SELECTIVE LOW-FREQUENCY RECOVERY FOR DIFFUSION GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance (CFG) improves conditional alignment in diffusion models, but high guidance scales can cause oversaturation, artifacts, and reduced sample diversity. Frequency-decoupled guidance (FDG) mitigates these effects by weakening low-frequency guidance while retaining stronger high-frequency guidance. However, uniformly suppressing the low-frequency residual can also attenuate information important for global structure and conditional consistency. Building on frequency-decoupled and projected guidance, we propose selective recovery guidance (SRG), a training-free method that restores low-frequency guidance according to its direction and denoising stage. SRG decomposes the low-frequency guidance residual into components parallel and orthogonal to the conditional low-frequency prediction. The method restores only the orthogonal component to a moderate guidance level while keeping the parallel component unchanged. This design preserves the projection of the conditional low-frequency prediction. A normalized log-SNR window confines the restoration to intermediate denoising stages, where structural predictions are reliable and global layout remains adjustable. The proposed method maintains original high-frequency guidance and requires no additional denoiser evaluations. Across five pretrained models, SRG improves conditional fidelity, prompt alignment and learned preference scores for text-to-image generation. On ImageNet and COCO 2017, it maintains FID, Precision and Recall comparable to FDG, preserving distributional quality and sample diversity. Ablations support the use of directional and temporal constraints with a moderate recovery strength. Selective low-frequency recovery improves the trade-off between conditional alignment and generation quality. Code is available at `https://anonymous.4open.science/r/SRG-6E6D`.

1 INTRODUCTION

Diffusion models are widely used for conditional image generation, enabled by advances in denoising, latent-space modeling, and Transformer architectures (Ho et al., 2020; Song et al., 2021b; Rombach et al., 2022; Peebles & Xie, 2023). Classifier-free guidance (CFG) combines conditional and unconditional predictions to strengthen conditioning during sampling (Ho & Salimans, 2022). Larger guidance scales often improve conditional alignment and visual detail, but can also cause oversaturation, artifacts, and reduced diversity, whereas smaller scales mitigate these effects at the cost of weaker conditioning (Sadat et al., 2025b). This trade-off motivates selective guidance: strengthening corrections that improve conditional generation while limiting those that harm sample quality or distributional coverage.

Existing guidance methods modify the conditioning signal, guidance geometry, or sampling schedule. CADS perturbs the conditioning signal to improve diversity under strong guidance (Sadat et al., 2024), while CFG++ constrains the guided update from a manifold perspective (Chung et al., 2025). APG decomposes the guidance residual into parallel and orthogonal components and further applies rescaling and momentum (Sadat et al., 2025a). Guidance Interval (GI) restricts the full guidance update to a selected noise-level range (Kynkäänniemi et al., 2024). Frequency-decoupled guidance (FDG) assigns separate scales to low- and high-frequency bands, using weaker low-frequency guidance to mitigate oversaturation and diversity loss while retaining stronger high-frequency guidance (Sadat et al., 2025b). However, its single low-frequency scale attenuates all directions within that band, including potentially useful structural corrections. We therefore investigate which attenuated

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.

parts should be recovered and when an additional orthogonal low-frequency update should be applied, while keeping high-frequency guidance unchanged.

Increasing low-frequency guidance can improve coarse object layouts but can also reintroduce oversaturation and visual artifacts (Sadat et al., 2025b). Uniform scaling treats all residual directions equally and cannot distinguish useful structural corrections from excessive amplification. Relative to the conditional low-frequency prediction, a parallel update changes the guided prediction's projection onto this direction, whereas an orthogonal update leaves the projection unchanged at a fixed denoising state. This characteristic allows us to exclusively restore the orthogonal residual component. Our residual analysis shows different temporal profiles across frequency bands, consistent with the coarse-to-fine denoising progression reported in prior work (Yang et al., 2023; Falck et al., 2025). Low-frequency predictions are less reliable at the initial high-noise steps, while late-stage updates have limited effect on global composition (Kynkäänniemi et al., 2024). Accordingly, we restrict orthogonal low-frequency recovery to a selected early-to-intermediate interval. Unlike existing global temporal strategies that regulate full guidance signals, our temporal constraint acts only on the supplementary low-frequency recovery term, leaving the original FDG low- and high-frequency guidance intact throughout the entire sampling process.

We propose selective recovery guidance (SRG), a training-free FDG extension implementing directional and temporal selection via additive correction. At each active sampling step, SRG adds the scaled orthogonal component of low-frequency guidance residual relative to conditional low-frequency prediction, preserving original parallel low-frequency and high-frequency guidance. Recovery is enabled within a fixed normalized log-SNR interval with strength bounded by high-frequency guidance scale. SRG reuses CFG-computed conditional and unconditional predictions and needs no extra training or denoiser forward passes. We evaluate SRG comprehensively on class-conditional and text-to-image generation for both distribution quality and prompt alignment. Experiments yield consistent gains in prompt alignment and learned-preference scores with competitive distributional fidelity and recall, achieving flexible fidelity-coverage trade-offs across generative backbones. Ablations verify the efficacy of our selective low-frequency recovery for conditional alignment.

The main contributions of this work are as follows: (i) we reformulate low-frequency guidance refinement as a selective-recovery paradigm rather than further suppression. We propose a direction-constrained rule that restores only the orthogonal component of the attenuated low-frequency residual, while preserving the original parallel low-frequency and high-frequency guidance; (ii) we bound the recovery strength by the frequency-scale gap and activate it within a normalized log-SNR interval; and (iii) we evaluate SRG across five pretrained models and both class-conditional and text-to-image tasks, showing improved average prompt alignment and learned preference scores with distributional metrics generally comparable to the frequency-decoupled baseline.

2 Related Work

Diffusion models learn data distributions by reversing a forward process that progressively corrupts samples with Gaussian noise (Ho et al., 2020; Song et al., 2021b). Latent-space formulations, Transformer architectures, and efficient samplers have improved diffusion modeling (Rombach et al., 2022; Peebles & Xie, 2023; Song et al., 2021a; Nichol & Dhariwal, 2021; Lu et al., 2022; 2025; Salimans & Ho, 2022). Guidance remains an important tool for improving sample fidelity and conditional alignment (Nichol et al., 2022).

Classifier guidance (CG) steers the reverse process using gradients from an auxiliary classifier (Dhariwal & Nichol, 2021), whereas classifier-free guidance (CFG) combines conditional and unconditional predictions without requiring an external classifier (Ho & Salimans, 2022; Bradley & Nakkiran, 2025). Although effective, large CFG scales can reduce diversity and cause oversaturation and other visual artifacts. CADS perturbs the conditioning signal to improve diversity under strong guidance (Sadat et al., 2024), while CFG++ constrains the guided update from a manifold perspective (Chung et al., 2025). APG decomposes the guidance residual into parallel and orthogonal components and further applies rescaling and momentum (Sadat et al., 2025a).

