CORRECTING GUIDED DIFFUSION TRAJECTORIES WITH SPECTRAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

The practical success of conditional image generation hinges on fine-grained differences in condition alignment and visual fidelity. Classifier-free guidance (CFG) is central to this success, but its lack of an explicit criterion makes it difficult to assess whether the guided trajectory is progressing as intended. To address this gap, we show that spectral alignment provides a principled criterion for understanding guidance behavior and improving guided diffusion sampling through adaptive correction. Our analysis identifies the spectra of intermediate states as an indicator of consistency with the expected spectral evolution of the forward process. Based on this observation, we introduce *Spectral Correction Guidance*, a method that corrects deviations from an analytic reference spectrum during sampling. The proposed method is training-free and applicable across diffusion backbones and conditional generation tasks without modifying the underlying model. Experiments demonstrate consistent gains in preference-based metrics over baseline guidance methods in text-to-image generation and improved generation quality over CFG on ImageNet. These improvements persist across a range of guidance scales and with fewer denoising steps. Our analyses and ablations provide insight into guidance behavior and how the proposed method affects generation quality.

1 INTRODUCTION

Denoising diffusion models (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2020a;b; Dhariwal & Nichol, 2021) have shown strong generative capabilities across diverse domains (Ho et al., 2022; Liu et al., 2023; Chi et al., 2025), including image synthesis (Nichol & Dhariwal, 2021; Karras et al., 2022; Peebles & Xie, 2023). They exhibit stable convergence during training while modeling the reverse of a forward noising process that perturbs clean data toward a noise distribution. The learned reverse process captures rich data distributions, enabling sample generation through denoising from high to low noise levels. Building on this capability, diffusion models support open-ended conditional generation tasks such as text-to-image synthesis (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2023; Chen et al., 2023; 2024; Xie et al., 2024; Esser et al., 2024).

Classifier-free guidance (CFG) (Ho & Salimans, 2022) has become a standard technique for steering sampling toward the target condition in diffusion models. As illustrated in Figures 1a and 1b, conditional sampling without guidance often fails to capture the target condition, whereas CFG improves both condition alignment and perceptual quality. These benefits depend on how guidance is applied along the denoising trajectory (Hong et al., 2023; Karras et al., 2024; Sadat et al., 2024). However, standard CFG uses a fixed guidance scale, limiting its ability to provide appropriate guidance at each denoising step (Kynkäänniemi et al., 2024).

Recent guidance methods aim to enhance the benefits of CFG while mitigating its adverse effects. They adjust the guidance strength across timesteps (Kynkäänniemi et al., 2024; Malarz et al., 2025; Jin et al., 2025a), modify the direction of guidance with geometric constraints (Sadat et al., 2025a; Chung et al., 2024; Kwon et al., 2025; Jin et al., 2025b), or modulate guidance within predefined frequency bands (Zhang et al., 2024; Sadat et al., 2025b; Song et al., 2026). However, these benefits may not extend beyond the settings these methods are designed to address, such as high-guidance regimes. In Section 6.1, experimental results show that these methods yield limited gains across a range of generation settings, where vanilla CFG remains a strong baseline.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

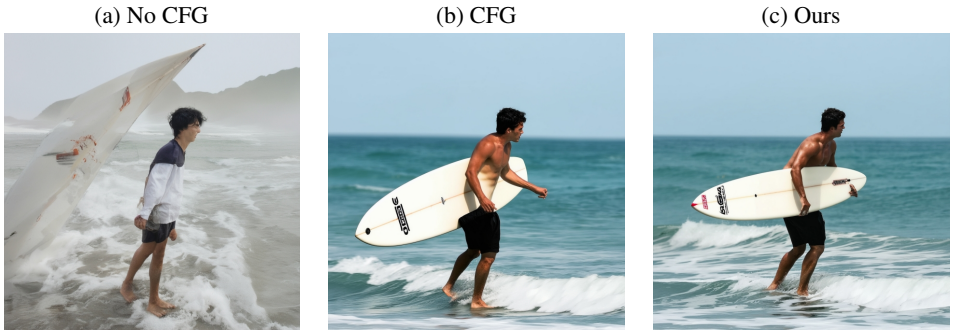

Figure 1: For the prompt "`a male surfer in black shorts carrying a white board`" on SD3.5 (Esser et al., 2024), sampling without CFG fails to form a reliable conditional image. CFG produces a more plausible image but introduces an artifact where the surfboard overlaps with the person. Our method better preserves the intended composition and improves visual coherence. It also yields the lowest spectral distance, which we analyze in Section 3.

In this paper, we argue that these limitations arise from reliance on constraints or suppression rules without a criterion to guide correction along the denoising trajectory. However, defining such a criterion is challenging in the conventional spatial domain, since coordinate-based representations encode local structures and reflect sample-specific characteristics. To address this difficulty, we identify alignment between the power spectra of guided intermediate states and the expected spectra under the forward process as a key indicator of effective guidance. This provides a complementary perspective, capturing global patterns of diffusion dynamics along the denoising trajectory.

Building on this observation, we propose *Spectral Correction Guidance*, a training-free method for correcting guided denoising trajectories according to the spectral dynamics of the diffusion process. We derive a criterion from the power-law spectral statistics of natural images (Ruderman, 1997), yielding a distributional reference that is not tied to individual samples or specific failure patterns. Based on this criterion, our method corrects deviations from the expected spectral behavior in intermediate states. The proposed correction enables adaptive adjustments across frequencies and timesteps without relying on a prescribed correction pattern.

We further validate our method through extensive experiments on diverse conditional image generation tasks. In text-to-image generation, the proposed correction improves preference-based metrics across multiple backbones and guidance settings, indicating better text alignment and visual quality. In class-conditional generation, our method achieves lower FID (Heusel et al., 2017) and higher Inception Score (Salimans et al., 2016) than CFG, even with fewer sampling steps. These results demonstrate that correcting spectral deviations during sampling can improve generation quality without additional training. Ablation studies examine how correction strength and the timestep range over which correction is applied affect generation outputs. We also discuss how the proposed method relates to existing guidance approaches from a spectral perspective.

2 Background

Denoising Diffusion Models. Diffusion models define a predefined forward noising process q and a learned reverse generative process p_θ (Luo, 2022; Lai et al., 2025). The forward process is given by $q(z_{1:T} \mid z_0) := \prod_{t=1}^{T} q(z_t \mid z_{t-1})$, where z denotes either a data sample or its latent representation. Under an appropriate noise schedule, the state at time t can be sampled from $q(z_t \mid z_0)$ as

$$z_t = \alpha_t z_0 + \sigma_t \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \tag{1}$$

where α_t and σ_t are time-dependent scaling factors. The forward process in Eq. 1 induces the marginal distribution

$$q(z_t) = \int q(z_t \mid z_0)\, q(z_0)\, dz_0, \tag{2}$$

where $q(z_0)$ denotes the data distribution. The learned reverse process p_θ parameterizes transitions $p_\theta(z_t \mid z_{t+1})$, which model the distribution of the less noisy state z_t given the noisier state z_{t+1}.

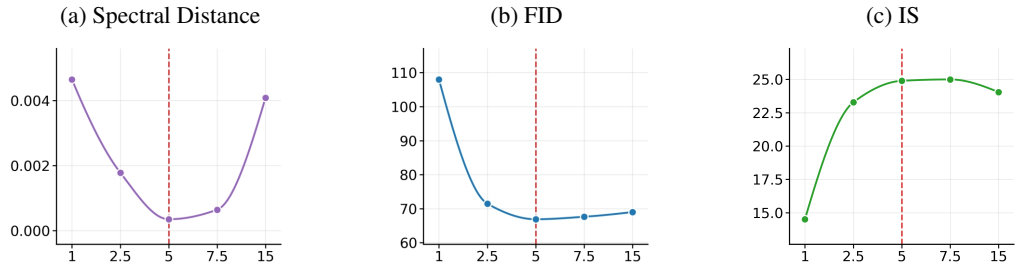

Figure 2: Effect of guidance scale on spectral distance defined in Eq. 6, FID, and IS. The x-axis is the guidance scale and the y-axis is the metric value. Guidance scale values are uniformly spaced for visual clarity. The red dashed line marks the default guidance scale.

Accordingly, the reverse transition defines the marginal at time t as

$$p_\theta(z_t) = \int p_\theta(z_t \mid z_{t+1})\, p_\theta(z_{t+1})\, dz_{t+1}, \tag{3}$$

where $p_\theta(z_t)$ denotes the distribution induced by reverse sampling at time t. By matching the reverse marginals $p_\theta(z_t)$ to the corresponding forward marginals $q(z_t)$, the diffusion model approximates the data distribution $q(z_0)$.

In practice, the generative process can be parameterized in different ways, including noise prediction (Ho et al., 2020), velocity prediction (Liu et al., 2022; Lipman et al., 2022), and score-based modeling (Song & Ermon, 2019; Song et al., 2020b). Across these variants, the models share the same objective of reversing the noising process, as formalized by the above notation.

Classifier-free Guidance (CFG). CFG is introduced as an alternative to classifier guidance (Dhariwal & Nichol, 2021), which uses the classifier gradient $\nabla_{z_t} \log p(c \mid z_t)$ to steer sampling toward the target condition c. With a noise prediction model ϵ_θ, classifier guidance modifies the denoising prediction as

$$\epsilon_{\mathrm{CG}}(z_t, t, c) = \epsilon_\theta(z_t, t) - s\, \sigma_t \nabla_{z_t} \log p(c \mid z_t), \tag{4}$$

where s is the classifier guidance scale. From Bayes' rule, the classifier gradient can be expressed as $\nabla_{z_t} \log p(c \mid z_t) = \nabla_{z_t} \log p(z_t \mid c) - \nabla_{z_t} \log p(z_t)$. Substituting this identity into Eq. 4 and using Tweedie's formula (Efron, 2011) to relate score terms to noise predictions yields the standard CFG form

$$\epsilon_{\mathrm{CFG}}(z_t, t, c) = \epsilon_\theta(z_t, t, \emptyset) + w\big(\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \emptyset)\big), \tag{5}$$

where w denotes the guidance scale and $\emptyset$ denotes the null condition.

Eq. 5 has been widely adopted in modern diffusion models due to its effectiveness in improving conditional generation quality. Nevertheless, determining appropriate guidance remains challenging in practice, as it remains difficult to assess whether the magnitude w and the direction $\epsilon_\theta(z_t, t, c) - \epsilon_\theta(z_t, t, \emptyset)$ of the guidance term are appropriate along the denoising trajectory.

3 Spectral Dynamics in the Diffusion Process

In this section, we examine how the guided diffusion trajectory evolves during sampling from a spectral perspective. To generate plausible samples from the data distribution, the reverse marginals $p_\theta(z_t)$ defined in Eq. 3 should remain close to the corresponding forward marginals $q(z_t)$ in Eq. 2 across timesteps. We thus examine the spectral statistics of intermediate states to assess this marginal alignment. The power spectra of intermediate states provide a tractable summary of how their energy is distributed across frequency components. Agreement between their expected power spectra is a necessary condition for matching the forward and reverse marginals.

We compare the expected power spectra of forward and reverse states across guidance scales under the same noise schedule, as detailed in Appendix E. Let $\hat{z}_t = \mathcal{F}(z_t)$ be the Fourier representation of z_t, and define its power spectrum as $P_t(\omega) = |\hat{z}_t(\omega)|^2$, where $\omega > 0$ denotes the radial frequency index.

