Guiding Diffusion Models with Semantically Degraded Conditions

Shilong Han* Yuming Zhang* Hongxia Wang†
College of Science, National University of Defense Technology
hanshilong20@nudt.edu.cn, zhangyuming@nudt.edu.cn, wanghongxia@nudt.edu.cn

Abstract

Classifier-Free Guidance (CFG) is a cornerstone of modern text-to-image models, yet its reliance on a semantically vacuous null prompt ($\varnothing$) generates a guidance signal prone to geometric entanglement. This is a key factor limiting its precision, leading to well-documented failures in complex compositional tasks. We propose Condition-Degradation Guidance (CDG), a novel paradigm that replaces the null prompt with a strategically degraded condition, $\mathbf{c}_{deg}$. This reframes guidance from a coarse "good vs. null" contrast to a more refined "good vs. almost good" discrimination, thereby compelling the model to capture fine-grained semantic distinctions. We find that tokens in transformer text encoders split into two functional roles: content tokens encoding object semantics, and context-aggregating tokens capturing global context. By selectively degrading only the former, CDG constructs $\mathbf{c}_{deg}$ without external models or training. Validated across diverse architectures including Stable Diffusion 3, FLUX, and Qwen-Image, CDG markedly improves compositional accuracy and text-image alignment. As a lightweight, plug-and-play module, it achieves this with negligible computational overhead. Our work challenges the reliance on static, information-sparse negative samples and establishes a new principle for diffusion guidance: the construction of adaptive, semantically-aware negative samples is critical to achieving precise semantic control. Code is available at https://github.com/Ming-321/Classifier-Degradation-Guidance.

1. Introduction

Diffusion Models (DMs) have become a dominant force in generative modeling [15, 37, 38], with latent diffusion [29] and transformers [26] continually advancing text-to-image synthesis. Pivotal to this progress is Classifier-Free Guidance (CFG) [14], which steers generation by extrapolating unconditional predictions toward conditional ones, and has become a cornerstone of modern text-to-image systems [6–8, 12, 17, 20, 25, 34, 44, 45].

While pivotal, CFG exhibits failure modes in compositional tasks—text rendering, complex attribute binding, and spatial relationships (Fig. 1). We argue this stems from the semantic poverty of $\emptyset$: the large gap between $\boldsymbol{c}$ and $\emptyset$ yields an entangled guidance signal that mixes content generation with style and structure [21, 31]. In contrast, a semantically close $\boldsymbol{c}_{\text{deg}}$ enables common-mode rejection—suppressing shared components to isolate pure semantic corrections, as validated in Sec. 4.

Existing methods to address CFG's limitations fall into two camps. *Process Rectification* methods [21, 31, 33, 36] retain $\boldsymbol{c}$ vs. $\emptyset$ but apply post-hoc corrections—treating symptoms, not causes. Meanwhile, *Negative Reframing* methods [1–3, 9, 19, 28, 30, 32] explore alternatives to $\emptyset$—using weak models, random perturbations, attention manipulations, or VLM-generated negatives—yet none exploit the inherent semantic structure within the prompt's own token embeddings. A critical question remains: *how to construct a negative condition that adaptively degrades the core semantics of the positive prompt in a principled manner?*

We address this challenge through a key structural observation: in transformer-based text encoders, token embeddings naturally divide into *content tokens* (encoding object-specific semantics) and *context-aggregating tokens* (encoding global compositional context). By selectively removing content tokens while preserving context-aggregating tokens, a strategy we call *stratified degradation*, we construct a degraded condition $\boldsymbol{c}_{\text{deg}}$ that retains the prompt's global semantic scaffold while losing fine-grained details, reframing guidance from "good vs. null" to "good vs. *almost good*". We instantiate this principle in Condition-Degradation Guidance (CDG), a lightweight, plug-and-play module (Fig. 1).

