DYNAMIC DRIFT-AWARE GUIDANCE FOR ACCELERATED GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance (CFG) enables high-fidelity generation in diffusion and flow-matching models but requires two inference passes in each sampling step because of the additional unconditional network pass. While recent acceleration heuristics skip unconditional passes using fixed timestep schedules or early termination, they ignore prompt-dependent trajectory curvature, leading to trajectory deviation, object omission, and semantic collapse. To resolve this problem, we propose Dynamic Drift-Aware Guidance (DrAG), a training-free framework that selectively skips redundant unconditional evaluations while closely approximating the original sampling trajectory. DrAG is developed on the key intuition that exact unconditional computation is not required at every step. Our first variant called DrAG-Spatial, continuously tracks the directional and magnitude drift of the conditional velocity, to bypass the unconditional pass whenever the trajectory remains stable. Instead of freezing unconditional predictions, DrAG-Spatial tracks the relative difference between the conditional and unconditional passes. By pairing this preserved difference with the freshly computed conditional velocity, DrAG-Spatial accurately extrapolates guidance at skipped steps, refreshing the unconditional state only when significant trajectory drift occurs. Further, to prevent high-frequency noise from triggering unnecessary unconditional evaluations, we introduce a second variant called DrAG-Spectral, which applies a Gaussian low-pass Fourier filter to isolate structural shift and guide compute allocation purely by meaningful trajectory evolution. Through extensive experiments across three T2I backbones (FLUX.1-dev, Qwen-Image, SD3) and two T2V models (Wan 2.1, HunyuanVideo), we demonstrate substantially reduced inference computation using DrAG-Spatial and DrAG-Spectral, while faithfully matching and being comparable to the full-guidance quality across established benchmarks (e.g. GenEval2, VBench). The source code is available at `https://anonymous.4open.science/r/DrAG-50D0`.

1 INTRODUCTION

Conditional generative models (Mirza & Osindero, 2014) aim to generate samples which look both realistic and consistent with a given condition, such as a class label, text prompt, or any other reference signal. However, the base pretrained model may not fully capture the specified conditioning, leading to outputs with weaker condition alignment or missing fine-grained details. Conditional guidance overcomes this limitation by directing the sampling trajectory towards regions that are more closely aligned with the specified condition, often improving sample fidelity and prompt alignment. Classifier-free guidance (CFG) (Ho & Salimans, 2021) has become a standard choice because it provides this control without requiring a separate classifier at inference time. Given an unconditional prediction v_u and a conditional prediction v_c, CFG forms the guided prediction as

$$v_{cfg} = v_u + s\,(v_c - v_u)\,, \tag{1}$$

where s controls the guidance strength. Larger s increases conditional influence but can reduce diversity or introduce artifacts (Ho & Salimans, 2021). CFG is therefore widely adopted in modern text-conditioned image and video generation (Rombach et al., 2022; Saharia et al., 2022; Ho et al., 2022).

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.

Despite its effectiveness, CFG roughly doubles the network evaluations at each sampling step by computing both conditional and unconditional predictions (Ho & Salimans, 2021; Meng et al., 2023). Thus, a T step guided sampler requires roughly $2T$ NFEs (neural function evaluations), adding substantial computational cost to the already expensive iterative sampling process (Lipman et al., 2022; Li et al., 2024), particularly for high-resolution image and video models.

Prior work has improved generative inference along several complementary directions. Fast samplers reduce the number of sampling steps without retraining (Song et al., 2021; Lu et al., 2022), while progressive and guidance distillation train faster student models (Salimans & Ho, 2022; Meng et al., 2023). These methods can provide large speedups, but reducing the number of sampling steps addresses a different source of cost from the two-branch computation performed by CFG and distillation introduces an additional training procedure. More recently, Momentum Guidance (Liao et al., 2026) estimates the unconditional contribution from previous velocity or momentum information, but its reported results indicate that unconditional evaluations often remain important for preserving generation quality. Further, Adaptive Guidance (Castillo et al., 2025) and Step-AG (Zhang & Wan, 2025) show that unconditional evaluations can be selectively omitted during parts of the sampling trajectory. This suggests that CFG need not be applied uniformly at every step. It also motivates a general, training-free mechanism that identifies unnecessary guidance computation, works with a pretrained model, and does not substantially alter the underlying generative process.

Towards this, we make following **Contributions**: we introduce **Dynamic Drift-Aware Guidance (DrAG)**, a training-free framework that substantially reduces the computational cost of CFG by dynamically bypassing redundant unconditional evaluations. Rather than relying on rigid schedule cutoffs (Zhang & Wan, 2025; Lee et al., 2025), DrAG introduces a new perspective by formulating CFG acceleration as a *guidance-gap reuse problem*. Specifically, DrAG decouples the conditional and unconditional velocities, identifies their displacement as the *guidance-gap*, which captures the conditional correction that drives the sampling trajectory. The key idea is to cache this gap and reuse it across sampling steps and refresh it only when the trajectory evolution indicates that the cached gap is no longer reliable. This transforms the question of *which timesteps to skip* into the more general question of *when the guidance-gap remains reusable*.

Based on this formulation, we develop two variants, **DrAG-Spatial** and **DrAG-Spectral**. To determine when the guidance-gap remains reusable, DrAG-Spatial measures the evolution of the conditional velocity in the spatial domain using directional cosine alignment and magnitude divergence. This enables adaptive identification of steps where the cached guidance-gap can be reused or should be recomputed. However, spatial directional and magnitude measurements can remain sensitive to high-frequency variations in the velocity field, potentially triggering unnecessary guidance-gap refreshes, since low-frequency components are commonly associated with coarse-scale structural information, whereas high-frequency components capture fine-scale variations (Xu et al., 2020). To address this limitation, we further introduce DrAG-Spectral, which measures trajectory deviation in the frequency domain using a Gaussian low-pass filter. By emphasizing low-frequency components, DrAG-Spectral reduces sensitivity to high-frequency variations while retaining the semantic structure of the sampling trajectory.

We provide a theoretical grounding for guidance-gap reuse by relating the approximation error to the observable conditional velocity. On the primary image-generation evaluation, we comprehensively compare DrAG with existing guidance-acceleration methods across FLUX.1-dev, Qwen-Image, and Stable Diffusion 3 using GenEval2 and MS-COCO, demonstrating substantial reductions in unconditional guidance evaluations while maintaining competitive generation quality.

Moreover, we examine the *transferability* of the same guidance-gap reuse principle beyond the primary image-generation setting. Without introducing task-specific modifications, we evaluate DrAG across video generation, few-step generation, and score/noise-based diffusion models, showing that the mechanism remains applicable across substantially different generative settings. Finally, we demonstrate that DrAG is not limited to inference or multi-step generative models. We extend it to one-step generative models, such as Generative Modeling via Drifting (Deng et al., 2026), where we show that DrAG reduces the computational overhead of training significantly.

2 Related Work

2.1 Conditional Guidance in Diffusion and Flow Models

Diffusion models were introduced as iterative generative models that learn to reverse a gradual noising process to recover data samples (Song & Ermon, 2019; Ho et al., 2020). Subsequent score-based generative modeling provided a continuous-time formulation based on stochastic differential equations (SDEs), offering a unified perspective on diffusion and score-based models (Lim et al., 2023). Conditional generation was first strengthened by classifier guidance (Dhariwal & Nichol, 2021), followed by CFG (Ho & Salimans, 2021), which jointly learns conditional and unconditional predictions and combines them at inference time. CFG has since become a standard mechanism for conditional generation in latent diffusion, text-to-image, and video models (Rombach et al., 2022; Saharia et al., 2022; Podell et al., 2024; Ho et al., 2022), and has also been adopted by flow-matching and rectified-flow models (Lipman et al., 2022; Esser et al., 2024; Liu et al., 2022).

Few works have explored the quality-diversity trade-off and trajectory distortion caused by strong, constant guidance. Condition-annealed diffusion sampling (Sadat et al., 2024) and limited-interval guidance (Kynkäänniemi et al., 2024) modify when or how strongly guidance is applied, while CFG++ (Chung et al., 2025) and CFG-Zero* (Fan et al., 2025) modify the guided trajectory to improve sampling behavior. These approaches primarily improve guidance by modifying its strength, application interval, or effect on the sampling trajectory. In contrast, our work focuses on the computational redundancy of repeatedly recomputing the unconditional branch and asks whether such evaluations are necessary at every sampling step.

2.2 Fast Sampling and Distillation With Guidance

A key research direction focuses on reducing the required sampling steps through more efficient discretization methods for the reverse-time SDE and probability-flow ODE. DDIM (Song et al., 2021) enables non-Markovian deterministic sampling. Further, DPM-Solver (Lu et al., 2022), DPM-Solver++ (Lu et al., 2023), and UniPC (Zhao et al., 2023) use higher-order or predictor–corrector formulations for accurate low-NFE sampling. EDM (Karras et al., 2022) further systematizes noise schedules, preconditioning, and solver choices. When these samplers are used with CFG, each guided sampling step still requires both conditional and unconditional network evaluations.

An alternative line of work uses distillation to learn models that generate samples in fewer steps. Progressive distillation (Salimans & Ho, 2022) repeatedly halves the number of sampling steps, and guidance distillation (Meng et al., 2023) incorporates the effect of CFG into the student model. Consistency Models (Song et al., 2023) and Latent Consistency Models (Luo et al., 2023) learn mappings that support one-step or few-step generation. Later approaches improve the few-step regime through adversarial (Sauer et al., 2024b), trajectory-consistency (Zheng et al., 2024), and distribution-matching (Sauer et al., 2024a; Yin et al., 2024) objectives. These approaches generally require additional training or optimization and often produce model-specific student models. More importantly, reducing the number of sampling steps does not remove the two-branch cost of CFG at the remaining steps. Our work addresses this complementary source of inference cost by reducing redundant unconditional evaluations within the existing sampling trajectory, without changing the pretrained model or sampling procedure.

2.3 Reducing NFE in Classifier-Free Guidance Evaluations

The research most relevant to our approach investigates the necessity of using both CFG branches throughout every timestep. Adaptive Guidance (AG) (Castillo et al., 2025) uses the alignment between conditional and guided predictions to determine when the additional branch can be omitted, while its LinearAG variant further approximates predictions using earlier evaluations. Step-AG (Zhang & Wan, 2025) restricts CFG to selected stages of the sampling trajectory based on temporal or signal-to-noise characteristics. Further, Tortoise and Hare Guidance (THG) (Lee et al., 2025) interprets CFG as a multirate ODE system and evaluates the guidance component on a coarser temporal grid. Taken together, these methods indicate that the contribution of guidance is time-dependent, while the unconditional branch contains significant redundancy across timesteps.

Existing approaches decide *when* to evaluate the unconditional branch through prediction alignment (Castillo et al., 2025), predefined trajectory stages (Zhang & Wan, 2025), or coarse temporal grids (Lee et al., 2025). In contrast, DrAG formulates the problem as *guidance-gap reuse*: rather than discarding the guidance correction when an unconditional evaluation is skipped, it reuses the previously computed conditional-unconditional gap together with the current conditional prediction. This formulation provides a theoretically grounded basis for determining when the cached guidance remains reusable by relating the refresh decision to the evolution of the sampling trajectory. Thus, we provide a training-free, adaptive and theoretically grounded mechanism for reducing redundant unconditional evaluations while retaining the original pretrained model and sampling procedure.

3 METHODOLOGY

Flow Matching and Classifier-Free Guidance. Flow Matching (FM) (Lipman et al., 2022) and Rectified Flow (RF) (Liu et al., 2022) generate samples along explicitly straightened probability paths by learning a deterministic ODE, rather than the stochastic sampling paths used by earlier score/noise-based diffusion models (Song & Ermon, 2019; Ho et al., 2020). This ODE transports a sample from a base noise distribution $x_0 \sim \mathcal{N}(0, I)$ to $x_1 \sim p_{data}$, where p_{data} is a target data distribution as: $\mathrm{d}x_t = v_\theta(x_t, t)\,\mathrm{d}t, t \in [0, 1]$ where $t = 0$ is pure noise, $t = 1$ is clean data, and v_θ is a learned velocity field (using a Diffusion Transformer (DiT)). Classifier-free guidance (CFG) (Ho & Salimans, 2021) steers this ODE towards a conditioning signal c by contrasting a conditional and an unconditional prediction. Writing $v_c(t) := v_\theta(x_t, t, c)$ and $v_u(t) := v_\theta(x_t, t, \varnothing)$, the guided velocity is

$$v_{CFG}(x_t, t) = v_u(t) + s\big(v_c(t) - v_u(t)\big), \qquad s \geq 1, \tag{2}$$

with guidance scale s. Because Eq.2 requires both $v_u(t)$ and $v_c(t)$ at every step, CFG doubles the number of network evaluations relative to unguided sampling.

Our goal is to reduce the inference cost of CFG without changing the sampling trajectory or the underlying flow model. A direct way to skip the unconditional branch would be to simply reuse a previously computed value of v_u at later steps; however, $v_u(t)$ depends on the current latent x_t, so freezing it at a stale value quickly becomes inaccurate as the trajectory evolves.

We propose Dynamic Drift-Aware Guidance (DrAG) based on the key insight that the unconditional velocity itself is not the quantity that alone determines the effect of guidance. Instead, CFG applies the conditioning through the difference between the conditional and unconditional velocity fields. We therefore define this difference as the *guidance-gap* at time t,

$$G(t) = v_c(t) - v_u(t). \tag{3}$$

Rather than recomputing this gap at every sampling step, DrAG treats it as a dynamic quantity that can be cached and reused over intervals where its evolution remains sufficiently stable. At each step, the conditional velocity is evaluated normally, while the cached guidance-gap provides the missing unconditional contribution. The central problem therefore becomes determining when the cached gap remains valid and when its drift requires a refresh. We address this problem through two complementary variants, **DrAG-Spatial** and **DrAG-Spectral**, which differ only in how they decide when the cached guidance-gap should be refreshed.

3.1 CACHING THE GAP

As defined in Eq.3, the guidance-gap captures the conditional correction applied by CFG. Using Eq.3, the guided velocity in Eq.2 can equivalently be written as

$$v_{CFG}(t) = v_c(t) + (s - 1)\, G(t). \tag{4}$$

This reformulation is the basis of DrAG: instead of recomputing $v_u(t)$ at every step, we can reuse a previously computed guidance gap as long as it is still an accurate estimate of the current one.

At an *anchor* timestep $t_a \in (0, 1)$, we evaluate both branches and store the resulting gap, $G(t_a) = v_c(t_a) - v_u(t_a)$. At a later timestep $t_i > t_a$, we evaluate only the conditional velocity $v_c(t_i)$ and approximate the guided velocity as:

$$\hat{v}_{CFG}(t_i) = v_c(t_i) + (s - 1)\, G(t_a). \tag{5}$$

The conditional branch is still evaluated at every step, so the model still responds to the current state x_{t_i}; only the unconditional evaluation is skipped.

3.2 Deciding When to Refresh the Cached Gap $G(t_a)$

The cached guidance-gap $G(t_a)$ remains useful as long as the conditional and unconditional velocities do not change substantially from the anchor timestep t_a to the current timestep t_i. We therefore characterize this change through the *velocity drift* between the anchor and current timestep. Specifically, we define the conditional and unconditional velocity drifts as

$$\Delta_c = \Delta_c(t_i, t_a) = v_c(t_i) - v_c(t_a),\ \Delta_u = \Delta_u(t_i, t_a) = v_u(t_i) - v_u(t_a). \tag{6}$$

These quantities measure how much the respective velocities have changed since the gap was last evaluated. The guidance-gap may also change as the sampling trajectory evolves. From Eq.3 and using Eq.6, the guidance-gap change between the anchor and current timestep can be written as

$$G(t_i) - G(t_a) = \Delta_c - \Delta_u. \tag{7}$$

We refer to this change in the guidance-gap as the *guidance-gap drift*. Directly evaluating the guidance-gap drift in Eq.7 would require computing $v_u(t_i)$, which defeats the purpose of avoiding the unconditional evaluation. In contrast, the conditional velocity $v_c(t_i)$ is already evaluated at every sampling step. DrAG therefore uses the observable conditional velocity drift Δ_c for approximating the evolution of the guidance-gap.

The motivation for this choice follows directly from the error introduced by reusing the cached gap. Combining equations 4 and 5, we obtain

$$\left\|\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\right\|_2 = (s-1)\left\|G(t_i) - G(t_a)\right\|_2. \tag{8}$$

Therefore, approximation error is directly determined by the amount of guidance-gap drift since the anchor. Since the unconditional drift Δ_u is unavailable when the unconditional branch is skipped, we relate the unobserved guidance-gap drift to the observable conditional velocity drift through the following assumption.

Assumption 1. *There exists a local constant $\kappa > 0$ such that $\|\Delta_u\|_2 \leq \kappa\|\Delta_c\|_2$. That is, over the interval $[t_a, t_i]$, the unconditional branch does not drift substantially faster than the conditional branch.*

This assumption is reasonable because v_c and v_u are not coming from independent networks, they were obtained using the same network v_θ evaluated on the same latent x_t and timestep t, differing only in the conditioning input (c versus $\varnothing$). Further, we empirically examine the validity of this coupling assumption, obtaining $\kappa \in [1.17, 1.22]$ across all individual models, as reported in Appendix B.1. Under Assumption 1, we obtain the following bound on the guidance-gap reuse error (proof in Appendix A.1).

Proposition 1 (Guidance-Gap Reuse Error)**.** *Under Assumption 1,*

$$\left\|\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\right\|_2 \leq (s-1)(1+\kappa)\,\|\Delta_c\|_2. \tag{9}$$

Proposition 1 provides the theoretical basis for DrAG: bounding the conditional drift $\|\Delta_c\|_2$ also bounds the local guidance error, making it a natural signal for determining when the cached guidance-gap can be reused or should be refreshed. We now discuss how the two DrAG variants, DrAG-Spatial and DrAG-Spectral, leverage Proposition 1 to determine when to skip an unconditional evaluation and when to refresh the cached guidance-gap.

3.3 DrAG-Spatial and DrAG-Spectral

DrAG-Spatial measures conditional velocity drift directly in the original velocity space using directional alignment $S_{\text{ang}}(t_i) = \langle v_c(t_i), v_c(t_a)\rangle / (\|v_c(t_i)\|_2 \|v_c(t_a)\|_2)$ and relative magnitude change $S_{\text{mag}}(t_i) = |\|v_c(t_i)\|_2 - \|v_c(t_a)\|_2| / \|v_c(t_a)\|_2$. The guidance-gap is refreshed once either quantity crosses a threshold:

$$S_{\text{ang}}(t_i) < \tau_{\text{sp}} \quad \text{or} \quad S_{\text{mag}}(t_i) > \gamma_{\text{sp}}. \tag{10}$$

S_{ang} and S_{mag} together decompose $\|\Delta_c\|_2$ into a directional and a magnitude component (Appendix A.2), so Eq.10 directly bounds $\|\Delta_c\|_2$ and, via Proposition 1, the guidance-gap reuse error:

Corollary 1 (DrAG-Spatial Guidance-Gap Reuse Error)**.** *Whenever DrAG-Spatial reuses the cached guidance-gap, $\left\|\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\right\|_2 \leq (s-1)(1+\kappa_{\text{sp}})\,\|v_c(t_a)\|_2 \sqrt{\gamma_{\text{sp}}^2 + 2(1+\gamma_{\text{sp}})(1-\tau_{\text{sp}})}$.*

Within a given anchor interval, the guidance scale s is fixed by the sampler, and $\|v_c(t_a)\|_2$ is likewise fixed, since it is computed once at the anchor step t_a and held constant until the gap is refreshed. Treating the local coupling constant κ_{sp} as fixed as well (Assumption 1), the bound in Corollary 1 reduces to a fixed multiplicative factor times a term that depends only on the thresholds τ_{sp} and γ_{sp}. Tightening thresholds shrinks the bound toward zero, recovering exact or close enough CFG, while loosening them increases the bound in exchange for few more unconditional evaluations. The full derivation of Corollary 1 is given in Appendix A.2.