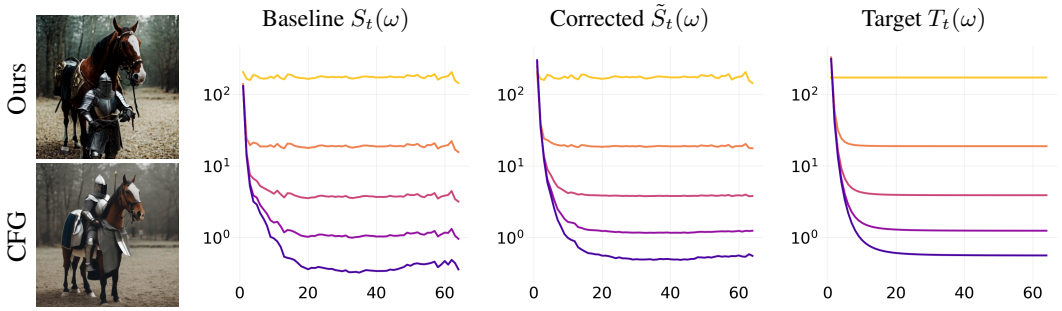

Figure 3: Qualitative and spectral comparison for the prompt "`a man holding a horse dressed like a knight`". Generated images are included for visual reference, and the power spectra are shown over denoising timesteps, with radial frequency bins on the x-axis and power on the y-axis. The corrected spectrum $\tilde{S}_t(\omega)$ follows the analytic target $T_t(\omega)$ more closely than the baseline spectrum $S_t(\omega)$, reducing artifacts and improving visual quality in the generated image.

The expected power spectra of forward and reverse states are $\mathbb{E}_{z_t\sim q(z_t)}[P_t(\omega)]$ and $\mathbb{E}_{z_t\sim p_\theta(z_t)}[P_t(\omega)]$, respectively. We quantify their discrepancy across timesteps and radial frequencies using the spectral distance defined as

$$D_{\text{spec}} = \mathbb{E}_{t,\omega}\left[\left(\log \mathbb{E}_{z_t\sim p_\theta(z_t)}[P_t(\omega)] - \log \mathbb{E}_{z_t\sim q(z_t)}[P_t(\omega)]\right)^2\right]. \tag{6}$$

Figure 2a shows the spectral distance at different guidance scales in Eq. 5. Moderate guidance reduces this distance relative to conditional sampling without guidance at scale 1. However, excessive guidance increases the spectral distance, indicating that the reverse trajectory deviates further from the expected spectral behavior. Figures 2b and 2c further show that spectral discrepancy is correlated with generation quality. Lower spectral discrepancy corresponds to lower FID and higher IS, indicating stronger generative performance. These observations suggest that spectral alignment can indicate whether guided sampling remains consistent with the expected diffusion dynamics. However, such corresponding forward states are unavailable during sampling, making it difficult to correct the denoising trajectory across timesteps and noise schedules.

4 Spectral Correction for Classifier-Free Guidance

To enable correction without access to corresponding forward states, we derive an analytic reference from the statistical structure of natural image spectra. Using the forward process in Eq. 1 and the spectral properties of Gaussian noise, we extend this reference to the expected power spectra of intermediate states along the forward diffusion trajectory.

Proposition 1. *Following the power-law model of natural image spectra (Ruderman, 1997), suppose that the expected power spectrum of z_0 takes the form*[1]

$$\mathbb{E}[P_0(\omega)] = C\omega^{-\beta} + c, \quad \omega > 0,\ C, \beta > 0,\ c \geq 0. \tag{7}$$

By definition, white Gaussian noise has a flat power spectrum,

$$\mathbb{E}[P_\epsilon(\omega)] = c_\epsilon, \quad \omega > 0,\ c_\epsilon \geq 0.$$

For the forward process $z_t = \alpha_t z_0 + \sigma_t \epsilon$ in Eq. 1, $z_0 \perp \epsilon$ and $\mathbb{E}[\epsilon] = 0$, so the cross term vanishes in expectation. Therefore, the expected power spectrum of z_t has the form

$$\mathbb{E}[P_t(\omega)] = \lambda_t \omega^{-\beta} + \gamma_t, \quad \omega > 0,\ \lambda_t, \gamma_t \geq 0,\ \beta > 0, \tag{8}$$

where λ_t and γ_t are determined by α_t, σ_t and the constants in Eq. 7.

Proposition 1 shows that the expected power spectrum at each timestep can be specified in an analytic functional form. This formulation allows us to compute the reference spectrum for assessing spectral alignment without simulating the corresponding forward states.

[1]We extend the standard power-law form with a constant offset c, where setting $c = 0$ recovers the conventional form.

Based on the criterion in Eq. 8, we formulate a spectral correction for guided denoising. Let $S_t(\omega)$ denote the power spectrum of the current reverse state z_t, and let $T_t(\omega)$ denote the expected power spectrum given by this criterion, which serves as the correction target at timestep t. We correct the current spectrum by interpolating between $S_t(\omega)$ and $T_t(\omega)$ in the log-spectral domain:

$$\log \tilde{S}_t(\omega) = (1-\lambda)\log S_t(\omega) + \lambda \log T_t(\omega), \tag{9}$$

where $\lambda \in [0,1]$ controls the correction strength. The use of log space accommodates the wide dynamic range of spectral power. Eq. 9 can be equivalently written as

$$\tilde{S}_t(\omega) = S_t(\omega)\left(\frac{T_t(\omega)}{S_t(\omega)}\right)^{\lambda}. \tag{10}$$

Eq. 10 shows that the correction rescales the current spectrum by the spectral ratio $R_t(\omega) = T_t(\omega)/S_t(\omega)$ with strength λ. This provides an adaptive correction that increases power when $R_t(\omega) > 1$ and decreases power when $R_t(\omega) < 1$.

Algorithm 1 summarizes how the correction is incorporated into the denoising process. The proposed correction preserves the phase of the current latent and modifies only its Fourier magnitude to match $\tilde{S}_t(\omega)$. This adjustment enables correction of spectral discrepancies while retaining the spatial information encoded in the Fourier phase. Figure 3 visualizes this correction process, showing that our method aligns the corrected spectrum $\tilde{S}_t(\omega)$ with the target $T_t(\omega)$ in both overall energy level and frequency-dependent shape. Our method therefore employs the expected spectral evolution of the forward process as a reference for correcting guided denoising trajectories.

In practice, we estimate the coefficients in Eq. 7 from the spectral statistics of real image samples (Lin et al., 2014) rather than through model-specific performance tuning. We perform this estimation in the z_0 space of each backbone without selecting samples to match its training data (Deng et al., 2009; Schuhmann et al., 2022). We compute radial power spectra by averaging two-dimensional Fourier power over radial frequency bins. Given the estimated coefficients, Eq. 8 determines the target spectrum $T_t(\omega)$ at each timestep through the corresponding noise schedule. To retain fine-grained frequency components (Karras et al., 2021), we relax the correction in the late denoising stage as discussed in Section 6.2. This procedure applies a common correction principle across diffusion models with different spectral statistics and noise schedules.

5 A Spectral View of Existing Guidance Methods

In this section, we discuss related guidance strategies. From a spectral perspective, we also examine how these strategies control distinct aspects of the guidance signal during denoising.

Scheduled guidance scaling. Scheduled guidance scaling replaces the constant guidance scale w with a predefined timestep-dependent schedule. Kynkäänniemi et al. (2024) restrict guidance to selected denoising intervals, while Malarz et al. (2025) and Jin et al. (2025a) adjust the guidance strength according to predefined functional forms. By Parseval's theorem, adjusting the guidance scale at each timestep also modulates the Fourier-domain energy of the guidance term. In this sense, our method provides fine-grained scaling through spectral correction at each timestep and frequency.

Geometry-aware guidance. Geometry-aware guidance methods constrain the magnitude (Jin et al., 2025b) or direction (Sadat et al., 2025a) of guided updates, or adjust sampling to reduce deviations from the data manifold (Chung et al., 2024; Kwon et al., 2025). Our method shares the goal of keeping guided trajectories consistent with the data distribution, but approaches it through spectral alignment with the forward marginals encountered during training.

Frequency-based guidance. Frequency-based methods modulate guidance using predefined frequency partitions (Zhang et al., 2024), applying different guidance strengths to separate bands (Sadat et al., 2025b) or suppressing selected frequencies (Song et al., 2026). Although these methods also consider spectral information, their adjustments follow band-specific rules. In contrast, our method adapts the correction at each frequency based on spectral deviations from an analytic reference.

6 Experiments and Results

Experimental setup. We evaluate our method on text-to-image and class-conditional image generation. For text-to-image generation, we compare against standard CFG (Ho & Salimans, 2022) and

Table 1: Quantitative comparison of guidance methods on text-to-image diffusion models, including HPSv3. Our method achieves the strongest preference-based performance across models. Among existing guidance strategies, standard CFG remains the most competitive baseline in most cases.

Model	w	Method	HPSv3↑	ImageReward↑	PickScore↑	HPSv2↑	CLIP-T↑	FID↓	sFID↓
SDXL	5.0	Ours	**8.437**	**0.731**	**22.43**	**0.279**	0.264	**17.02**	**47.94**
		CFG	8.096	0.700	22.41	0.274	**0.265**	17.62	49.27
		CFG++	7.962	0.684	22.36	0.273	**0.265**	17.46	49.11
		TV-CFG	6.952	0.514	22.03	0.263	0.264	18.94	50.14
		LF-CFG	7.045	0.566	21.95	0.267	0.262	18.00	52.75
PixArt-α	4.5	Ours	**9.957**	**0.876**	**22.63**	**0.299**	0.260	**27.95**	66.35
		CFG	9.571	0.845	22.55	0.298	0.260	28.30	66.43
		CFG++	6.664	0.346	21.69	0.263	0.245	28.27	56.85
		TV-CFG	9.510	0.864	22.43	0.298	**0.261**	29.97	66.61
		LF-CFG	3.371	0.120	20.97	0.246	0.245	33.77	**54.13**
SD3.5	4.5	Ours	**9.471**	**0.986**	**22.44**	**0.284**	0.268	20.24	49.07
		CFG	9.406	0.979	22.41	**0.284**	**0.269**	20.31	48.79
		CFG++	-4.826	-1.082	19.70	0.174	0.229	74.49	66.35
		TV-CFG	8.618	0.812	22.06	0.275	0.266	**15.98**	44.22
		LF-CFG	8.631	0.889	22.13	0.275	0.268	19.93	**12.61**

representative methods from Section 5: TV-CFG (Jin et al., 2025a), CFG++ (Chung et al., 2024), and LF-CFG (Song et al., 2026). We use 10K prompts sampled from the COCO validation set (Lin et al., 2014). To assess generality across backbones, we conduct experiments with SDXL (Podell et al., 2023), PixArt-α (Chen et al., 2023), and SD3.5 (Esser et al., 2024). We evaluate each model at its default guidance scale and at a higher guidance scale. For class-conditional generation, we use DiT-XL/2 (Peebles & Xie, 2023) on ImageNet (Deng et al., 2009) and evaluate robustness across different numbers of sampling steps. We generate 50 samples per ImageNet class, resulting in 50K samples in total. Unless otherwise specified, all methods use the same prompts, labels, sampling budget, and guidance scales, and follow the recommended settings for each method.