Stratified degradation rests on the role of context-aggregating tokens: padding and special tokens that originally lack intrinsic semantics but acquire rich global context through attention. This is not a quirk of specific architectures but a fundamental property of transformer encoders; we confirm generality across diverse architectures

* These authors contributed equally.
† Corresponding author.

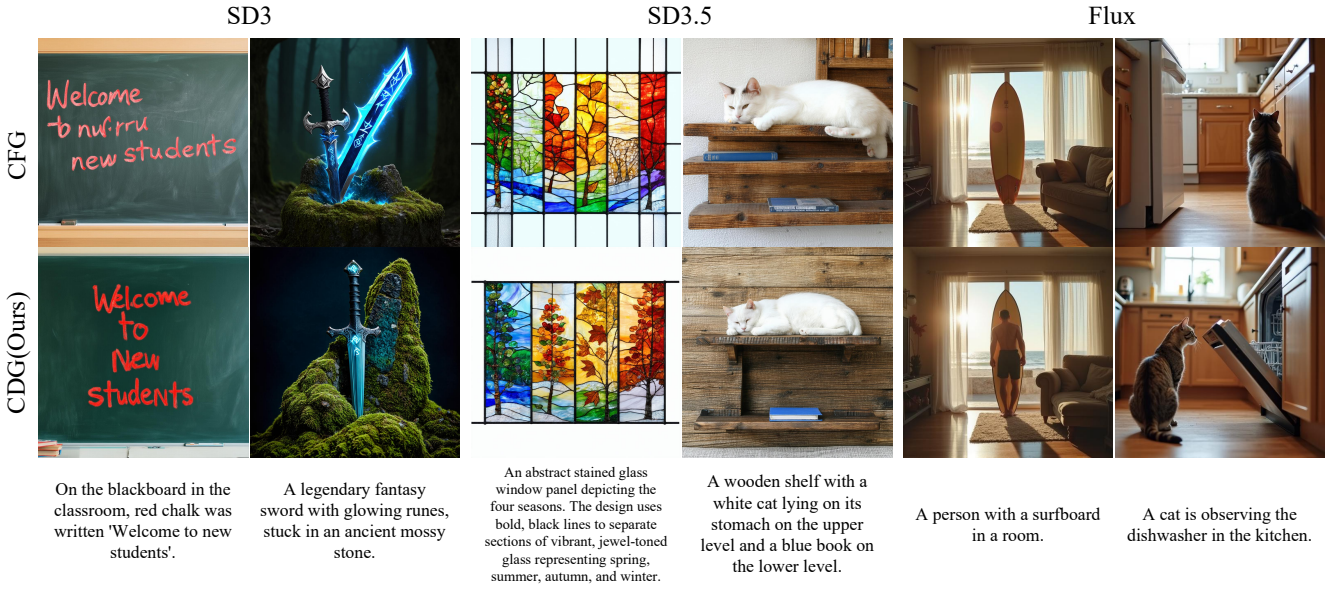

Figure 1. **Qualitative comparison between Classifier-Free Guidance (CFG) and our Condition-Degradation Guidance (CDG) across three state-of-the-art models (SD3, SD3.5, and Flux).** These examples demonstrate CDG's superior capability in handling complex compositional prompts where CFG often fails. CDG consistently outperforms CFG in accurate text rendering, precise spatial relationships and attribute binding, as well as complex object interactions.

(Sec. 6.2).

We validate CDG on Stable Diffusion 3, SD3.5, FLUX.1-dev, and Qwen-Image, demonstrating consistent improvements over baselines on FID, CLIP Score, VQA Score, and GenAI-Bench compositional reasoning with minimal overhead.

In summary, our contributions are:

- We reveal a functional dichotomy in transformer text encoders between *content tokens* and *context-aggregating tokens*, and propose *stratified degradation* as a principled strategy for constructing semantically degraded negative conditions.
- Based on this finding, we introduce Condition-Degradation Guidance (CDG), a lightweight, training-free, plug-and-play module requiring no external models or additional training.
- Extensive experiments across diverse models (SD3, SD3.5, FLUX.1-dev, Qwen-Image) validate CDG, providing geometric evidence for superior signal orthogonality and demonstrating consistent metric improvements with negligible overhead.