Recent work has examined the role of frequency components in CFG. FDG separates the guidance residual into low- and high-frequency components, using a lower guidance scale for the low-

frequency component. Its experiments associate stronger low-frequency guidance with reduced diversity and oversaturation, whereas high-frequency guidance improves visual fidelity (Sadat et al., 2025b). FDG keeps these band-wise guidance scales fixed throughout the sampling trajectory.

Guidance is not equally useful at all denoising stages (Papalampidi et al., 2025). Guidance Interval reports that guidance can be harmful at very high noise levels, is most useful over an intermediate interval, and is largely unnecessary near the end of sampling (Kynkäänniemi et al., 2024). The denoising trajectory also has a frequency-dependent progression. Spectral analysis shows that diffusion models recover different frequency components at different timesteps (Yang et al., 2023), while spectral analyses identify an implicit coarse-to-fine bias in diffusion generation (Rissanen et al., 2023). A spectral analysis of the forward process attributes this ordering to the faster SNR degradation of high-frequency components, which leads to low-frequency structure being recovered before high-frequency detail in the reverse process (Falck et al., 2025). These observations suggest that frequency-specific guidance need not be applied uniformly over the entire denoising trajectory.

FDG also demonstrates compatibility with APG through an extension in its appendix, where orthogonal projection is incorporated into the frequency-decoupled framework(Sadat et al., 2025b). However, this combination continues to operate under FDG's deliberately low low-frequency guidance scale, with APG regulating an already attenuated update; consequently, useful low-frequency conditional information remains suppressed. SRG instead introduces an explicit recovery mechanism that selectively restores this information through an orthogonal low-frequency correction, while constraining its direction, magnitude, and denoising stage to avoid uniformly strengthening low-frequency guidance.

3 BACKGROUND

Diffusion models Let $\mathbf{x} \sim p_{\mathrm{data}}(\mathbf{x})$ be a data sample and $t \in [0,1]$ a continuous time variable. Following the continuous-time formulation of diffusion models (Karras et al., 2022), the forward process is

$$\mathbf{z}_t = \mathbf{x} + \sigma(t)\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $\sigma(t)$ is a time-dependent noise schedule satisfying $\sigma(0) = 0$ and $\sigma(1) = \sigma_{\max}$ (Lin et al., 2024). The corresponding probability-flow ODE is

$$\mathrm{d}\mathbf{z}_t = -\dot{\sigma}(t)\sigma(t)\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)\,\mathrm{d}t, \tag{2}$$

where $p_t(\mathbf{z}_t)$ is the distribution of noisy samples at time t. Sampling integrates this ODE backward from a high-noise state. Since the score $\nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t)$ is unknown, it is approximated by a denoiser $D_\theta(\mathbf{z}_t, t)$ trained to recover the clean sample. For conditional generation, the denoiser additionally depends on a condition $\mathbf{y}$, such as a text prompt, yielding $D_\theta(\mathbf{z}_t, t, \mathbf{y})$.

Classifier-free guidance Classifier-free guidance (CFG) combines conditional and unconditional denoiser predictions (Ho & Salimans, 2022). Let $\mathbf{y}_{\mathrm{null}} = \varnothing$ denote the null condition. CFG computes

$$\widehat{D}_{\mathrm{CFG}}(\mathbf{z}_t, t, \mathbf{y}) = D_\theta(\mathbf{z}_t, t, \mathbf{y}_{\mathrm{null}}) + w\left[D_\theta(\mathbf{z}_t, t, \mathbf{y}) - D_\theta(\mathbf{z}_t, t, \mathbf{y}_{\mathrm{null}})\right], \tag{3}$$

where w denotes the guidance scale. When $w = 1$, the guided prediction reduces to the original conditional prediction, while $w > 1$ further amplifies the effect of the condition. For notational simplicity, we denote the conditional and unconditional predictions by $D_{c,t} = D_\theta(\mathbf{z}_t, t, \mathbf{y})$ and $D_{u,t} = D_\theta(\mathbf{z}_t, t, \mathbf{y}_{\mathrm{null}})$, respectively, and define their difference as $\Delta D_t = D_{c,t} - D_{u,t}$, which we refer to as the CFG guidance residual. With this notation, CFG can be written compactly as $\widehat{D}_{\mathrm{CFG},t} = D_{u,t} + w\Delta D_t$.

Frequency decompositions Multi-level frequency decompositions, such as Laplacian pyramids (Burt & Adelson, 1983) and wavelet transforms (Mallat, 1989), are commonly used to separate an image into different frequency bands. Low-frequency components capture global structure, object placement, geometry, and color distribution, whereas high-frequency components describe edges, textures, and fine details (Phung et al., 2023; Huang et al., 2024).

Frequency-decoupled guidance Frequency-decoupled guidance (FDG) (Sadat et al., 2025b) extends CFG by applying separate guidance scales to low- and high-frequency components. Let $\psi[\cdot]$ denote a linear and invertible frequency transformation, with $\psi_{\mathrm{low}}[\cdot]$ and $\psi_{\mathrm{high}}[\cdot]$ denoting the corresponding components. Since ψ is invertible, $\widehat{D}_{\mathrm{CFG},t} = \psi^{-1}\Big(\psi[\widehat{D}_{\mathrm{CFG},t}]\Big)$.

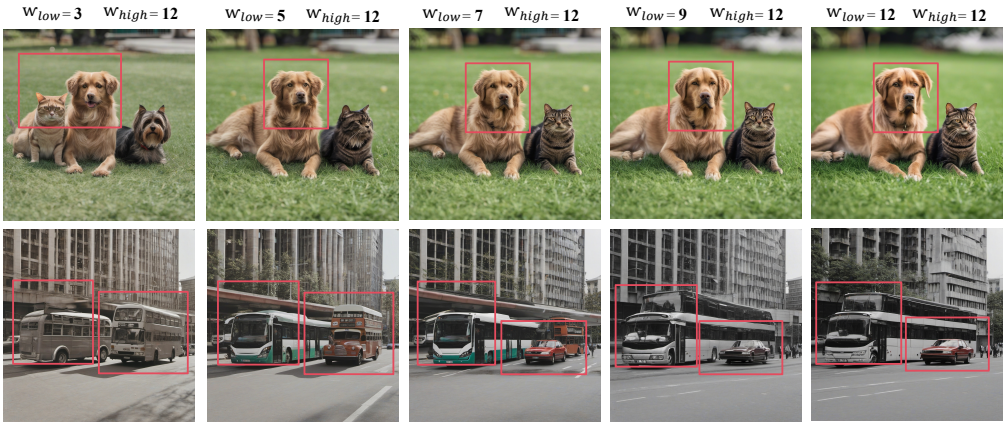

Figure 1: Illustration of gradually increasing low-frequency guidance under fixed high-frequency guidance. Red boxes highlight regions to inspect distortions, oversaturation and changes. Low-frequency recovery improves global structure yet may introduce oversaturation and visual artifacts.



Figure 2: Directional analysis of low-frequency recovery. From left to right: FDG, full low-frequency recovery, parallel-only recovery, and Ours with orthogonal recovery. Orthogonal recovery preserves structure without the amplification observed with full or parallel recovery.

