, r, s\}$. Since the number of evaluation steps during inference is not fixed, the resulting model $S(t, f)$ needs to capture the expected spectral characteristics of $\hat{x}_{0|t}$ at any timestep.

3.3 Guidance-based Spectral Mismatch Calibration

Having established the target spectrum, we now describe how to steer the inference trajectory toward it. We frame this as a guidance problem, adapting Diffusion Posterior Sampling (DPS) [3] to enforce spectral alignment at each denoising step. This inference-time approach serves as a practical add-on for existing pretrained models.

Guidance Loss. At each reverse diffusion step, we compute $\hat{x}_{0|t}$ via Tweedie's formula (Eq. (3)), extract its RAPS $y_t(f)$, and compare against the target spectrum $S(t, f)$. The guidance loss is defined as:

$$\mathcal{L}(x_t, t) = \frac{1}{N}\sum_{c=1}^{N}\sum_{k=1}^{K}\big(\log_{10} y_t(f_k^c) - \log_{10} S(t, f_k^c)\big)^2, \tag{9}$$

where the quadratic loss is summed over K frequency bins and averaged over N channels. $f_k \in (0, 0.5]$ represents the center frequency of the k-th bin. We use an unweighted MSE loss in log-spectrum space for simplicity and stability, and leave frequency-dependent weighting for future work.

Asymmetric Penalty. During the fitting process, we observed that the sample-wise median of the spectrum $\log_{10} y_\tau(f)$ from the forward process tends to be larger than the corresponding mean and exhibits positive skewness at each frequency f. This indicates that $\log_{10} y_t(f)$–an observation during inference–falling too far below the target is undesirable, while exceeding the target is comparatively benign (see the Appendix for visualization).

This observation leads us to introduce an asymmetric penalty, which penalizes cases where the current spectrum falls below the target more strongly:

$$\phi_a(x) = \begin{cases} x & \text{if } x \geq 0, \\ a \cdot x & \text{if } x < 0, \end{cases} \tag{10}$$

Algorithm 1 Spectral Alignment (SPA) for Conditional Sampling

Require: Pretrained model ϵ_θ, target spectrum model $S(t, f)$, guidance strength η, penalty intensity a, text embedding c, null embedding $\emptyset$, and CFG scale w
Notation: $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator.
1: Sample initial noise $x_T \sim \mathcal{N}(0, I)$
2: **for** $t = T, T-1, \ldots, 1$ **do**
3: $\quad \epsilon_c, \epsilon_\emptyset \leftarrow \epsilon_\theta(x_t, t, c),\ \epsilon_\theta(x_t, t, \emptyset)$
4: $\quad \epsilon^{\mathrm{CFG}} \leftarrow \epsilon_\emptyset + w(\epsilon_c - \epsilon_\emptyset)$ ▷ *Classifier-Free Guidance*
5: $\quad \hat{x}_{0|t} \leftarrow (x_t - \sigma_t \,\mathrm{sg}[\epsilon^{\mathrm{CFG}}])/\sqrt{\bar{\alpha}_t}$ ▷ *Tweedie's formula*
6: $\quad y_t \leftarrow \mathrm{RadialAvg}(|\mathbf{F}\,\hat{x}_{0|t}|^2)$ ▷ *Extract RAPS*
7: $\quad \mathcal{L}_{\mathrm{spec}} \leftarrow \sum_{k=1}^{K} \big(\phi_a(\log_{10} y_t(f_k) - \log_{10} S(t, f_k))\big)^2$ ▷ *Asymmetric Loss*
8: $\quad x_t \leftarrow x_t - \eta\, \nabla_{x_t} \mathcal{L}_{\mathrm{spec}}$ ▷ *Spectral Alignment*
9: $\quad x_{t-1} \leftarrow \mathrm{Denoise}(x_t, \epsilon^{\mathrm{CFG}}, t)$ ▷ *Sampler step (DDIM, Euler, etc.)*
10: **end for**
11: **return** x_0

where $a > 1$ controls the penalty intensity. The modified loss becomes:

$$\mathcal{L}_{\mathrm{spec}}(x_t, t) = \frac{1}{N} \sum_{c=1}^{N} \sum_{k=1}^{K} \big(\phi_a(\log_{10} y_t(f_k^c) - \log_{10} S(t, f_k^c))\big)^2 \tag{11}$$

Adaptation to LDMs. The method we described above can be applied to latent diffusion models without modification. Our assumption so far is that the power spectrum of the intermediate variable x_t follows a power law. VAEs of latent diffusion models are known to preserve the spatial information from the pixel space [26]. In practice, the power spectrum of z_t can be well fitted by our parametric power-law model. Visualization of the fitting can be found in the Appendix.