Evaluation metrics. For text-to-image generation, we assess human preference using HPSv3 (Ma et al., 2025), ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), and HPSv2 (Wu et al., 2023), text-image alignment using CLIP-T (Radford et al., 2021), and distributional matching using FID (Heusel et al., 2017) and sFID (Nash et al., 2021). We complement these metrics with a human preference study. For class-conditional generation, we follow the ADM evaluation protocol (Dhariwal & Nichol, 2021) and report FID, sFID, Inception Score (IS) (Salimans et al., 2016), Precision, and Recall to evaluate generation quality and distributional coverage across sampling budgets measured by the number of function evaluations (NFEs).

6.1 Experimental Results

Text-to-image generation. Table 1 presents quantitative results comparing our method with guidance baselines under their standard guidance settings. Standard CFG outperforms the other baseline methods on most preference-based metrics across the three backbones, indicating limited gains from existing guidance strategies at standard guidance scales. Our method improves on this strong baseline, achieving the highest HPSv3, ImageReward, and PickScore across all three models, along with the best or joint-best HPSv2 scores. Even on SD3.5, where standard CFG already achieves strong preference scores, our method yields additional gains. For distributional metrics, our method achieves the lowest FID and sFID on SDXL and the lowest FID on PixArt-α. Although TV-CFG and LF-CFG achieve lower FID and sFID on SD3.5, this comes with lower preference scores.

Figure 4 provides qualitative support for the trends in Table 1. In the SDXL examples, most methods produce plausible images, although LF-CFG shows weaker prompt alignment. CFG introduces a visible artifact around the front part of the airplane, which is alleviated by our method and CFG++. For PixArt-α, although other guidance variants provide some improvements, our method shows clearer gains in overall appearance and brightness balance. In the SD3.5 examples, CFG++ produces over-smoothed images, consistent with its lower quantitative scores. The remaining methods achieve similar overall quality, but our method produces more distinct fine-grained details.

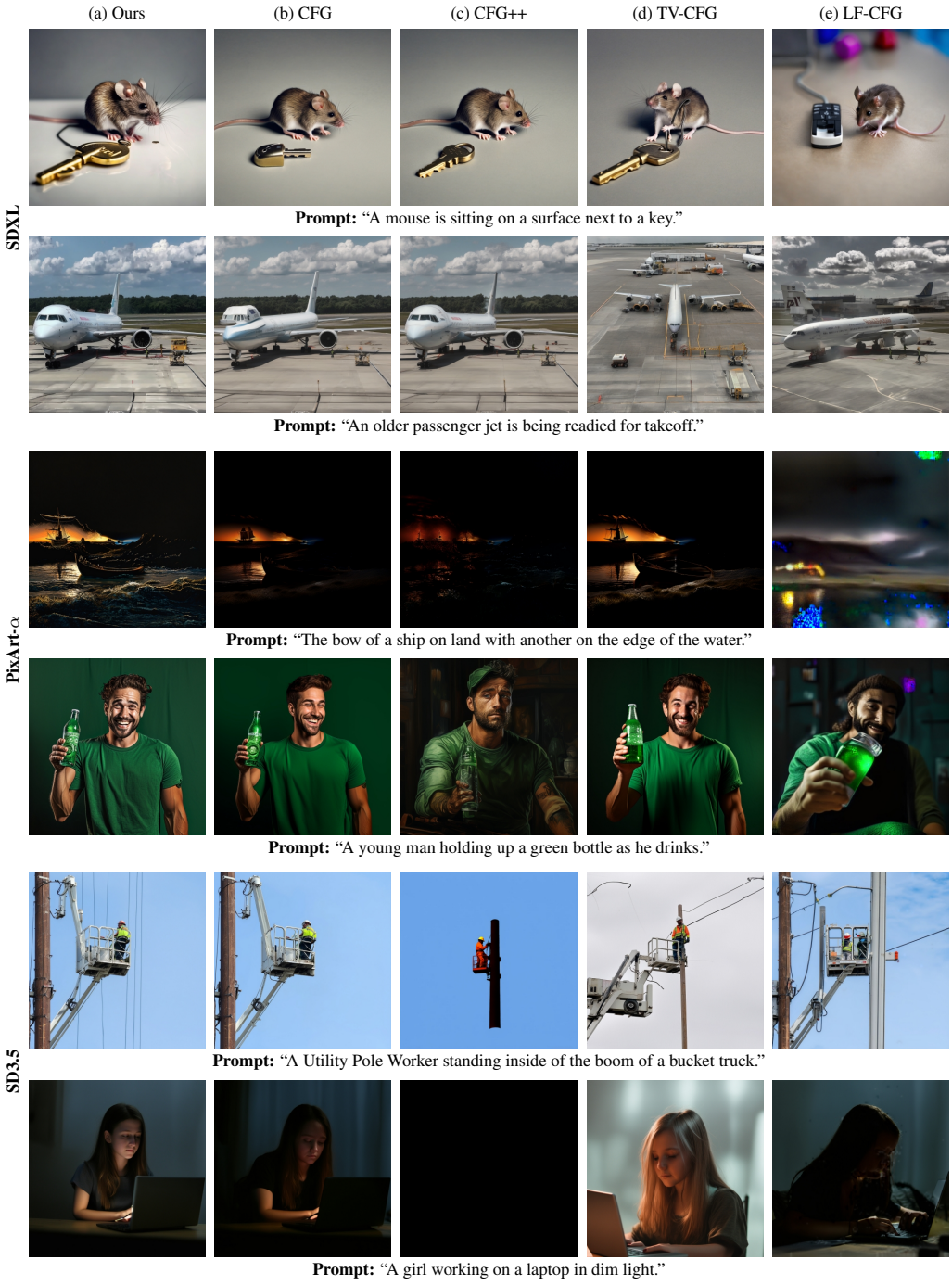

Figure 4: Qualitative comparison of guidance methods across text-to-image diffusion models. Our method better preserves the intended composition and improves visual coherence.

Table 2: Quantitative comparison between CFG and our method at higher guidance scales. Our method improves preference-based metrics across all three backbones.

Model	w	Method	ImageReward↑	PickScore↑	HPSv2↑	CLIP-T↑	FID↓	sFID↓
SDXL	10.0	Ours	**0.860**	**22.61**	**0.290**	**0.268**	**19.01**	**48.88**
		CFG	0.842	22.60	0.288	**0.268**	19.49	50.03
PixArt-α	9.0	Ours	**0.871**	**22.41**	**0.298**	**0.261**	**32.00**	72.09
		CFG	0.844	22.37	0.296	**0.261**	32.19	**71.59**
SD3.5	14.0	Ours	**1.055**	**22.28**	**0.292**	**0.270**	**22.71**	**12.93**
		CFG	0.657	21.70	0.266	0.267	24.17	21.86

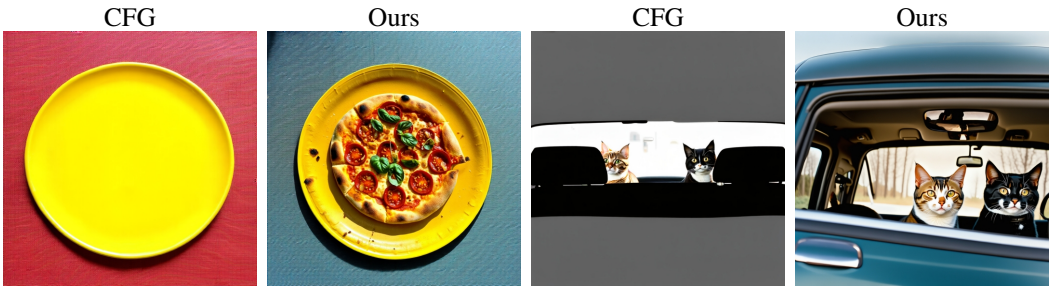

Figure 5: Qualitative comparison between CFG and our method on SD3.5 at a high guidance scale of $w = 14$. Our method corrects over-saturated and over-simplified outputs produced by CFG.

Furthermore, we evaluate our method under different guidance regimes. Table 2 demonstrates that our method improves over CFG at higher guidance scales. The improvement is clear for SD3.5, where our method maintains strong image quality and alignment while CFG becomes more sensitive to the increased guidance scale. As shown in Figure 5, the larger guidance scale can lead to over-saturated images or weaker prompt following, whereas our method mitigates these effects and recovers more faithful compositions. Table 9 also shows that our method improves under lower guidance, while existing guidance methods show limited gains. These results indicate that the proposed correction provides stable guidance control across models and guidance scales. We provide a human preference study, multi-seed experiments and runtime measurements in the appendix.

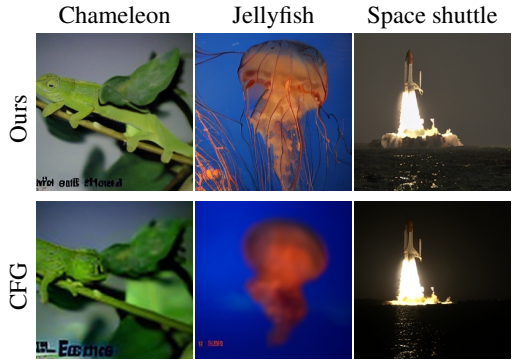

Figure 6: Qualitative examples for ImageNet class-conditional generation with DiT-XL/2.

Table 3: Class-conditional ImageNet generation results with DiT-XL/2. Results include the default 250-step setting and reduced sampling budgets measured by the number of function evaluations (NFE).

NFE	Method	FID↓	sFID↓	Precision↑	Recall↑	IS↑
250	Ours	**2.62**	**4.92**	**0.842**	0.550	**62.92**
	CFG	2.89	7.42	0.823	**0.565**	61.77
100	Ours	**2.78**	**5.65**	**0.843**	0.553	**63.13**
	CFG	3.07	7.64	0.827	**0.557**	61.84
50	Ours	**2.88**	**6.29**	**0.840**	0.552	**62.52**
	CFG	3.09	7.54	0.825	**0.554**	61.54
25	Ours	**3.07**	**6.78**	**0.830**	0.545	**61.47**
	CFG	3.20	7.43	0.824	**0.548**	60.79

Class-conditional image generation. To evaluate our method across conditional generation tasks, we apply the proposed correction to class-conditional generation with DiT-XL/2 on ImageNet. As shown in Table 3, our method achieves strong performance with the default 250 denoising steps. In particular, it reduces sFID from 7.42 to 4.92 and improves FID, Precision, and IS. Figure 6 shows that these quantitative gains are reflected in image quality. In the chameleon example, the proposed correction leads to a clearer depiction of the target class. The improvement is also visible in the jellyfish and space shuttle examples, where our method produces brighter and sharper images than the blurrier or darker CFG samples.

The proposed method maintains strong performance across sampling budgets, with NFEs ranging from 250 to 25. Even with 50 NFEs, our method achieves an FID of 2.88 and an sFID of 6.29, matching the FID and improving upon the sFID of CFG with 250 NFEs. At 25 NFEs, our method remains competitive with CFG using 100 NFEs. Recall is slightly lower than CFG across budgets, indicating a trade-off between coverage and the observed gains in precision and fidelity. These results demonstrate that the proposed correction improves generation quality across sampling budgets and extends to conditional generation tasks beyond text-to-image synthesis.

6.2 Ablation Studies

Here, we study how correction strength affects generation results and examine how to make effective use of the proposed correction in the denoising process.



Prompt: "A bathroom that is empty and not in use"

Figure 7: Ablation on the correction strength λ. Increasing λ improves visual fidelity and prompt alignment up to a point, but excessive correction introduces artifacts.



Prompt: "A young girl pretending to be a pirate" **Prompt:** "a person standing on a large green field"

Figure 8: Ablation on the relaxation parameter ρ. Smaller ρ applies correction for longer during denoising, while larger ρ relaxes it earlier. Intermediate relaxation yields the best balance between spectral correction and artifact reduction.