FDG applies separate guidance scales to the two frequency components, yielding

$$\begin{aligned} \widehat{D}^{\text{low}}_{\text{FDG},t} &= \psi_{\text{low}}[D_{u,t}] + w_{\text{low}}\left(\psi_{\text{low}}[D_{c,t}] - \psi_{\text{low}}[D_{u,t}]\right), \\ \widehat{D}^{\text{high}}_{\text{FDG},t} &= \psi_{\text{high}}[D_{u,t}] + w_{\text{high}}\left(\psi_{\text{high}}[D_{c,t}] - \psi_{\text{high}}[D_{u,t}]\right). \end{aligned} \tag{4}$$

The two components are recombined with ψ^{-1} to obtain the final guided prediction.

4 Motivation and Method

4.1 Motivation

Although FDG uses separate scales for low- and high-frequency guidance, it treats the entire low-frequency band uniformly. Since low-frequency components encode critical global semantics such as object layout and coarse geometry, suppressing low-frequency guidance may discard useful conditional information. Increasing low-frequency guidance strength with high-frequency guidance fixed, we observe richer global structures and spatial relationships relative to vanilla FDG sampling (Figure 1). Yet excessive low-frequency recovery produces color oversaturation and implausible artifacts, consistent with FDG's prior observations (Sadat et al., 2025b). This dual property — beneficial structural cues alongside undesired amplification effects — motivates our exploration of selective low-frequency recovery.

Directional low-frequency recovery The preceding analysis shows that the low-frequency guidance attenuated by FDG contains both useful structural information and undesirable amplification effects. We therefore examine how this residual should be selected for recovery. At sampling step

t, FDG uses the low-frequency residual $\Delta D_t^{\text{low}} = \psi_{\text{low}}[D_{c,t}] - \psi_{\text{low}}[D_{u,t}]$. Let $C_{L,t} = \psi_{\text{low}}[D_{c,t}]$ denote the corresponding conditional low-frequency prediction, and let $\Delta D_t^{\text{low},\parallel}$ and $\Delta D_t^{\text{low},\perp}$ denote the components of the residual parallel and orthogonal to $C_{L,t}$, respectively. For a recovery coefficient $A_t = Ag(u_t)$,

$$\begin{aligned} \left\langle C_{L,t} + A_t \Delta D_t^{\text{low},\parallel}, C_{L,t} \right\rangle &= \|C_{L,t}\|_2^2 + A_t \left\langle \Delta D_t^{\text{low},\parallel}, C_{L,t} \right\rangle, \\ \left\langle C_{L,t} + A_t \Delta D_t^{\text{low},\perp}, C_{L,t} \right\rangle &= \|C_{L,t}\|_2^2, \end{aligned} \tag{5}$$

where $\left\langle \Delta D_t^{\text{low},\perp}, C_{L,t} \right\rangle = 0$. Thus, parallel recovery changes the projection onto the conditional low-frequency prediction, whereas orthogonal recovery leaves this projection unchanged while modifying complementary directions. This provides a geometric criterion for selecting the part of the attenuated FDG residual to recover. As shown in Figure 2, restoring the complete low-frequency residual improves structural consistency but also introduces strong oversaturation. Recovering only the parallel component produces a similar amplification effect in this example, whereas recovering only the orthogonal component alleviates oversaturation while preserving the structural benefits of low-frequency guidance.

Stage-dependent frequency guidance dynamics Diffusion sampling follows a coarse-to-fine trajectory. Early steps establish global structure and semantic layout, whereas later steps refine local appearance and texture (Jin et al., 2025). Low- and high-frequency guidance may therefore have different temporal profiles. We measure the relative contribution of each frequency band at every sampling step. For $f \in \{\text{low}, \text{high}\}$, define

$$r_f(u_t) = \frac{\left\| \Delta D_t^f \right\|_2}{\left\| \Delta D_t^{\text{low}} \right\|_2 + \left\| \Delta D_t^{\text{high}} \right\|_2}, \tag{6}$$

where ΔD_t^f denotes the guidance residual in frequency band f. Thus, $r_f(u_t)$ is the per-step share of the residual norm contributed by frequency band f at normalized log-SNR u_t. As shown in Appendix Figure 6, low-frequency guidance dominates during global-structure formation, while high-frequency guidance becomes increasingly critical in the late refinement stage. Accordingly, low-frequency recovery should match the effective temporal range of low-frequency residuals. Performing recovery at the initial high-noise stage may amplify noise and propagate visual artifacts, yet excluding these initial steps does not require discarding the entire early denoising phase. Delaying recovery strictly to the middle stage would discard valuable structural corrections. Meanwhile, recovery near the final low-noise stage yields negligible improvements for global layout. These observations motivate our stage-selective recovery window.

4.2 Selective Recovery Guidance

We retain FDG as the base guidance and add only the orthogonal component of the low-frequency residual within a selected early-to-intermediate denoising interval. The high-frequency guidance remains unchanged.

Orthogonal directional decomposition Using the quantities defined above, we compute the projection coefficient of the low-frequency residual onto the conditional low-frequency prediction:

$$\alpha_t = \begin{cases} \dfrac{\left\langle \Delta D_t^{\text{low}}, C_{L,t} \right\rangle}{\|C_{L,t}\|_2^2}, & \|C_{L,t}\|_2 > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

The residual components are $\Delta D_t^{\text{low},\parallel} = \alpha_t C_{L,t}$ and $\Delta D_t^{\text{low},\perp} = \Delta D_t^{\text{low}} - \Delta D_t^{\text{low},\parallel}$. By construction, $\langle \Delta D_t^{\text{low},\perp}, C_{L,t} \rangle = 0$. Only the orthogonal component is used for additional recovery; the parallel component remains part of the original FDG update.

Temporal activation Let $\ell_t = \log \text{SNR}_t$ and let u_t be its normalized value over the sampling schedule:

$$u_t = \frac{\ell_t - \ell_{\min}}{\ell_{\max} - \ell_{\min}}. \tag{8}$$

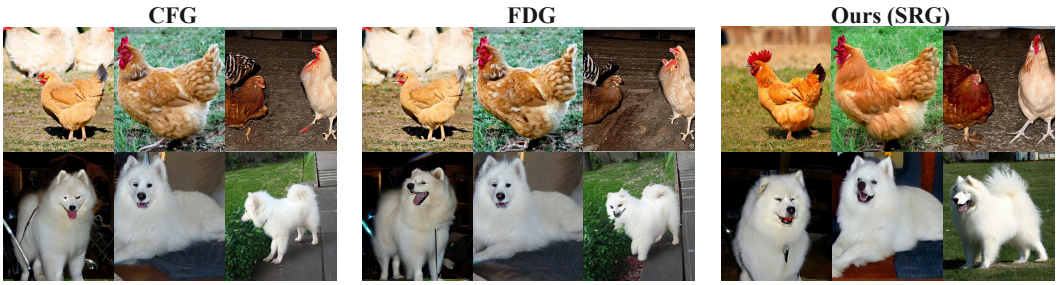

Figure 3: Class-conditional results on ImageNet with EDM2-S. CFG uses $w = 3$, while FDG and SRG use $(w_{\text{low}}, w_{\text{high}}) = (1, 3)$. Our method improves global structure while preserving diversity.