2. Related Work

Our work is situated within a broad research effort to enhance text-to-image generation. We contextualize CFG refinements [14], then focus on paradigms moving beyond the null prompt.

Refinement of the CFG Framework. Many methods focus on refining how the standard $\boldsymbol{c}$ vs $\emptyset$ guidance is applied, via geometric corrections (APG [31]) or SVD (TCFG [21]). While impactful, this research retains the foundational reliance on the semantically poor null condition.

Beyond the Null Prompt. Another paradigm moves beyond $\emptyset$ entirely, reframing guidance as a contrast between a "good" prediction and a "degraded" one. These methods can be distinguished by the source of degradation:

- **Model Level.** One approach uses a separate, typically weaker model to provide the negative signal, such as in Autoguidance [19] and Weak-to-Strong Diffusion [3]—effective but requires external model tuning.
- **Internal Mechanism Level.** Another direction perturbs the model's internal representations during the forward pass, including perturbing attention matrices (PAG [1]) or smoothing energy curvature (SEG [16]). These methods manipulate the model's computational flow to generate an implicit negative signal. As they operate on a different principle from input-level modifications, they are largely orthogonal to our approach and can potentially be combined.
- **Input Level.** The most direct strategy is to degrade the conditioning signal $\boldsymbol{c}$ itself. Methods in this category include using random prompts (ICG [32]), adding unstructured Gaussian noise (CADS [30]), spatially varying negatives (SFG [2]), and VLM-generated negatives (DNP [9]). However, these approaches either remain semantically blind or require expensive external models,

without exploiting the inherent semantic structure within the prompt's own token embeddings.

Our Approach. To address this gap, our Condition-Degradation Guidance (CDG) exploits the inherent functional dichotomy in transformer text encoders between content tokens and context-aggregating tokens. Through *stratified degradation*—selectively removing content tokens while preserving context-aggregating tokens—CDG constructs a semantically degraded condition that retains global compositional context. This creates a precise "good" vs. "almost good" contrast, directly improving guidance quality without external models or blind perturbations.

3. Background

Denoising Diffusion. Denoising diffusion generates samples from a data distribution p_{data} by reversing a process that gradually corrupts the data with Gaussian noise. This process yields smoothed densities $p(\boldsymbol{x};\sigma) = p_{\text{data}}(\boldsymbol{x}) * \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, indexed by the noise level σ. The generation process is then defined by a probability flow ODE [18, 35, 38] that evolves samples from pure noise back towards the data distribution:

$$d\boldsymbol{x}_\sigma = -\sigma \nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma;\sigma) d\sigma. \tag{1}$$

The core of this process is the score function, $\nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma;\sigma)$, which directs the update.

The score is approximated by a neural network $D_\theta(\boldsymbol{x}_\sigma;\sigma)$ parameterized by weights θ. While this network can be parameterized in various ways (e.g., to predict noise), we follow the formulation where it is trained as a denoiser to predict the clean sample $\boldsymbol{x}_0$ from a noised input $\boldsymbol{x}_\sigma$:

$$\theta^* = \arg\min_\theta \mathbb{E}_{\boldsymbol{x}_0,\sigma,\boldsymbol{\epsilon}} \left[\| D_\theta(\boldsymbol{x}_0 + \sigma\boldsymbol{\epsilon};\sigma) - \boldsymbol{x}_0 \|^2 \right]. \tag{2}$$

The expectation is taken over $\boldsymbol{x}_0 \sim p_{\text{data}}$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\sigma \sim p_{\text{train}}$, where p_{train} governs the noise level distribution during training. Given D_θ, we can estimate of the score function $\nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma;\sigma) \approx (D_\theta(\boldsymbol{x}_\sigma;\sigma) - \boldsymbol{x}_\sigma)/\sigma^2$. For conditional generation, the network is trained with an additional conditioning input $\boldsymbol{c}$ (e.g., a text embedding), becoming $D_\theta(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c})$, which provides an estimate of the conditional score $\nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma|\boldsymbol{c};\sigma)$.