Effect of correction strength. The ablation shows that an appropriate correction strength improves visual fidelity and prompt alignment. As shown in Figure 7b, a small correction leaves the output close to standard guidance with little visible effect, suggesting that weak spectral modification has a limited effect on the guided trajectory. At moderate strengths, Figure 7c shows richer colors and clearer details, indicating that stronger correction better restores spectral components affected by guidance. However, excessive correction introduces artifacts and degrades perceptual quality as shown in Figure 7d. These results suggest that the correction strength controls the balance between correcting deviations from the expected spectrum and preserving individual sample characteristics.

Correction relaxation. The previous ablation shows that increasing correction strength improves fidelity and prompt alignment but can introduce artifacts. We thus study how to relax the correction across the denoising process to retain its benefits while reducing artifacts. Motivated by the role of late denoising steps in fine-grained detail formation (Karras et al., 2021), we vary the relaxation parameter ρ, which controls when the frequency-domain constraint is relaxed during denoising. Figure 8 shows that applying correction until the end can retain artifacts, whereas applying it only in the early stage weakens its effect. An intermediate relaxation point provides the best balance between spectral correction and natural detail formation. We therefore adopt this setting in our final method. The relaxation analysis clarifies where spectral correction improves generation quality and where its continued application can degrade quality (Kynkäänniemi et al., 2024; Jin et al., 2025a).

7 Conclusion

In this paper, we demonstrated that spectral alignment can serve as a practical criterion for understanding guidance behavior and improving guided diffusion sampling through principled correction. This perspective provides a basis for assessing whether guided trajectories remain consistent with the spectral dynamics induced by the forward process. We derived an analytic reference from natural image statistics and used it to correct spectral deviations along guided denoising trajectories. Building on this criterion, Spectral Correction Guidance is a training-free method applicable to diverse diffusion backbones and conditional generation tasks without modifying the underlying model. Experiments show that the proposed method improves generation quality across model architectures, guidance strengths, and reduced sampling budgets, while prior guidance strategies provide limited gains in common guidance settings. This work highlights how diffusion dynamics and domain-specific statistics can be combined to establish a principled criterion for trajectory correction.

REFERENCES

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

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 *Forty-first international conference on machine learning*, 2024.

Alexandre Galashov, Ashwini Pokle, Arnaud Doucet, Arthur Gretton, Mauricio Delbracio, and Valentin De Bortoli. Learn to guide your diffusion model. In *International Conference on Learning Representations*, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7462–7471, 2023.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. *arXiv preprint arXiv:2509.22007*, 2025a.

Cheng Jin, Zhenyu Xiao, Chutao Liu, and Yuantao Gu. Angle domain guidance: Latent diffusion requires rotation rather than extrapolation. In *International Conference on Machine Learning*. PMLR, 2025b.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *Advances in Neural Information Processing Systems*, 37:52996–53021, 2024.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023.

Mingi Kwon, Shin seong Kim, Jaeseok Jeong, Yi Ting Hsiao, and Youngjung Uh. Tcfg: Tangential damping classifier-free guidance. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2620–2629, 2025.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Chieh-Hsin Lai, Yang Song, Dongjun Kim, Yuki Mitsufuji, and Stefano Ermon. The principles of diffusion models. *arXiv preprint arXiv:2510.21890*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095, October 2025.

Dawid Malarz, Artur Kasymov, Maciej Zieba, Jacek Tabor, and Przemysław Spurek. Classifier-free guidance with adaptive scaling. *arXiv preprint arXiv:2502.10574*, 2025.

SG Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

Charlie Nash, Jacob Menick, Sander Dieleman, and Peter W Battaglia. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Pinelopi Papalampidi, Olivia Wiles, Ira Ktena, Aleksandar Shtedritski, Emanuele Bugliarello, Ivana Kajic, Isabela Albuquerque, and Aida Nematzadeh. Dynamic classifier-free diffusion guidance via online feedback. In *International Conference on Learning Representations*, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Daniel L. Ruderman. Origins of scaling in natural images. *Vision Research*, 1997.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. *arXiv preprint arXiv:2407.02687*, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M Weber. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales. *arXiv preprint arXiv:2506.19713*, 2025b.

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.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, 2016.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Kaiyu Song, Baoshun Tong, and Hanjiang Lai. Rethinking oversaturation in classifier-free guidance via low frequency. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10107–10111. IEEE, 2026.

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

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

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.

Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 15903–15935, 2023.

Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024.

APPENDIX

A REPRODUCIBILITY STATEMENT

Section 4 presents the formulation of the proposed method, and Algorithm 1 summarizes the correction procedure. Appendix F provides the proof of Proposition 1. Appendix E details spectral coefficient estimation, hyperparameter settings, experimental configurations, and computational resources to facilitate reproduction of the results.

B LLM USAGE STATEMENT

We used generative AI tools in auxiliary roles to support and polish the writing of the manuscript and to assist with implementing and debugging experimental code. All AI-assisted outputs were reviewed by the authors.

C ETHICS STATEMENT

Generative modeling can support creative and practical applications, including image editing, design, and content creation. At the same time, improvements in generation quality can increase the risk of misuse, including harmful synthetic content. We therefore emphasize the need for responsible deployment policies.

D LIMITATIONS AND FUTURE WORK

The proposed method demonstrates that spectral alignment based on power spectra can provide a useful criterion for guiding diffusion trajectories and improving generation quality. However, this criterion only captures marginal frequency-domain statistics and does not fully characterize the generated distribution. Matching power spectra can align the overall spectral energy across frequencies, but it does not account for phase information or spatial structure. Future work could extend spectral correction beyond power spectra by using wavelet representations (Mallat, 1989) to capture localized frequency structure and by incorporating constraints that account for phase and individual sample structure. Another limitation is that our correction is applied with a shared rule across samples and conditions. In practice, the amount and type of guidance needed may depend on the input condition, prompt, or class. Guidance strategies that adapt to the input condition (Galashov et al., 2026; Papalampidi et al., 2026) suggest a promising direction for extending spectral correction to account for differences across samples and conditions. Combining spectral alignment with condition-aware guidance control could enable more flexible adjustment of the denoising trajectory.

E ADDITIONAL IMPLEMENTATION DETAILS

Baseline implementation details. We implement each baseline according to its recommended setting whenever applicable.[2] For TV-CFG, we set the time-varying guidance schedule such that it peaks at the midpoint of the denoising process and has the same average guidance scale as the CFG guidance scale in the corresponding experiment. For LF-CFG, we follow the original guideline and implement the low-frequency correction using upsampling and downsampling with an $8\times$ scaling ratio and a low-frequency attenuation factor of $\rho = 0.5$. For CFG++, we use a guidance scale of $w/12.5$ based on the approximate scale correspondence reported in prior work (Chung et al., 2024; Jin et al., 2025b). Prompts are shown exactly as used during generation, including typographical errors in the original COCO captions.

Implementation details of our method. We apply the proposed correction at every denoising step, as described in Algorithm 1. For text-to-image generation, we use 64 radial bins. The correction

[2]Backbone models: stabilityai/stable-diffusion-xl-base-1.0, PixArt-alpha/PixArt-XL-2-1024-MS, stabilityai/stable-diffusion-3.5-medium, facebook/DiT-XL-2-256, JiT-L/16, and SiT-XL/2-256.

Algorithm 1 Spectral Correction Guidance

Require: condition y, CFG scale w, denoising steps T, correction strength λ, target spectra $\{T_t(\omega)\}_{t=0}^{T-1}$, radial bin count B

Initialize latent $z_T \sim \mathcal{N}(0, I)$
for $t = T, \dots, 1$ **do**
  $\epsilon_\varnothing \leftarrow \epsilon_\theta(z_t, t, \varnothing), \quad \epsilon_y \leftarrow \epsilon_\theta(z_t, t, y)$
  $\hat{\epsilon}_t \leftarrow \epsilon_\varnothing + w(\epsilon_y - \epsilon_\varnothing)$
  $z_{t-1} \leftarrow \mathrm{Update}(z_t, \hat{\epsilon}_t, t)$
  $\hat{z}_{t-1} \leftarrow \mathcal{F}(z_{t-1})$
  Compute radial power spectrum $S_{t-1}(\omega_k)$ from $\hat{z}_{t-1}$ for $k = 1, \dots, B$
  Compute corrected power spectrum:

$$\tilde{S}_{t-1}(\omega_k) = S_{t-1}(\omega_k) \left(\frac{T_{t-1}(\omega_k) + \varepsilon}{S_{t-1}(\omega_k) + \varepsilon} \right)^{\lambda}, \quad k = 1, \dots, B$$

  Compute corrected Fourier magnitude:

$$M_{t-1}(\omega_k) = \sqrt{\tilde{S}_{t-1}(\omega_k)}, \quad k = 1, \dots, B$$

  Set the Fourier magnitude to M_{t-1} while preserving the phase of $\hat{z}_{t-1}$:

$$|\hat{\tilde{z}}_{t-1}| = M_{t-1}, \quad \angle\hat{\tilde{z}}_{t-1} = \angle\hat{z}_{t-1}$$

  $z_{t-1} \leftarrow \mathcal{F}^{-1}(\hat{\tilde{z}}_{t-1})$
end for
return $\mathrm{Decode}(z_0)$

strength is set to $\lambda = 0.05$ for SDXL and PixArt-α, and $\lambda = 0.1$ for SD3.5. The relaxation factor is set to $\rho = 0.5$ for SDXL and SD3.5. For PixArt-α at the higher guidance scale, we use $\lambda = 0.1$ and $\rho = 0.5$, reflecting the need for stronger correction under stronger guidance. We estimate the target spectrum using the same inference-time sampling schedule as the one used for generation. For the DiT-XL/2 class-conditional experiments, we use 32 radial bins and set $\lambda = 0.01$ across all sampling budgets. We add a small constant ε to the numerator and denominator of the spectrum ratio for numerical stability.

To model the z_0 space of each model, we encode COCO val 50K images (Lin et al., 2014) into the corresponding latent space and fit the power-law curve based on Eq. 7. We note that this fitting step is not essential to the method. It is a practical choice for setting the reference spectrum in our experiments. Since the reference spectrum is parameterized by only three scalar values, the corresponding power-law parameters can also be tuned without using external data.

For the conceptual validation in Section 3, we use 1,000 image–caption pairs from COCO val 50K. We use SDXL with 50 denoising steps, resolution 1024×1024, and fp16 inference. For each sample, we encode the real image into the SDXL VAE latent space and measure the spectrum along the forward noising trajectory. We then run reverse denoising from random Gaussian noise using the corresponding caption as the prompt, and measure the spectrum of the generated trajectory. We evaluate CFG scales $\{2.5, 5.0, 7.5, 15.0\}$, together with unconditional and conditional baselines. To obtain the reported curve, we first average the spectra over the 1,000 samples for each trajectory type, and then compute the mean squared error between the log-power spectra. Thus, the comparison is performed between aggregate spectra rather than as a pairwise per-sample distance.

Our method is training-free and does not require additional model training. All reported experiments were conducted on NVIDIA RTX 3090 GPUs. For class-conditional ImageNet experiments, generating 50K samples was parallelized over 15 GPUs and took approximately 8 hours. For text-to-image experiments, generating 10K samples was parallelized over 5 GPUs and took approximately 12 hours.