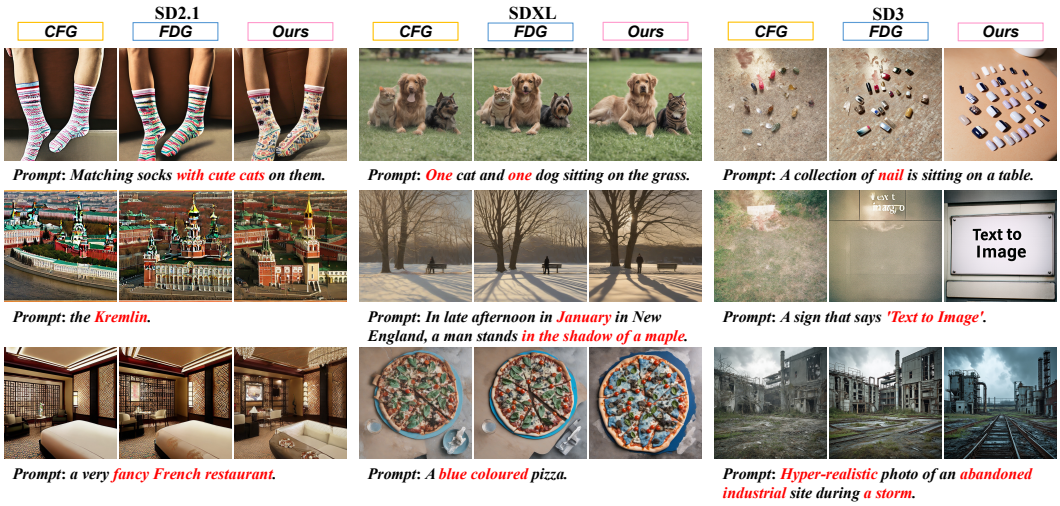

Figure 4: Qualitative comparison on DrawBench using Stable Diffusion 2.1, Stable Diffusion XL, and Stable Diffusion 3. For each model, the columns show outputs from CFG, FDG and SRG, from left to right, for the same prompts.

Recovery is activated with $g(u_t) = \mathbf{1}(u_t \in \mathcal{W})$, where $\mathcal{W}$ is a fixed early-to-intermediate interval. Let A denote the recovery strength, with $0 \le A \le w_{\text{high}} - w_{\text{low}}$.

Guided prediction The frequency-domain SRG predictions are

$$\begin{aligned} \widehat{D}^{\text{low}}_{\text{SRG},t} &= \psi_{\text{low}}[D_{u,t}] + w_{\text{low}} \Delta D^{\text{low}}_t + A\, g(u_t) \Delta D^{\text{low},\perp}_t, \\ \widehat{D}^{\text{high}}_{\text{SRG},t} &= \psi_{\text{high}}[D_{u,t}] + w_{\text{high}} \Delta D^{\text{high}}_t. \end{aligned} \tag{9}$$

The two components are recombined through the inverse frequency transform:

$$\widehat{D}_{\text{SRG},t} = \psi^{-1}\left[\widehat{D}^{\text{low}}_{\text{SRG},t}, \widehat{D}^{\text{high}}_{\text{SRG},t}\right]. \tag{10}$$

When $g(u_t) = 0$, the update reduces to FDG. Otherwise, the method adds only the orthogonal low-frequency component while preserving the original high-frequency guidance.

5 Experiments

Experimental Setup We evaluate our method on class-conditional and text-to-image generation tasks. For class-conditional generation, we evaluate EDM2-S (Karras et al., 2024) and DiT-XL/2 (Peebles & Xie, 2023) on ImageNet (Russakovsky et al., 2015; Deng et al., 2009). For text-to-image generation, we use Stable Diffusion 2.1 (Rombach et al., 2022), Stable Diffusion XL (Podell et al., 2024), and Stable Diffusion 3 (Esser et al., 2024). The COCO 2017 (Lin et al., 2014) is used

Table 1: Distributional comparison of CFG, APG, FDG, and SRG on ImageNet and COCO 2017.

Dataset	Model	Method	FID ↓	Precision ↑	Recall ↑
ImageNet	EDM2-S (Karras et al., 2024)	CFG	11.67	0.843	0.518
		APG	7.64	0.845	0.636
		FDG	**7.51**	0.808	**0.662**
		SRG(Ours)	7.61	**0.881**	0.651
	DiT-XL/2 (Peebles & Xie, 2023)	CFG	12.58	0.783	0.476
		APG	10.22	0.802	**0.601**
		FDG	9.96	0.843	0.541
		SRG(Ours)	**9.62**	**0.874**	0.575
COCO 2017	SD2.1 (Rombach et al., 2022)	CFG	23.38	**0.667**	0.473
		APG	**21.72**	0.663	0.493
		FDG	21.96	0.664	**0.512**
		SRG(Ours)	21.96	**0.667**	0.511
	SDXL (Podell et al., 2024)	CFG	25.97	0.624	0.489
		APG	**25.55**	0.624	0.486
		FDG	25.68	0.621	0.491
		SRG(Ours)	25.68	**0.627**	**0.500**
	SD3 (Esser et al., 2024)	CFG	30.93	0.757	0.384
		APG	28.72	0.774	0.388
		FDG	**28.10**	0.779	**0.400**
		SRG(Ours)	28.31	**0.784**	0.398

for distributional evaluation, while prompt-conditioned generation is evaluated on DrawBench (Saharia et al., 2022), Parti Prompts (Yu et al., 2022), and HPS Prompts (Wu et al., 2023). All methods use the same pretrained checkpoint and sampling configuration within each model and task. We compare against CFG (Ho & Salimans, 2022), APG (Sadat et al., 2025a) and FDG (Sadat et al., 2025b). FDG is the main baseline, while APG provides an additional geometry-based reference. For APG, we use the corresponding CFG guidance scale together with its model-specific parameter settings. The SRG recovery strength is $A = \eta(w_{\text{high}} - w_{\text{low}})$ with $\eta = \frac{2}{3}$, and recovery is activated within $0.15 < u_t < 0.70$. Detailed settings and sampling configurations are provided in Appendix B.

Evaluation Metrics Our metrics assess whether recovery introduces saturation or diversity loss and whether it improves generation quality and conditional alignment. For class-conditional generation, we report FID (Heusel et al., 2017), precision, and recall (Kynkäänniemi et al., 2019; Sajjadi et al., 2018). FID measures the distributional gap, Precision reflects sample fidelity, and Recall measures distribution coverage. For text-to-image generation, we report ImageReward (Xu et al., 2023), HPSv2 (Wu et al., 2023), PickScore (Kirstain et al., 2023), and CLIP Score (Hessel et al., 2021; Radford et al., 2021). ImageReward, HPSv2, and PickScore measure preference-based generation quality, while CLIP Score measures prompt-image alignment.