Classifier-Free Guidance. Due to approximation errors inherent in finite-capacity networks, generated images often fail to match the fidelity of the training data.

A widely adopted technique to counteract this is Classifier-Free Guidance (CFG) [14], which enhances sample quality by extrapolating from an unconditional prediction towards a conditional one:

$$\begin{aligned} D_\theta^{\text{CFG}}(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}) &= D_\theta(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}) \\ &\quad + (w-1)(D_\theta(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}) - D_\theta(\boldsymbol{x}_\sigma;\sigma,\emptyset)), \end{aligned} \tag{3}$$

where $w > 1$ is the guidance scale.

Recalling the equivalence between the denoiser and the score function, we can rewrite the above equation as:

$$\begin{aligned} \nabla_{\boldsymbol{x}_\sigma} \log p_w(\boldsymbol{x}_\sigma|\boldsymbol{c};\sigma) &= \nabla_{\boldsymbol{x}_\sigma} \log p(\boldsymbol{x}_\sigma|\boldsymbol{c};\sigma) \\ &\quad + (w-1)\nabla_{\boldsymbol{x}_\sigma} \log \frac{p(\boldsymbol{x}_\sigma|\boldsymbol{c};\sigma)}{p(\boldsymbol{x}_\sigma|\emptyset;\sigma)}. \end{aligned} \tag{4}$$

4. Understanding CDG: A Geometric Perspective

We propose Condition-Degradation Guidance (CDG), which replaces the semantically distant null condition $\emptyset$ in CFG with a semantically degraded condition $\boldsymbol{c}_{\text{deg}}$ that is close to the original prompt $\boldsymbol{c}$:

$$\begin{aligned} D_\theta^{\text{CDG}}(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}) &= D_\theta(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}) \\ &\quad + (w-1)(D_\theta(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}) - D_\theta(\boldsymbol{x}_\sigma;\sigma,\boldsymbol{c}_{\text{deg}})). \end{aligned} \tag{5}$$

This reframes guidance from a coarse "good vs. null" contrast to a refined "good vs. almost good" discrimination. But why does this substitution improve guidance quality? We hypothesize that semantically distant contrasts ($\boldsymbol{c}$ vs. $\emptyset$) may produce guidance signals that interfere with the primary denoising direction, while semantic differencing ($\boldsymbol{c}$ vs. $\boldsymbol{c}_{\text{deg}}$) achieves better decoupling. To test this hypothesis, we analyze the geometric properties of the guidance signals.

Analytical Framework. Our analysis builds on the manifold hypothesis [4, 11], which posits that high-dimensional natural image data resides on low-dimensional manifolds. The diffusion process can be viewed as an evolution across a series of progressively smoother manifolds $\mathcal{M}_t$. Recent theoretical and empirical work [21, 39] has demonstrated that the score function $\nabla \log p_t(z_t)$ aligns with the manifold's normal space $\mathcal{N}_{z_t}(\mathcal{M}_t)$, which governs the primary denoising direction, throughout the reverse process. Following [21], we apply SVD to conditional predictions $\{\boldsymbol{\varepsilon}_{\boldsymbol{c}}\}$ across diverse prompts from MS-COCO 2017 [23] to approximate the principal denoising subspace $\mathcal{S}_{\boldsymbol{c}}(t)$ at each timestep t.

To quantify the geometric relationship between guidance signals and $\mathcal{S}_{\boldsymbol{c}}(t)$, we introduce two metrics, where $\mathcal{S}_g$ denotes the subspace spanned by the guidance signal $\Delta\boldsymbol{\varepsilon}$:

- **Geometric Decoupling** measures orthogonality between $\mathcal{S}_g$ and $\mathcal{S}_{\boldsymbol{c}}$:

$$\text{Decoupling}(\mathcal{S}_g, \mathcal{S}_{\boldsymbol{c}}) = \frac{1}{k}\sum_{i=1}^{k} \sin^2(\theta_i), \tag{6}$$

where θ_i is the i-th principal angle between the two subspaces and k is the subspace dimension; values approaching 1 indicate near-perfect orthogonality.