F PROOF OF PROPOSITION 1

Proof. Starting from the forward process in Eq. 1,

$$z_t = \alpha_t z_0 + \sigma_t \epsilon,$$

we apply the Fourier transform. By linearity,

$$\hat{z}_t(\omega) = \alpha_t \hat{z}_0(\omega) + \sigma_t \hat{\epsilon}(\omega).$$

Since the power spectrum is defined as $P_t(\omega) = |\hat{z}_t(\omega)|^2$, we have

$$P_t(\omega) = \left|\alpha_t \hat{z}_0(\omega) + \sigma_t \hat{\epsilon}(\omega)\right|^2.$$

Expanding the squared magnitude gives

$$P_t(\omega) = \alpha_t^2 P_0(\omega) + \sigma_t^2 P_\epsilon(\omega) + \alpha_t \sigma_t C_t(\omega),$$

where $C_t(\omega)$ denotes the cross term between $\hat{z}_0(\omega)$ and $\hat{\epsilon}(\omega)$. Taking expectation on both sides yields

$$\mathbb{E}[P_t(\omega)] = \alpha_t^2 \mathbb{E}[P_0(\omega)] + \sigma_t^2 \mathbb{E}[P_\epsilon(\omega)] + \alpha_t \sigma_t \mathbb{E}[C_t(\omega)].$$

Since z_0 and ϵ are independent and $\mathbb{E}[\epsilon] = 0$, the cross term vanishes in expectation. Therefore,

$$\mathbb{E}[P_t(\omega)] = \alpha_t^2 \mathbb{E}[P_0(\omega)] + \sigma_t^2 \mathbb{E}[P_\epsilon(\omega)]. \tag{11}$$

Substituting the assumed form of the clean-image spectrum,

$$\mathbb{E}[P_0(\omega)] = C\omega^{-\beta} + c,$$

and the flat spectrum of white Gaussian noise,

$$\mathbb{E}[P_\epsilon(\omega)] = c_\epsilon,$$

into Eq. 11, we obtain

$$\mathbb{E}[P_t(\omega)] = \alpha_t^2 (C\omega^{-\beta} + c) + \sigma_t^2 c_\epsilon.$$

Expanding the terms gives

$$\mathbb{E}[P_t(\omega)] = \alpha_t^2 C\omega^{-\beta} + \alpha_t^2 c + \sigma_t^2 c_\epsilon.$$

Finally, defining

$$\lambda_t := \alpha_t^2 C, \qquad \gamma_t := \alpha_t^2 c + \sigma_t^2 c_\epsilon,$$

gives

$$\mathbb{E}[P_t(\omega)] = \lambda_t \omega^{-\beta} + \gamma_t.$$

This completes the proof. □

G Computational Cost

To better understand the computational overhead of the proposed correction, we compare the peak GPU memory usage and per-image generation time of our method with those of standard CFG over 100 images. As shown in Table 4, peak memory usage remains comparable. The proposed correction requires no backpropagation or additional denoiser evaluations. It adds approximately 1.7 seconds per image on SDXL and 2.6 seconds on SD3.5. The reduced-NFE results in Table 3 suggest that reducing the number of denoising steps could offset this overhead while maintaining generation quality.

Table 4: Computational cost of standard CFG and our method, measured over 100 images.

Model	Method	Peak VRAM (GB)	Time (s/image)
SDXL	Ours	12.70	15.442
	CFG	12.84	13.775
SD3.5	Ours	12.68	25.236
	CFG	12.67	22.628

H Human Evaluation

We conduct a human preference study comparing our method with standard CFG on SDXL. We randomly sample 50 prompts and ask 20 participants to compare the generated images in terms of visual quality and prompt alignment. For each prompt, we randomize the left-right placement of the two methods to reduce positional bias. As shown in Table 5, our method receives 595 out of 1,000 votes for visual quality and 588 out of 1,000 votes for prompt alignment, corresponding to win rates of 59.50% and 58.80%, respectively. Standard CFG already produces high-quality images, yet participants prefer the outputs of our method on both criteria. These gains come from sampling-time correction without additional training or changes to the underlying model.

Table 5: Human preference study comparing our method with CFG on SDXL. Each criterion receives 1,000 votes from 20 participants evaluating 50 prompts.

	Ours	CFG	Win Rate (Ours)
Visual Quality	**595**	405	**59.50%**
Prompt Alignment	**588**	412	**58.80%**

I Robustness Across Random Seeds

To assess robustness to random seed variation, we compare our method with CFG on SDXL across 10 seeds using 1K randomly sampled prompts per seed. We report the mean and standard deviation of each metric across seeds. As shown in Table 6, our method achieves higher mean HPSv3, ImageReward, PickScore, and HPSv2 scores and lower mean FID than CFG. CFG obtains slightly better mean CLIP-T and sFID scores. The gains in HPSv3, ImageReward, and HPSv2 are large relative to the observed standard deviations, supporting the robustness of these improvements across random seeds.

Table 6: Multi-seed comparison of our method and CFG on SDXL using 1K prompts per seed. Results are reported as mean $\pm$ standard deviation across 10 random seeds.

Method	HPSv3↑	ImageReward↑	PickScore↑	HPSv2↑	CLIP-T↑	FID↓	sFID↓
Ours	$\mathbf{8.375 \pm 0.045}$	$\mathbf{0.7368 \pm 0.0088}$	$\mathbf{22.463 \pm 0.021}$	$\mathbf{0.2800 \pm 0.0004}$	0.2651 ± 0.0004	$\mathbf{66.572 \pm 0.557}$	344.398 ± 0.434
CFG	8.068 ± 0.040	0.7020 ± 0.0137	22.449 ± 0.014	0.2750 ± 0.0003	$\mathbf{0.2660 \pm 0.0005}$	67.201 ± 0.504	$\mathbf{343.793 \pm 0.861}$

J Comparison with Output-Only Correction

To examine whether the gains can be explained by spectral correction of the final output alone, we compare our method with an output-only correction baseline on SDXL using 1K randomly

sampled prompts. This baseline applies spectral correction only after generation, treating it as a post-processing step. As shown in Table 7, output-only correction yields modest improvements over CFG in preference-based metrics. Our method achieves larger gains in HPSv3, ImageReward, PickScore, and HPSv2, together with lower FID and sFID. These results support the benefit of correcting spectral deviations during sampling beyond the improvements obtained by adjusting the final spectrum alone.

Table 7: Comparison with output-only spectral correction on SDXL using 1K randomly sampled prompts.

Method	HPSv3↑	ImageReward↑	PickScore↑	HPSv2↑	CLIP-T↑	FID↓	sFID↓
Ours	**8.437**	**0.738**	**22.499**	**0.280**	0.265	**66.818**	**345.093**
Output-only correction	8.218	0.710	22.482	0.276	**0.267**	67.557	345.246
CFG	8.104	0.708	22.473	0.275	**0.267**	67.430	345.259

K Non-Photorealistic and Stylized Generation

To evaluate our method on stylized generation, we append "anime illustration" or "oil painting" to each of 1K prompts and generate images using SDXL. As shown in Table 8, our method improves HPSv3, ImageReward, PickScore, and HPSv2 over standard CFG for both styles. CLIP-T improves for anime illustration and remains unchanged at the reported precision for oil painting. These findings are consistent with the motivation for correction relaxation discussed in Section 6.2. The correction constrains spectral deviations during earlier sampling stages and relaxes toward the end, reducing spectral constraints on the formation of style-specific details. The results support the applicability of the proposed correction beyond standard photorealistic generation.

Table 8: Comparison of our method and CFG on SDXL for stylized generation using 1K prompts per style.

Style	Method	HPSv3↑	ImageReward↑	PickScore↑	HPSv2↑	CLIP-T↑
Anime illustration	Ours	**8.065**	**1.045**	**23.357**	**0.317**	**0.337**
	CFG	7.995	1.011	23.331	0.314	0.336
Oil painting	Ours	**8.646**	**1.016**	**22.822**	**0.300**	**0.320**
	CFG	8.473	1.011	22.775	0.294	**0.320**

L Additional Experiments across Guidance Settings

Section 6.1 evaluates text-to-image generation under the standard and high guidance settings. We further examine whether the proposed correction remains effective when the guidance scale is reduced. Table 9 reports the results on SDXL at $w = 2.5$, which is half of the standard guidance scale used in our main experiments. Our method achieves the best performance on all metrics except sFID, improving preference-based scores and FID over standard CFG. In contrast, existing guidance correction methods often degrade performance compared to CFG in this low-guidance regime. This suggests that correction strategies designed mainly to suppress over-guidance artifacts do not effectively address settings where the guidance signal is weak, whereas our spectral correction remains beneficial across guidance settings.

Figure 9 shows qualitative examples under reduced guidance. Generation at this guidance scale is challenging, and the images are not always fully satisfactory across methods. Nevertheless, our method shows relatively improved prompt alignment and visual coherence. For example, it better reflects the racket held in one hand in the first prompt and the feeding interaction with the giraffe in the second prompt. These results further support the robustness of the proposed correction across different guidance scales.

Table 9: Quantitative comparison under reduced guidance on SDXL. At half of the standard guidance scale ($w = 2.5$), our method improves most metrics, while existing guidance correction methods provide limited benefit.

Setting	Method	ImageReward↑	PickScore↑	HPSv2↑	CLIP-T↑	FID↓	sFID↓
$w = 2.5$	Ours	**0.391**	**21.86**	**0.252**	**0.257**	**18.28**	52.82
	CFG	0.362	21.83	0.249	**0.257**	18.76	**52.26**
	CFG++	0.329	21.77	0.247	**0.257**	19.30	52.75
	TV-CFG	0.104	21.43	0.235	0.253	24.85	56.01
	LF-CFG	0.022	21.23	0.232	0.248	26.42	60.18

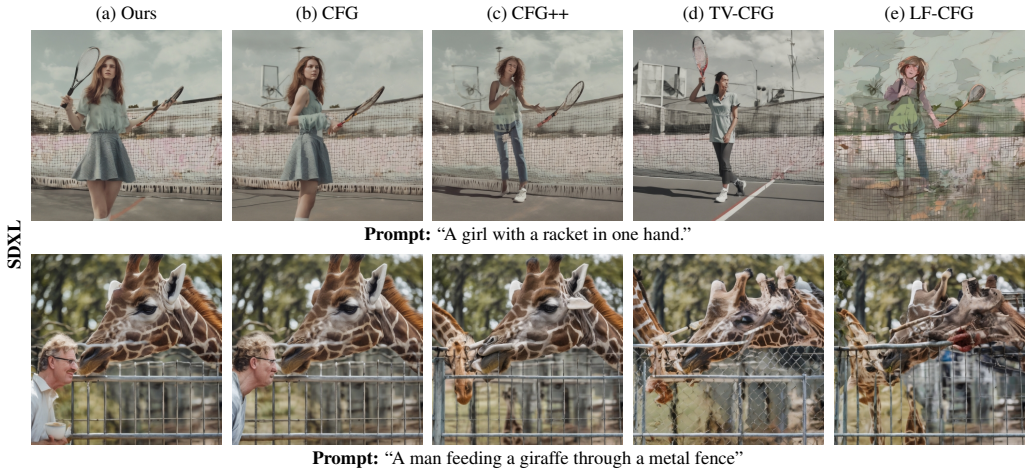

Figure 9: Qualitative comparison under reduced guidance on SDXL. At half of the standard guidance scale ($w = 2.5$), our method better preserves prompt-relevant content and visual coherence, while existing guidance correction methods provide limited qualitative improvement.

M Additional Experiments on Diverse Backbone Models

We further evaluate whether the proposed spectral correction can be applied beyond the DiT setting studied in Section 6.1. To this end, we conduct additional experiments on other class-conditional image generation backbones, including SiT and JiT.