A limitation of measuring drift directly on v_c is that this signal also carries high-frequency fluctuations consisting fine details or noise-level variation that can shift S_{ang} and S_{mag} even when the guidance-gap itself has not meaningfully changed. Such fluctuations can trigger unnecessary gap refreshes, thereby diminishing the computational benefits of guidance-gap reuse. DrAG-Spectral addresses this by basing the refresh decision on the low-frequency content of v_c instead, which is less sensitive to fine-grained fluctuations.

Let $\mathcal{F}$ denote the discrete Fourier transform and $\mathcal{M}$ a fixed Gaussian low-pass filter mask that emphasizes low frequencies while suppressing high ones. We define the frequency-weighting operator $W_\sigma(x) := (\mathcal{M} \odot \mathcal{F})(x)$, where $\odot$ denotes Hadamard multiplication, which suppresses high-frequency content while retaining low-frequency structure. The relative spectral drift is then

$$d_i^{\text{spec}} = \frac{\|W_\sigma(v_c(t_i) - v_c(t_a))\|_2}{\|W_\sigma(v_c(t_a))\|_2}, \qquad \text{and guidance-gap refresh if } d_i^{\text{spec}} > \gamma_{\text{spec}}, \tag{11}$$

with γ_{spec} a hyperparameter tuned via ablation. As shown in Appendix A.3, d_i^{spec} decomposes into the same directional and magnitude terms as $S_{\text{ang}}/S_{\text{mag}}$ in DrAG-Spatial, only evaluated after low-pass filtering. Under a corresponding coupling assumption similar to Assumption 1 in frequency domain, this again bounds the guidance-gap reuse error via Proposition 1:

Corollary 2 (DrAG-Spectral Guidance-Gap Reuse Error)**.** *Whenever DrAG-Spectral reuses the cached guidance-gap,* $\left\|W_\sigma\big(\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\big)\right\|_2 \le (s-1)(1+\kappa_\sigma)\,\gamma_{\text{spec}}\,\|W_\sigma(v_c(t_a))\|_2$.

Unlike the spatial bound, this error is linear in the single threshold γ_{spec}. Within a given anchor interval, the scaling factor s, the coupling constant κ_σ (Assumption 2 in Appendix A.3), and $\|W_\sigma(v_c(t_a))\|_2$ are fixed, since the latter is computed once at the anchor step t_a and held constant until the gap is refreshed. Tightening $\gamma_{\text{spec}} \to 0$ approximates CFG close enough, while loosening it increases the bound linearly in exchange for few more unconditional evaluations. The full derivation of Corollary 2 is given in Appendix A.3. Both variants DrAG-Spatial and DrAG-Spectral share a similarly structured sampling loop discussed in Appendix A.4.

4 Experiments

4.1 Image Generative Models

We evaluate the quality–efficiency trade-off of DrAG on three text-to-image models, FLUX.1-dev, Qwen-Image, and Stable Diffusion 3 (SD3), using the GenEval2 benchmark and complementary distributional metrics (FID, KID, etc.). All experiments are implemented in PyTorch and conducted on NVIDIA A100 GPU with 80 GB memory. We compare DrAG-Spatial and DrAG-Spectral against Baseline CFG and three existing CFG acceleration methods: Step-AG (Zhang & Wan, 2025), Autoguidance (AG) (Castillo et al., 2025), and Tortoise-and-Hare Guidance (THG) (Lee et al., 2025). GenEval2 evaluates compositional generation across all 800 prompts along five dimensions: object presence, counting, attribute binding, spatial positioning, and verb interaction. Hyperparameters for both DrAG variants are selected from the model-specific ablation studies reported in Tables 17–19 in Appendix D.2, and the resulting full evaluation is reported in Table 1.

Table 1 shows that the quality-efficiency trade-off varies across models and guidance strategies. On FLUX.1-dev, DrAG-Spatial achieves an overall score (column 8 in Table 1) of 26.04, close to the Baseline CFG score of 26.11, while reducing NFE by 24.5%. DrAG-Spectral further reduces NFE by 39.0% with a score of 25.55, compared with 24.79 for THG at a 33.3% reduction. On Qwen-Image, DrAG-Spectral achieves a score of 38.08 at 36.9 NFE, compared with 38.30 for Baseline CFG at 60.0 NFE, while outperforming AG (37.86) and THG (37.90) in overall quality. DrAG-Spatial achieves a score of 37.35 with a 36.8% compute reduction. On SD3, DrAG-Spatial slightly

improves the overall score from 20.37 to 20.68 while reducing NFE by 26.1%, whereas DrAG-Spectral achieves a comparable score of 20.35 with a larger 31.4% reduction. Overall, the results show that the two DrAG variants provide complementary quality-efficiency operating points across different generative models.

Table 1: Quantitative Results on GenEval2 Benchmark

Model	Method	Object (↑)	Count (↑)	Attribute (↑)	Position (↑)	Verb (↑)	Overall (↑)	Avg NFE (↓)	Compute Reduction (↑)
FLUX	Baseline CFG	91.41	55.20	75.95	42.34	26.19	26.11	60.0	0.0%
FLUX	Step-AG	**91.04**	55.06	74.74	41.82	26.08	**26.10**	45.0	25.0%
FLUX	THG	89.82	54.68	74.95	40.47	25.42	24.79	40.0	33.3%
FLUX	AG	89.72	55.14	72.99	41.33	25.04	25.92	39.0	35.0%
FLUX	DrAG-Spatial	90.89	**55.39**	**75.65**	**41.96**	26.05	26.04	45.3	24.5%
FLUX	DrAG-Spectral	89.34	54.68	72.49	41.55	**26.28**	25.55	**36.6**	**39.0%**
Qwen	Baseline CFG	98.57	69.87	86.86	59.05	38.15	38.30	60.0	0.0%
Qwen	Step-AG	**98.56**	**69.66**	86.31	59.37	**40.05**	**39.15**	45.0	25.0%
Qwen	THG	98.30	69.29	**86.67**	58.82	39.46	37.90	40.0	33.3%
Qwen	AG	98.03	69.29	86.50	59.41	37.71	37.86	**35.1**	**41.6%**
Qwen	DrAG-Spatial	97.43	69.12	82.50	59.20	37.36	37.35	37.9	36.8%
Qwen	DrAG-Spectral	98.02	69.14	83.47	**59.57**	37.41	38.08	36.9	38.5%
SD3	Baseline CFG	88.59	48.55	68.58	37.68	11.13	20.37	56.0	0.0%
SD3	Step-AG	87.04	**47.98**	67.01	37.75	10.99	19.19	42.0	25.0%
SD3	THG	83.95	44.51	63.09	35.41	13.40	17.42	38.0	32.1%
SD3	AG	76.44	43.29	60.56	34.01	12.81	16.68	**33.0**	**41.1%**
SD3	DrAG-Spatial	**87.60**	47.71	**67.30**	**39.79**	**13.77**	**20.68**	41.4	26.1%
SD3	DrAG-Spectral	85.36	46.26	66.06	38.62	11.92	20.35	38.4	31.4%

However, directly comparing the individual GenEval2 dimensions together with compute reduction makes the overall quality-efficiency trade-off difficult to summarize. For example, a method may achieve a higher overall compositional score while providing relatively small computational savings, whereas another method may obtain slightly lower quality but substantially reduce the number of guidance evaluations. Therefore, we formulate a compact quality-efficiency ranking that jointly considers compositional quality and computational reduction.

Quality-Efficiency Ranking: For each model, we first rank the acceleration methods according to their overall GenEval2 score $R_Q(m)$ and their compute reduction $R_C(m)$. In both cases, Rank 1 denotes the best-performing method. We then define the *Quality-Efficiency Rank* as:

$$R_{\mathrm{QE}}(m) = \lambda R_Q(m) + (1-\lambda) R_C(m). \tag{12}$$

Rather than assigning equal importance to quality and computational efficiency, we use $\lambda = 0.40$. We assign a slightly larger weight to computational efficiency because the primary objective of the proposed method is to reduce redundant unconditional CFG evaluations while preserving the compositional behavior of the original guided trajectory. The $40/60$ of λ split therefore reflects the intended objective of the evaluation. Importantly, the same weighting is applied uniformly to all models and methods. The resulting score therefore provides a common measure for comparing the quality-efficiency trade-off across all models.

Table 2: Quality–efficiency ranking on GenEval2. Each cell reports Quality Rank / Compute Rank (Weighted Rank). Rank 1 denotes the best performance, and lower weighted rank is better.

Model	**Step-AG**	**THG**	**AG**	**DrAG-Spatial**	**DrAG-Spectral**
FLUX	1/4 (2.80)	5/3 (3.80)	3/2 (2.40)	2/5 (3.80)	**4/1 (2.20)**
Qwen	1/5 (3.40)	3/4 (3.60)	4/1 (2.20)	5/3 (3.80)	**2/2 (2.00)**
SD3	3/5 (4.20)	4/2 (2.80)	5/1 (2.60)	1/4 (2.80)	**2/3 (2.60)**
Average Weighted Rank	3.47	3.40	2.40	3.47	**2.27**

Table 2 provides a unified view of the quality-efficiency trade-off. DrAG-Spectral achieves the lowest weighted rank on FLUX.1-dev, Qwen-Image, and SD3, with scores of 2.20, 2.00, and 2.60, respectively, and the lowest average rank of 2.27. This reflects its ability to maintain competitive compositional quality while consistently reducing unconditional evaluations. On Qwen-Image, for example, it ranks second in both quality and compute reduction, yielding 2.00, compared with AG's first and fourth ranks, respectively. Notably, the trend is preserved also under equal weighting ($\lambda = 0.5$), where DrAG-Spectral achieves the lowest average rank of 2.33 over AG with avg. rank of 2.5.

The two DrAG variants provide complementary operating points. DrAG-Spatial offers more conservative gap reuse, particularly on FLUX.1-dev and SD3, while DrAG-Spectral enables more aggressive reuse by suppressing sensitivity to high-frequency velocity variations. Thus, spatial gating prioritizes conservative guidance-gap reuse, whereas spectral gating favors greater computational reduction when the low-frequency trajectory remains stable.

Adaptive Guidance Allocation. The main distinction between DrAG and fixed-schedule methods is how unconditional evaluations are allocated. While fixed-schedule methods predefine the steps at which the unconditional branch is evaluated, DrAG dynamically determines whether the cached guidance-gap remains reusable based on the observed conditional velocity drift. Figure 1 visualizes this adaptive allocation, showing more frequent unconditional evaluations during periods of larger velocity changes and greater guidance-gap reuse during stable regions. More details are provided in Appendix B.10.

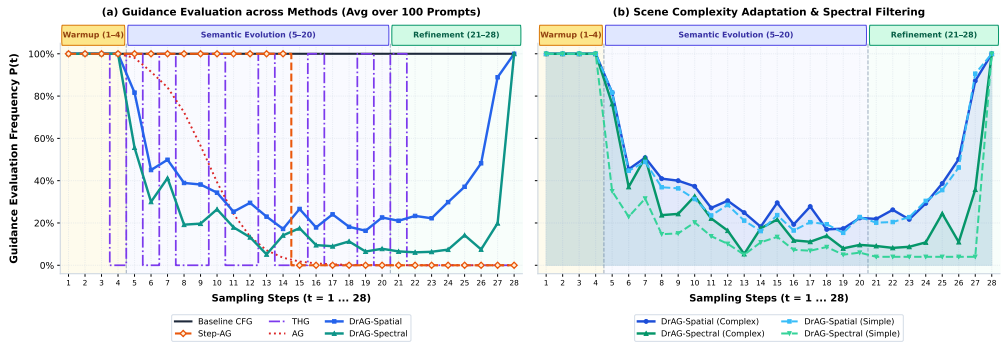

Figure 1: Empirical guidance evaluation frequency across sampling steps on the 100 GenEval-2 prompts using SD3 model. (a) Guidance evaluation frequency averaged across all prompts under identical sampling conditions ($N = 28$). (b) Scene-complexity analysis showing that DrAG dynamically allocates more guidance evaluations to complex multi-object scenes during semantic evolution (steps 5–20), while DrAG-Spectral suppresses the increase in late-step evaluations, observed with DrAG-Spatial during refinement.

Qualitative Results: Figure 2 presents qualitative comparisons across FLUX.1-dev, Qwen-Image, and SD3 for complex compositional prompts, where static or aggressive guidance skipping can affect fine-grained attributes. On SD3, for "Four Purple Lions", Baseline CFG, Step-AG, and AG generate only three lions, with AG also showing degraded facial details, whereas DrAG-Spatial and DrAG-Spectral retain guidance during the relevant generation stage and produce all four lions with clearer anatomical details. On Qwen-Image, for "Two Glass Suitcases", AG produces less transparent suitcases, while both DrAG variants preserve glass transparency and background visibility. On FLUX.1-dev, for "Five Blue Giraffes", AG and THG produce smoother outputs with less distinct spot patterns, whereas DrAG-Spectral preserves these fine-grained details despite substantially reduced NFE. These examples qualitatively illustrate that dynamic guidance-gap reuse reduces unconditional evaluations while preserving fine-grained attributes in compositional generation.

Additional Distributional Evaluation: Additional experiments on MS-COCO (Lin et al., 2014), including FID, KID, CMMD, and CLIP metrics, are reported in Appendix D.1. These results provide an additional evaluation of distributional fidelity and text alignment at a larger prompt set.

Generalization Studies: The main evaluation of DrAG focuses on image generation, where we provide comprehensive comparisons with existing guidance methods. To examine the transferability of the proposed guidance-gap reuse principle beyond this primary setting, we additionally evaluate DrAG without modifying its core mechanism across video generation, few-step generation, and score-based diffusion. For video generation, we apply DrAG-Spectral to Wan 2.1 (Wan et al., 2025) and HunyuanVideo (Kong et al., 2024). Using the hyperparameters selected from the ablations in Appendix D.4, DrAG-Spectral reduces NFE by 37.1% on Wan 2.1 and 36.8% on HunyuanVideo,

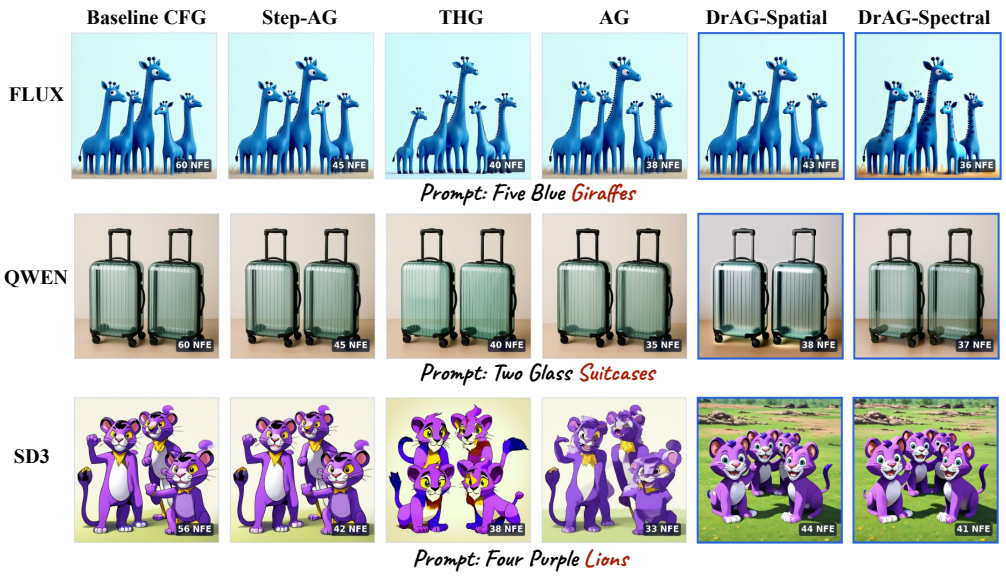

Figure 2: Qualitative comparison across text-to-image backbones: FLUX.1-dev (row 1), Qwen-Image (row 2), and Stable Diffusion 3 (row 3).

with total VBench scores changing from 0.6906 to 0.6874 and from 0.6465 to 0.6388, respectively (Appendix D.3, Table 20). We further evaluate guidance-gap reuse in the few-step regime using SDXL-Lightning (Lin et al., 2024) with its 8-step sampling setting. DrAG-Spatial and DrAG-Spectral reduce average NFE by 31.3% and 30.3%, respectively, while maintaining metrics close to baseline CFG (Appendix B.5, Table 6). Finally, we evaluate DrAG on Playground v2.5 (Li et al., 2024), an EDM-based score model (Karras et al., 2022) using DPM-Solver++ (Lu et al., 2023). DrAG-Spatial and DrAG-Spectral reduce average NFE by 18.9% and 20.9%, respectively, with DrAG-Spectral reducing NFE from 50.00 to 39.56 while achieving an FID of 26.48 compared with 27.20 for baseline CFG (Appendix B.6, Table 8). These results provide transferability evidence across substantially different generative settings, without introducing task-specific modifications to the core DrAG formulation.

Extension to One-Step Generative Models. We further investigate DrAG-Spectral beyond iterative diffusion and flow-matching models to a one-step generative model namely Generative Modeling via Drifting (Deng et al., 2026), where CFG redundancy arises during training rather than inference. By monitoring low-frequency spectral changes and reusing cached features, DrAG-Spectral skips redundant negative feature-encoder evaluations, reducing them by 57.2% and training latency by 42.3%, while achieving the comparable generation quality. Detailed implementation and results are provided in Appendix C.

5 Conclusion

We presented Dynamic Drift-Aware Guidance (DrAG), a training-free framework for reducing the computational cost of classifier-free guidance by dynamically reusing the conditional-unconditional guidance-gap. We introduced two variants: DrAG-Spatial, which measures directional and magnitude changes in velocity space, and DrAG-Spectral, which uses frequency-weighted drift to reduce sensitivity to high-frequency variations. Experiments across text-to-image models show that DrAG substantially reduces guidance evaluations while preserving generation quality. We further demonstrate its applicability to text-to-video, few-step and noise/score-based generative models, and extend the same principle to training in One-Step Generative Modeling via Drifting, where caching unconditional feature representations reduces training-time computation while maintaining comparable generation quality. Overall, DrAG provides a simple, theoretically motivated approach for dynamically allocating guidance computation according to the evolution of the generative process.