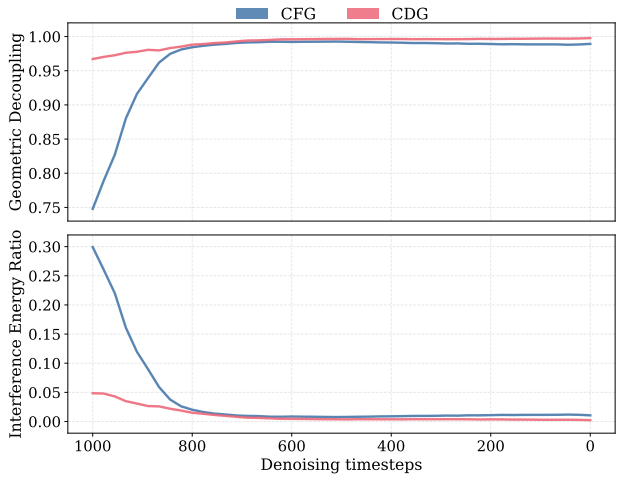

Figure 2. **CDG synthesizes a geometrically superior guidance signal compared to CFG.** (Top) Geometric Decoupling: CDG maintains near-perfect orthogonality throughout generation, while CFG suffers from significant early-stage entanglement. (Bottom) Interference Energy Ratio: CDG exhibits minimal interference, in stark contrast to CFG's substantial energy waste in misaligned directions. Together, these analyses demonstrate that CDG's guidance signal is structurally cleaner and more efficient from its inception, explaining its enhanced compositional control.

- **Interference Energy Ratio** measures the fraction of guidance energy projected onto $\mathcal{S}_{\boldsymbol{c}}(t)$:

$$\text{Interference}(\Delta\boldsymbol{\varepsilon}) = \frac{\|P_{\mathcal{S}_{\boldsymbol{c}}(t)}\Delta\boldsymbol{\varepsilon}\|_F^2}{\|\Delta\boldsymbol{\varepsilon}\|_F^2}, \tag{7}$$

  where $P_{\mathcal{S}_{\boldsymbol{c}}(t)}$ denotes the orthogonal projection onto $\mathcal{S}_{\boldsymbol{c}}(t)$; lower values indicate less interference with denoising.

We evaluate these metrics on CFG and CDG across multiple diffusion timesteps, and the results in Fig. 2 reveal a striking divergence.

As Fig. 2 shows, CDG maintains near-perfect orthogonality with minimal interference throughout generation, directly supporting our hypothesis. This aligns with Sadat et al. [31], who observed that perpendicular guidance components improve quality while parallel ones introduce artifacts.

Why does CDG achieve this? We attribute it to a common-mode rejection effect. As semantic neighbors, $\boldsymbol{c}$ and $\boldsymbol{c}_{\text{deg}}$ share similar normal components; their difference $\Delta\boldsymbol{\varepsilon}_{\text{CDG}} \propto \nabla_{z_t} \log \frac{p_t(z_t|\boldsymbol{c})}{p_t(z_t|\boldsymbol{c}_{\text{deg}})}$ cancels these, leaving primarily semantic distinctions. In contrast, CFG's semantically distant contrast ($\boldsymbol{c}$ vs. $\emptyset$) cannot achieve this cancellation, leading to entangled signals that conflate correction with denoising.

We hypothesize that the effectiveness of this "common-mode rejection" may stem from ensuring that $\boldsymbol{c}_{\text{deg}}$ preserves the "global context" shared with $\boldsymbol{c}$ (the common mode) while removing "specific semantics" (the correction signal). As we demonstrate in Sec. 5 (see Fig. 4), such a decoupling is achievable through our stratified degradation strategy. In Sec. 6.3.2 and Sec. 6.3.3, we will observe asymmetric patterns in experimental results that are consistent with this geometric interpretation.