SiT. We first evaluate our method on SiT-XL/2 (Ma et al., 2024). We use 32 radial bins for power spectrum estimation, with $\lambda = 0.005$ and $\rho = 0.5$. Figure 10 shows qualitative results. The results suggest that our method can improve generation quality also on SiT. Our method produces a more object-centric image for the European fire salamander class and reveals a clearer object shape for the Indian cobra class.

JiT. We also evaluate our method on JiT-L/16 (Li & He, 2025). We use 128 radial bins for power spectrum estimation, with $\lambda = 0.5$ and $\rho = 0.5$. Figure 11 presents the qualitative comparison. In the water ouzel example, our method produces a sharper image. In the iguana example, class-specific visual details are more visible. However, we observe that JiT exhibits larger sample-level variation than the other backbones. Consistent with the ablation results in Section 6.2, the effect of stronger correction is not uniform across samples, as it can introduce artifacts in some cases while producing only marginal changes in others. This behavior may be related to JiT operating directly in the raw-pixel space, where spectral statistics can exhibit greater sample-level variability. The observation suggests that, although our method is motivated by spectral assumptions in pixel space, compressed latent spaces may provide a more stable domain for applying the correction by reducing sample-level spectral variability.



Figure 10: Qualitative comparison on SiT for class-conditional generation. Compared to CFG, our method yields clearer structures aligned with the target class and better preserves class-relevant visual attributes.

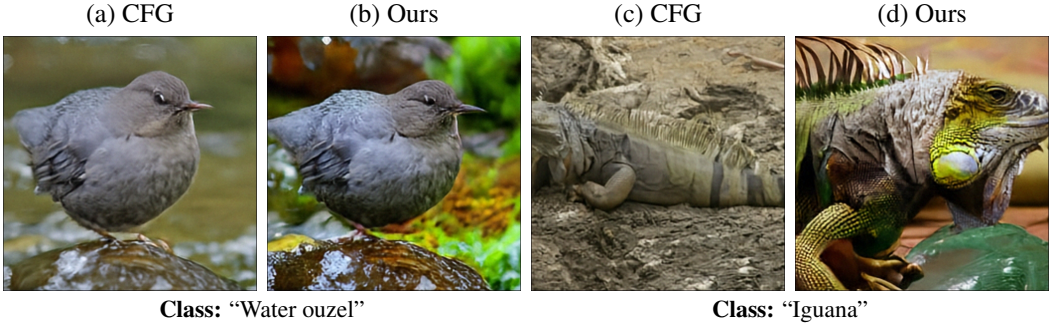

Figure 11: Qualitative comparison on JiT for class-conditional generation. Our method improves image sharpness and fine visual details, producing more recognizable samples than CFG.

N Additional Analysis of Power Spectrum Modeling

Our method extends a spectral principle from pixel space to latent space. To empirically validate this assumption, we examine whether the power-law behavior of natural image spectra is preserved after encoding images into the latent spaces of the models used in our experiments. Figure 12 shows the mean power spectrum of encoded COCO val 50K images (Lin et al., 2014) in the latent space of each model, along with the fitted curve based on Eq. 7. The fitted curves achieve $R^2 > 0.99$ across models, indicating that the z_0 spaces of the image generation models considered in our experiments can be well characterized by this power-law structure. This result provides empirical justification for applying the proposed spectral correction in these spaces.

We also analyze the effect of binning resolution when estimating the power spectrum for the proposed correction. The binning resolution determines how many frequency components are grouped into each radial bin. To make the effect of this design choice more visible, we vary the number of radial bins under a relatively strong correction strength of $\lambda = 0.5$. Figure 13 shows the qualitative results. Coarse binning, such as 8 radial bins, covers a broad range of frequencies within each bin and mainly produces global changes, such as changes in overall brightness. By contrast, finer binning, such as 128 radial bins, produces more noticeable changes in image structure. These results suggest that binning resolution is one of the design choices that can affect the behavior of spectral correction.

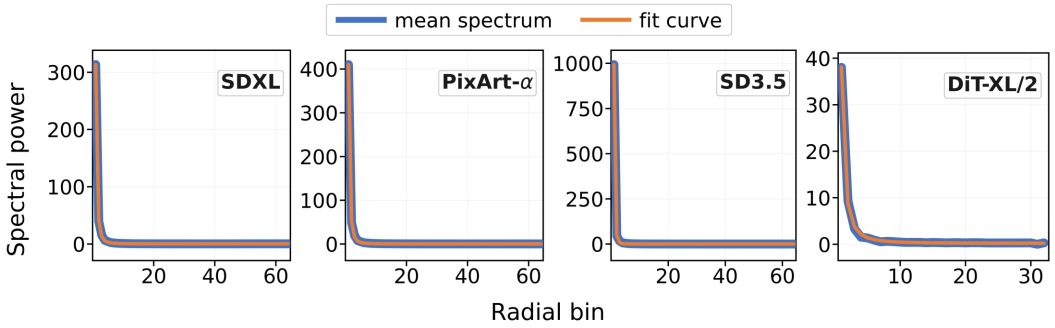

Figure 12: Power spectrum of z_0 from baseline diffusion models with fitted curves. Although the spectral scale varies across models, the spectra consistently follow a power-law statistical structure. The fit based on Eq. 7 achieves $R^2 > 0.99$ across models.



Figure 13: Ablation on the number of radial bins for power spectrum estimation on SDXL. We use a relatively strong correction strength of $\lambda = 0.5$ to better highlight the effect of binning.

O Additional Analysis of Correction Relaxation

We further analyze the effect of the relaxation factor ρ used in the proposed correction. Figure 14 shows the results on SDXL, where we vary ρ and evaluate PickScore, FID, and sFID. The value $\rho = 0.5$ provides the most balanced results across the three metrics. This provides quantitative support for the role of correction relaxation in balancing alignment and fidelity. We also observe that the benefits of relaxation are often more apparent in visual quality than in metrics, with corrected samples showing a more natural appearance and fewer visible artifacts.

P Additional Qualitative Results

Figures 15, 16, 17, 18, 19, and 20 show additional qualitative results on SDXL, PixArt-α, and SD3.5. Figures 21, 22, 23, 24, 25, and 26 show additional results under high guidance. Across these examples, our method produces more favorable outputs than the baselines. It mitigates artifacts observed in CFG, improves brightness and sharpness, and better preserves text-aligned visual content. Some outputs remain visually close to those of CFG. When intermediate spectra are already close to the analytic reference, our method applies only small spectral adjustments, helping preserve the

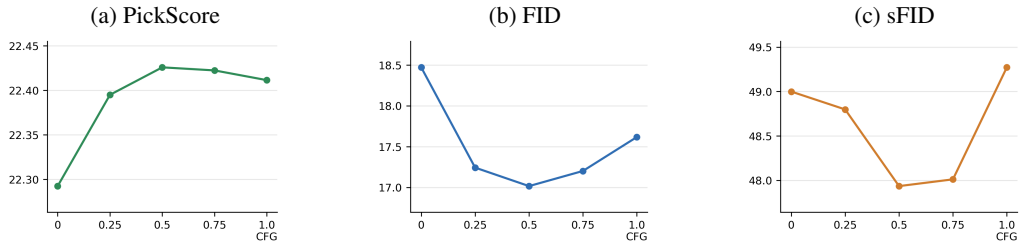

Figure 14: Effect of the relaxation factor ρ on PickScore, FID, and sFID. The x-axis is the relaxation factor and the y-axis is the metric value. Overall, $\rho = 0.5$ yields the most balanced performance across the three metrics.

quality already achieved by CFG. We recommend zooming in on the figures to examine fine-grained differences in visual quality and text alignment.

Analysis of Failure Cases. Despite the overall effectiveness of the proposed correction, the additional qualitative results also reveal two common failure patterns. These include cases where the correction only partially mitigates artifacts observed in CFG, and cases where improvements in visual quality or prompt alignment are accompanied by undesired changes in other aspects of the image. First, in the first example of Figure 16, our method produces a cat with a more natural composition, but the generated cat still exhibits an abnormal tail structure. Similarly, in the fifth example of Figure 18, the face of the person near the center is improved, while some details such as the hands remain unnatural. In these examples, our method retains much of the quality already achieved by CFG while refining some fine-grained details, though certain artifacts persist. In the last example of Figure 19, our method improves visual quality over CFG, but some artifacts remain. Second, some cases show a trade-off between different aspects of generation quality. In the last example of Figure 22, the overall color tone is improved, but the number of people increases, making the result less consistent with the text condition. In the last example of Figure 24, the luggage shape becomes more natural, but the interaction between objects is weakened during the correction. These examples suggest that, while the proposed method can improve generation quality in many cases, there remains room for further improving its robustness.

Q Additional Discussion

The proposed method improves guided generation by normalizing spectral deviations from the expected denoising dynamics. This correction keeps the guided trajectory closer to the spectral statistics induced by the forward process, which can improve image quality and prompt alignment. At the same time, the correction does not provide explicit feedback toward a desired semantic direction. As a result, its ability to control generation toward a specific target remains limited. More direct semantic control may require additional learning signals, such as external rewards or task-specific supervision, but the use of such signals introduces separate challenges. These observations suggest that more fundamental improvements may require aligning the denoising dynamics from the training stage. This direction calls for a deeper understanding of the representation space and generation dynamics formed by diffusion models.