AI USE STATEMENT

In this work, we used generative AI tools to assist with debugging code files during the setup of code pipelines. Additionally, AI tools were used to correct grammatical errors while writing the paper. All AI-assisted contributions, including code corrections and textual suggestions, were thoroughly reviewed by the authors. We take full responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we have provided an anonymous code repository link in the abstract for review. In addition, we include detailed pseudocode in the supplement to facilitate replication of our methods and results.

REFERENCES

Angela Castillo, Jonas Kohler, Juan C Pérez, Juan Pablo Pérez, Albert Pumarola, Bernard Ghanem, Pablo Arbeláez, and Ali Thabet. Adaptive guidance: Training-free acceleration of conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1962–1970, 2025.

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*, 2025. URL https://openreview.net/forum?id=E77uvbOTtp.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

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

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models, 2025. URL https://arxiv.org/abs/2503.18886.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Amita Kamath, Kai-Wei Chang, Ranjay Krishna, Luke Zettlemoyer, Yushi Hu, and Marjan Ghazvininejad. Geneval 2: Addressing benchmark drift in text-to-image evaluation. *arXiv preprint arXiv:2512.16853*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=k7FuTOWMOc7.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

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 *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nAIhvNy15T.

Yunghee Lee, Byeonghyun Pak, Junwha Hong, and Hoseong Kim. Tortoise and hare guidance: Accelerating diffusion model inference with multirate integration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=3cYcUmcDhU.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Runlong Liao, Jian Yu, Baiyu Su, Chi Zhang, Lizhang Chen, and Qiang Liu. Momentum guidance: Plug-and-play guidance for flow models. 2026. URL https://arxiv.org/abs/2602.20360.

Sungbin Lim, Eunbi Yoon, Taehyun Byun, Taewon Kang, Seungwoo Kim, Kyungjae Lee, and Sungjoon Choi. Score-based generative modeling through stochastic evolution equations in hilbert spaces. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

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

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.

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 *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc., 2022. ISBN 9781713871088.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-solver++: Fast solver for guided sampling of diffusion probabilistic models, 2023. URL https://openreview.net/forum?id=4vGwQqviud5.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference, 2023.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014. URL https://arxiv.org/abs/1411.1784.

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 *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

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*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022. doi: 10.52202/068431-2643. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song and Stefano Ermon. *Generative modeling by estimating gradients of the data distribution*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, 2023.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Hanyang Wang, Yiyang Liu, Jiawei Chi, Fangfu Liu, Ran Xue, and Yueqi Duan. Cfg-ctrl: Control-based classifier-free diffusion guidance, 2026. URL https://arxiv.org/abs/2603.03281.

Ke Xu, Xin Yang, Baocai Yin, and Rynson W.H. Lau. Learning to restore low-light images via decomposition-and-enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024.

Huixuan Zhang and Xiaojun Wan. How much to guide: Revisiting adaptive guidance in classifier-free guidance text-to-vision diffusion models. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pp. 1–8, 2025.

Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: a unified predictor-corrector framework for fast sampling of diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, 2023.

Jianbin Zheng, Minghui Hu, Zhongyi Fan, Chaoyue Wang, Changxing Ding, Dacheng Tao, and Tat-Jen Cham. Trajectory consistency distillation, 2024.

A Additional Details on Dynamic Drifting-Aware Guidance

A.1 Proof of Proposition 1

From Eq.4 and Eq.5, the reuse error is exactly

$$\hat{v}_{CFG}(t_i) - v_{CFG}(t_i) = (s-1)\big[G(t_a) - G(t_i)\big], \tag{13}$$

which gives Eq.8 in the main paper after taking norms.

To relate this to the observable conditional drift Δ_c, recall from Eq.7 that $G(t_i) - G(t_a) = \Delta_c - \Delta_u$. By the triangle inequality, for L2 norm $\|\cdot\|_2$,

$$\|G(t_i) - G(t_a)\|_2 \le \|\Delta_c\|_2 + \|\Delta_u\|_2. \tag{14}$$

Applying Assumption 1, $\|\Delta_u\|_2 \le \kappa\|\Delta_c\|_2$, gives

$$\|G(t_i) - G(t_a)\|_2 \le (1+\kappa)\|\Delta_c\|_2, \tag{15}$$

and combining this with Eq.8 yields Proposition 1:

$$\|\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\|_2 \le (s-1)(1+\kappa)\|\Delta_c\|_2. \tag{16}$$

DrAG-Spatial and DrAG-Spectral instantiate Proposition 1.

A.2 DrAG-Spatial

Let

$$r_i = \frac{\|v_c(t_i)\|_2}{\|v_c(t_a)\|_2}, \tag{17}$$

and let θ_i denote the angle between $v_c(t_i)$ and $v_c(t_a)$.

Then, by the definition of the angle between two vectors,

$$\langle v_c(t_i), v_c(t_a)\rangle = \|v_c(t_i)\|_2\|v_c(t_a)\|_2\cos\theta_i. \tag{18}$$

Moreover, from 3.3,

$$\cos\theta_i = S_{\text{ang}}(t_i). \tag{19}$$

Expanding the squared norm of the difference $v_c(t_i) - v_c(t_a)$,

$$\|v_c(t_i) - v_c(t_a)\|_2^2 = \|v_c(t_i)\|_2^2 - 2\langle v_c(t_i), v_c(t_a)\rangle + \|v_c(t_a)\|_2^2. \tag{20}$$

Substituting $\langle v_c(t_i), v_c(t_a)\rangle = \|v_c(t_i)\|_2\|v_c(t_a)\|_2\cos\theta_i$ and dividing both sides by $\|v_c(t_a)\|_2^2$ gives

$$\frac{\|v_c(t_i) - v_c(t_a)\|_2^2}{\|v_c(t_a)\|_2^2} = r_i^2 - 2r_i\cos\theta_i + 1. \tag{21}$$

Adding and subtracting $2r_i$ on the right-hand side separates this into a magnitude term and a directional term:

$$r_i^2 + 1 - 2r_i\cos\theta_i = \underbrace{(r_i^2 - 2r_i + 1)}_{(r_i-1)^2} + \underbrace{2r_i - 2r_i\cos\theta_i}_{2r_i(1-\cos\theta_i)}.$$

Hence Eq.21 is equivalent to

$$\begin{aligned}\frac{\|v_c(t_i) - v_c(t_a)\|_2^2}{\|v_c(t_a)\|_2^2} &= r_i^2 + 1 - 2r_i\cos\theta_i \\ &= (r_i-1)^2 + 2r_i(1-\cos\theta_i).\end{aligned} \tag{22}$$

By this rearrangement, the relative conditional drift splits into a magnitude term $(r_i-1)^2$ and a directional term $2r_i(1-\cos\theta_i)$, matching the two quantities S_{mag} and S_{ang} that DrAG-Spatial monitors.

If the reuse condition of Eq.10 holds (the gap is *not* refreshed), then $|r_i - 1| \le \gamma_{\mathrm{sp}}$ and $\cos\theta_i \ge \tau_{\mathrm{sp}}$. Substituting these into Eq.22 bounds the conditional drift whenever the cached gap is reused:

$$\|v_c(t_i) - v_c(t_a)\|_2 \le \|v_c(t_a)\|_2 \sqrt{\gamma_{\mathrm{sp}}^2 + 2(1+\gamma_{\mathrm{sp}})(1-\tau_{\mathrm{sp}})}. \tag{23}$$

DrAG-Spatial corresponds to Assumption 1 and Proposition 1; write κ_{sp} for the resulting coupling constant. Substituting Eq.23 into Proposition 1 gives Corollary 1, stated in Section 3.3, which shows how the thresholds τ_{sp} and γ_{sp} directly control the worst-case reuse error whenever DrAG-Spatial reuses the cached gap.

$$\boxed{\left\|\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\right\|_2 \le (s-1)(1+\kappa_{\mathrm{sp}})\,\|v_c(t_a)\|_2 \sqrt{\gamma_{\mathrm{sp}}^2 + 2(1+\gamma_{\mathrm{sp}})(1-\tau_{\mathrm{sp}})}} \tag{24}$$

A.3 DRAG-SPECTRAL

We define the frequency-weighting operator

$$W_\sigma(x) := (\mathcal{M} \odot \mathcal{F})(x), \tag{25}$$

Define $z_i = W_\sigma(v_c(t_i))$, $z_a = W_\sigma(v_c(t_a))$, and

$$r_i^\sigma = \frac{\|z_i\|_2}{\|z_a\|_2}, \qquad \cos\theta_i^\sigma = \frac{\mathrm{Re}\langle z_i, z_a\rangle}{\|z_i\|_2\|z_a\|_2}. \tag{26}$$

By the same expansion used in Eq.22, the spectral drift of Eq.11 satisfies

$$\left(d_i^{\mathrm{spec}}\right)^2 = (r_i^\sigma - 1)^2 + 2r_i^\sigma(1 - \cos\theta_i^\sigma), \tag{27}$$

which is exactly the spatial decomposition of Eq.22 applied to z_i, z_a instead of $v_c(t_i), v_c(t_a)$.

To bound the reuse error, we refine Assumption 1 for the frequency-weighted setting.

Assumption 2 (Spectral Branch Coupling)**.** *Following Assumption 1, we assume the local branch-coupling condition also holds after applying the frequency-weighting operator W_σ (Eq.25): there exists a local constant $\kappa_\sigma \ge 0$ such that*

$$\|W_\sigma(\Delta_u)\|_2 \le \kappa_\sigma \|W_\sigma(\Delta_c)\|_2. \tag{28}$$

As in Assumption 1, this reflects that v_c and v_u come from the same underlying network and differ only in the conditioning input; since W_σ is a fixed linear operator applied identically to both branches, it does not introduce independent frequency-domain dynamics, so the two branches are expected to remain coupled after frequency weighting, with κ_σ capturing the local coupling strength in this weighted space.

Following the same steps as Appendix A.1 with $W = W_\sigma$, Assumption 2 gives

$$\|W_\sigma(G(t_i) - G(t_a))\|_2 \le (1+\kappa_\sigma)\|W_\sigma(\Delta_c)\|_2. \tag{29}$$

If the cached gap is reused, $d_i^{\mathrm{spec}} \le \gamma_{\mathrm{spec}}$,

so $\|W_\sigma(\Delta_c)\|_2 \le \gamma_{\mathrm{spec}}\|W_\sigma(v_c(t_a))\|_2$, and therefore

$$\|W_\sigma(G(t_i) - G(t_a))\|_2 \le (1+\kappa_\sigma)\,\gamma_{\mathrm{spec}}\,\|W_\sigma(v_c(t_a))\|_2. \tag{30}$$

Using Eq.8, Proposition 1 under Assumption 2 gives Corollary 2, stated in Section 3.3.

$$\boxed{\left\|W_\sigma\big(\hat{v}_{CFG}(t_i) - v_{CFG}(t_i)\big)\right\|_2 \le (s-1)(1+\kappa_\sigma)\,\gamma_{\mathrm{spec}}\,\|W_\sigma(v_c(t_a))\|_2} \tag{31}$$

A.4 FULL PSEUDOCODE

Algorithms 1 and 2 give the complete procedure for DrAG-Spatial and DrAG-Spectral, respectively.

Algorithm 1 DrAG-Spatial

Require: Pre-trained model v_θ, sampling points $\{t_i\}_{i=0}^{T-1}$, guidance scale s, angular threshold τ_{sp}, magnitude threshold γ_{sp}
Ensure: Generated sample x_0

Initialize x_{t_0}
$v_c^a \leftarrow v_\theta(x_{t_0}, t_0, c)$
$v_u^a \leftarrow v_\theta(x_{t_0}, t_0, \varnothing)$
$G^a \leftarrow v_c^a - v_u^a$
$v_{\text{cfg}} \leftarrow v_c^a + (s-1)G^a$
Update x_{t_0} using the sampler
for $i = 1, \ldots, T-1$ **do**
  $v_c^i \leftarrow v_\theta(x_{t_i}, t_i, c)$
  $S_{\text{ang}} \leftarrow \dfrac{\langle v_c^i, v_c^a \rangle}{\|v_c^i\|_2 \|v_c^a\|_2}$
  $S_{\text{mag}} \leftarrow \dfrac{\left|\|v_c^i\|_2 - \|v_c^a\|_2\right|}{\|v_c^a\|_2}$
  if $S_{\text{ang}} < \tau_{\text{sp}}$ **or** $S_{\text{mag}} > \gamma_{\text{sp}}$ **then**
    $v_u^i \leftarrow v_\theta(x_{t_i}, t_i, \varnothing)$
    $G^a \leftarrow v_c^i - v_u^i$
    $v_c^a \leftarrow v_c^i$
  end if
  $v_{\text{cfg}} \leftarrow v_c^i + (s-1)G^a$
  Update x_{t_i} using the sampler
end for
return x_0

Algorithm 2 DrAG-Guidance

Require: Pre-trained Model v_θ, sampling points $\{t_i\}_{i=0}^{T-1}$, guidance scale s, spectral operator W_σ, spectral threshold γ_{spec}
Ensure: Generated sample x_0

Initialize x_{t_0}
$v_c^a \leftarrow v_\theta(x_{t_0}, t_0, c)$
$v_u^a \leftarrow v_\theta(x_{t_0}, t_0, \varnothing)$
$G^a \leftarrow v_c^a - v_u^a$
$v_{\text{cfg}} \leftarrow v_c^a + (s-1)G^a$
Update x_{t_0} using the sampler
for $i = 1, \ldots, T-1$ **do**
  $v_c^i \leftarrow v_\theta(x_{t_i}, t_i, c)$
  $z_i \leftarrow W_\sigma(v_c^i)$
  $z_a \leftarrow W_\sigma(v_c^a)$
  $d_i^{\text{spec}} \leftarrow \dfrac{\|z_i - z_a\|_2}{\|z_a\|_2}$
  if $d_i^{\text{spec}} > \gamma_{\text{spec}}$ **then**
    $v_u^i \leftarrow v_\theta(x_{t_i}, t_i, \varnothing)$
    $G^a \leftarrow v_c^i - v_u^i$
    $v_c^a \leftarrow v_c^i$
  end if
  $v_{\text{cfg}} \leftarrow v_c^i + (s-1)G^a$
  Update x_{t_i} using the sampler
end for
return x_0

B Additional Experiments and Ablation Results

B.1 Theoretical Grounding: Bounded Conditional–Unconditional Drift Coupling

The cached guidance gap $G(t_a)$ remains valid as long as the conditional and unconditional velocity fields do not change substantially from the anchor timestep t_a to the current timestep t_i. We characterize this change through the velocity drift between the anchor and current timestep:

$$\Delta_c = \Delta_c(t_i, t_a) = v_c(t_i) - v_c(t_a), \qquad \Delta_u = \Delta_u(t_i, t_a) = v_u(t_i) - v_u(t_a). \tag{32}$$

In Assumption 1, DrAG assumes that, over an interval (t_a, t_i), the evolution of the unconditional velocity field remains coupled to the observable evolution of the conditional field:

$$\|\Delta_u(t_i, t_a)\|_2 \le \kappa \|\Delta_c(t_i, t_a)\|_2, \tag{33}$$

where κ denotes the conditional-unconditional drift coupling coefficient.

Under this coupling relationship, bounding the observable conditional drift directly bounds the unobserved unconditional drift. In particular, if

$$\|\Delta_c(t_i, t_a)\|_2 \le \delta, \tag{34}$$

then

$$\|\Delta_u(t_i, t_a)\|_2 \le \kappa\delta. \tag{35}$$

Thus, the observable conditional velocity drift provides a direct proxy for the evolution of the uncomputed unconditional field. This provides the theoretical motivation for using conditional drift to determine when the cached guidance gap should be retained or refreshed. The empirical validity and cross-architecture magnitude of κ are examined below.

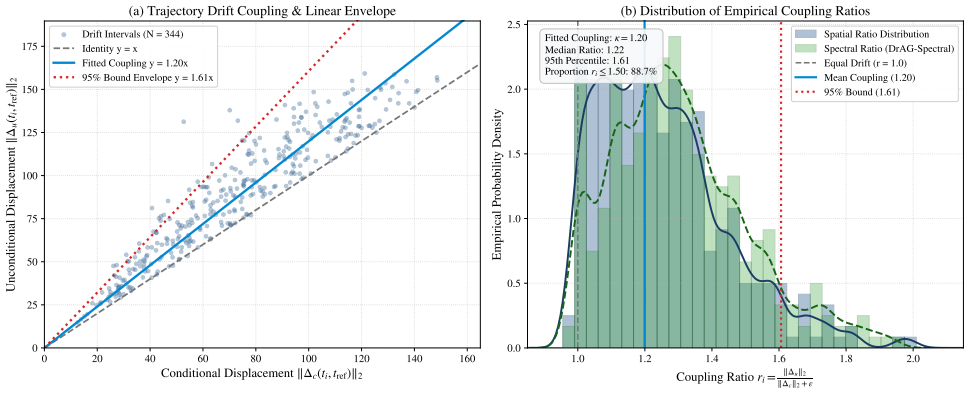

Figure 3: Empirical validation of conditional–unconditional drift coupling for Stable Diffusion 3 ($N = 344$ intervals). (a) Joint distribution of observable conditional drift $\|\Delta_c\|_2$ and uncomputed unconditional drift $\|\Delta_u\|_2$ across dynamic inference intervals, with fitted linear regression slope $\kappa_{\text{fit}} = 1.20$ and empirical 95th-percentile envelope $\kappa_{0.95} = 1.61$. (b) Empirical probability densities of the coupling ratio r_i comparing spatial velocity drift (solid blue) and spectral structural drift after Gaussian low-pass filtering (dashed green).

B.1.1 Cross-Architecture Empirical Validation

To empirically examine the conditional-unconditional drift coupling and characterize the corresponding value of κ, we conduct an offline diagnostic evaluation across three generative architectures: Stable Diffusion 3, FLUX.1-dev, and Qwen-Image. The evaluation comprises 1,090 discrete inference intervals collected using an identical set of 20 prompts and random seeds 42–61. During this offline evaluation, trajectories are governed by DrAG's adaptive skipping; at each bypassed step t_i relative to its anchor t_a, we evaluate the ground-truth unconditional velocity field $v_u(t_i)$ solely to measure the uncomputed drift Δ_u against the observable conditional drift Δ_c.