Overall Algorithm: Algorithm 1 summarizes the complete sampling procedure with CFG. Our spectral alignment is applied after the CFG combination step (line 4) and before the standard denoising update (line 9). Thus, the method is generally compatible with most CFG variants [4, 14, 27, 35] and other noise prediction modifications, since it operates on the combined prediction ϵ^{CFG} without any assumptions about how it was obtained.

Computational Efficiency. The additional cost of our method is minimal. Spectrum extraction via FFT has complexity $\mathcal{O}(D \log D)$ for D-dimensional data, and radial averaging is $\mathcal{O}(D)$. The gradient computation involves only elementwise operations and an inverse FFT for the backpropagation through the FFT, since it only involves the loss function, RAPS, and Tweedie's formula (lines 5–7 of Algorithm 1). Because the gradient from ϵ^{CFG} is not used, no additional neural network evaluations or backpropagation steps are required for guidance. In practice, the per-step overhead is less than 5% of the total sampling time (see Section 4.2).

4 Experiments

4.1 Experimental Setup

Models. We evaluate our method across a wide range of diffusion architectures: unconditional/class-conditional pixel-space models (DDPM [10] on CelebA-HQ [12], ADM on ImageNet 256×256 [5]), text-to-image latent diffusion models (Stable Diffusion 2.0, SDXL [24]), and flow-matching models (SD3.5 medium, FLUX.1 [dev]). For SDXL, we used the first stage without the refiner. We used the DDPM sampler for ADM, which is the default in the official implementation, and the DDIM sampler for the other models. We used 50 denoising steps for DDPM, 100 steps for ADM, and 30 steps for the other models.

Pre-calculation of target spectra. We fit the parametric spectrum model (Section 3.2) using 10K samples. The model's training dataset was used for DDPM and ADM (CelebA-HQ and ImageNet, respectively), and the LAION-Aesthetics V2 dataset [28] was used for text-to-image models. The computational cost of this process is comparable to the scale of calculating metrics such as FID, and this process is only required once per model. We confirmed that the coefficient of determination ($\mathcal{R}^2$ score) exceeds 0.9 in most cases (see the Appendix for details).

Evaluation Metrics. For unconditional and class-conditional models (DDPM, ADM), we report FID and KID as primary metrics, which reliably measure distributional fidelity in unconditional settings. We used clean-fid [23] to compute these metrics. We additionally report Density and Coverage [20], which are updated versions of Precision and Recall. For text-to-image models (SD2.0, SDXL, SD3.5, FLUX), we adopt HPSv3 and ImageReward (reward-model-based metrics) [18, 36] as primary indicators, since FID is known to correlate negatively with human preference for these models, especially at high CFG scales [24]. We also report the CLIP score (ViT-G/14) as a measure of text alignment. 5K prompts from the MS COCO validation set.

Baselines. We compare against the following noise correction methods: (1) ϵ-rescaling [21], which rescales the predicted noise to match the expected norm; (2) time-shift sampling [15], which adjusts the noise schedule at inference time; and (3) wavelet-based frequency regulation [37], which applies frequency-band reweighting.

Hyperparameters were determined as follows. For (1), we used 1.004, following their ImageNet experiment. For SD2.0 and SDXL, we evaluated a candidate set $\{1.003, 1.005, 1.007\}$ and selected 1.005. For (2), the default values were used. The coefficient for (3) depends on the NFE and image resolution, which complicates the hyperparameter search. They reported the unimodality of FID with respect to the hyperparameters. Furthermore, since the qualitative effectiveness of this method becomes more prominent as the coefficient increases, we adopted the largest possible value within the range where no visual artifacts were observed.

Specifically, we applied $(w^l, w^h) = (1.0005, 1.005)$ for ADM, $(1.003, 1.005)$ for SD2.0, and $(1.002, 1.001)$ for SDXL. Note that (1) has negligible computational overhead, (2) requires an additional timestep search (as detailed in Appendix F of their original work), and (3) involves a wavelet transform.