5. Method

To construct $\boldsymbol{c}_{\text{deg}}$ as introduced in Sec. 4, we analyze the model's internal information flow to identify and strategically degrade the most semantically important tokens in the original prompt $\boldsymbol{c}$. Our complete pipeline is illustrated in Fig. 3.

To validate the content/context-aggregating dichotomy that motivates stratified degradation, we employ Weighted PageRank (WPR) as an analytical tool for quantifying token importance. Unlike standard approaches that aggregate cross-attention scores—which can paradoxically assign higher importance to context-aggregating tokens (see Appendix A)—we model the token relationships as a graph, as shown in Fig. 3 (a–c). Specifically, at a designated block λ_{block} (detailed below), we extract the self-attention map $A \in \mathbb{R}^{N\times N}$, where N is the sequence length. The tokens serve as the graph's nodes, while the attention weights in A provide the edge weights. We then apply the Weighted PageRank (WPR) algorithm [40, 42] to this attention-weighted graph to compute a final importance score vector $\boldsymbol{s} \in \mathbb{R}^N$ (Fig. 3 (d)). The core iterative update of WPR is defined as:

$$\boldsymbol{s}^{(k+1)} = \frac{A^T\boldsymbol{s}^{(k)}}{\|A^T\boldsymbol{s}^{(k)}\|_1}, \tag{8}$$

where $\boldsymbol{s}^{(k)}$ is the importance score vector at the k-th iteration, and the process is repeated until convergence. Details are provided in Appendix A.

As shown in Fig. 4, WPR reveals a clear importance dichotomy: content tokens exhibit substantially higher importance scores than context-aggregating tokens. This separation is intuitive: content tokens (e.g., "minecraft", "cooking") from meaningful text carry specific, fine-grained semantics; context-aggregating tokens lack explicit content before encoding. Despite this, they absorb contextual information through the encoder, carrying coarse-grained global semantics (as verified in Sec. 6.3.3). The dichotomy revealed by WPR informs our design of the Stratified Degradation strategy: rather than degrading blindly, such as setting global degradation ratios, we exploit this structure to design a degradation path that prioritizes content tokens over context-aggregating tokens.

As shown in Fig. 3(e), we partition the set of token indices $\mathcal{T}$ into a content set $\mathcal{T}_{\text{content}}$ and a context-aggregating set $\mathcal{T}_{\text{CtxAgg}}$. We then introduce two key hyperparame-

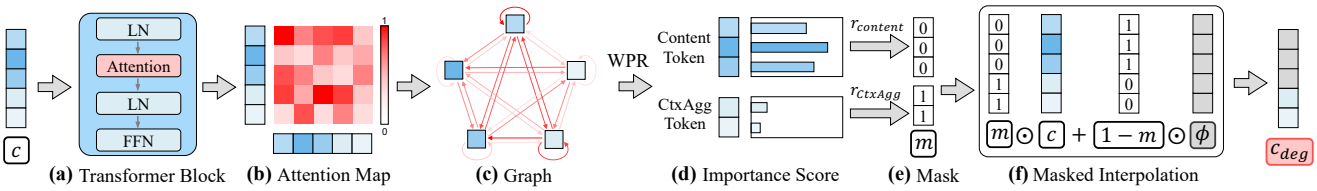

Figure 3. **An illustration of our proposed pipeline for constructing the semantically degraded condition, c_{deg}.** The process begins with attention graph extraction (a–c), where the self-attention map (b) from a transformer block (a) is modeled as a graph (c). Next, the Weighted PageRank (WPR) algorithm is applied to compute an importance score for each token (d). Following our Stratified Degradation strategy, these scores are used to generate a binary mask $\boldsymbol{m}$ (e). Finally, the mask facilitates the construction of $\boldsymbol{c}_{\text{deg}}$ via masked interpolation (f) between the original condition $\boldsymbol{c}$ and the null condition $\emptyset$.