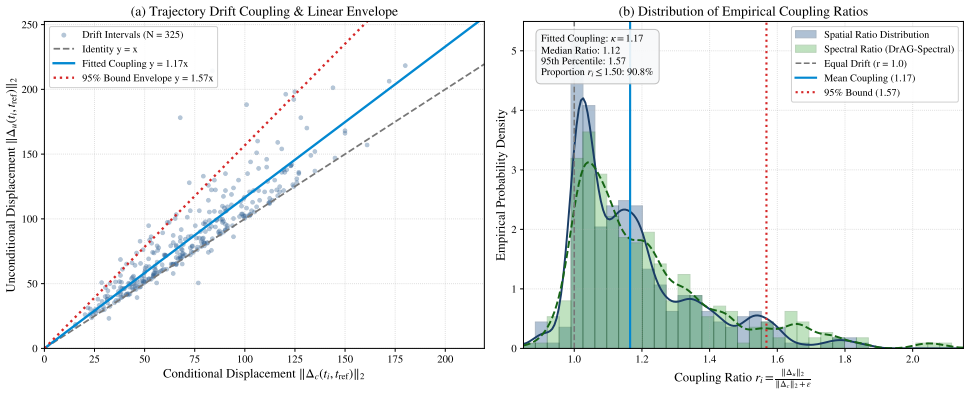

Figure 4: Empirical validation of conditional–unconditional drift coupling for FLUX.1-dev ($N = 325$ intervals). (a) Joint distribution of $\|\Delta_c\|_2$ versus $\|\Delta_u\|_2$ alongside fitted slope $\kappa_{\text{fit}} = 1.17$ and 95th-percentile envelope $\kappa_{0.95} = 1.57$. (b) Empirical distributions of the coupling ratio r_i for spatial drift (solid blue) and low-frequency spectral drift (dashed green).

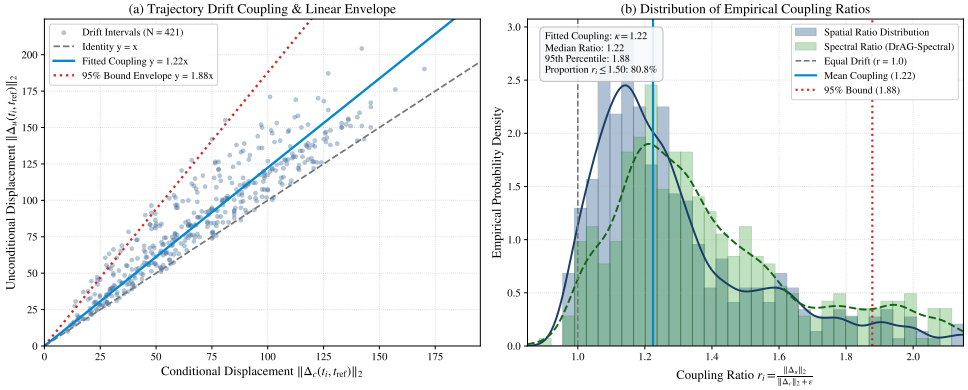

Figure 5: Empirical validation of conditional–unconditional drift coupling for Qwen-Image ($N = 421$ intervals). (a) Joint distribution of $\|\Delta_c\|_2$ versus $\|\Delta_u\|_2$ displaying fitted slope $\kappa_{\text{fit}} = 1.22$ and 95th-percentile envelope $\kappa_{0.95} = 1.88$. (b) Empirical distributions of the coupling ratio r_i for spatial drift (solid blue) and low-frequency spectral drift (dashed green).

To rigorously examine both spatial and spectral gating variants, we evaluate the coupling ratio under both formulations:

$$r_i^{\text{spatial}} = \frac{\|\Delta_u(t_i, t_a)\|_2}{\|\Delta_c(t_i, t_a)\|_2 + \epsilon}, \qquad r_i^{\text{spectral}} = \frac{\|W_\sigma(\Delta_u(t_i, t_a))\|_2}{\|W_\sigma(\Delta_c(t_i, t_a))\|_2 + \epsilon}, \tag{36}$$

where W_σ denotes the frequency-weighting operator defined in equation 25, which filters high-frequency noise and isolates structural manifold dynamics.

The diagnostic measurements are illustrated in Figures 3, 4, and 5. In each figure, panel (a) plots the joint distribution of $\|\Delta_c\|_2$ versus $\|\Delta_u\|_2$ across all sampled intervals, with the fitted regression line κ_{fit} and the empirical 95th-percentile envelope $\kappa_{0.95}$. Across all three architectures, the data points remain clustered along the diagonal, confirming that unconditional drift scales proportionally with conditional drift. Panel (b) depicts the empirical probability density functions of the coupling ratio r_i: the solid blue curve shows the spatial drift ratio distribution, while the dashed green curve shows the spectral drift ratio distribution. For all three backbones, low-frequency spectral filtering yields a concentrated distribution around $r_i \approx 1.0$, demonstrating that structural velocity drift exhibits even tighter coupling than raw spatial drift (validating the Assumption 2).

The quantitative statistics are summarized in Table 3. Across the three architectures, the fitted coupling coefficients remain within a remarkably narrow range: $\kappa_{\text{fit}} = 1.20$ for Stable Diffusion 3,

1.17 for FLUX.1-dev, and 1.22 for Qwen-Image. The aggregate fitted coefficient is approximately 1.20, while the empirical 95th-percentile bound remains below 1.71. These observations indicate that unconditional velocity displacement is of comparable magnitude to conditional displacement across all evaluated intervals.

Table 3: Empirical validation statistics for the drift coupling ratio $r_i = \|\Delta_u\|_2/(\|\Delta_c\|_2 + \epsilon)$ across 1,090 discrete inference intervals evaluated on identical prompt distributions (20 shared prompts, seeds 42–61).

Model Backbone	Intervals (N)	Mean Ratio	Median (p_{50})	Fitted Slope κ_{fit}	95% Bound ($\kappa_{0.95}$)	$r_i \leq 1.50$
Stable Diffusion 3	344	1.26	1.22	1.20	1.61	88.7%
FLUX.1-dev	325	1.18	1.12	1.17	1.57	90.8%
Qwen-Image	421	1.29	1.22	1.22	1.88	80.8%
Aggregate	1,090	1.25	1.19	≈ 1.20	≤ 1.71	86.2%

The coupling-ratio statistics exhibit a consistent pattern across all backbones. The mean ratio ranges from 1.18 to 1.29, while the median ranges from 1.12 to 1.22. The unconditional displacement is moderately larger than the conditional displacement in a fraction of observed intervals, which supports adopting a conservative coupling coefficient $\kappa > 1$ rather than assuming identical magnitudes. Overall, 86.2% of the measured intervals satisfy $r_i \leq 1.50$, with 88.7% for Stable Diffusion 3, 90.8% for FLUX.1-dev, and 80.8% for Qwen-Image. Together with the fitted slopes and percentile envelopes, these results provide empirical support for the local conditional-unconditional drift coupling used in Eq. 33.

The total number of measured intervals differs across architectures despite using identical prompts and seeds (344, 325, and 421 intervals for Stable Diffusion 3, FLUX.1-dev, and Qwen-Image, respectively). This variation is an expected consequence of differing default step budgets ($T = 28$ for SD3 versus $T = 30$ for FLUX and Qwen) combined with DrAG's adaptive trajectory monitoring.

In summary, the cross-architecture empirical analysis supports the core premise of Assumption 1 and 2, introduced in DrAG: observable changes in the conditional velocity field remain boundedly coupled with changes in the unconditional field across both spatial and spectral domains. This coupling validates conditional drift as an effective proxy for deciding when the cached guidance gap can be safely retained or refreshed.

B.2 Large-Scale Generation Evaluation

To further evaluate the stability of our results at a larger generation scale, we conduct an additional experiment using 50,000 MS-COCO prompts with SD3 model. The same sampling configuration and guidance scale ($s = 7.5$) are used for all methods. The quantitative results are reported in Table 4.

Table 4: Quantitative results on 50K MS-COCO samples using Stable Diffusion 3.

Model	**Method**	**CFG Scale**	**FID (↓)**	**KID (↓)**	**CMMD (↓)**	**CLIP (↑)**	**NFE (↓)**	**Compute Reduction (↑)**
SD3	Baseline CFG	7.5	24.11	0.0084	0.0006	0.7922	56.00	0.00%
SD3	DrAG-Spatial	7.5	23.86	0.0082	0.0006	0.7915	40.62	27.46%
SD3	DrAG-Spectral	7.5	23.60	0.0077	0.0006	0.7903	37.94	32.25%

The large-scale evaluation follows the same trend observed in our main experiments. Compared with baseline CFG, both DrAG variants reduce the number of function evaluations while maintaining competitive generation quality. DrAG-Spectral achieves the lowest FID of 23.60 and KID of 0.0077, while reducing the inference compute by 32.25%. DrAG also achieves a 27.46% compute reduction with an FID of 23.86. These results on 50K generated samples further support the stability of the proposed guidance-gap reuse strategy at a larger evaluation scale.

B.3 Longer Sampling Schedules

We further evaluate DrAG with a longer 50-step sampling schedule using 500 generated images per configuration. As shown in Table 5, DrAG continues to reduce the number of guidance evaluations

while maintaining comparable generation quality. On FLUX, DrAG-Spectral reduces NFE by 40.1% with a slight improvement in FID and KID. On Qwen, DrAG-Spatial achieves a 38.2% reduction in NFE while improving FID from 105.23 to 100.11. On SD3, DrAG-Spectral reduces NFE by 36.8% and improves FID, KID, and CMMD over the baseline.

Table 5: Fifty-step generation results on 500 images.

Model	**Method**	**CFG Scale**	**FID (↓)**	**KID (↓)**	**CMMD (↓)**	**CLIP (↑)**	**NFE (↓)**	**Compute Reduction (↑)**
FLUX	Baseline CFG	2.0	92.54	0.015	0.0010	0.784	100.0	0.0%
FLUX	DrAG-Spatial	2.0	92.35	0.015	0.0010	0.783	69.7	30.3%
FLUX	DrAG-Spectral	2.0	92.27	0.014	0.0010	0.783	59.9	40.1%
Qwen	Baseline CFG	4.0	105.23	0.023	0.0012	0.787	100.0	0.0%
Qwen	DrAG-Spatial	4.0	100.11	0.018	0.0011	0.785	61.8	38.2%
Qwen	DrAG-Spectral	4.0	101.44	0.020	0.0011	0.787	60.0	40.0%
SD3	Baseline CFG	7.5	97.08	0.017	0.0008	0.793	100.0	0.0%
SD3	DrAG-Spatial	7.5	95.96	0.016	0.0008	0.794	66.9	33.1%
SD3	DrAG-Spectral	7.5	94.71	0.015	0.0007	0.792	63.2	36.8%

B.4 Experiments on Few-Step Models(SDXL-Lightning)

To demonstrate the impact of our method on few-step generative models, we evaluate it using SDXL-Lightning (Lin et al., 2024). We first conduct an ablation study to select suitable hyper parameters, Table 7), followed by evaluation using MS-COCO dataset (Lin et al., 2014). The results are presented in Table 6. Compared with the standard CFG, both DrAG-Spatial and DrAG-Spectral substantially reduce the average NFE, yielding up to 31.3% compute reduction while maintaining comparable generation quality across all scores.

Table 6: Quantitative Results on MS-COCO 5K-Prompts with SDXL-Lightning (8-Step Generation)

Method	Hyperparameters	FID (↓)	KID (↓)	CMMD (↓)	CLIP (↑)	Avg NFE (↓)	Compute Reduction (↑)
Baseline CFG	Standard (8 steps)	26.71	0.0055	0.00080	0.7899	16.00	0.0%
DrAG-Spatial	$\tau_{\text{sp}} = 0.90,\ \gamma_{\text{sp}} = 0.20$	27.15	0.0057	0.00077	0.7852	**10.98**	**31.3**%
DrAG-Spectral	$\sigma = 20.0,\ \gamma_{\text{spec}} = 0.40$	27.14	0.0058	0.00077	0.7853	<u>11.14</u>	<u>30.3</u>%

B.5 Ablation on Few-Step Models(SDXL-Lightning)

Table 7 studies the hyperparameter sensitivity of DrAG on SDXL-Lightning using an 8-step distilled sampling schedule (16.0 baseline NFE). Since few-step models take larger steps along the trajectory, they have less room for error. Very strict thresholds, such as $\gamma \leq 0.10$ for spectral gating, trigger guidance at most steps and provide less than 3% compute savings. In contrast, relaxed drift thresholds allow DrAG to skip redundant evaluations while maintaining the sampling trajectory.

For DrAG-Spatial, $\tau = 0.90$ and $\gamma = 0.20$ provide a good balance between quality and efficiency, reducing NFE to 11.0 (31.2% reduction). At the same time, KID improves from 0.0058 to 0.0039, while CLIP remains at 0.8035. For DrAG-Spectral, $\sigma = 20.0$ and $\gamma = 0.40$ achieve the same 31.2% compute reduction with a slightly lower KID of 0.0037. We use these two configurations for the full 5,000-prompt MS-COCO evaluation in Table 6, where they achieve 31.3% and 30.3% compute reductions, respectively, while maintaining comparable quality to baseline CFG.

Table 7: Ablation Analysis on SDXL-Lightning across Gating Thresholds

Configuration	FID (↓)	KID (↓)	CMMD (↓)	CLIP Score (↑)	Avg NFE (↓)	Compute Reduction (↑)
Baseline CFG (8 Steps)	294.25	0.0058	0.0009	0.8072	16.0	0.0%
DrAG-Spatial ($\tau = 0.99, \gamma = 0.05$)	298.65	0.0048	0.0009	0.8061	12.9	19.4%
DrAG-Spatial ($\tau = 0.99, \gamma = 0.10$)	297.01	0.0046	0.0009	0.8028	12.9	19.4%
DrAG-Spatial ($\tau = 0.99, \gamma = 0.20$)	297.01	0.0040	0.0009	0.8028	12.9	19.4%
DrAG-Spatial ($\tau = 0.95, \gamma = 0.05$)	298.71	0.0050	0.0009	0.8065	12.4	22.5%
DrAG-Spatial ($\tau = 0.98, \gamma = 0.05$)	298.84	0.0056	0.0009	0.8064	12.4	22.5%
DrAG-Spatial ($\tau = 0.98, \gamma = 0.10$)	296.73	0.0036	0.0009	0.8026	12.0	25.0%
DrAG-Spatial ($\tau = 0.98, \gamma = 0.20$)	296.73	0.0040	0.0009	0.8026	12.0	25.0%
DrAG-Spatial ($\tau = 0.95, \gamma = 0.10$)	299.52	0.0048	0.0009	0.8037	11.6	27.5%
DrAG-Spatial ($\tau = 0.95, \gamma = 0.20$)	297.82	0.0042	0.0009	0.8025	11.3	29.4%
DrAG-Spatial ($\tau = 0.95, \gamma = 0.30$)	297.82	0.0037	0.0009	0.8025	11.3	29.4%
DrAG-Spatial ($\tau = 0.95, \gamma = 0.40$)	297.82	0.0034	0.0009	0.8025	11.3	29.4%
DrAG-Spatial ($\tau = 0.90, \gamma = 0.20$)	297.57	0.0039	0.0009	0.8035	11.0	31.2%
DrAG-Spatial ($\tau = 0.90, \gamma = 0.30$)	297.57	0.0040	0.0009	0.8035	11.0	31.2%
DrAG-Spatial ($\tau = 0.90, \gamma = 0.40$)	297.57	0.0037	0.0009	0.8035	11.0	31.2%
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.05$)	294.98	0.0058	0.0009	0.8071	16.0	0.0%
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.05$)	294.98	0.0073	0.0009	0.8071	16.0	0.0%
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.10$)	294.19	0.0062	0.0009	0.8071	15.7	1.9%
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.10$)	294.15	0.0056	0.0009	0.8073	15.6	2.5%
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.20$)	295.25	0.0044	0.0010	0.8032	13.7	14.4%
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.20$)	296.26	0.0020	0.0009	0.8029	13.4	16.2%
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.20$)	297.43	0.0036	0.0009	0.8069	12.9	19.4%
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.20$)	295.01	0.0045	0.0009	0.8046	12.6	21.3%
DrAG-Spectral ($\sigma = 50.0, \gamma = 0.20$)	298.04	0.0043	0.0009	0.8024	12.2	23.8%
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.30$)	298.71	0.0027	0.0009	0.8038	12.0	25.0%
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.30$)	299.98	0.0032	0.0009	0.8043	11.9	25.6%
DrAG-Spectral ($\sigma = 50.0, \gamma = 0.30$)	299.51	0.0036	0.0009	0.8033	11.7	26.9%
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.40$)	297.88	0.0043	0.0009	0.8029	11.2	30.0%
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.40$)	297.53	0.0037	0.0009	0.8032	11.0	31.2%
DrAG-Spectral ($\sigma = 50.0, \gamma = 0.40$)	297.57	0.0045	0.0009	0.8035	11.0	31.2%

Further, Figure 6 shows the visual superiority of DrAG-Spectral under an exact iso-compute constraint (16 NFE) on SDXL-Lightning. In few-step distillation, standard CFG is strictly limited to

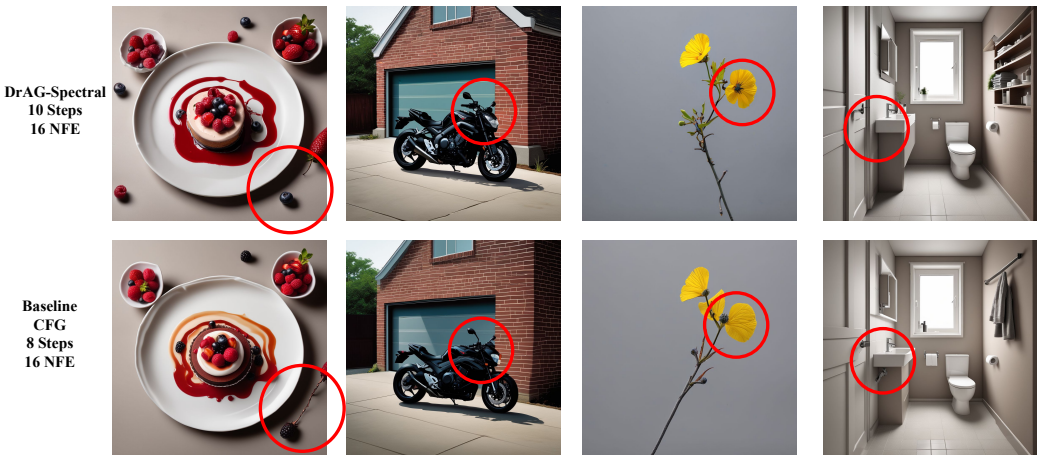

Figure 6: Qualitative Iso-Compute comparison on SDXL-Lightning under an identical budget of 16 NFE.

8 integration steps (2 NFE per step), forcing large discrete numerical jumps that induce geometric warping and structural collapse. By dynamically skipping redundant unconditional evaluations, DrAG-Spectral reallocates the saved compute budget to execute 10 discrete solver steps within the identical 16 NFE budget. As highlighted in the red annotations, this additional integration granularity delivers marked structural improvements as shown in the Figure 6.

B.6 Generalization to Score-Based Diffusion Models(Playground v2.5)

Unlike flow-matching models, which typically follow near-linear probability paths, continuous-time score-based models follow curved ODE trajectories due to their non-linear noise schedules. To test whether our method generalizes beyond rectified flows and first-order Euler solvers, we evaluate it on Playground v2.5 (Li et al., 2024), which uses the EDM formulation (Karras et al., 2022) together with the higher-order DPM-Solver++ (Lu et al., 2023).