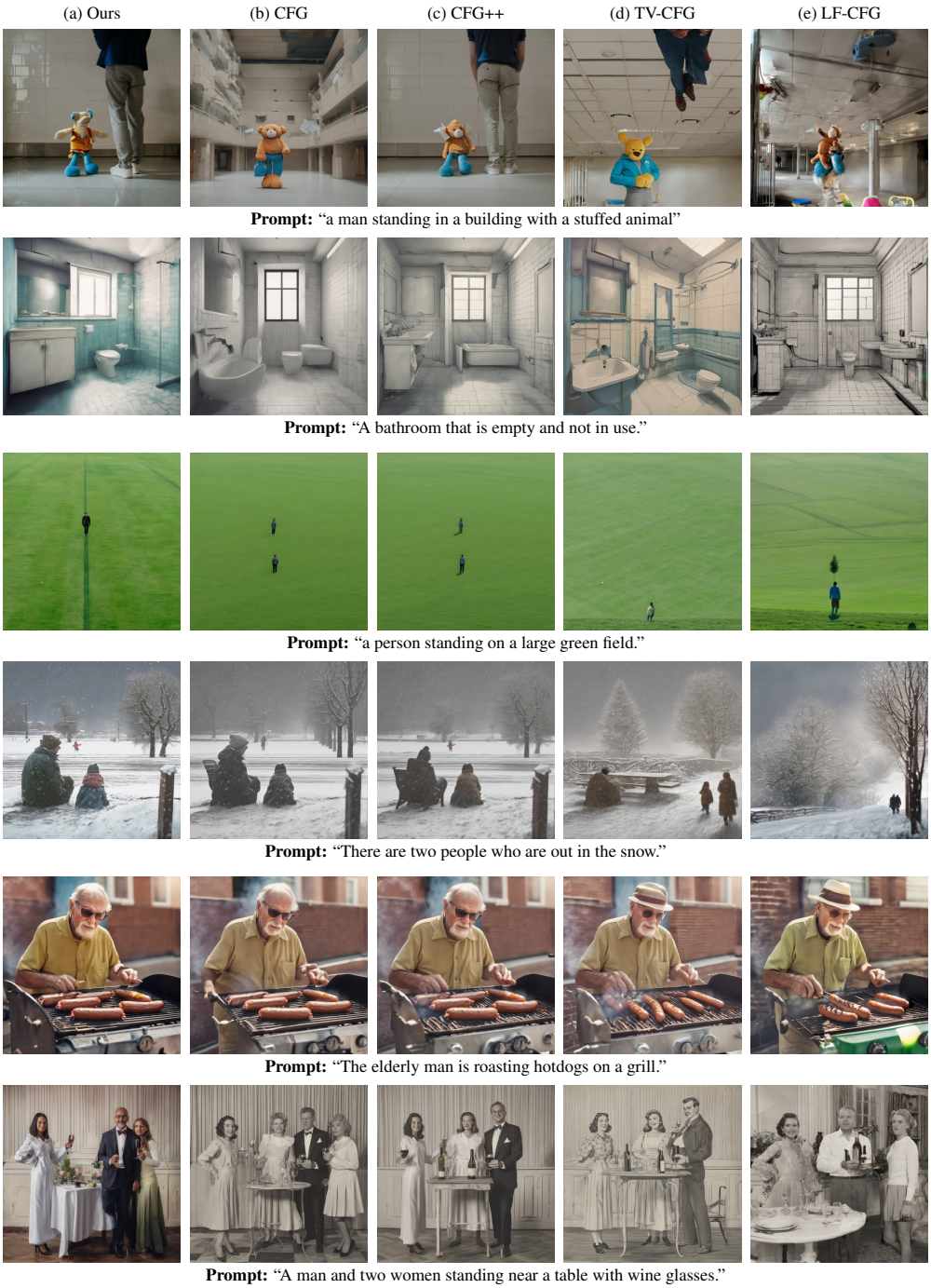

Figure 15: Additional qualitative results on SDXL.

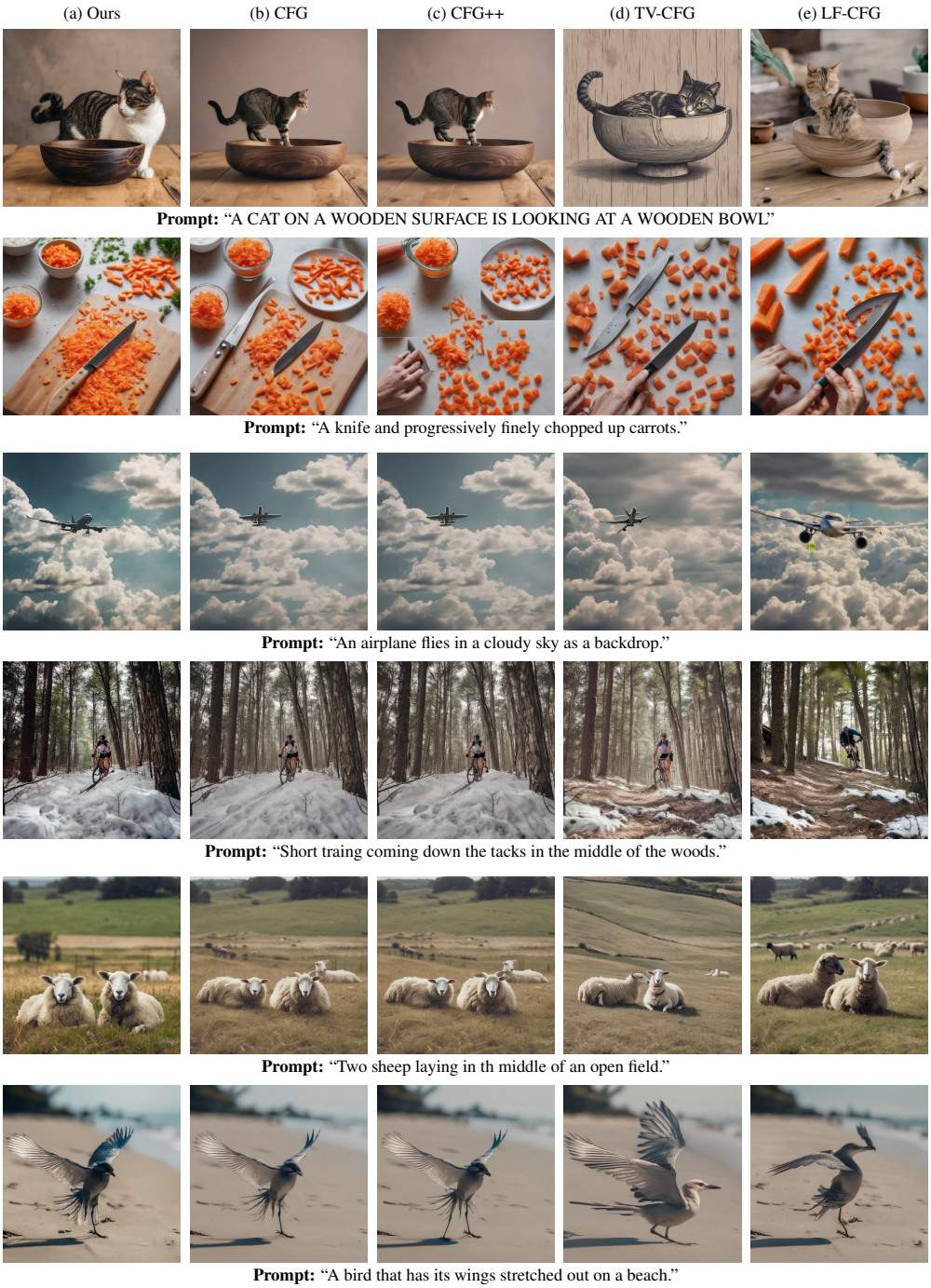

Figure 16: Additional qualitative results on SDXL.

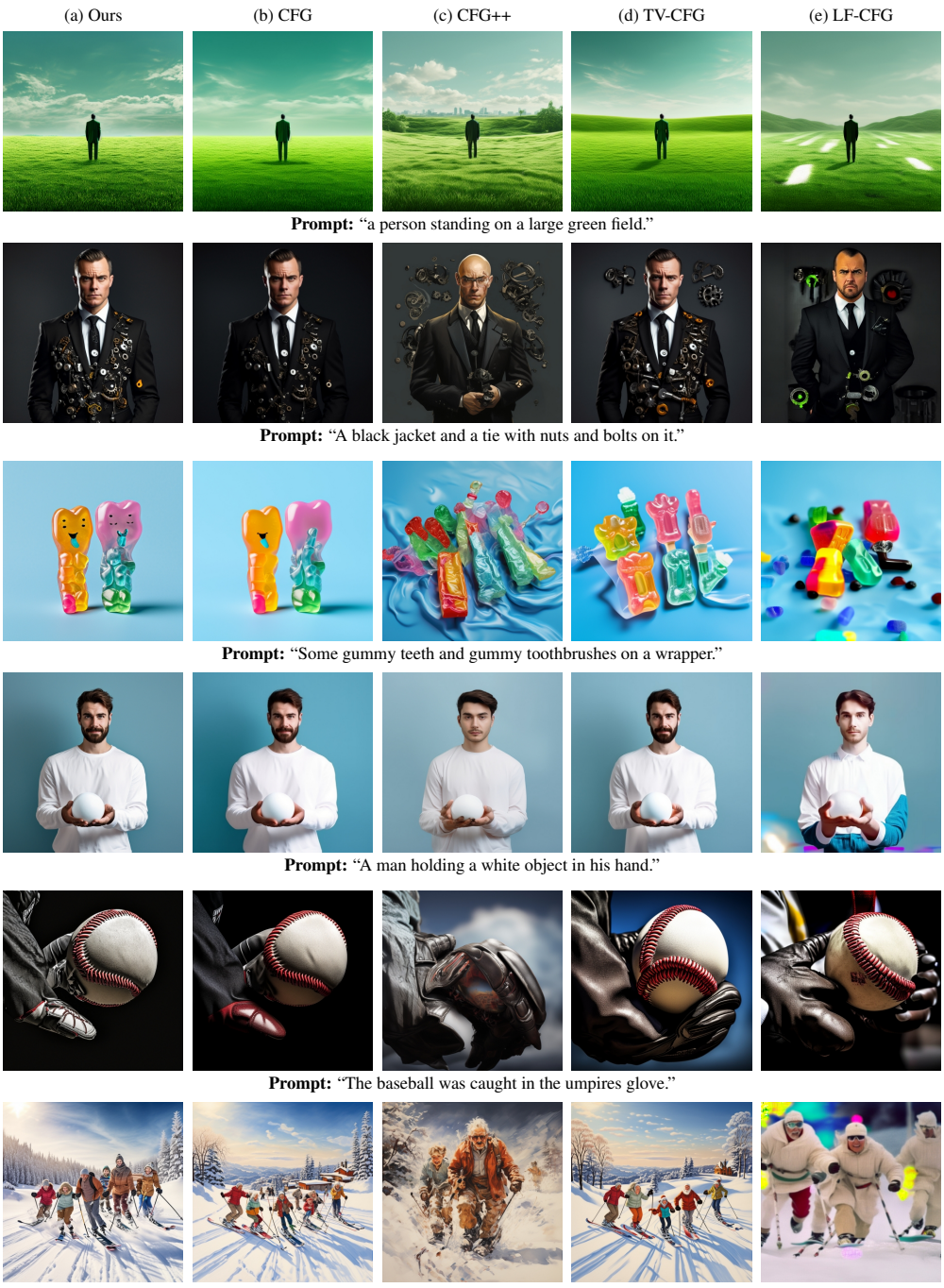

Figure 17: Additional qualitative results on PixArt-α.