SPA Hyperparameters. The hyperparameters used for our proposed method are as follows. DDPM: $(\eta, a) = (0.01, 1.0)$, ADM: $(0.05, 2.0)$, SD2.0 and SDXL: $(0.2, 5.0)$, SD3.5: $(0.75, 3.0)$, FLUX.1 [dev]: $(0.2, 2.0)$ for CFG scale $w = 2.5$ and $(0.2, 1.0)$ for $w = 3.5$. Here, η is the guidance strength and a is the asymmetric penalty intensity. To manage computational costs, we first performed a preliminary qualitative assessment to identify promising parameter ranges, followed by a focused quantitative evaluation to finalize these values. Although SPA introduces two additional hyperparameters, they typically require tuning only once per model architecture. Notably, models with similar designs (e.g., SD2.0 and SDXL) converged to identical optimal values despite being searched independently, suggesting the robustness and transferability of these parameters.

4.2 Quantitative Results

Unconditional and Class-Conditional Generation We evaluated our method on unconditional/class-conditional pixel-space models. Since all baselines are implemented for ADM, we conducted the comparison on ADM (ImageNet 256×256). The results are shown in Table 1. On ADM, our method consistently outperforms the other methods. Density/Coverage are also improved, even though SPA guides intermediate variables toward the average spectrum. This indicates that SPA has a benefits in terms of generation diversity. The time-shift sampler did not perform well in our experiments. This is likely due to the difference in image resolution, as they only experimented with resolutions up to 128×128.

We also observe consistent improvement on DDPM. However, the performance gain is slightly smaller than that on ADM, and simple $\epsilon-$rescaling works slightly better. This is likely because our method involves single-step model prediction in the target spectrum fitting stage. The target may be inaccurate when the single-step prediction itself is not accurate. We observed that the DDPM model did not perform well at predicting $\hat{x}_{0|t}$ even from a ground-truth (forward process) training data x_t, which was not the case with the other models.

Text-to-Image Generation Next, we evaluated text-to-image models. Although baseline methods were designed for ADM, we re-implemented them for SD2.0 and SDXL with minimal modifications. Table 2 shows the quantitative results.

SPA also performs well across various T2I models. Note that scores from reward models behave more like an ordinal scale than an interval or ratio scale. With that in mind, we conclude that SPA performs best among these baselines, especially on SDXL (approximately a 4.5% gain). CLIP Scores appear to be saturated, but SPA does not degrade text alignment.

Table 1: Results on unconditional and class-conditional generation.

Model	Method	FID ↓	$\text{KID}_{(\times 1000)}$ ↓	Density ↑	Coverage ↑
ADM (ImageNet)	Vanilla	9.29	19.48	1.133	0.5939
	ϵ-rescaling	8.00	11.01	1.146	0.5976
	Time-shift	15.35	77.47	0.761	0.4321
	Wavelet reg.	8.35	12.80	1.122	0.5929
	SPA (Ours)	**7.81**	**10.25**	**1.226**	**0.6184**
DDPM (CelebA)	Vanilla	49.44	43.57	0.8186	**0.0169**
	ϵ-rescaling	**48.53**	**40.98**	**0.8791**	**0.0169**
	Time-shift	90.19	46.65	0.8440	0.0038
	Wavelet reg.	49.46	43.31	0.8182	0.0168
	SPA (Ours)	48.63	41.36	0.8344	0.0167

SPA also works well with both flow-matching models. For FLUX.1, SPA achieves decent performance with true CFG (the scores in parentheses), which is the standard two-pass CFG without distillation, but the score improvement with guidance distillation (without parentheses) is relatively small. In the following section, we discuss the results for FLUX.1 [dev] in detail.

Discussion on Guidance-Distilled Models. FLUX.1 [dev] is a guidance-distilled model [19], trained to internalize the effect of CFG. We observe that SPA provides statistically significant improvements at lower guidance scales ($w = 2.5$, win-rate $53.3 \pm 1.5\%$, $p = 2.3 \times 10^{-5}$), while the effect diminishes at higher scales ($w = 3.5$). This suggests that guidance distillation may partially address spectral mismatch during training but does not fully eliminate it, particularly at lower guidance scales.

Notably, SPA shows stronger improvements on lower-quality samples. For images with HPSv3 scores in the bottom 20%, the win-rate increases to $54.1 \pm 3.3\%$ ($p = 0.02$) even at $w = 3.5$. The average score gain for the bottom 5% of samples is 0.70 ($w = 2.5$) and 0.19 ($w = 3.5$). This indicates that SPA serves as a useful refinement mechanism that selectively improves samples suffering from spectral distortion.