Table 8: Quantitative Results on MS-COCO 5K-Prompts with Playground v2.5

Model	Guidance Method	Avg. NFE ↓	Compute Reduction ↑	FID ↓	KID ($\times 10^3$) ↓	CMMD ↓	CLIP ↑
Playground v2.5	Baseline CFG	50.00	0.0%	27.20	6.28	0.00098	0.7775
Playground v2.5	DrAG-Spatial	40.54	18.9%	26.77	6.14	0.00093	**0.7780**
Playground v2.5	**DrAG-Spectral**	**39.56**	**20.9%**	**26.48**	**6.05**	**0.00086**	0.7760

As shown in Table 8 on MS-COCO, DrAG-Spatial and DrAG-Spectral reduce the computation by 18.9% (40.54 NFE) and 20.9% (39.56 NFE), respectively. Interestingly, reducing the number of evaluations also improves the distributional metrics. DrAG-Spectral reduces FID from 27.20 to 26.48, lowers KID to 0.00605, and improves CMMD by 12.2% (from 0.00098 to 0.00086), while maintaining similar text alignment (CLIP ≈ 0.776). These results show that unconditional evaluations can still be redundant in curved EDM trajectories, and our DrAG can help reduce over-guidance while saving computation.

B.7 Ablation on Score-Based Models(Playground v2.5)

Table 9 evaluates DrAG on Playground v2.5 using 25 inference steps and a baseline of 50.0 NFE. We study the sensitivity to the main hyperparameters under the EDM score-based sampling schedule. Unlike linear flow-matching trajectories, score-based diffusion follows a more curved trajectory. Very strict thresholds, such as $\tau_{\text{sp}} \geq 0.98$, trigger guidance too often and provide less than 9% compute savings. Relaxing the drift threshold allows DrAG to skip more redundant evaluations while maintaining sample quality.

For DrAG-Spatial, $\tau_{\text{sp}} = 0.93$ and $\gamma_{\text{sp}} = 0.15$ provide a good balance between quality and efficiency, reducing NFE to 40.64 (18.7% reduction) while improving FID from 204.71 to 201.61 and achieving a CLIP score of 0.7833. For DrAG-Spectral, $\sigma = 1.0$ and $\gamma_{\text{spec}} = 0.20$ give the best overall results, reducing NFE to 39.60 (20.8% reduction) with the lowest FID (200.48) and CMMD (0.00098). We use these configurations for the full 5,000-prompt evaluation in Table 8, where they achieve 18.9% and 20.9% NFE reductions, respectively, while maintaining competitive distributional quality. These quantitative improvements are also reflected in the qualitative results shown in Figure 7, where the proposed configurations preserve visual quality while reducing the number of function evaluations.

Table 9: Ablation Analysis on Playground v2.5 (50 Prompts) across Gating Configurations

Configuration	Avg NFE (↓)	Compute Reduction (↑)	CLIP Score (↑)	FID (↓)	KID (↓)	CMMD (↓)
Baseline CFG (50 Steps)	50.00	0.0%	0.7764	204.71	0.00908	0.00115
DrAG-Spatial ($\tau = 0.99$)	47.12	5.8%	0.7754	203.89	0.00731	0.00114
DrAG-Spatial ($\tau = 0.99, \gamma = 0.10$)	47.16	5.7%	0.7760	203.39	0.00767	0.00115
DrAG-Spatial ($\tau = 0.99, \gamma = 0.15$)	47.12	5.8%	0.7754	203.89	0.00763	0.00114
DrAG-Spatial ($\tau = 0.99, \gamma = 0.20$)	47.12	5.8%	0.7754	203.88	0.00814	0.00114
DrAG-Spatial ($\tau = 0.98$)	45.74	8.5%	0.7761	203.05	0.00836	0.00114
DrAG-Spatial ($\tau = 0.98, \gamma = 0.10$)	45.86	8.3%	0.7747	202.15	0.00786	0.00114
DrAG-Spatial ($\tau = 0.98, \gamma = 0.15$)	45.74	8.5%	0.7761	203.05	0.00797	0.00114
DrAG-Spatial ($\tau = 0.98, \gamma = 0.20$)	45.74	8.5%	0.7761	203.18	0.00799	0.00114
DrAG-Spatial ($\tau = 0.97$)	43.96	12.1%	0.7808	204.74	0.00953	0.00110
DrAG-Spatial ($\tau = 0.97, \gamma = 0.10$)	44.86	10.3%	0.7788	203.24	0.00837	0.00110
DrAG-Spatial ($\tau = 0.97, \gamma = 0.15$)	43.96	12.1%	0.7808	204.74	0.00940	0.00110
DrAG-Spatial ($\tau = 0.97, \gamma = 0.20$)	43.68	12.6%	0.7808	204.71	0.00965	0.00110
DrAG-Spatial ($\tau = 0.95$)	41.64	16.7%	0.7800	204.10	0.00848	0.00110
DrAG-Spatial ($\tau = 0.95, \gamma = 0.10$)	42.56	14.9%	0.7781	204.55	0.00849	0.00111
DrAG-Spatial ($\tau = 0.95, \gamma = 0.15$)	41.64	16.7%	0.7800	204.10	0.00836	0.00110
DrAG-Spatial ($\tau = 0.95, \gamma = 0.20$)	41.54	16.9%	0.7797	204.77	0.00843	0.00109
DrAG-Spatial ($\tau = 0.93$)	40.64	18.7%	0.7833	201.61	0.00877	0.00108
DrAG-Spatial ($\tau = 0.93, \gamma = 0.10$)	41.30	17.4%	0.7791	202.70	0.00896	0.00108
DrAG-Spatial ($\tau = 0.93, \gamma = 0.15$)	40.64	18.7%	0.7833	201.61	0.00869	0.00108
DrAG-Spatial ($\tau = 0.93, \gamma = 0.20$)	40.36	19.3%	0.7788	203.10	0.00815	0.00109
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.08$)	44.26	11.5%	0.7835	205.18	0.00833	0.00106
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.08$)	44.80	10.4%	0.7781	204.86	0.00837	0.00111
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.08$)	45.02	10.0%	0.7776	203.96	0.00815	0.00111
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.10$)	43.04	13.9%	0.7802	207.86	0.00930	0.00108
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.10$)	43.32	13.4%	0.7816	205.73	0.00864	0.00107
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.10$)	43.52	13.0%	0.7780	205.10	0.00845	0.00109
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.12$)	42.30	15.4%	0.7804	206.47	0.00891	0.00106
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.12$)	42.66	14.7%	0.7782	208.16	0.00904	0.00109
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.12$)	42.74	14.5%	0.7815	206.69	0.00933	0.00109
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.15$)	41.18	17.6%	0.7844	201.29	0.00895	0.00102
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.15$)	41.54	16.9%	0.7819	201.89	0.00748	0.00106
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.15$)	41.78	16.4%	0.7826	203.41	0.00734	0.00105
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.20$)	39.60	20.8%	0.7803	200.48	0.00935	0.00098
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.20$)	39.94	20.1%	0.7795	202.50	0.00987	0.00101
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.20$)	40.08	19.8%	0.7785	202.63	0.00944	0.00102



Figure 7: Qualitative Results of Playground v2.5

B.8 Trajectory Stabilization via Momentum Guidance and Sliding Mode Control

While Dynamic Drift-Aware Guidance (DrAG) adaptively skips conditional evaluations based on spatial and spectral divergence, alternating between guided updates and unconditional extrapolations may introduce high-frequency perturbations in the velocity trajectory. We therefore evaluate two complementary trajectory-stabilization mechanisms:

1. **Momentum Guidance (MG) (Liao et al., 2026):** Smooths consecutive guided velocities using $m_t = \beta m_{t-1} + (1-\beta)v_t^{\text{guided}}$.
2. **Sliding Mode Control (SMC-CFG) (Wang et al., 2026):** Suppresses trajectory chattering through the sliding manifold $s(t) = (\epsilon_t^{\text{guided}} - \epsilon_{t-1}^{\text{guided}}) + \lambda\epsilon_{t-1}^{\text{guided}}$ and switching control $u_{\text{sw}} = -K\,\text{sign}(s)$.

Table 10 reports results on 500 MS-COCO validation samples. Two distinct regimes emerge. The effectiveness of the stabilization mechanism depends on the guidance regime. For linear flow models under low guidance scales, MG consistently improves distribution fidelity with negligible computational overhead: on FLUX, it reduces FID from 93.66 to 92.12 and KID from 14.55 to 13.70×10^{-3}, while improving CLIP from 0.7771 to 0.7789 with nearly unchanged NFE ($36.59 \rightarrow 36.79$); Qwen exhibits the same trend, with MG improving all four quality metrics. In contrast, under the high-guidance regime of SD3 Medium ($w = 7.5$), SMC provides stronger stabilization, achieving the best FID (91.94), KID (13.87×10^{-3}), CMMD (0.639×10^{-3}), and CLIP (0.7985), albeit with higher NFE (44.06). Overall, MG is effective as a lightweight smoother for flow-matching models, while SMC is better suited to high-guidance dynamics.

Overall, both mechanisms can be combined with DrAG-Spectral, but their benefits are regime-dependent: MG provides lightweight velocity smoothing for flow-matching models, whereas SMC offers better stabilization under high-guidance dynamics.

Table 10: Ablation of Trajectory Stabilization Mechanisms on MS-COCO (500 Samples)

Model	Stabilization Variant	Avg NFE (↓)	FID (↓)	KID ($\times10^3$, ↓)	CMMD ($\times10^3$, ↓)	CLIP Score (↑)
FLUX.1-dev	DrAG-Spectral (Base)	36.59	93.66	14.55	0.967	0.7771
	+Momentum Guidance (MG)	36.79	**92.12**	**13.70**	**0.958**	0.7789
	+Sliding Mode Control (SMC)	36.79	93.04	15.03	0.979	**0.7801**
Qwen-Image	DrAG-Spectral (Base)	36.44	99.50	19.44	1.118	0.7869
	+Momentum Guidance (MG)	37.01	**98.59**	**18.01**	**1.090**	**0.7877**
	+Sliding Mode Control (SMC)	38.30	100.12	19.84	1.145	0.7869
SD3 Medium	DrAG-Spectral (Base)	37.91	92.72	14.09	0.770	0.7919
	+Momentum Guidance (MG)	38.03	93.38	14.14	0.741	0.7891
	+Sliding Mode Control (SMC)	44.06	**91.94**	**13.87**	**0.639**	**0.7985**

B.9 Analysis of Generation Diversity

To examine whether compute skipping affects generation diversity, we evaluate the perceptual diversity of samples generated from different random seeds. We use Learned Perceptual Image Patch Similarity (LPIPS) to measure the pairwise perceptual distance between samples generated from the same prompt.

We consider five diverse semantic categories and generate ten samples using different random seeds for each prompt. The average pairwise LPIPS is computed as

$$\text{Diversity}(p) = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} \text{LPIPS}(I_{p,i}, I_{p,j}), \tag{37}$$

where $S = 10$, resulting in 45 pairwise comparisons per prompt. Higher values indicate greater perceptual variation across different random seeds.

Table 11: Generation diversity measured using pairwise intra-prompt LPIPS (↑).

Model Architecture	Inference Steps	Baseline CFG	DrAG-Spectral	Diversity Gain (Δ)
FLUX.1-dev	30	0.6304	0.6487	+0.0183 (+2.90%)
Qwen-Image	30	0.5715	0.6193	+0.0478 (+8.36%)
Stable Diffusion 3	28	0.6834	0.6945	+0.0111 (+1.62%)

Table 11 shows that DrAG-Spectral maintains, and in these experiments increases, perceptual diversity across random seeds. On FLUX.1-dev, the average LPIPS increases from 0.6304 to 0.6487, corresponding to a 2.90% gain. For Qwen-Image, diversity increases from 0.5715 to 0.6193 (+8.36%), while Stable Diffusion 3 shows an increase from 0.6834 to 0.6945 (+1.62%).

These results indicate that reducing the frequency of unconditional evaluations does not lead to a loss of sample diversity. The spectral drift criterion allows the cached guidance-gap to be reused when the observed velocity remains stable, while the guidance is refreshed when the drift becomes significant. This adaptive reuse preserves the overall conditional trajectory while allowing different random initializations to retain their sample-specific variations. The qualitative examples in Fig. 8 further illustrate this behavior, where DrAG-Spectral produces diverse realizations of both prompts across different seeds.

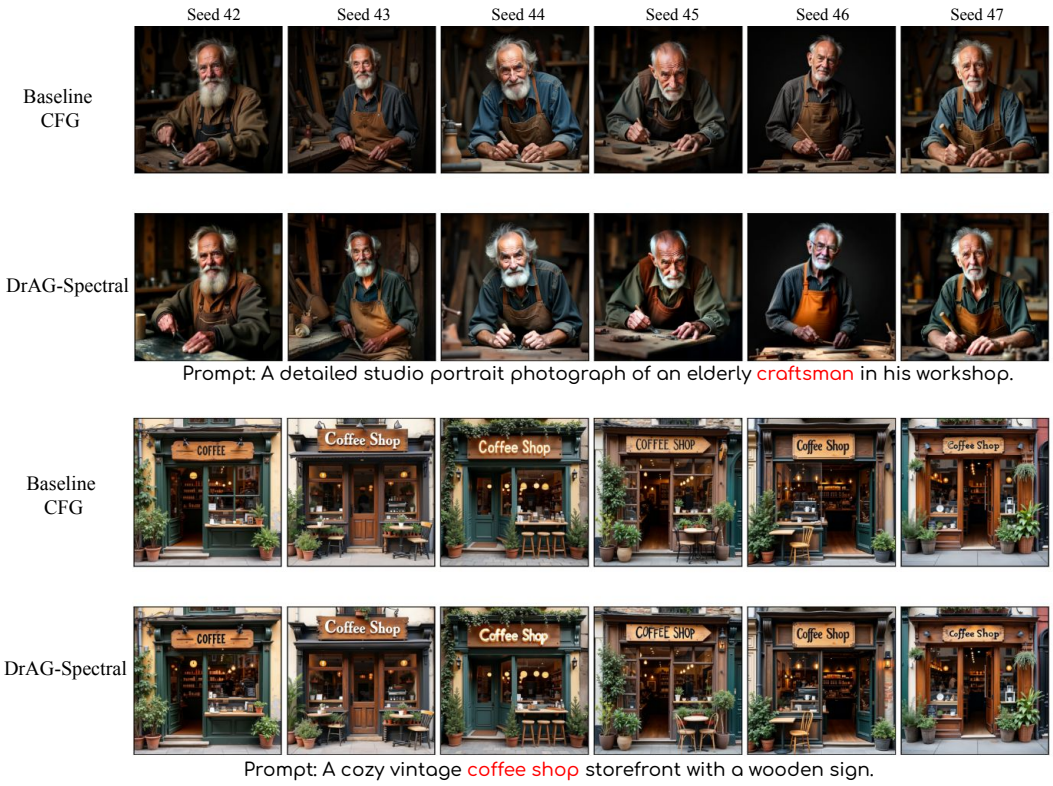

Figure 8: Qualitative comparison of generation diversity across different random seeds on FLUX.

B.10 Mechanistic Analysis of Dynamic Compute Allocation

To examine how DrAG dynamically allocates guidance computation according to scene complexity, we conduct a controlled study on SD3 model using 100 GenEval-2 prompts, consisting of 50 simple single-object scenes and 50 complex multi-object compositions. Figure 9 shows the guidance evaluation frequency across the sampling trajectory, while Table 12 reports the corresponding compositional scores.

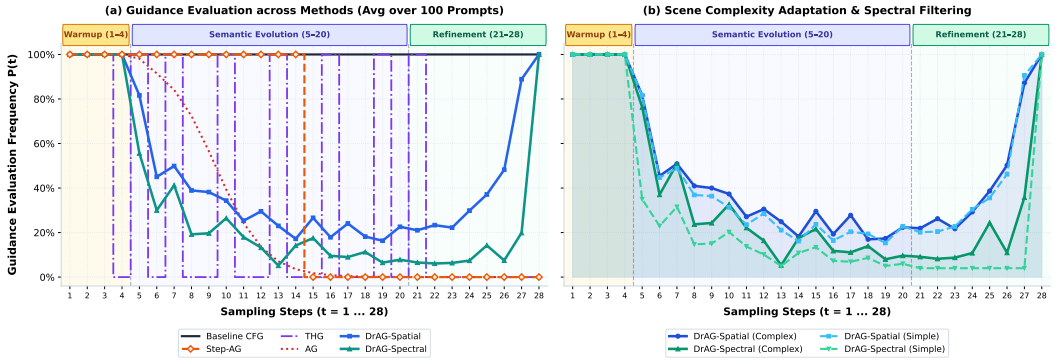

Figure 9: Empirical guidance evaluation frequency across sampling steps on the 100 GenEval-2 prompts. (a) Guidance evaluation frequency averaged across all prompts under identical sampling conditions ($N = 28$). (b) Scene-complexity analysis showing that DrAG dynamically allocates more guidance evaluations to complex multi-object scenes during semantic evolution (steps 5–20), while DrAG-Spectral suppresses the late-step evaluation increase observed with DrAG-Spatial during refinement.

Figure 9(a) shows that the compared methods follow substantially different compute-allocation patterns. Step-AG and THG use predefined evaluation schedules, while AG rapidly reduces guidance evaluations after the early sampling stages. In contrast, DrAG adapts the evaluation frequency to the observed conditional velocity drift. During the semantic evolution stage, DrAG retains more guidance evaluations when the trajectory changes rapidly and skips evaluations when the velocity remains relatively stable. This adaptive behavior allows DrAG-Spatial and DrAG-Spectral to achieve overall accuracies of 30.59 and 32.87, respectively, while reducing the inference cost by 26.1% and 31.4%.

Figure 9(b) further shows that the allocation depends on scene complexity. For simple prompts, both DrAG variants maintain relatively low guidance frequency during the semantic evolution stage, indicating that the cached guidance-gap can be reused for longer intervals. In contrast, complex multi-object prompts exhibit consistently higher guidance frequency during steps 6–21, reflecting the larger conditional-velocity changes required to establish multiple objects, attributes, and spatial relationships. Thus, DrAG automatically allocates more computation to trajectories that exhibit greater observed drift.

Table 12: GenEval-2 compositional benchmark scores across all sub-skills on the controlled 100-prompt empirical subset (50 simple single-object and 50 complex multi-object prompts) for Stable Diffusion 3 ($N = 28$).

Method	Avg NFE (↓)	Reduction (%)	Overall (↑)	Object (↑)	Attribute (↑)	Count (↑)	Position (↑)
Baseline CFG	56.0	0.0%	32.06	87.70	70.78	50.60	41.19
Step-AG	42.0	25.0%	28.65	86.63	68.89	49.74	38.41
THG	38.0	32.1%	26.73	83.33	64.60	44.91	31.04
AG	33.0	41.1%	25.20	72.33	62.27	43.54	30.30
DrAG-Spatial	41.4	26.1%	30.59	86.50	70.01	49.58	47.38
DrAG-Spectral	38.4	31.4%	32.87	85.87	68.32	48.51	41.89

The refinement stage further highlights the difference between the spatial and spectral criteria. DrAG-Spatial exhibits an increasing guidance frequency toward the final sampling steps, reaching approximately 52% at step 28 for complex prompts. DrAG-Spectral largely avoids this late-stage increase, maintaining a guidance frequency of approximately 8% at step 28 before the final evaluation. This behavior is consistent with the motivation for spectral drift measurement: frequency-weighted comparison reduces the sensitivity of the refresh criterion to high-frequency variations in the conditional velocity during late-stage refinement. Consequently, DrAG-Spectral achieves a 31.4% reduction in inference cost while obtaining the highest overall accuracy of 32.87 on this controlled subset.