Qualitative results We first examine sample diversity. Figure 3 shows that our method preserves substantial variation across generated samples without an obvious loss of diversity. This behavior is consistent with the distributional results reported below. Figure 4 compares conditional generation across SD2.1, SDXL, and SD3. The comparison covers prompts involving spatial relations, object layouts, and scene-level composition. Our method better preserves global structures than FDG while maintaining comparable local appearance.

Quantitative Results Table 1 report distributional results on ImageNet and COCO 2017 for CFG, APG, FDG, and SRG. Relative to FDG and APG, our method achieves comparable FID, Precision, and Recall, while improving some of these metrics in several settings. These results indicate that selective low-frequency recovery preserves the distributional behavior of FDG.

Table 2 reports results on DrawBench, Parti Prompts, and HPS Prompts. APG modifies guidance geometry without separating frequency bands; SRG retains strong high-frequency guidance and selectively restores the orthogonal low-frequency residual. Under the evaluated settings, SRG scores higher than APG on all four metrics. Relative to FDG, SRG also improves every reported preference and prompt-alignment score. On DrawBench with SD3, ImageReward rises from 0.695 to 0.931. This substantial gain demonstrates the effectiveness of the proposed recovery strategy under this setting. Selective orthogonal low-frequency recovery improves conditional quality while maintaining distributional metrics comparable to FDG.

Table 2: Prompt-conditioned evaluation of CFG, FDG, APG and SRG on DrawBench, Parti Prompts, and HPS Prompts. Higher scores are better for all metrics.

Benchmark	Model	Method	ImageReward ↑	HPSv2 ↑	PickScore ↑	CLIP Score ↑
DrawBench	SD2.1	CFG	-0.174	0.243	21.04	25.28
		APG	-0.832	0.211	20.03	21.78
		FDG	-0.091	0.246	20.86	25.80
		SRG(Ours)	**0.035**	**0.254**	**21.05**	**26.37**
	SDXL	CFG	0.282	0.237	21.93	26.45
		APG	0.281	0.237	21.92	26.45
		FDG	0.508	0.258	22.15	27.90
		SRG(Ours)	**0.671**	**0.267**	**22.37**	**28.34**
	SD3	CFG	0.083	0.237	21.55	25.93
		APG	-0.174	0.221	21.17	24.71
		FDG	0.695	0.276	22.11	28.08
		SRG(Ours)	**0.931**	**0.291**	**22.41**	**28.60**
Parti Prompts	SD2.1	CFG	0.029	0.250	21.12	26.05
		APG	-0.810	0.214	19.91	22.36
		FDG	0.146	0.255	20.98	26.44
		SRG(Ours)	**0.232**	**0.258**	**21.12**	**26.81**
	SDXL	CFG	0.453	0.249	22.04	26.81
		APG	0.078	0.228	21.64	25.93
		FDG	0.696	0.265	22.19	27.72
		SRG(Ours)	**0.697**	**0.277**	**22.31**	**28.37**
	SD3	CFG	0.505	0.253	21.86	26.30
		APG	0.212	0.223	21.34	25.43
		FDG	0.968	0.288	22.29	27.75
		SRG(Ours)	**1.154**	**0.298**	**22.43**	**28.22**
HPS Prompts	SD2.1	CFG	0.002	0.240	20.72	27.34
		APG	-0.782	0.202	19.34	22.41
		FDG	0.142	0.246	20.65	28.00
		SRG(Ours)	**0.243**	**0.251**	**20.77**	**28.32**
	SDXL	CFG	0.481	0.255	22.00	28.81
		APG	0.155	0.233	21.53	27.83
		FDG	0.736	0.275	22.29	29.68
		SRG(Ours)	**0.931**	**0.287**	**22.59**	**30.31**
	SD3	CFG	0.483	0.256	21.71	26.72
		APG	0.187	0.293	21.28	25.77
		FDG	0.991	0.295	22.19	28.84
		SRG(Ours)	**1.113**	**0.305**	**22.25**	**29.11**

Prompt-level Consistency of Improvements To assess whether the gains in Table 2 extend across prompts, we compute paired score differences between SRG and FDG for each DrawBench prompt using SD3. Figure 5 shows the resulting distributions for ImageReward, HPSv2, PickScore, and CLIP Score; positive values favor SRG. The mean difference is positive for all four metrics, and the corresponding 95% bootstrap confidence intervals exclude zero. These results support positive average gains across the evaluated prompt set.

6 Ablation Studies

We examine three core design choices: recovery window, strength, and direction. All ablations are conducted on SD2.1 with the same sampling configuration. We evaluate prompt alignment on DrawBench (Saharia et al., 2022) and distributional quality on COCO 2017 (Lin et al., 2014).

Temporal recovery window Full-trajectory and early-stage recovery improve prompt alignment but increase FID and decrease Recall. Middle-only recovery yields smaller alignment gains than the selected window, while late recovery remains close to FDG. These results support retaining recovery after the initial high-noise phase and into the intermediate stage, balancing alignment gains against distributional quality.

Recovery strength We compare three strengths, $A = \eta(w_{\text{high}} - w_{\text{low}})$, with $\eta \in \{\frac{2}{3}, 1, \frac{4}{3}\}$. These yield orthogonal low-frequency guidance scales below, equal to, and above w_{high}, respectively. Increasing the recovery strength improves some alignment metrics but gradually degrades FID and Recall. We use $\eta = \frac{2}{3}$ by default, which provides alignment gains while keeping the distributional metrics close to FDG.

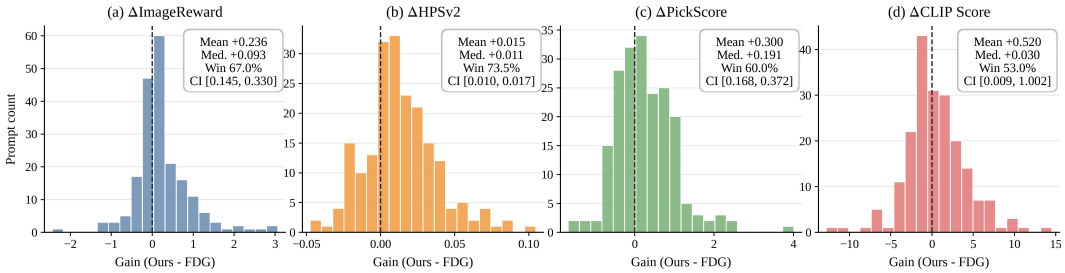

Figure 5: Per-prompt score differences between our method and FDG on DrawBench with SD3. Positive values indicate improvements over FDG. The markers and intervals denote the mean differences and 95% bootstrap confidence intervals, respectively.

Table 3: Ablation studies on SD2.1. We evaluate conditional alignment on DrawBench and distributional quality on COCO 2017. The shaded rows denote the configurations adopted in our method.