Computational Overhead We evaluated the computational overhead of SPA on SDXL. We compared the time consumed for (A) the standard denoising step (lines 3–4 in Algorithm 1) with (B) the additional steps required for SPA (lines 5–8). The average time for (A) was 2.47 s and for (B) was 0.0952 s; thus, the overhead is only +3.86%. Since the FFT is the bottleneck, SPA works efficiently with latent diffusion models.

4.3 Qualitative Results

Figure 3 shows results for SD2.0 and SDXL with the baseline methods listed in Section 4.1. In our experiments, SPA tends to refine flawed object shapes, as

Table 2: Quantitative results on text-to-image generation. We primarily evaluate image quality using HPSv3 and ImageReward, and additionally report CLIP Score (in gray) to indicate text alignment is preserved. w denotes a CFG scale. † The scores in parentheses uses true CFG (no distillation) with a "de-distilled" checkpoint [22]. See Section 4.1 for metric selection.

Model	Method	HPSv3 ↑	ImageReward ↑	CLIP Score ↑
SD2.0 (w=7.5)	Vanilla	7.043	0.393	0.333
	ϵ-rescaling	7.186	0.415	0.333
	Time-shift	7.101	0.407	0.333
	Wavelet reg.	7.193	**0.429**	0.332
	SPA (Ours)	**7.238**	0.421	0.333
SDXL (w=7.5)	Vanilla	8.426	0.791	0.341
	ϵ-rescaling	8.528	0.800	0.340
	Time-shift	8.352	0.795	0.340
	Wavelet reg.	8.576	0.807	0.340
	SPA (Ours)	**8.829**	**0.829**	0.341
SD3.5 (w=5)	Vanilla	9.782	0.939	0.335
	SPA (Ours)	**9.975**	**0.976**	0.334
FLUX.1 (w=2.5)	Vanilla	12.53 (11.55†)	1.044 (0.989)	0.328 (0.332)
	SPA (Ours)	**12.57** (**11.76**)	**1.054** (**1.005**)	0.328 (0.332)
FLUX.1 (w=3.5)	Vanilla	12.47 (11.94)	1.070 (1.044)	0.330 (0.334)
	SPA (Ours)	**12.48** (**12.02**)	1.070 (**1.052**)	0.330 (0.334)

shown in the SD2.0 example in this figure. When the vanilla inference is already decent, SPA keeps the content as is, while slightly enhancing the color and detail, as seen in the SDXL sample.

Figure 4 shows results from FLUX.1 [dev]. This model has strong baseline performance; however, the model still occasionally generates unnatural structures. In such cases, SPA can refine the images. This behavior is practically useful for making adjustments after fixing the layout (via seed and prompts). Note that this is consistent with the statistical results in Section 4.2. We show five examples in the figure. The first two images contain defective objects (paddle/bat), while the remaining images have unnatural object structures.

Why Does SPA Improve Object Shapes? Interestingly, SPA can refine object shapes in addition to statistical image properties (e.g., lighting, color distribution). We hypothesize two possible reasons for this phenomenon. (A) By fixing the SNR at each timestep, diffusion models can better utilize signal components from earlier steps, resulting in refined object shapes. (B) In latent diffusion models, VAEs are known to have disentangled channel representations (colors, object shape, layout, lighting, etc.) [34]; therefore, aligning these channels can alter the overall content.