C ACCELERATING TRAINING OF GENERATIVE DRIFTING MODELS VIA DRAG-SPECTRAL GUIDANCE

C.1 BACKGROUND: GENERATIVE MODELING VIA DRIFTING

Generative Modeling via Drifting (GMD) (Deng et al., 2026) introduces an alternative paradigm where generation is learned by evolving a generated particle distribution toward a target data distribution through a continuous vector field. Unlike diffusion and flow-matching models that learn time-dependent velocity fields and rely on multi-step numerical ODE solvers, GMD directly trains a one-step generator f_θ mapping random noise to target samples. For a noise vector $\epsilon \sim p_\epsilon$, class condition c, and classifier-free guidance scale α, the generator produces:

$$x = f_\theta(\epsilon, c, \alpha), \tag{38}$$

where x denotes the generated sample and q_θ denotes the resulting distribution.

C.1.1 DRIFTING FIELD FORMULATION

The central formulation defines a vector field that simultaneously attracts generated samples toward the target data distribution and repels them from rival modes. Given a positive distribution p and a negative distribution q, the drifting field is defined as:

$$V_{p,q}(x) = V_p^+(x) - V_q^-(x), \tag{39}$$

where the positive attraction and negative repulsion terms are given by:

$$V_p^+(x) = \frac{1}{Z_p(x)} \mathbb{E}_{y^+ \sim p} \left[k(x, y^+)(y^+ - x) \right], \tag{40}$$

$$V_q^-(x) = \frac{1}{Z_q(x)} \mathbb{E}_{y^- \sim q} \left[k(x, y^-)(y^- - x) \right]. \tag{41}$$

Here, $k(\cdot, \cdot)$ denotes a interaction kernel between particles, while Z_p and Z_q provide normalization. At equilibrium, the net drifting field approaches zero, locking the generated distribution onto the target data manifold. For class-conditional generation, the positive distribution is conditioned on the target class, $p(y \mid c)$, while the negative distribution comprises generated samples q_θ to prevent mode collapse.

C.1.2 FEATURE-SPACE DRIFTING AND CLASSIFIER-FREE GUIDANCE

Rather than computing particle interactions solely in pixel space, GMD formulates interactions within the representations of a pre-trained feature encoder $\phi(\cdot)$. Generated, positive, and negative samples are projected into feature space:

$$F_{\text{gen}} = \phi(X_{\text{gen}}), \qquad F_{\text{pos}} = \phi(X_{\text{pos}}), \qquad F_{\text{neg}} = \phi(X_{\text{neg}}). \tag{42}$$

Classifier-free guidance is embedded into this objective by augmenting the negative distribution with real images sampled from the unconditional data distribution $p_{\text{data}}(\cdot)$, scaled by guidance parameter α. This forces the generator to separate class-specific modes from generic real-world distributions during one-step inference.

C.1.3 THE TRAINING-TIME COMPUTATIONAL BOTTLENECK

During training, every optimization step requires the feature encoder $\phi(\cdot)$ to process generated samples, positive class-conditional samples, and unconditional negative samples. While this work avoids disk loading bottlenecks by storing raw reference images in an in-memory queue, those images must still be passed through the heavy ViT encoder on every iteration. With DrAG-Spectral, we take the crucial next step by caching the extracted feature representations directly, eliminating the GPU forward-pass computation entirely.

There is an important distinction between caching raw images and caching their feature representations. While sample queues alleviate data loading, they do not reduce neural network FLOPs. The unconditional negative branch alone accounts for nearly 50% of the feature extraction workload, forming the primary computational bottleneck in GMD training.

C.2 DrAG-Spectral Guidance for Generative Drifting

As the core contribution of this work, DrAG shows that unconditional evaluations at every discrete step are redundant in iterative diffusion models. Extending this insight beyond sampling, we observe that in single-step generative drifting, inference latency is already compressed to a single forward pass, making inference-time skipping inapplicable. Instead, we observe that the computational redundancy of unconditional guidance shifts directly into the training loop.

During training, the generator evolves incrementally across consecutive gradient updates. Once the low-frequency structural layout of the generated particles aligns with target class features, repeatedly passing fresh unconditional negative samples through the ViT encoder yields negligible directional innovation while consuming half the training FLOPs.

To eliminate this bottleneck, we propose DrAG-Spectral Guidance for training one-step generative models. The framework consists of two core components:

- Spectral Gating: A frequency-domain criterion that monitors when low-frequency structures have stabilized.
- Negative-Feature Caching: Bypassing the negative feature encoder and reusing cached representations when the trajectory is stable.

C.2.1 Spectral Gating

To determine whether negative feature extraction can be safely bypassed, DrAG-Spectral measures the low-frequency spectral shift between generated and positive feature representations. Let $\phi(\cdot)$ denote the feature representation, $\mathcal{F}(\cdot)$ the two-dimensional discrete Fourier transform (DFT), and $\mathcal{M}$ a Gaussian low-pass frequency mask. The normalized spectral discrepancy is defined as:

$$\Delta_{\text{spectral}} = \frac{\|\mathcal{M} \odot \mathcal{F}\left(\phi(X_{\text{gen}})\right) - \mathcal{M} \odot \mathcal{F}\left(\phi(X_{\text{pos}})\right)\|_2}{\|\mathcal{M} \odot \mathcal{F}\left(\phi(X_{\text{pos}})\right)\|_2}. \tag{43}$$

When $\Delta_{\text{spectral}} < \tau$, the generated particles are structurally aligned with the target class distribution in the low-frequency domain. In this regime, the repulsive force from unconditional data is locally stationary, making fresh negative feature extraction redundant.

C.2.2 Direct Negative-Feature Caching

Accordingly, DrAG-Spectral adopts a two-state execution policy:

- Full Step ($\Delta_{\text{spectral}} \geq \tau$): Negative samples are encoded through $\phi(\cdot)$, and the resulting representation is stored as the latest cached representation:

$$F_{\text{neg}} = \phi(X_{\text{neg}}), \qquad F_{\text{neg}}^{\text{last}} \leftarrow F_{\text{neg}}. \tag{44}$$

- Cached Step ($\Delta_{\text{spectral}} < \tau$): Forward passes through the negative encoder are bypassed, and the cached representation is directly reused in the drift loss:

$$\mathcal{L}_{\text{drift}} = \mathcal{D}_{\text{drift}}\left(\phi(X_{\text{gen}}), \phi(X_{\text{pos}}), F_{\text{neg}}^{\text{last}}\right). \tag{45}$$

Rather than removing unconditional guidance (which causes class boundary collapse), DrAG-Spectral preserves continuous out-of-class repulsion while entirely eliminating ViT forward passes on cached iterations.

C.2.3 Execution Policy and Stability Constraints

To ensure continuous convergence and prevent feature staleness, DrAG-Spectral enforces two safety mechanisms:

1. Initial Warm-up: For the first $t_{\text{warmup}} = 1,000$ steps, full steps are enforced unconditionally to establish clean initial representations.
2. Bounded Consecutive Skips: To prevent cached features from diverging from the evolving generator, at most $K_{\max} = 5$ consecutive cached steps are permitted before forcing a full step refresh.

Formally, the execution decision at optimization step t is:

$$\text{Step Policy} = \begin{cases} \text{Cached Step} & \text{if } \Delta_{\text{spectral}} < \tau \ \wedge \ k_{\text{consec}} < K_{\max} \ \wedge \ t \geq t_{\text{warmup}}, \\ \text{Full Step} & \text{otherwise.} \end{cases} \tag{46}$$

C.3 Experimental Evaluation

We evaluate DrAG-Spectral on class-conditional ImageNet 256×256 synthesis. Models are trained for 400,000 steps. Generation quality (FID and Inception Score) is evaluated on 50,000 generated samples across all 1,000 classes.

C.3.1 Ablation Study: Calibrating τ

To rigorously calibrate the gating threshold τ, we conduct a controlled 5,000-step ablation sweep across $\tau \in [0.735, 0.830]$ under identical initialization seeds (Table 13). Conservative thresholds ($\tau = 0.735$) limit skips to 41.2%, incurring unnecessary compute without quality gains (FID $=$ 122.32). Conversely, aggressive thresholds ($\tau \geq 0.800$) skip excessively ($> 74\%$), leading to stale negative repulsion and degraded synthesis (FID > 129.5). Optimal performance is attained at $\tau = 0.740$, achieving the lowest FID of 119.44 while safely bypassing 49.8% of negative feature passes. We adopt $\tau = 0.740$ for full run.

Table 13: Ablation on Spectral Gating Threshold τ (For 5,000-Step). Evaluated across 10,000 samples.

Gating Threshold (τ)	Negative Skip Rate (%)	FID (10k) ↓	Inception Score ↑
$\tau = 0.735$	41.2%	122.32	12.87 ± 0.28
$\tau = 0.740$	49.8%	119.44	12.27 ± 0.24
$\tau = 0.745$	58.6%	122.54	11.79 ± 0.22
$\tau = 0.800$	74.3%	129.59	10.96 ± 0.19
$\tau = 0.810$	79.1%	137.78	11.02 ± 0.21
$\tau = 0.830$	84.5%	131.05	10.53 ± 0.18

C.3.2 Quantitative Results on Full Run

Table 14 reports the 50,000-sample benchmark across guidance scales $\alpha \in \{1.0, 1.5, 2.0, 2.5\}$. At the standard guidance scale ($\alpha = 1.5$), DrAG-Spectral achieves an FID of 35.65 and an Inception Score of 41.30 ± 0.89, compared with 34.31 FID and 42.40 ± 1.11 for GMD. Interestingly, across all guidance scales, DrAG-Spectral preserves consistent CFG scaling, with FID improving from 39.24 to 32.20 and IS increasing from 36.66 to 46.76. At $\alpha = 2.5$, the FID gap is only 1.19 points (32.20 vs. 31.01) and lower IS variability, with ± 0.65 compared with ± 0.94 for the GMD at $\alpha = 2.5$, indicating stable generation quality despite the use of feature caching.

Table 14: Quantitative Results on ImageNet-256 (50,000 Samples, 400,000 Steps Training)

CFG Scale (α)	Method	FID (50k) ↓	Inception Score (IS) ↑	Skip Rate (%)
1.0	GMD	37.06	38.83 ± 0.62	0.0%
	DrAG-Spectral	39.24	36.66 ± 0.72	57.2%
1.5	GMD	34.31	42.40 ± 1.11	0.0%
	DrAG-Spectral	35.65	41.30 ± 0.89	57.2%
2.0	GMD	32.17	45.98 ± 0.98	0.0%
	DrAG-Spectral	33.56	44.40 ± 0.84	57.2%
2.5	GMD	31.01	47.64 ± 0.94	0.0%
	DrAG-Spectral	32.20	46.76 ± 0.65	57.2%

C.4 Computational Impact

As summarized in Table 15, DrAG-Spectral skips 228,880 negative Vision Transformer passes across 400,000 training steps, achieving a 57.2% skip rate. This reduces the average per-step train-

ing latency from 0.6747s to 0.3891s (a 42.3% acceleration), cutting total wall-clock time from 74.96 hours to 43.24 hours, saving 31.72 hours ($\sim$1.32 days) and 126.8 cluster GPU-hours. Crucially, because the feature encoder is utilized strictly during training, the final generator retains pure one-step real-time inference capability.

Table 15: Training Compute Saving

Training Metric	GMD	DrAG-Spectral (Ours)	Net Savings
Total Optimization Steps	400,000	400,000	Matched Convergence Volume
Negative ViT Passes Skipped	0 / 400,000 (0.0%)	228,880 / 400,000 (57.2%)	228,880 ViT Passes Bypassed
Average Step Latency	0.6747 s/step	0.3891 s/step	42.3% Latency Reduction (1.73$\times$ Faster)
Total Wall-Clock Duration	74.96 hours (3.12 days)	43.24 hours (1.80 days)	31.72 Hours Saved ($\sim$1.32 Days)
Final FID at CFG 1.5 (50k)	34.31	35.65	Minimal Gap (-1.34 points)

In conclusion, DrAG-Spectral Guidance resolves the primary computational bottleneck of training-time classifier-free guidance in single-step generative drift models. By unifying spectral manifold gating with negative feature caching, our method cuts training compute by over 42% while fully preserving class boundaries and sample generation fidelity.

D IMAGE AND VIDEO GENERATIVE MODELS

D.1 IMAGE GENERATIVE MODELS RESULT ON MS-COCO DATASET

To complement the compositional evaluation on GenEval2, we additionally evaluate DrAG on MS-COCO using standard distributional and image-text alignment metrics, including FID, KID, CMMD, and CLIP. We compare DrAG-Spatial and DrAG-Spectral with Baseline CFG, Step-AG, AG, and THG across FLUX.1-dev, Qwen-Image, and Stable Diffusion 3. The corresponding results are reported in Table 16.

The MS-COCO results show that DrAG achieves substantial reductions in NFE while maintaining distributional metrics in a range comparable to the evaluated acceleration methods. For example, DrAG-Spectral reduces NFE by 39.30% for FLUX, 39.30% for Qwen, and 32.30% for SD3. However, the individual FID, KID, CMMD, and CLIP metrics exhibit mixed behavior across models and acceleration methods. For example, AG achieves the lowest FID for Qwen and SD3, while DrAG-Spectral provides the largest compute reduction for FLUX and Qwen. Similarly, different methods achieve the best values across KID, CMMD, and CLIP. This makes the distributional metrics difficult to interpret as a single measure of whether the semantic behavior of the original guided trajectory has been preserved.

More importantly, these metrics do not directly evaluate the fine-grained compositional properties targeted by our guidance acceleration setting, such as object counting, attribute binding, spatial relationships, and verb interactions. In contrast, GenEval2 provides dimension-specific evaluation of these properties and therefore offers a more direct assessment of whether guidance-gap reuse preserves the behavior of the original guided generation process. We therefore use GenEval2 as the primary evaluation benchmark in the main paper, while the MS-COCO results are provided here as complementary distributional evidence.

Table 16: Quantitative Results on MS-COCO

Model	Method	CFG Scale	FID (↓)	KID (↓)	CMMD (↓)	CLIP (↑)	NFE (↓)	Compute Reduction (↑)
FLUX	Baseline CFG	2.0	28.77	0.0091	0.00080	0.7783	60.00	0.00%
FLUX	Step-AG	2.0	**28.94**	0.0092	0.00081	0.7774	45.00	25.00%
FLUX	AG	2.0	29.28	0.0097	0.00082	0.7746	36.45	39.25%
FLUX	THG	2.0	30.81	0.0101	0.00084	0.7731	40.00	33.33%
FLUX	DrAG-Spatial	2.0	29.91	0.0092	**0.00080**	**0.7775**	44.13	26.40%
FLUX	DrAG-Spectral	2.0	29.66	**0.0090**	**0.00080**	**0.7749**	**36.40**	**39.30%**
Qwen	Baseline CFG	4.0	34.04	0.0116	0.00110	0.7932	60.00	0.00%
Qwen	Step-AG	4.0	34.82	0.0120	0.00110	**0.7952**	45.00	25.00%
Qwen	AG	4.0	**33.51**	0.0120	0.00101	0.7948	**35.00**	**41.67%**
Qwen	THG	4.0	34.61	0.0115	0.00105	0.7944	40.00	33.33%
Qwen	DrAG-Spatial	4.0	33.57	**0.0111**	**0.00100**	0.7942	36.96	38.40%
Qwen	DrAG-Spectral	4.0	33.64	0.0114	**0.00100**	0.7935	36.40	39.30%
SD3	Baseline CFG	7.5	26.92	0.0079	0.00060	0.7926	56.00	0.00%
SD3	Step-AG	7.5	26.60	0.0083	**0.00051**	**0.7955**	42.00	25.00%
SD3	AG	7.5	**24.65**	**0.0072**	0.00053	0.7801	**33.00**	**41.07%**
SD3	THG	7.5	26.40	0.0073	0.00052	0.7914	38.00	32.14%
SD3	DrAG-Spatial	7.5	26.79	0.0077	0.00060	0.7919	40.61	27.50%
SD3	DrAG-Spectral	7.5	26.33	0.0073	0.00060	0.7905	37.92	32.30%

D.2 ABLATION ON IMAGE GENERATIVE MODELS

To select the best hyper parameters for DrAG-Spatial and DrAG-Spectral, we perform grid-search ablations on Stable Diffusion 3, FLUX.1-dev, and Qwen-Image using a 50-prompt validation set. For spectral gating, we evaluate Gaussian bandwidths $\sigma \in \{1.0, 3.0, 5.0, 10.0, 20.0\}$ with drift tolerances $\gamma \in [0.01, 0.30]$. Whereas for spatial gating, we sweep directional thresholds $\tau \in [0.90, 0.995]$.

As shown in Tables 17 18 19, the ablations reveal two main trends across the three models. First, the best spectral bandwidth depends on the architecture. For Stable Diffusion 3, a mid-band filter ($\sigma = 5.0$, $\gamma = 0.10$) gives a good balance between quality and efficiency, reducing NFE from

56.0 to 37.90 (32.3% reduction) while improving FID from 199.89 to 196.38 and KID from 0.0138 to 0.0116. In contrast, FLUX.1-dev and Qwen-Image perform best with $\sigma = 1.0$ and $\gamma = 0.20$, reaching 36.52 NFE (39.1% reduction) and 36.94 NFE (38.4% reduction), respectively.

Second, spatial gating works best with moderate directional thresholds. On SD3, $\tau = 0.995$ and $\gamma = 0.05$ achieve 40.56 NFE while maintaining high generation fidelity. On FLUX, $\tau = 0.99$ and $\gamma = 0.15$ reduce compute by 26.0% with an improved FID of 198.72. On Qwen, $\tau = 0.95$ and $\gamma = 0.10$ preserve strong CLIP alignment while reducing NFE by 38.0%. Relaxing the threshold further to $\tau = 0.90$ introduces noticeable trajectory drift, making $\tau \in [0.95, 0.995]$ the more stable operating range.

Based on these results, we use the following settings for the full 5,000-prompt and GenEval2 evaluation:

1. **FLUX.1-dev:** Spatial ($\tau = 0.99$, $\gamma = 0.15$), Spectral ($\sigma = 1.0$, $\gamma = 0.20$).
2. **Qwen-Image:** Spatial ($\tau = 0.95$, $\gamma = 0.10$), Spectral ($\sigma = 1.0$, $\gamma = 0.20$).
3. **Stable Diffusion 3:** Spatial ($\tau = 0.995$, $\gamma = 0.05$), Spectral ($\sigma = 5.0$, $\gamma = 0.10$).