	DrawBench				COCO 2017		
Setting	CLIP Score ↑	ImageReward ↑	PickScore ↑	HPSv2 ↑	FID ↓	Precision ↑	Recall ↑
(a) Temporal recovery window							
FDG	25.80	-0.091	20.86	0.246	**21.96**	0.664	**0.512**
Full	26.21	**0.037**	**21.05**	**0.254**	22.93	**0.672**	0.487
Early	**26.39**	0.025	21.04	**0.254**	22.85	0.667	0.491
Middle	25.98	-0.075	20.96	0.247	22.02	0.662	0.498
Late	25.86	-0.086	20.85	0.246	**21.96**	0.663	0.507
SRG ($0.15 < u_t < 0.70$)	26.37	0.035	**21.05**	**0.254**	**21.96**	0.667	0.511
(b) Recovery strength							
FDG	25.80	-0.091	20.86	0.246	**21.96**	0.664	**0.512**
Below HF ($\eta = 2/3$)	26.37	**0.035**	21.05	0.254	**21.96**	**0.667**	0.511
Equal to HF ($\eta = 1$)	26.22	0.032	21.05	**0.256**	22.60	0.663	0.493
Above HF ($\eta = 4/3$)	**26.41**	**0.035**	**21.06**	**0.256**	22.77	0.665	0.482
(c) Recovery direction							
FDG	25.80	-0.091	20.86	0.246	**21.96**	0.664	**0.512**
Full LF	**26.37**	**0.052**	21.04	**0.254**	22.72	**0.670**	0.482
Parallel-only	25.89	-0.036	20.89	0.250	22.43	0.661	0.485
SRG	**26.37**	0.035	**21.05**	**0.254**	**21.96**	0.667	0.511

Recovery direction We compare full low-frequency recovery, parallel-only recovery, and SRG under the same sampling protocol. Full recovery improves alignment but substantially degrades FID and Recall. Parallel-only recovery provides limited alignment gains and also reduces distributional quality. Our method preserves the alignment improvements while maintaining FID and Recall close to those of FDG, yielding a better overall trade-off.

Overall, the ablations support stage-selective orthogonal recovery with a moderate strength: the temporal window limits unnecessary updates, the recovery strength controls their magnitude, and the orthogonal direction preserves the distributional behavior of FDG.

7 Conclusion

Our analysis suggests that attenuated low-frequency guidance contains useful structural information, motivating selective recovery rather than uniform suppression. We present SRG, a training-free extension of FDG that restores only the orthogonal low-frequency residual within a selected early-to-intermediate interval, leaving high-frequency guidance unchanged. Across five pretrained models, SRG improves text-to-image alignment and learned preference scores while maintaining FID, Precision, and Recall generally comparable to FDG. Ablations support combining orthogonal recovery with temporal selection and moderate strength to balance conditional generation quality and distributional coverage.

AI USE STATEMENT

Generative AI tools were used to assist with literature searching, code modification, language polishing, and iterative revision of the manuscript. These tools provided auxiliary support during the research and writing process. All scientific content, experimental decisions, code changes, and reported results were reviewed and finalized by the authors. The modified code was tested for correctness, and the authors take full responsibility for the final manuscript, implementation, experimental results, and conclusions.

ETHICS STATEMENT

This work uses publicly available datasets and pretrained diffusion models. It does not involve human participants, private or identifiable personal data, or new data collection. We are not aware of additional ethical concerns raised by the methods or experiments in this work.

REPRODUCIBILITY STATEMENT

We describe the model architectures, datasets, frequency decomposition, guidance formulation, recovery direction, temporal window, evaluation metrics, and sampling settings in the main paper and appendix. All reported results use fixed pretrained checkpoints and deterministic evaluation protocols. The implementation and configuration files will be released to support reproduction of the experiments.

REFERENCES

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=zrWNtzSZsf.

Peter J. Burt and Edward H. Adelson. The laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, 31(4):532–540, 1983. URL https://doi.org/10.1109/TCOM.1983.1095851.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: manifold-constrained classifier free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. IEEE Computer Society, 2009. URL https://doi.org/10.1109/CVPR.2009.5206848.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12606–12633. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/esser24a.html.

Fabian Falck, Teodora Pandeva, Kiarash Zahirnia, Rachel Lawrence, Richard E. Turner, Edward Meeds, Javier Zazo, and Sushrut Karmalkar. A fourier space perspective on diffusion models. *CoRR*, abs/2505.11278, 2025. URL https://doi.org/10.48550/arXiv.2505.11278.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pp. 7514–7528. Association for Computational Linguistics, 2021. URL https://doi.org/10.18653/v1/2021.emnlp-main.595.

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 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022. URL https://doi.org/10.48550/arXiv.2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html.

Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XII*, volume 15070 of *Lecture Notes in Computer Science*, pp. 196–212. Springer, 2024. URL https://doi.org/10.1007/978-3-031-73254-6_12.

Cheng Jin, Qitan Shi, and Yuantao Gu. Stage-wise dynamics of classifier-free guidance in diffusion models. *CoRR*, abs/2509.22007, 2025. URL https://doi.org/10.48550/arXiv.2509.22007.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/a98846e9d9cc01cfb87eb694d946ce6b-Abstract-Conference.html.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 24174–24184. IEEE, 2024. URL https://doi.org/10.1109/CVPR52733.2024.02282.

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. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 3929–3938, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html.

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. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural*

Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/dd540e1c8d26687d56d296e64d35949f-Abstract-Conference.html.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 5392–5399. IEEE, 2024. URL https://doi.org/10.1109/WACV57701.2024.00532.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/260a14acce2a89dad36adc8eefe7c59e-Abstract-Conference.html.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Mach. Intell. Res.*, 22(4):730–751, 2025. URL https://doi.org/10.1007/s11633-025-1562-4.

Stéphane Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(7):674–693, 1989. URL https://doi.org/10.1109/34.192463.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8162–8171. PMLR, 2021. URL http://proceedings.mlr.press/v139/nichol21a.html.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804. PMLR, 2022. URL https://proceedings.mlr.press/v162/nichol22a.html.

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. *CoRR*, abs/2509.16131, 2025. URL https://doi.org/10.48550/arXiv.2509.16131.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4172–4182. IEEE, 2023. URL https://doi.org/10.1109/ICCV51070.2023.00387.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 10199–10208. IEEE, 2023. URL https://doi.org/10.1109/CVPR52729.2023.00983.

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. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

Severi Rissanen, Markus Heinonen, and Arno Solin. Generative modelling with inverse heat dissipation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=4PJUBT9f2Ol.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 10674–10685. IEEE, 2022. URL https://doi.org/10.1109/CVPR52688.2022.01042.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. URL https://doi.org/10.1007/s11263-015-0816-y.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

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, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=e2ONKX6qzJ.

Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M. Weber. Guidance in the frequency domain enables high-fidelity sampling at low CFG scales. *CoRR*, abs/2506.19713, 2025b. URL https://doi.org/10.48550/arXiv.2506.19713.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, S. Sara Mahdavi, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 5234–5243, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/f7696a9b362ac5a51c3dc8f098b73923-Abstract.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=PxTIG12RRHS.

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. *CoRR*, abs/2306.09341, 2023. URL https://doi.org/10.48550/arXiv.2306.09341.

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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/33646ef0ed554145eab65f6250fab0c9-Abstract-Conference.html.

Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22552–22562. IEEE, 2023. URL https://doi.org/10.1109/CVPR52729.2023.02160.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=AFDcYJKhND.