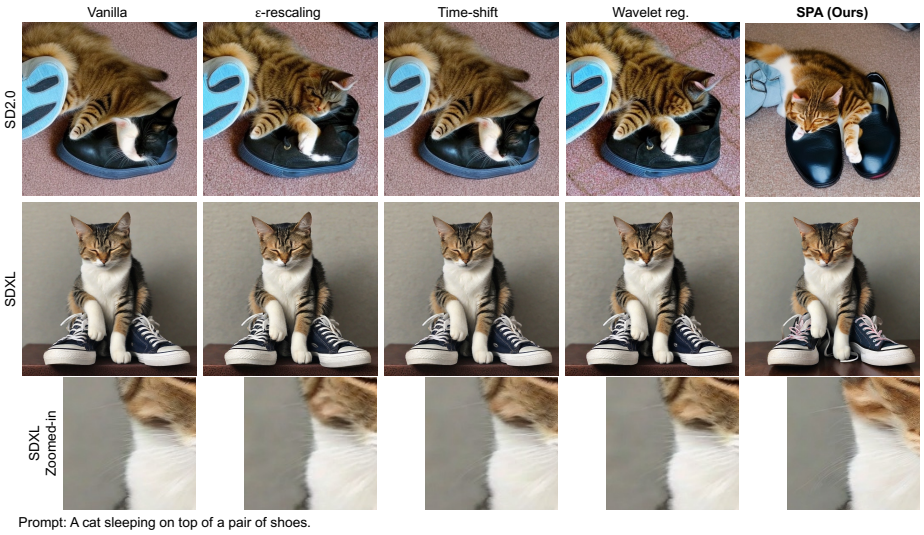

Fig. 3: Qualitative comparison of generated images with baseline methods. The baseline methods are enumerated in Section 4.1. SPA successfully improved the sample quality and fixed the flawed object shapes in the SD2.0 sample. On the SDXL sample, SPA refined the edge definition (bottom row, zoomed), while baseline methods added frequency artifacts to the background.

5 Related Work

5.1 Exposure Bias

The concept of exposure bias was originally identified in autoregressive sequence generation, particularly in neural machine translation and language modeling [1, 25]. In teacher forcing, models are trained to predict each token given ground-truth previous tokens, but at inference time, they must condition on their own predictions. This train-inference mismatch causes error accumulation, as models encounter inputs from a distribution different from that seen during training.

Diffusion models exhibit a structurally identical problem. Ning *et al.* [21] explicitly identified this connection and proposed ϵ*-rescaling* to correct noise-level mismatch during inference. Li *et al.* [16] analyzed error accumulation using an analytical model. Subsequent work has explored complementary corrections, as introduced in Section 4.1.

5.2 Guidance-based Correction Methods

Fundamentally, exposure bias stems from prediction errors in single-step predictions. Modifications to Classifier-Free Guidance (CFG) [11] aim to reduce such errors [4,14,27,35]. Although CFG is known as a truncation and error-correction method, it can also be a source of inference error, which often leads to oversaturation, unnatural contrast, and loss of fine details.

Fig. 4: Qualitative results from FLUX.1 [dev]. Despite the strong baseline performance of FLUX.1 [dev], it occasionally generates physically implausible structures (e.g., malformed objects, unnatural compositions). SPA selectively refines such failure cases without degrading already high-quality samples. This behavior is particularly valuable in practical deployment where consistent quality is required.

Discriminator Guidance [13] refines model scores using a specially trained discriminator. Although it is a powerful way to refine predictions during inference, it requires unstable adversarial training and incurs additional inference costs for the discriminator at every denoising step. In the context of training-free guidance, MPGD [8] proposes to alleviate *manifold error* using backpropagation through autoencoders, but this also requires additional inference costs. Note that our method is technically not mutually exclusive with these approaches, although empirical validation of their combination remains for future work.

6 Conclusion

We proposed Spectral Alignment (SPA) for diffusion models, which alleviates spectral mismatch and refines image quality during inference with minimal computational overhead. We demonstrated the effectiveness of SPA across various models, from classic pixel-space models to a modern flow-matching models. In addition, the concept of spectral mismatch that we introduced can serve as a lens for further improvements in model training, where inspection tools are valuable.

Limitations. The current formulation of our method assumes a single target spectrum as a prior; however, the optimal spectral prior may vary depending on the target data distribution. For example, illustrations and medical images exhibit different frequency properties. For conditional generation, it may be possible to use different target spectra depending on the conditioning information. In this paper, we did not explore alternative guidance methods or loss weighting strategies. Although we adopted a simple DPS-based guidance approach, there is room to explore more sophisticated methods. Regarding loss weighting, one could design a weighting schedule for $\mathcal{L}_{spec}$ across frequency bands and timesteps.

7 Acknowledgements

We would like to thank Naoki Murata for carefully reading the manuscript and giving us valuable feedback.