Figures 11 12 10 provide more qualitative examples across different prompts.

Table 17: Ablation Analysis on Stable Diffusion 3 (50 Prompts) across Guidance Configurations

Configuration	Avg NFE (↓)	Compute Reduction (↑)	FID (↓)	KID (↓)	CMMD (↓)	CLIP Score (↑)
Baseline CFG (56 Steps)	56.00	0.0%	199.89	0.0138	0.00081	0.8118
DrAG-Spatial ($\tau = 0.995, \gamma = 0.01$)	42.00	25.0%	197.67	0.0129	0.00079	0.8102
DrAG-Spatial ($\tau = 0.995, \gamma = 0.05$)	40.56	27.6%	198.17	0.0124	0.00078	0.8132
DrAG-Spatial ($\tau = 0.995, \gamma = 0.10$)	40.56	27.6%	198.17	0.0128	0.00078	0.8132
DrAG-Spatial ($\tau = 0.995, \gamma = 0.15$)	40.56	27.6%	198.17	0.0127	0.00078	0.8132
DrAG-Spatial ($\tau = 0.995, \gamma = 0.20$)	40.56	27.6%	198.17	0.0125	0.00078	0.8132
DrAG-Spatial ($\tau = 0.99, \gamma = 0.01$)	41.40	26.1%	197.63	0.0130	0.00075	0.8081
DrAG-Spatial ($\tau = 0.99, \gamma = 0.05$)	39.14	30.1%	200.62	0.0127	0.00076	0.8110
DrAG-Spatial ($\tau = 0.99, \gamma = 0.10$)	39.10	30.2%	200.56	0.0129	0.00076	0.8109
DrAG-Spatial ($\tau = 0.99, \gamma = 0.15$)	39.10	30.2%	200.56	0.0126	0.00076	0.8109
DrAG-Spatial ($\tau = 0.99, \gamma = 0.20$)	39.10	30.2%	200.56	0.0123	0.00076	0.8109
DrAG-Spatial ($\tau = 0.98, \gamma = 0.01$)	40.68	27.4%	197.56	0.0137	0.00081	0.8051
DrAG-Spatial ($\tau = 0.98, \gamma = 0.05$)	37.98	32.2%	203.54	0.0126	0.00078	0.8062
DrAG-Spatial ($\tau = 0.98, \gamma = 0.10$)	37.82	32.5%	203.44	0.0126	0.00078	0.8064
DrAG-Spatial ($\tau = 0.98, \gamma = 0.15$)	37.82	32.5%	203.44	0.0126	0.00078	0.8064
DrAG-Spatial ($\tau = 0.98, \gamma = 0.20$)	37.82	32.5%	203.44	0.0123	0.00078	0.8064
DrAG-Spatial ($\tau = 0.95, \gamma = 0.01$)	39.94	28.7%	204.46	0.0135	0.00080	0.7978
DrAG-Spatial ($\tau = 0.95, \gamma = 0.05$)	36.66	34.5%	204.17	0.0126	0.00079	0.7946
DrAG-Spatial ($\tau = 0.95, \gamma = 0.10$)	36.38	35.0%	204.52	0.0126	0.00079	0.7936
DrAG-Spatial ($\tau = 0.95, \gamma = 0.15$)	36.36	35.1%	204.51	0.0127	0.00079	0.7938
DrAG-Spatial ($\tau = 0.95, \gamma = 0.20$)	36.36	35.1%	204.51	0.0126	0.00079	0.7938
DrAG-Spatial ($\tau = 0.90, \gamma = 0.01$)	39.70	29.1%	203.76	0.0126	0.00078	0.7954
DrAG-Spatial ($\tau = 0.90, \gamma = 0.05$)	35.92	35.9%	211.51	0.0177	0.00087	0.7870
DrAG-Spatial ($\tau = 0.90, \gamma = 0.10$)	35.40	36.8%	212.30	0.0182	0.00088	0.7819
DrAG-Spatial ($\tau = 0.90, \gamma = 0.15$)	35.38	36.8%	212.25	0.0183	0.00088	0.7818
DrAG-Spatial ($\tau = 0.90, \gamma = 0.20$)	35.38	36.8%	212.25	0.0178	0.00088	0.7818
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.01$)	51.36	8.3%	200.95	0.0150	0.00081	0.8114
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.05$)	39.42	29.6%	201.86	0.0117	0.00075	0.8156
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.10$)	36.60	34.6%	201.98	0.0123	0.00080	0.8098
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.20$)	34.62	38.2%	213.38	0.0179	0.00084	0.7919
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.30$)	34.02	39.2%	212.43	0.0202	0.00096	0.7745
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.01$)	54.04	3.5%	200.02	0.0138	0.00081	0.8116
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.05$)	40.68	27.4%	200.67	0.0151	0.00078	0.8127
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.10$)	37.36	33.3%	202.53	0.0120	0.00077	0.8074
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.20$)	35.00	37.5%	207.32	0.0149	0.00082	0.7981
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.30$)	34.16	39.0%	209.39	0.0177	0.00089	0.7793
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.01$)	55.04	1.7%	200.10	0.0138	0.00081	0.8114
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.05$)	41.44	26.0%	200.52	0.0145	0.00078	0.8146
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.10$)	37.90	32.3%	196.38	0.0116	0.00077	0.8108
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.20$)	35.36	36.9%	207.14	0.0138	0.00080	0.7977
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.30$)	34.32	38.7%	211.76	0.0167	0.00087	0.7839
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.01$)	55.66	0.6%	200.04	0.0142	0.00081	0.8117
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.05$)	42.66	23.8%	201.72	0.0134	0.00081	0.8117
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.10$)	38.52	31.2%	198.30	0.0120	0.00076	0.8077
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.20$)	35.98	35.8%	204.82	0.0137	0.00080	0.8005
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.30$)	34.70	38.0%	211.65	0.0167	0.00086	0.7851
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.01$)	55.90	0.2%	199.88	0.0131	0.00081	0.8118
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.05$)	43.58	22.2%	200.68	0.0129	0.00081	0.8118
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.10$)	38.98	30.4%	199.38	0.0121	0.00075	0.8092
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.20$)	36.42	35.0%	204.38	0.0125	0.00079	0.8025
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.30$)	35.20	37.1%	211.26	0.0164	0.00086	0.7887

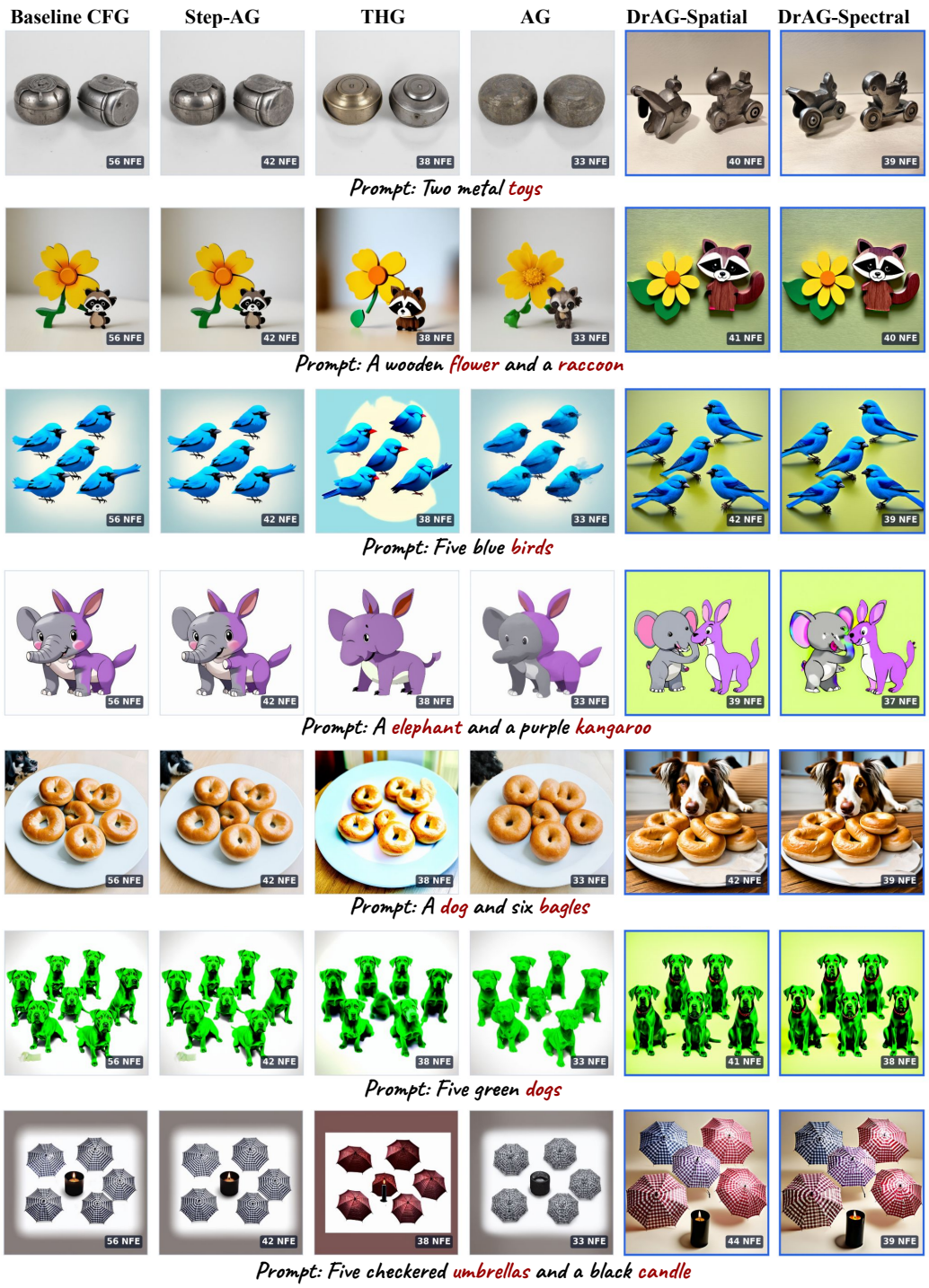

Figure 10: Qualitative comparison on Stable Diffusion 3

Table 18: Ablation Analysis on FLUX.1-dev (50 Prompts) across Guidance Configurations

Configuration	Avg NFE (↓)	Compute Reduction (↑)	FID (↓)	KID (↓)	CMMD (↓)	CLIP Score (↑)
Baseline CFG (60 Steps)	60.00	0.0%	199.30	0.0138	0.00097	0.7919
DrAG-Spatial ($\tau = 0.995, \gamma = 0.01$)	48.16	19.7%	199.20	0.0143	0.00097	0.7902
DrAG-Spatial ($\tau = 0.995, \gamma = 0.05$)	47.46	20.9%	199.63	0.0148	0.00098	0.7908
DrAG-Spatial ($\tau = 0.995, \gamma = 0.10$)	47.46	20.9%	199.63	0.0143	0.00098	0.7908
DrAG-Spatial ($\tau = 0.995, \gamma = 0.15$)	47.46	20.9%	199.63	0.0146	0.00098	0.7908
DrAG-Spatial ($\tau = 0.995, \gamma = 0.20$)	47.46	20.9%	199.63	0.0149	0.00098	0.7908
DrAG-Spatial ($\tau = 0.99, \gamma = 0.01$)	46.30	22.8%	198.99	0.0154	0.00098	0.7901
DrAG-Spatial ($\tau = 0.99, \gamma = 0.05$)	44.40	26.0%	198.72	0.0142	0.00097	0.7906
DrAG-Spatial ($\tau = 0.99, \gamma = 0.10$)	44.40	26.0%	198.72	0.0151	0.00097	0.7906
DrAG-Spatial ($\tau = 0.99, \gamma = 0.15$)	44.40	26.0%	198.72	0.0137	0.00097	0.7906
DrAG-Spatial ($\tau = 0.99, \gamma = 0.20$)	44.40	26.0%	198.72	0.0144	0.00097	0.7906
DrAG-Spatial ($\tau = 0.98, \gamma = 0.01$)	45.36	24.4%	199.36	0.0143	0.00098	0.7908
DrAG-Spatial ($\tau = 0.98, \gamma = 0.05$)	42.62	29.0%	199.25	0.0152	0.00097	0.7903
DrAG-Spatial ($\tau = 0.98, \gamma = 0.10$)	42.62	29.0%	199.25	0.0153	0.00097	0.7903
DrAG-Spatial ($\tau = 0.98, \gamma = 0.15$)	42.62	29.0%	199.25	0.0150	0.00097	0.7903
DrAG-Spatial ($\tau = 0.98, \gamma = 0.20$)	42.62	29.0%	199.25	0.0144	0.00097	0.7903
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.01$)	57.80	3.7%	200.10	0.0136	0.00097	0.7929
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.05$)	44.42	26.0%	199.20	0.0139	0.00098	0.7914
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.10$)	39.90	33.5%	199.75	0.0155	0.00098	0.7901
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.20$)	36.52	39.1%	196.48	0.0148	0.00098	0.7878
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.30$)	35.60	40.7%	199.33	0.0135	0.00100	0.7861
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.01$)	59.04	1.6%	199.19	0.0137	0.00098	0.7913
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.05$)	46.58	22.4%	198.16	0.0151	0.00098	0.7909
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.10$)	41.22	31.3%	201.14	0.0143	0.00096	0.7906
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.20$)	37.16	38.1%	198.79	0.0144	0.00097	0.7864
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.30$)	35.90	40.2%	199.62	0.0133	0.00099	0.7881
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.01$)	59.48	0.9%	199.26	0.0139	0.00098	0.7913
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.05$)	47.92	20.1%	197.85	0.0142	0.00098	0.7905
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.10$)	42.04	29.9%	198.70	0.0150	0.00097	0.7909
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.20$)	37.70	37.2%	198.89	0.0147	0.00097	0.7887
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.30$)	36.16	39.7%	198.86	0.0147	0.00099	0.7878
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.01$)	59.78	0.4%	199.31	0.0140	0.00098	0.7913
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.05$)	49.50	17.5%	197.94	0.0139	0.00097	0.7903
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.10$)	43.38	27.7%	199.05	0.0145	0.00097	0.7918
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.20$)	38.50	35.8%	198.22	0.0147	0.00098	0.7883
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.30$)	36.54	39.1%	197.67	0.0140	0.00099	0.7877
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.01$)	59.90	0.2%	199.26	0.0133	0.00097	0.7918
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.05$)	51.36	14.4%	198.99	0.0136	0.00098	0.7901
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.10$)	44.54	25.8%	198.42	0.0154	0.00098	0.7920
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.20$)	39.54	34.1%	198.81	0.0147	0.00098	0.7878
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.30$)	37.12	38.1%	196.84	0.0145	0.00097	0.7872

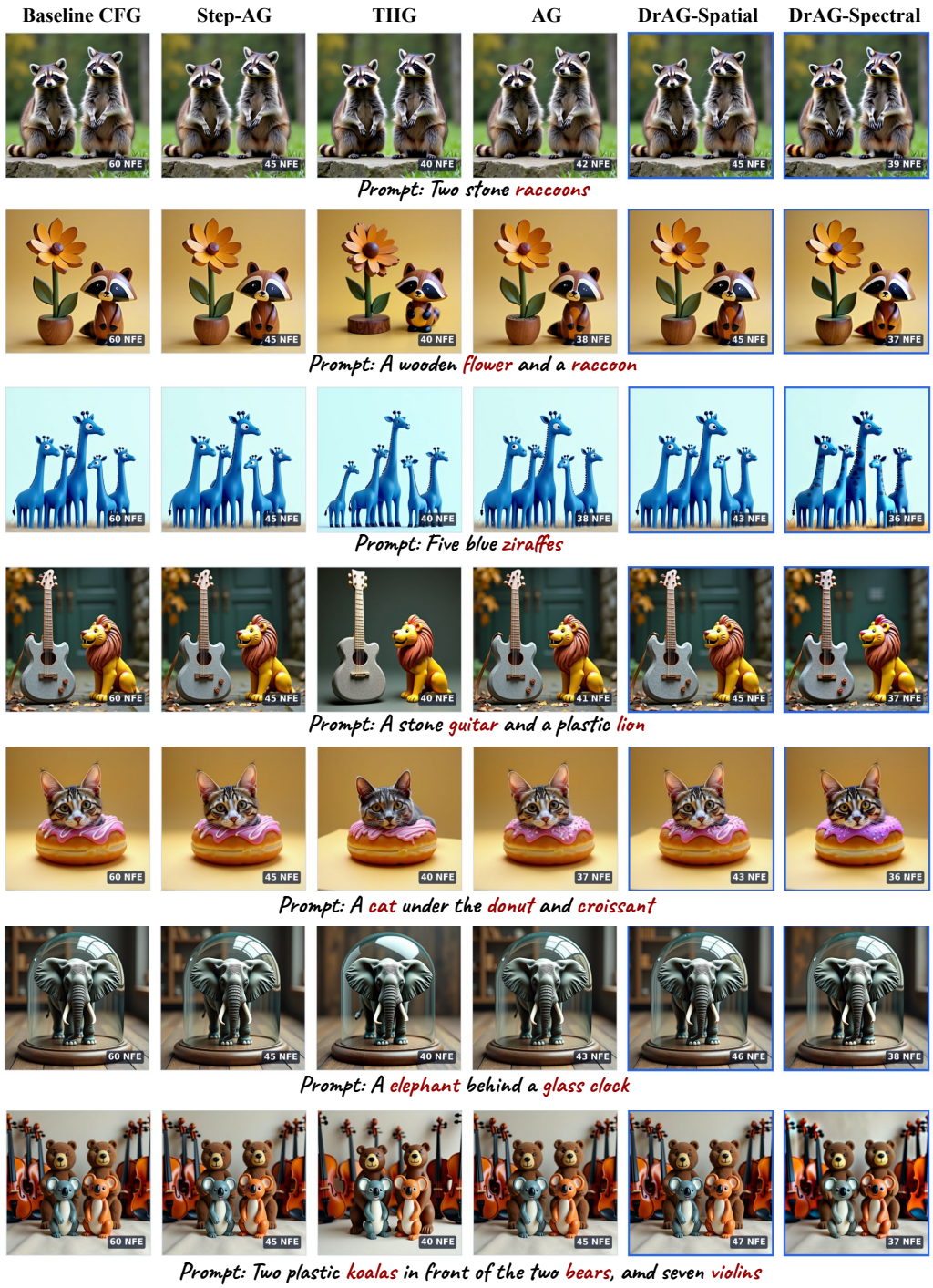

Figure 11: Qualitative comparison on FLUX.1-dev.