A SUPPLEMENTARY ANALYSIS

A.1 DIRECTIONAL RECOVERY

We use the decomposition defined in Section 4.2, where $C_{L,t} = \psi_{\text{low}}[D_{c,t}]$ and $\langle \Delta D_t^{\text{low},\perp}, C_{L,t} \rangle = 0$. For $C_{L,t} \neq 0$, the directional projection is

$$\text{proj}_{C_{L,t}}(X) = \frac{\langle X, C_{L,t} \rangle}{\|C_{L,t}\|_2^2} C_{L,t}. \tag{11}$$

Since $\Delta D_t^{\text{low},\parallel} = \alpha_t C_{L,t}$, parallel and orthogonal recovery satisfy

$$\begin{aligned} \text{proj}_{C_{L,t}} \left(C_{L,t} + A\Delta D_t^{\text{low},\parallel} \right) &= (1 + A\alpha_t) C_{L,t}, \\ \text{proj}_{C_{L,t}} \left(C_{L,t} + A\Delta D_t^{\text{low},\perp} \right) &= C_{L,t}. \end{aligned} \tag{12}$$

A nonzero parallel correction changes the projection, whereas an orthogonal correction preserves it. This does not preserve the total norm:

$$\left\| C_{L,t} + A\Delta D_t^{\text{low},\perp} \right\|_2^2 = \|C_{L,t}\|_2^2 + A^2 \left\| \Delta D_t^{\text{low},\perp} \right\|_2^2. \tag{13}$$

When $C_{L,t} = 0$, the projection is undefined; the convention $\alpha_t = 0$ gives $\Delta D_t^{\text{low},\perp} = \Delta D_t^{\text{low}}$.

SRG adds $Ag(u_t)\Delta D_t^{\text{low},\perp}$ to the low-frequency FDG prediction. Therefore,

$$\begin{aligned} &\left\langle \widehat{D}_{\text{SRG},t}^{\text{low}} - \widehat{D}_{\text{FDG},t}^{\text{low}}, C_{L,t} \right\rangle \\ &\quad = Ag(u_t) \left\langle \Delta D_t^{\text{low},\perp}, C_{L,t} \right\rangle = 0. \end{aligned} \tag{14}$$

This preserves the guided prediction's projection at a fixed denoising state, not its subsequent trajectory. The effective orthogonal coefficient is $w_{\text{low}} + Ag(u_t)$, which does not exceed w_{high} under the prescribed bound on A. This bounds the coefficient rather than the recovery-vector norm. The parallel coefficient and high-frequency guidance remain unchanged.

A.2 NORMALIZED LOG-SNR AND TEMPORAL ACTIVATION

For $\mathbf{z}_t = \mathbf{x} + \sigma_t \boldsymbol{\epsilon}$, with independent unit-variance Gaussian noise, define $\text{SNR}_t = s_x^2/\sigma_t^2$, where s_x^2 is a fixed positive reference variance for the clean data or latents. Under the unit-variance convention, $\text{SNR}_t = 1/\sigma_t^2$. More generally, for $\mathbf{z}_t = a_t \mathbf{x} + b_t \boldsymbol{\epsilon}$, we have $\text{SNR}_t = a_t^2 s_x^2 / b_t^2$ and effective noise scale $\sigma_t = b_t/a_t$ for $a_t > 0$.

Let $\ell_t = \log \text{SNR}_t$ and let $\mathcal{T}_{\text{fin}}$ contain the denoiser evaluation times with finite, positive SNR. We define

$$\begin{aligned} \ell_{\min} = \min_{\tau \in \mathcal{T}_{\text{fin}}} \ell_\tau, \qquad \ell_{\max} = \max_{\tau \in \mathcal{T}_{\text{fin}}} \ell_\tau, \\ u_t = \frac{\ell_t - \ell_{\min}}{\ell_{\max} - \ell_{\min}}. \end{aligned} \tag{15}$$

The bounds come from the finite sampling schedule, not theoretical endpoints, with $\ell_{\max} > \ell_{\min}$. The constant $\log s_x^2$ cancels, so any fixed positive reference variance yields the same u_t.

Zero-noise and exact pure-noise endpoints are excluded from these bounds and assigned $u_t = 1$ and $u_t = 0$, respectively, for gating. Neither activates recovery. A terminal state without a denoiser evaluation receives no correction.

For additive-noise samplers, σ_t is the actual noise scale. For variance-preserving schedules, $\sigma_t = \sqrt{(1 - \bar{\alpha}_t)/\bar{\alpha}_t}$. For the linear flow interpolation $\mathbf{z}_\tau = (1 - \tau)\mathbf{x} + \tau\boldsymbol{\epsilon}$, $\sigma_\tau = \tau/(1 - \tau)$. These conversions use the actual schedule after timestep shifting, rather than raw timestep indices.

Recovery uses the fixed normalized log-SNR interval $0.15 < u_t < 0.70$. It is not data-adaptive, and the same interval can contain different numbers of denoiser evaluations across models and samplers.

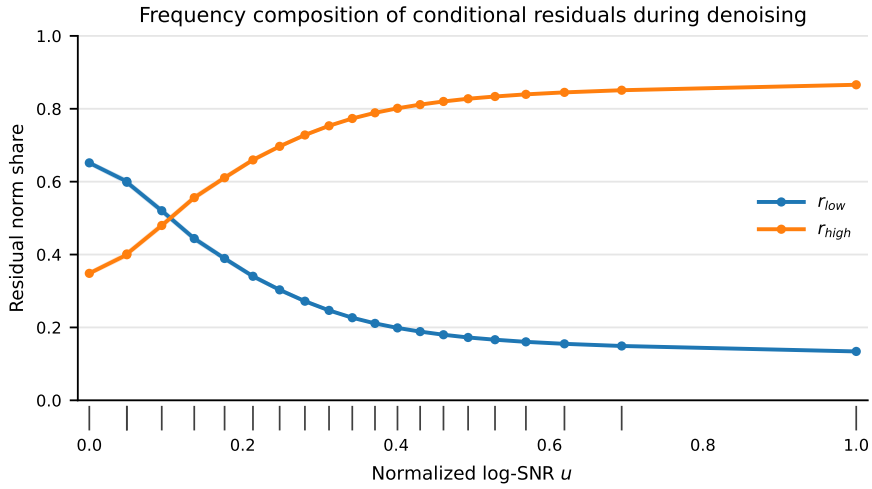

Figure 6: Relative low- and high-frequency guidance residual norms during denoising. The low-frequency share decreases as normalized log-SNR increases, while the high-frequency share increases.

A.3 Temporal frequency analysis

We characterize the residual composition using the per-step norm share

$$r_f(u_t) = \frac{\|\Delta D_t^f\|_2}{\|\Delta D_t^{\text{low}}\|_2 + \|\Delta D_t^{\text{high}}\|_2}, \qquad f \in \{\text{low}, \text{high}\}. \tag{16}$$

The shares sum to one whenever the denominator is nonzero. They describe relative residual magnitudes, not cumulative contributions or changes in image quality.