References

1. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1. p. 1171–1179. NIPS'15, MIT Press, Cambridge, MA, USA (2015)
2. BlackForestLabs: https://huggingface.co/black-forest-labs/FLUX.1-dev
3. Chung, H., Kim, J., Mccann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=OnD9zGAGT0k
4. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: CFG++: Manifold-constrained classifier free guidance for diffusion models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=E77uvbOTtp
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
6. Dhariwal, P., Nichol, A.Q.: Diffusion models beat GANs on image synthesis. In: Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems (2021), https://openreview.net/forum?id=AAWuCvzaVt
7. Field, D.J.: Relations between the statistics of natural images and the response properties of cortical cells. J. Opt. Soc. Am. A **4**(12), 2379–2394 (Dec 1987). https://doi.org/10.1364/JOSAA.4.002379, https://opg.optica.org/josaa/abstract.cfm?URI=josaa-4-12-2379
8. He, Y., Murata, N., Lai, C.H., Takida, Y., Uesaka, T., Kim, D., Liao, W.H., Mitsufuji, Y., Kolter, J.Z., Salakhutdinov, R., Ermon, S.: Manifold preserving guided diffusion. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=o3BxOLoxm1
9. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6629–6640. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc., Red Hook, NY, USA (2020)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
12. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of GANs for improved quality, stability, and variation. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=Hk99zCeAb

13. Kim, D., Kim, Y., Kwon, S.J., Kang, W., Moon, I.C.: Refining generative process with discriminator guidance in score-based diffusion models. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
14. Kwon, M., Kim, S.S., Jeong, J., Hsiao, Y.T., Uh, Y.: TCFG: Tangential Damping Classifier-free Guidance . In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2620–2629. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2025). https://doi.org/10.1109/CVPR52734.2025.00250, https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00250
15. Li, M., Qu, T., Yao, R., Sun, W., Moens, M.F.: Alleviating exposure bias in diffusion models through sampling with shifted time steps. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=ZSD3MloKe6
16. Li, Y., van der Schaar, M.: On error propagation of diffusion models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=RtAct1E2zS
17. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PqvMRDCJT9t
18. Ma, Y., Wu, X., Sun, K., Li, H.: Hpsv3: Towards wide-spectrum human preference score. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15086–15095 (October 2025)
19. Meng, C., Gao, R., Kingma, D.P., Ermon, S., Ho, J., Salimans, T.: On distillation of guided diffusion models. In: NeurIPS 2022 Workshop on Score-Based Methods (2022), https://openreview.net/forum?id=6QHpSQt6VR-
20. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: Proceedings of the 37th International Conference on Machine Learning. ICML'20, JMLR.org (2020)
21. Ning, M., Li, M., Su, J., Salah, A.A., Ertugrul, I.O.: Elucidating the exposure bias in diffusion models. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=xEJMoj1SpX
22. nyanko7: Flux-dev-de-distill, https://huggingface.co/nyanko7/flux-dev-de-distill
23. Parmar, G., Zhang, R., Zhu, J.Y.: On Aliased Resizing and Surprising Subtleties in GAN Evaluation . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11400–11410. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01112, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01112
24. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: Kim, B., Yue, Y., Chaudhuri, S., Fragkiadaki, K., Khan, M., Sun, Y. (eds.) International Conference on Learning Representations. vol. 2024, pp. 1862–1874 (2024), https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf
25. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016), http://arxiv.org/abs/1511.06732
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10674–10685.

IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.01042, https://doi.ieeecomputersociety.org/10.1109/CVPR52688.2022.01042
27. Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=e2ONKX6qzJ
28. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: an open large-scale dataset for training next generation image-text models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc., Red Hook, NY, USA (2022)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP
30. Spector, A., Grant, F.S.: Statistical models for interpreting aeromagnetic data. Geophysics **35**(2), 293–302 (04 1970). https://doi.org/10.1190/1.1440092, https://doi.org/10.1190/1.1440092
31. StabilityAI: https://stability.ai/news/stable-diffusion-v2-release
32. StabilityAI: https://huggingface.co/stabilityai/stable-diffusion-3.5-medium
33. van der Schaaf, A., van Hateren, J.: Modelling the power spectra of natural images: Statistics and information. Vision Research **36**(17), 2759–2770 (1996). https://doi.org/https://doi.org/10.1016/0042-6989(96)00002-8, https://www.sciencedirect.com/science/article/pii/0042698996000028
34. Vass, T.A.: Explaining the sdxl latent space (2024), https://huggingface.co/blog/TimothyAlexisVass/explaining-the-sdxl-latent-space, accessed: 2026-03-04
35. WANG, X., Dufour, N., Andreou, N., CANI, M.P., Abrevaya, V.F., Picard, D., Kalogeiton, V.: Analysis of classifier-free guidance weight schedulers. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=SUMtDJqicd
36. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: learning and evaluating human preferences for text-to-image generation. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 15903–15935 (2023)
37. Yu, M., Zhan, K.: Frequency regulation for exposure bias mitigation in diffusion models. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 10370–10378. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3755608, https://doi.org/10.1145/3746027.3755608