Table 19: Ablation Analysis on Qwen-Image (50 Prompts) across Guidance Configurations

Configuration	Avg NFE (↓)	Compute Reduction (↑)	FID (↓)	KID (↓)	CMMD (↓)	CLIP Score (↑)
Baseline CFG (60 Steps)	60.00	0.0%	207.85	0.0206	0.00108	0.7932
DrAG-Spatial ($\tau = 0.995, \gamma = 0.01$)	42.76	28.7%	205.30	0.0190	0.00111	0.7921
DrAG-Spatial ($\tau = 0.995, \gamma = 0.05$)	40.98	31.7%	204.46	0.0180	0.00112	0.7913
DrAG-Spatial ($\tau = 0.995, \gamma = 0.10$)	40.96	31.7%	204.50	0.0192	0.00112	0.7914
DrAG-Spatial ($\tau = 0.995, \gamma = 0.15$)	40.96	31.7%	204.50	0.0182	0.00112	0.7914
DrAG-Spatial ($\tau = 0.995, \gamma = 0.20$)	40.96	31.7%	204.50	0.0195	0.00112	0.7914
DrAG-Spatial ($\tau = 0.99, \gamma = 0.01$)	42.02	30.0%	203.96	0.0190	0.00111	0.7944
DrAG-Spatial ($\tau = 0.99, \gamma = 0.05$)	39.40	34.3%	200.44	0.0158	0.00111	0.7908
DrAG-Spatial ($\tau = 0.99, \gamma = 0.10$)	39.20	34.7%	200.63	0.0155	0.00111	0.7908
DrAG-Spatial ($\tau = 0.99, \gamma = 0.15$)	39.20	34.7%	200.63	0.0163	0.00111	0.7908
DrAG-Spatial ($\tau = 0.99, \gamma = 0.20$)	39.20	34.7%	200.63	0.0150	0.00111	0.7908
DrAG-Spatial ($\tau = 0.98, \gamma = 0.01$)	41.68	30.5%	203.17	0.0179	0.00110	0.7947
DrAG-Spatial ($\tau = 0.98, \gamma = 0.05$)	38.62	35.6%	199.59	0.0154	0.00110	0.7911
DrAG-Spatial ($\tau = 0.98, \gamma = 0.10$)	38.24	36.3%	199.11	0.0161	0.00110	0.7911
DrAG-Spatial ($\tau = 0.98, \gamma = 0.15$)	38.24	36.3%	199.11	0.0151	0.00110	0.7911
DrAG-Spatial ($\tau = 0.98, \gamma = 0.20$)	38.24	36.3%	199.11	0.0158	0.00110	0.7911
DrAG-Spatial ($\tau = 0.95, \gamma = 0.01$)	41.14	31.4%	202.64	0.0167	0.00106	0.7943
DrAG-Spatial ($\tau = 0.95, \gamma = 0.05$)	37.90	36.8%	198.71	0.0161	0.00105	0.7931
DrAG-Spatial ($\tau = 0.95, \gamma = 0.10$)	37.22	38.0%	198.78	0.0172	0.00105	0.7936
DrAG-Spatial ($\tau = 0.95, \gamma = 0.15$)	37.16	38.1%	198.70	0.0169	0.00105	0.7938
DrAG-Spatial ($\tau = 0.95, \gamma = 0.20$)	37.16	38.1%	198.70	0.0162	0.00105	0.7938
DrAG-Spatial ($\tau = 0.90, \gamma = 0.01$)	41.02	31.6%	201.99	0.0154	0.00106	0.7934
DrAG-Spatial ($\tau = 0.90, \gamma = 0.05$)	36.94	38.4%	197.91	0.0169	0.00116	0.7893
DrAG-Spatial ($\tau = 0.90, \gamma = 0.10$)	36.70	38.8%	198.62	0.0171	0.00118	0.7895
DrAG-Spatial ($\tau = 0.90, \gamma = 0.15$)	36.60	39.0%	198.54	0.0169	0.00118	0.7894
DrAG-Spatial ($\tau = 0.90, \gamma = 0.20$)	36.40	39.3%	198.61	0.0168	0.00118	0.7894
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.01$)	57.08	4.9%	207.24	0.0196	0.00108	0.7927
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.05$)	42.36	29.4%	207.40	0.0202	0.00111	0.7917
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.10$)	39.10	34.8%	204.22	0.0195	0.00109	0.7915
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.20$)	36.94	38.4%	196.62	0.0147	0.00106	0.7915
DrAG-Spectral ($\sigma = 1.0, \gamma = 0.30$)	36.00	40.0%	198.39	0.0155	0.00106	0.7927
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.01$)	59.16	1.4%	207.72	0.0202	0.00108	0.7929
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.05$)	43.70	27.2%	205.64	0.0196	0.00111	0.7924
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.10$)	39.84	33.6%	205.53	0.0184	0.00111	0.7931
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.20$)	37.52	37.5%	200.51	0.0153	0.00108	0.7915
DrAG-Spectral ($\sigma = 3.0, \gamma = 0.30$)	36.40	39.3%	198.67	0.0152	0.00105	0.7929
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.01$)	59.38	1.0%	207.62	0.0209	0.00108	0.7930
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.05$)	44.28	26.2%	204.37	0.0194	0.00111	0.7929
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.10$)	40.20	33.0%	204.32	0.0209	0.00111	0.7925
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.20$)	37.82	37.0%	200.78	0.0160	0.00108	0.7925
DrAG-Spectral ($\sigma = 5.0, \gamma = 0.30$)	36.74	38.8%	198.79	0.0160	0.00106	0.7933
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.01$)	59.66	0.6%	207.46	0.0198	0.00108	0.7931
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.05$)	45.20	24.7%	204.48	0.0194	0.00110	0.7910
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.10$)	40.66	32.2%	203.17	0.0190	0.00110	0.7935
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.20$)	38.08	36.5%	200.97	0.0164	0.00109	0.7922
DrAG-Spectral ($\sigma = 10.0, \gamma = 0.30$)	36.94	38.4%	198.99	0.0158	0.00106	0.7925
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.01$)	59.80	0.3%	207.70	0.0203	0.00108	0.7933
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.05$)	45.60	24.0%	204.21	0.0187	0.00110	0.7909
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.10$)	41.06	31.6%	203.92	0.0210	0.00111	0.7921
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.20$)	38.38	36.0%	199.17	0.0157	0.00109	0.7922
DrAG-Spectral ($\sigma = 20.0, \gamma = 0.30$)	37.22	38.0%	199.99	0.0159	0.00106	0.7930

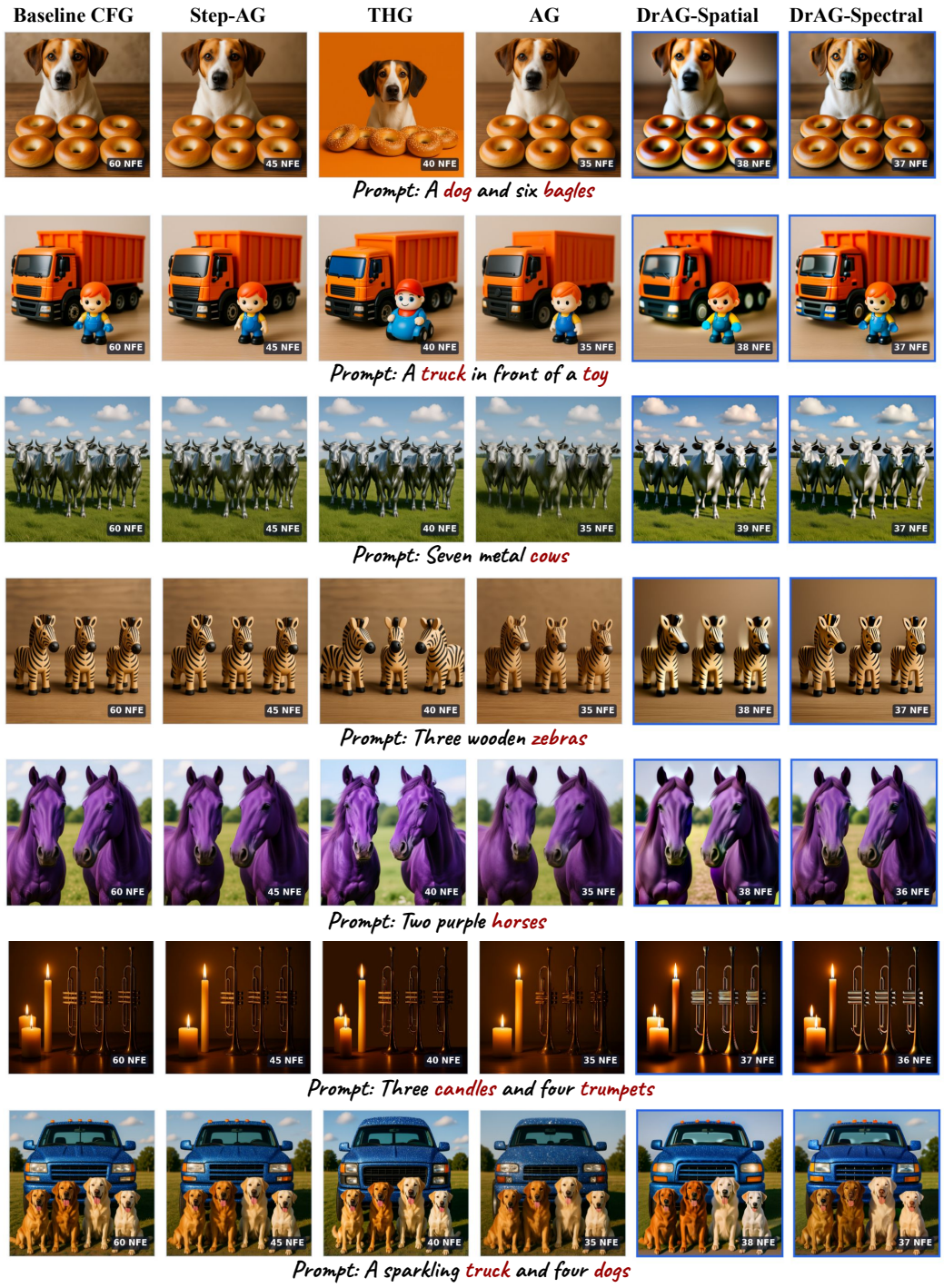

Figure 12: Qualitative comparison on Qwen.

D.3 Video Generative Models

Extending our approach to video generation, we evaluate HunyuanVideo (Kong et al., 2024) and Wan 2.1 (Wan et al., 2025) using the VBench evaluation (Huang et al., 2024) suite. Video models are intrinsically more computationally expensive, making NFE reduction highly desirable. We employ VBench's fine-grained evaluation dimensions to assess the generated videos in terms of Video Quality and Semantic Alignment, covering 16 dimensions in total.

Table 20: Quantitative Results on VBench Benchmark

Model	Method	Avg NFE (↓)	Compute Reduction (↑)	Video Quality (7 dims) (↑)	Semantic Alignment (9 dims) (↑)	Total VBench (16 dims) (↑)
Wan 2.1	Baseline CFG	100.00	0.0%	0.8472	0.5688	0.6906
	DrAG-Spectral ($\sigma = 1.0$)	**62.88**	**37.1%**	0.8456	0.5643	0.6874
HunyuanVideo	Baseline CFG	60.00	0.0%	0.8219	0.5101	0.6465
	DrAG-Spectral ($\sigma = 3.0$)	**37.94**	**36.8%**	0.8190	0.4987	0.6388

To determine the optimal hyperparameters for the DrAG-Spectral threshold (σ) and magnitude threshold (γ), we conduct ablation studies on both models, as reported in D.4, Tables 21 and 22. Using the selected configurations, Table 20 presents the performance of DrAG-Spectral on the full VBench benchmark, while Figure 13 visualizes the results as a radar plot. The results demonstrate that DrAG-Spectral substantially reduces the computational cost while preserving the overall quality of the generated videos. For Wan 2.1, DrAG-Spectral reduces the average NFE from 100.00 to 62.88, corresponding to a 37.1% reduction in computation. Despite this reduction, the Video Quality score decreases only marginally from 0.8472 to 0.8456, while Semantic Alignment changes from 0.5688 to 0.5643. Consequently, the overall VBench score remains comparable.

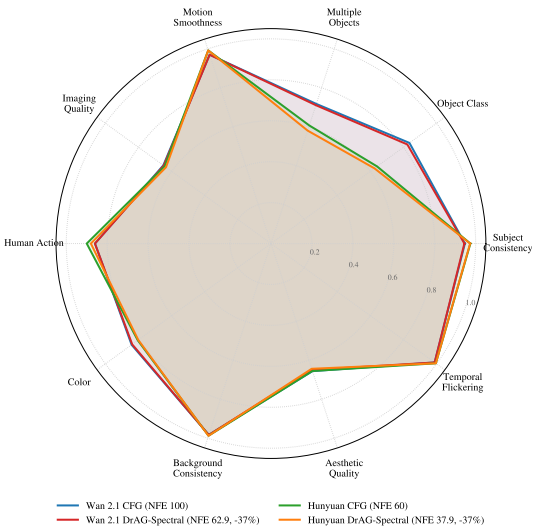

Figure 13: VBench Evaluation Radar Plot

A similar trend is observed for HunyuanVideo. DrAG-Spectral reduces the average NFE from 60.00 to 37.94, yielding a 36.8% reduction in computation. The Video Quality score decreases from 0.8219 to 0.8190, while the Semantic Alignment score changes from 0.5101 to 0.4987. The resulting Total VBench score remains comparable. These results indicate that DrAG-Spectral achieves substantial computational savings with relatively small changes in both perceptual video quality and semantic alignment. Overall, the results support the effectiveness of DrAG-Spectral as a compute-efficient acceleration strategy for video generation, maintaining performance close to the baseline while requiring considerably fewer function evaluations.

D.4 Ablation on Video Generative Models

Tables 21 and 22 evaluate DrAG-Spectral on 10 diverse prompts covering physical dynamics, camera motion, and multi-entity interactions. We measure Aesthetic Quality, Temporal Flickering, Subject Consistency, and Motion Smoothness using VBench.

For Wan 2.1 at its native guidance scale of 5.0, $\sigma = 1.0$ and $\gamma = 0.05$ provide the closest match to full CFG. The method achieves Aesthetic Quality of 0.6598 (vs. 0.6633 for CFG), Temporal Flickering of 0.9718 (vs. 0.9726), and Motion Smoothness of 0.9845 (vs. 0.9852), while reducing compute by 36.0% to approximately 64 NFE. Increasing γ to 0.10 gives only a further 3% reduction but lowers Aesthetic Quality to 0.6465. Similarly, using $\sigma = 3.0$ reduces Subject Consistency from

0.9369 to 0.9313. We therefore use $\sigma = 1.0$ and $\gamma = 0.05$ for the full VBench evaluation, where DrAG achieves 62.88 NFE (37.1% reduction) with minimal quality loss.

For HunyuanVideo at CFG scale 2.0, a slightly broader passband performs better. The setting $\sigma = 3.0$ and $\gamma = 0.05$ gives the best overall quality among the accelerated variants, with Aesthetic Quality of 0.6377 (vs. 0.6418 for CFG), Temporal Flickering of 0.9879 (vs. 0.9883), and Motion Smoothness of 0.9929 (vs. 0.9932), while reducing compute by 36.7% to approximately 38 NFE. Using $\sigma = 1.0$ lowers Aesthetic Quality to 0.6319, while increasing γ to 0.10 reduces it further to 0.6190. We therefore use $\sigma = 3.0$ and $\gamma = 0.05$ for the full VBench evaluation, where DrAG achieves 37.94 NFE (36.8% reduction).

Table 21: DrAG-Spectral Ablation for Wan 2.1

Method	Hyperparameters	Avg NFE	Compute Reduction (↑)	Aesthetic Quality (↑)	Temporal Flickering (↑)	Subject Consistency (↑)	Motion Smoothness (↑)
Baseline CFG	Standard	100	0.00%	0.6633	0.9726	0.9385	0.9852
DrAG-Spectral	$\sigma = 1.0, \gamma = 0.05$	∼64	36.00%	0.6598	0.9718	0.9369	0.9845
	$\sigma = 1.0, \gamma = 0.10$	∼61	39.00%	0.6465	0.9695	0.9342	0.9831
	$\sigma = 3.0, \gamma = 0.05$	∼65	35.00%	0.6559	0.9718	0.9313	0.9849
	$\sigma = 3.0, \gamma = 0.10$	∼61	39.00%	0.6525	0.9700	0.9325	0.9836

Table 22: DrAG-Spectral Ablation for HunyuanVideo

Method	Hyperparameters	Avg NFE	Compute Reduction (↑)	Aesthetic Quality (↑)	Temporal Flickering (↑)	Subject Consistency (↑)	Motion Smoothness (↑)
Baseline CFG	Standard	60	0.00%	0.6418	0.9883	0.9522	0.9932
DrAG-Spectral	$\sigma = 1.0, \gamma = 0.05$	∼37	38.33%	0.6319	0.9878	0.9525	0.9929
	$\sigma = 1.0, \gamma = 0.10$	∼36	40.00%	0.6197	0.9877	0.9552	0.9924
	$\sigma = 3.0, \gamma = 0.05$	∼38	36.67%	0.6377	0.9879	0.9509	0.9929
	$\sigma = 3.0, \gamma = 0.10$	∼36	40.00%	0.6190	0.9875	0.9564	0.9924

To rigorously assess temporal stability, subject consistency, motion smoothness, and aesthetic quality under varied dynamics, our 10-prompt video ablation evaluates the following diverse test scenarios:

P1. (Wildlife & Fur Dynamics) A tiny red fox walking through a misty bamboo forest at sunrise, soft golden light filtering through the leaves, cinematic camera slowly tracking sideways, realistic fur movement and gentle wind in the vegetation.

P2. (Urban Rain & Reflections) A futuristic Tokyo street at night during heavy rain, neon signs reflecting on wet pavement, a lone person walking with an umbrella, cinematic handheld camera following from behind, realistic rain and reflections.

P3. (High-Speed Aerial Motion) A majestic eagle taking off from a rocky mountain cliff, wings spreading in slow motion as dust and small stones scatter, dramatic clouds in the background, cinematic wide-angle shot.

P4. (Underwater Light Refraction) A colorful underwater coral reef with a sea turtle swimming peacefully past vibrant fish, sunlight rays moving through the water, gentle floating particles, smooth cinematic camera movement.

P5. (Macro Fluid Dynamics) A steaming cup of coffee on a wooden table beside a rainy window, visible steam curling upward, raindrops sliding down the glass, warm morning light, slow macro camera push-in, photorealistic.

P6. (Kinetic Historical Motion) A medieval knight riding a black horse across a grassy valley, cape flowing strongly in the wind, distant castle on a hill, dramatic overcast sky, cinematic tracking shot with realistic motion.

P7. (Metamorphic Fantasy) A paper airplane flying through a child's bedroom and suddenly transforming into a glowing origami bird, warm afternoon sunlight, whimsical magical atmosphere, smooth camera following the transformation.

P8. (Crowded Ambient Dynamics) A busy street food market in Mumbai during the evening, vendors cooking over hot flames, steam rising from food stalls, colorful lights and people moving naturally, energetic cinematic camera pan.

P9. (Zero-Gravity Physics) A small astronaut floating inside a futuristic space station, gently pushing off a wall and drifting toward a large window overlooking Earth, subtle zero-gravity motion, cinematic lighting, realistic details.

P10. (Landscape Depth & Flora) A field of lavender flowers moving in waves under a summer breeze, a lone wooden windmill in the distance, golden-hour sunlight, butterflies passing through the foreground, slow cinematic camera dolly forward.