Figure 6 shows a larger low-frequency share earlier in denoising than during later refinement. This pattern motivates retaining recovery after the initial high-noise phase rather than postponing it entirely to the middle stage. Initial predictions may be unreliable, while late low-frequency updates may have limited influence on global composition. SRG therefore uses an early-to-intermediate interval that excludes the initial and final phases.

B Experimental details

B.1 Evaluation datasets and sample counts

Within each model–dataset pair, all methods use the same class-label or prompt list. ImageNet evaluation uses 10,000 generated images per method. For COCO 2017 (Lin et al., 2014), we use 5,000 validation image–caption records and generate one image per record. For all models, we select the first caption associated with each image. Prompt-conditioned evaluation uses 200 DrawBench prompts (Saharia et al., 2022), 1,632 Parti Prompts (Yu et al., 2022), and 3,200 HPS Prompts (Wu et al., 2023). Each prompt produces one image per method. The HPS prompt set covers anime, concept art, paintings, and photography. For each model–dataset pair, all methods share identical input conditions for fair comparison.

B.2 Model and sampling configurations

We use the pretrained Stable Diffusion 2.1 Base, Stable Diffusion XL Base 1.0, and Stable Diffusion 3 Medium models for text-to-image generation. Inference uses FP16 with an empty negative prompt. Table 5 summarizes the sampling settings. The cor-

Table 4: Evaluation sample counts per method and model. For ImageNet, 1000 class labels are used, yielding 10000 generated images.

Dataset	Text prompts	Generated images
ImageNet	1,000	10,000
COCO 2017	5,000	5,000
DrawBench	200	200
Parti Prompts	1,632	1,632
HPS Prompts	3,200	3,200

Table 5: Text-to-image sampling configurations. Steps denotes the requested number of inference steps.

Model	Scheduler	Steps	Resolution
SD2.1	PNDM	20	512×512
SDXL	Euler	20	1024×1024
SD3	FlowMatch Euler	28	1024×1024

responding scheduler classes are `PNDMScheduler`, `EulerDiscreteScheduler`, and `FlowMatchEulerDiscreteScheduler`. The SD2.1 schedule contains 21 denoiser evaluation timesteps for 20 requested steps because one timestep is repeated. SDXL and SD3 use 20 and 28 evaluation timesteps, respectively.

B.3 Frequency decomposition and prediction space

The text-to-image implementations use a single-level, two-band Gaussian–Laplacian decomposition. A 5×5 Gaussian filter with standard deviation 1 is followed by factor-two downsampling. The SD2.1 and SDXL implementations use stride-two subsampling, while the SD3 implementation uses stride-two average pooling. The low-frequency tensor is reconstructed at the original resolution using bilinear interpolation with `align_corners=False`. The high-frequency component is the residual $\psi_{\text{high}}[X] = X - \psi_{\text{low}}[X]$.

The same decomposition is applied to $\Delta D_t = D_{c,t} - D_{u,t}$ and the conditional prediction used to obtain $C_{L,t}$. Guidance operates in clean-sample latent prediction space. Native noise or velocity outputs are converted to this representation before guidance and converted back afterward. SD3 uses the corresponding conversion between clean-sample predictions and rectified-flow velocity. The scheduler update rule is unchanged.

Inner products and squared norms are computed independently for each sample over all channel and spatial dimensions:

$$\langle X, Y \rangle_b = \sum_{c,h,w} X_{b,c,h,w} Y_{b,c,h,w}. \tag{17}$$

The batch dimension is excluded. The implementations add 10^{-8} to the projection denominator for numerical stability. SD2.1 and SDXL perform projection and recovery calculations in FP32 and cast the guided prediction back to the sampler dtype afterward.

B.4 Guidance parameters

Table 6 summarizes the guidance parameters for both evaluations. CFG and APG use w, while FDG and SRG share $(w_{\text{low}}, w_{\text{high}})$. SRG uses $A = \eta(w_{\text{high}} - w_{\text{low}})$ with $\eta = 2/3$ and the common recovery interval $0.15 < u_t < 0.70$. APG uses $\eta_{\text{APG}} = 0$, rescaling threshold r, and momentum coefficient β. The APG rescaling and momentum parameters for EDM2-S, DiT-XL/2, SD2.1, and SDXL follow APG (Sadat et al., 2025a). For SD3, we adopt the SDXL values; these are not an SD3-specific configuration reported in the APG paper.

Table 6: Guidance parameters for distributional and prompt-conditioned evaluation. We abbreviate w_{low} and w_{high} as w_L and w_H. The recovery strength A applies to SRG; r and β apply to APG. In (a), EDM2-S and DiT-XL/2 use ImageNet, and the remaining models use COCO 2017. Panel (b) applies to DrawBench, Parti Prompts, and HPS Prompts.

(a) Distributional evaluation.

Model	w	w_L	w_H	A	r	β
EDM2-S	3	1	3	$4/3$	2.5	-0.75
DiT-XL/2	2	1	2	$2/3$	5.0	-0.50
SD2.1	7	3	7	$8/3$	7.5	-0.75
SDXL	10	5	10	$10/3$	15.0	-0.50
SD3	7	3	7	$8/3$	15.0	-0.50

(b) Prompt-conditioned evaluation.

Model	w	w_L	w_H	A	r	β
SD2.1	3	3	12	6	7.5	-0.75
SDXL	3	3	12	6	15.0	-0.50
SD3	1.5	1.5	12	7	15.0	-0.50

Table 7: Recovery schedules in the temporal ablation.

Setting	Recovery activation
Full trajectory	Throughout sampling
Early	First third of sampling nodes
Middle	Middle third of sampling nodes
Late	Final third of sampling nodes
SRG	$0.15 < u_t < 0.70$

B.5 Ablation settings

All ablations use SD2.1. Distributional evaluation on COCO 2017 (Lin et al., 2014) uses $(w_{\text{low}}, w_{\text{high}}) = (3, 7)$; prompt-conditioned evaluation on DrawBench (Saharia et al., 2022) uses $(3, 12)$.

The temporal ablation divides the sampling nodes, ordered from high to low noise, into three consecutive groups: Early, Middle, and Late. The boundaries are placed at one-third and two-thirds of the sampling sequence and mapped to normalized log-SNR using Eq. equation 15. These groups therefore need not have equal widths in normalized log-SNR. SRG instead uses the fixed interval $0.15 < u_t < 0.70$.

The strength ablation evaluates $\eta \in \{2/3, 1, 4/3\}$. These correspond to $A \in \{8/3, 4, 16/3\}$ on COCO 2017 and $A \in \{6, 9, 12\}$ on DrawBench. During recovery, the effective orthogonal coefficient $w_{\text{low}} + A$ is below, equal to, or above w_{high}, respectively. The $\eta = 4/3$ setting intentionally exceeds the default bound to examine stronger recovery.

The direction ablation compares FDG, full low-frequency recovery, parallel-only recovery, and SRG. Their additional recovery components are zero, ΔD_t^{low}, $\Delta D_t^{\text{low},\parallel}$, and $\Delta D_t^{\text{low},\perp}$, respectively. This comparison examines the effect of recovery direction on alignment and distributional quality.