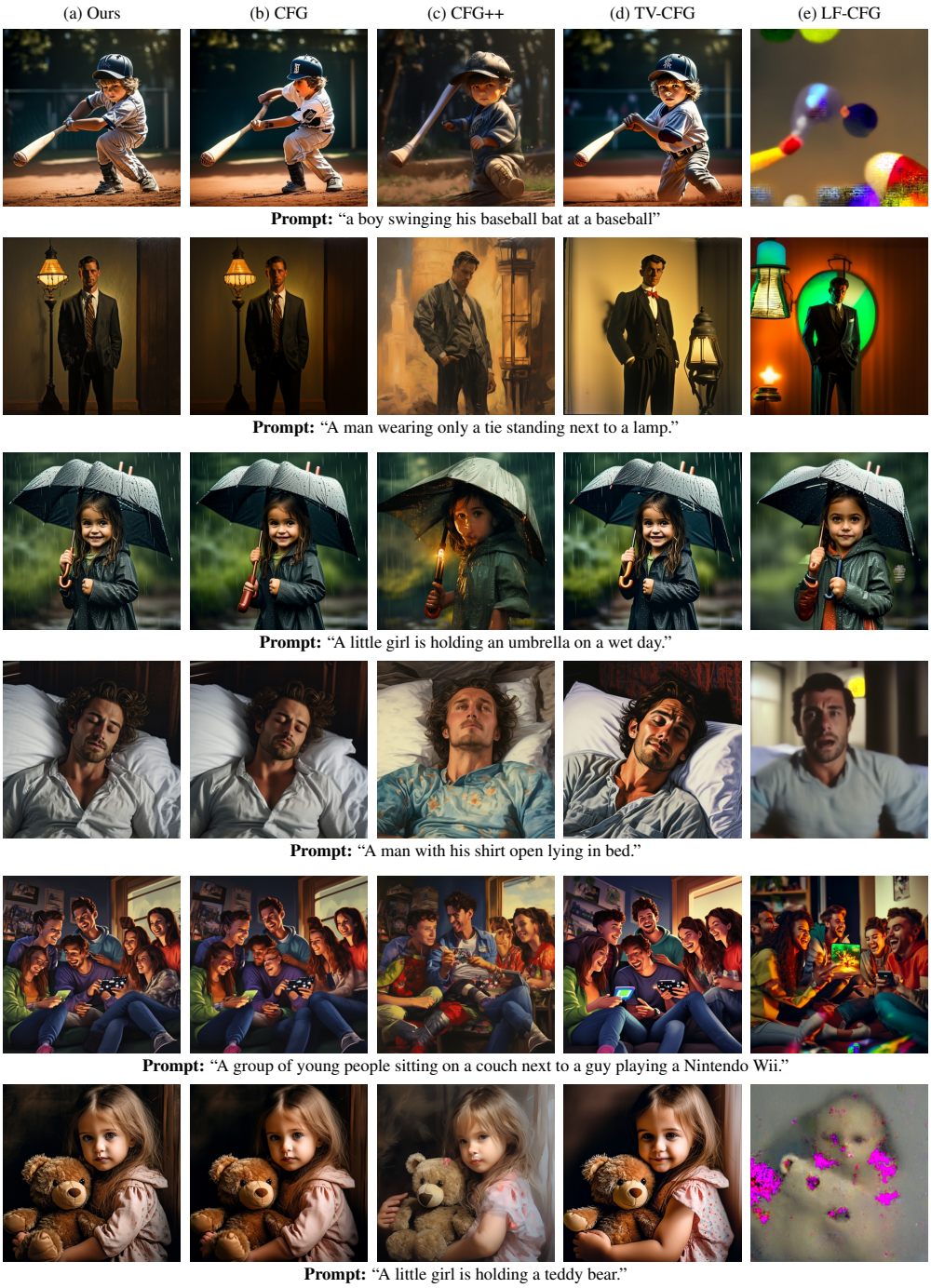

Figure 18: Additional qualitative results on PixArt-α.

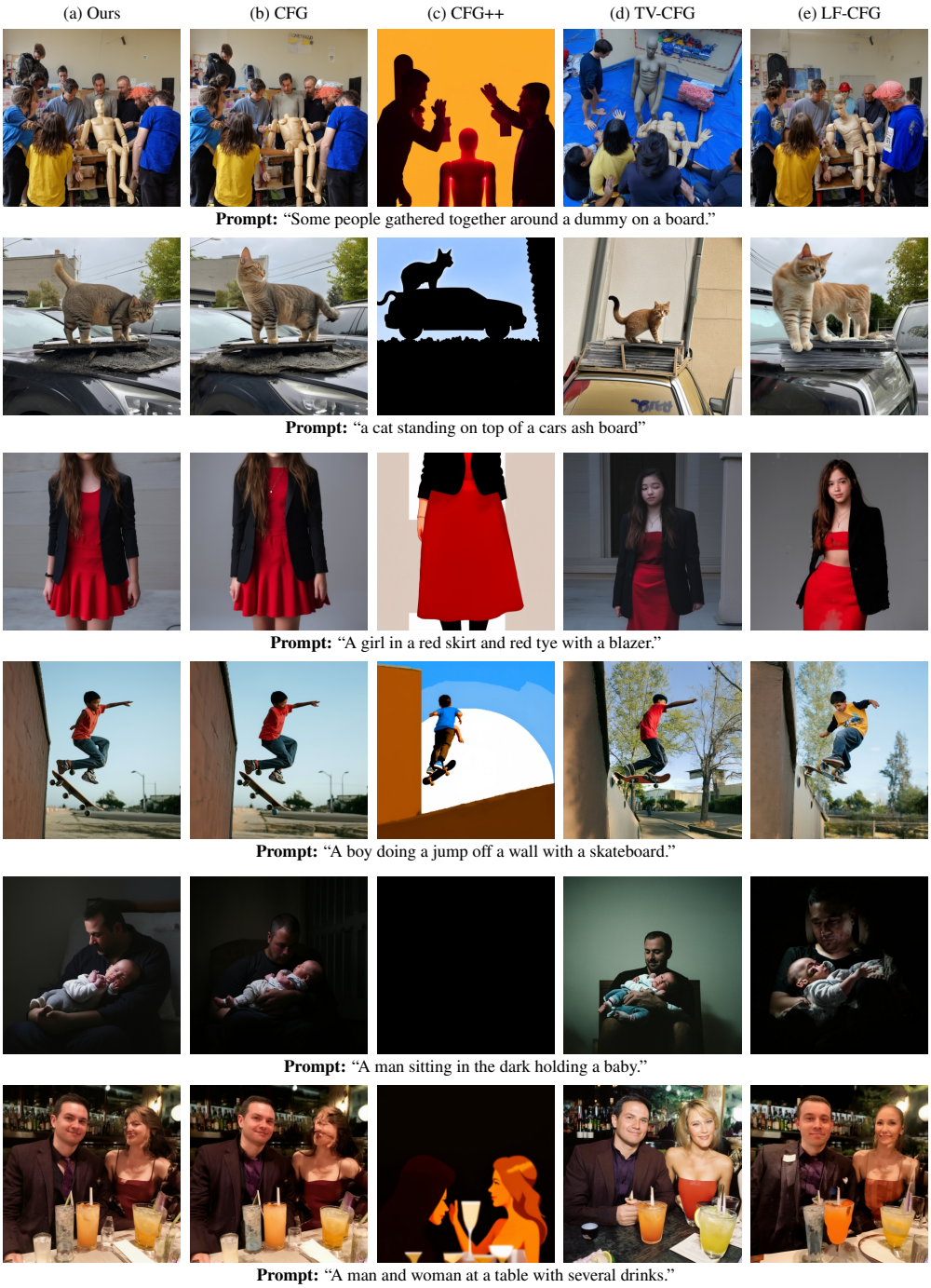

Figure 19: Additional qualitative results on SD3.5.

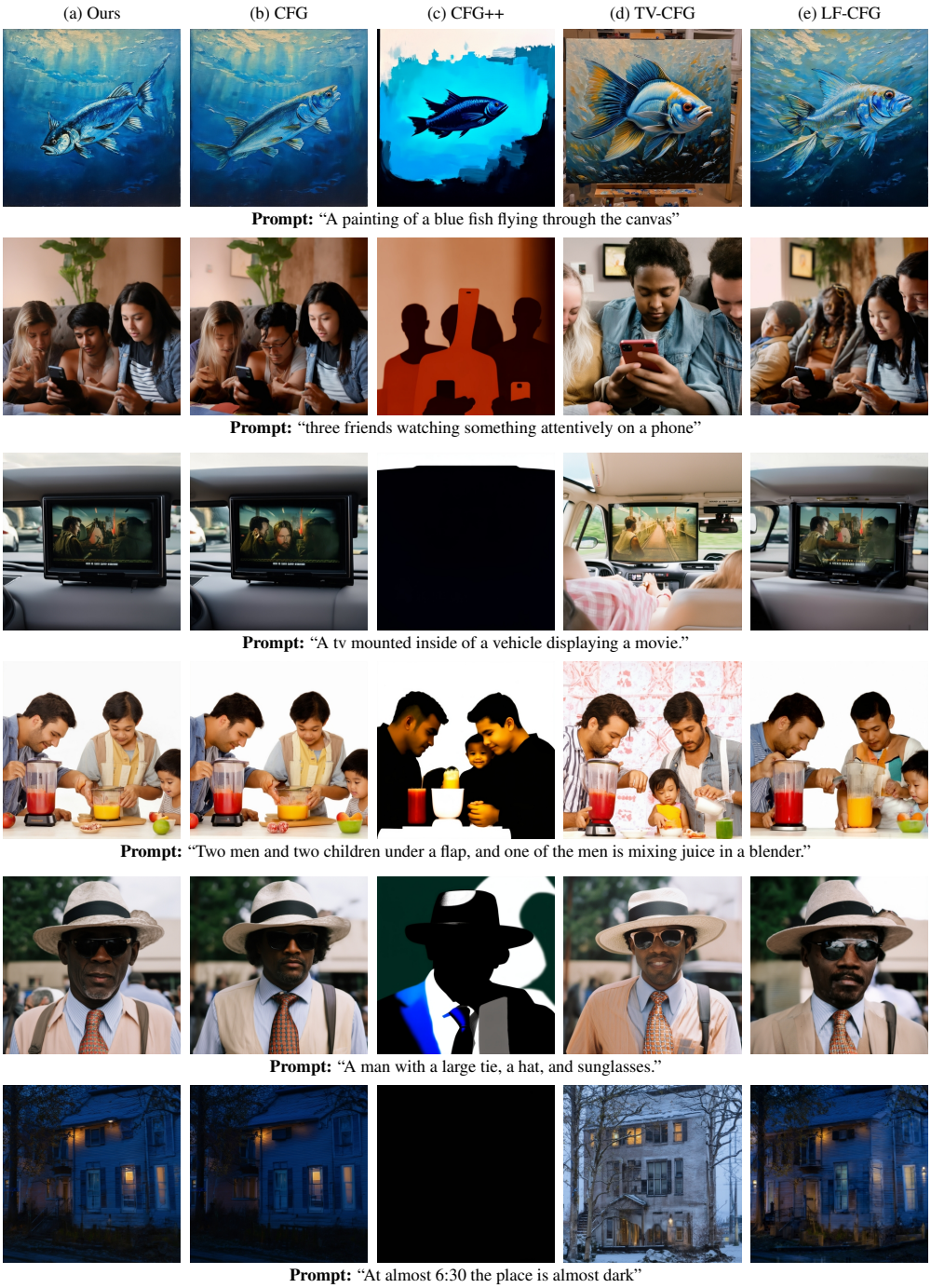

Figure 20: Additional qualitative results on SD3.5.

(a) Ours (b) CFG



Prompt: "An airplane parked at a airport terminal under gray skies."



Prompt: "A person sitting on a surfboard on a beach."



Prompt: "A young man skate boarding on a curbed edge in a parking lot."

Figure 21: Additional qualitative results under high guidance on SDXL.

(a) Ours (b) CFG



Prompt: "a soccer player kicks a ball directly at the goalkeeper."



Prompt: "A woman is typing on a laptop on a wooden table."



Prompt: "Four men with surf boards standing in the front of the beach side."

Figure 22: Additional qualitative results under high guidance on SDXL.

(a) Ours (b) CFG



Prompt: "A man smiles wearing a helmet and goggles."



Prompt: "A man putting on a polka dot tie in the bathroom."



Prompt: "A woman is throwing a frisbee on the beach."

Figure 23: Additional qualitative results under high guidance on PixArt-α.

(a) Ours (b) CFG



Prompt: "A red and black truck with a man riding on back of it."



Prompt: "Hot dog Philly cheese style next to very large order of fries."



Prompt: "Woman grabbing her luggage at an airport terminal."

Figure 24: Additional qualitative results under high guidance on PixArt-α.

(a) Ours (b) CFG



Prompt: "A cat is wearing a plastic baseball cap."



Prompt: "A woman on a bench plays an accordion as a man looks on."



Prompt: "two people sitting next to each other on folding chairs on a tennis court"

Figure 25: Additional qualitative results under high guidance on SD3.5.

(a) Ours (b) CFG



Prompt: "A couple is sitting at a table with food."



Prompt: "Pink roses in a little polka-dotted boot-shaped vase"



Prompt: "A woman in white shirt sitting at couch with a laptop."

Figure 26: Additional qualitative results under high guidance on SD3.5.