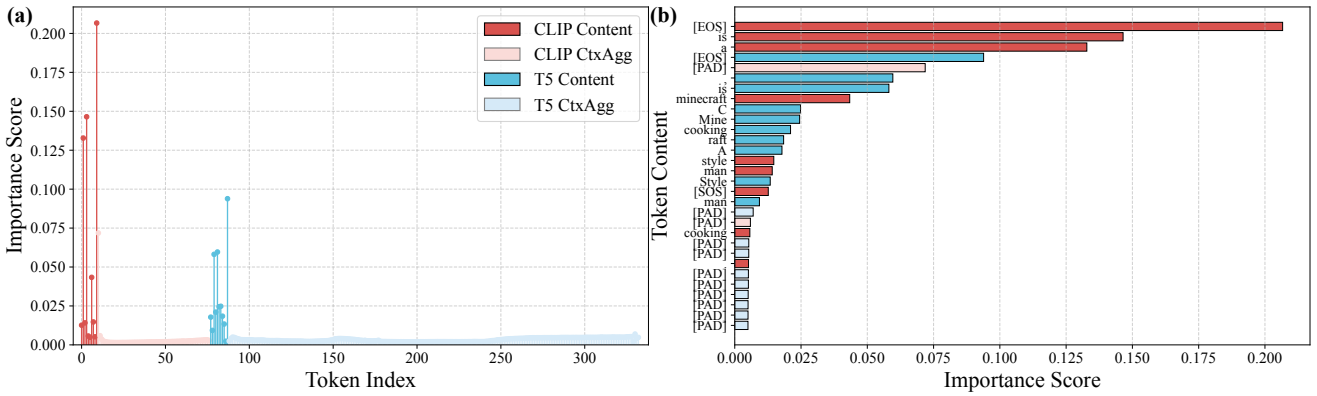

Figure 4. **WPR reveals a clear importance dichotomy between content and context-aggregating tokens:** content tokens carry fine-grained semantics while context-aggregating tokens carry coarse-grained semantics, as exemplified by the prompt "A man is cooking, MineCraft Style." (a) The stem plot shows that high importance scores (red for CLIP, cyan for T5) are almost exclusively concentrated on semantic content tokens. (b) The ranked list confirms that the top tokens ("minecraft", "cooking", "man") are almost all content-related. This dichotomy motivates our Stratified Degradation strategy, which first degrades content tokens and then context-aggregating tokens for controllable semantic degradation.

ters, $r_{\text{content}} \in [0,1]$ and $r_{\text{CtxAgg}} \in [0,1]$, which represent the desired replacement ratios for the content and context-aggregating tokens, respectively. We parameterize these two ratios through a single unified Degradation Ratio $R_{\text{deg}} \in [0,2]$, which maps to the per-type ratios via:

$$r_{\text{content}} = \min(R_{\text{deg}}, 1.0), \quad r_{\text{CtxAgg}} = \max(R_{\text{deg}} - 1.0, 0). \tag{9}$$

This formulation ensures that semantically important content tokens are degraded before context-aggregating tokens. Based on these ratios, the number of top-ranked tokens to replace from each subset is determined: $k_{\text{content}} = \lfloor r_{\text{content}} \cdot |\mathcal{T}_{\text{content}}| \rfloor$ and $k_{\text{CtxAgg}} = \lfloor r_{\text{CtxAgg}} \cdot |\mathcal{T}_{\text{CtxAgg}}| \rfloor$. The final binary replacement mask $\boldsymbol{m} \in \{0,1\}^N$ is then defined for each token i as:

$$m_i = \begin{cases} 0 & \text{if } i \in \mathcal{T}_{\text{content}} \text{ and } \text{rank}_i \le k_{\text{content}}, \\ 0 & \text{if } i \in \mathcal{T}_{\text{CtxAgg}} \text{ and } \text{rank}_i \le k_{\text{CtxAgg}}, \\ 1 & \text{otherwise}, \end{cases} \tag{10}$$

where rank_i is the rank of token i within its respective subset, determined by sorting the corresponding importance scores s_i in descending order. A smaller rank value (e.g., 1) thus signifies higher importance.

Through this design, $R_{\text{deg}} = 1.0$ represents a natural "semantic boundary" that separates two degradation regimes:

- $R_{\text{deg}} \in [0, 1.0]$: removes content tokens (fine-grained semantics),
- $R_{\text{deg}} \in (1.0, 2.0]$: removes context-aggregating tokens (coarse-grained semantics).

This dichotomy-driven formulation provides an interpretable control space. As we show in Sec. 6.3.2 (Fig. 5) and Sec. 6.3.3 (Fig. 6), $R_{\text{deg}} = 1.0$ serves as a robust default across models—balancing multiple metrics while offering computational efficiency (no WPR computation needed at this boundary). Users can adjust around $R_{\text{deg}} = 1.0$ to fine-tune style-alignment trade-offs.

Using the final mask $\boldsymbol{m}$, we construct the degraded condition $\boldsymbol{c}_{\text{deg}}$ by performing a masked interpolation between the original condition $\boldsymbol{c}$ and the null condition $\emptyset$, as shown in Fig. 3 (f):

$$\boldsymbol{c}_{\text{deg}} = \boldsymbol{m} \odot \boldsymbol{c} + (1 - \boldsymbol{m}) \odot \emptyset, \tag{11}$$

where $\odot$ denotes the element-wise Hadamard product. To make this degradation adaptive, we introduce an intervention block index, λ_{block}, to specify from which transformer block the attention map is extracted. At each step, reaching λ_{block} triggers mask construction via Eq. (10), applying $\boldsymbol{c}_{\text{deg}}$ (Eq. (11)) for all subsequent blocks within that step.

For computational efficiency, we compute the mask $\boldsymbol{m}$ only once at the first denoising step and reuse it throughout generation (Tab. 4), introducing minimal overhead with negligible performance impact.

6. Experiments

6.1. Experimental Setup

Base Model. Our experiments are built upon four text-to-image diffusion models: Stable Diffusion 3 Medium (SD3) [10], Stable Diffusion 3.5 (SD3.5) [10], FLUX.1-dev (FLUX.1) [5], and Qwen-Image [41]. All models are based on Transformer architectures at their core.

Dataset. We use 5,000 captions from the MS-COCO 2017 validation set [23] for comprehensive assessment.

Evaluation Metrics. To evaluate the performance of text-to-image models, we adopt four key metrics: FID [13], Aesthetic Score [43], CLIP Score [27], and VQA Score [24]. These metrics primarily assess performance from two aspects: image quality and text-image alignment.

Additionally, we report results on GenAI-Bench [22], a compositional reasoning benchmark covering diverse basic and advanced skills, to evaluate complex compositional capabilities.

More details of models, datasets, hyperparameters, and metrics are provided in Appendix C.1–C.3.

6.2. Comparison with Baselines

We compare CDG against baselines including CFG [14], CADS [30], ICG [32], PAG [1], SEG [16], SFG [2], and DNP [9] under identical evaluation settings.

Quantitative Results. As shown in Tab. 1, CDG consistently improves over CFG on all four backbones, achieving best or near-best results in most categories.

Cross-Model Analysis. Tab. 1 reveals CDG's improvements are more pronounced on SD3/SD3.5 than on FLUX.1, consistent with their training paradigms: FLUX.1 employs *Guidance Distillation* [5], reducing dependence on inference-time guidance. To further validate generalization, we include Qwen-Image, which uses special tokens (<|im_end|>) instead of padding as context aggregators. CDG consistently improves over CFG on this architecture (Tabs. 1 and 2), confirming that stratified degradation generalizes beyond padding-token architectures to any model with content/context-aggregating token structure.

Qualitative Results. Fig. 1 presents representative comparisons across three models, illustrating typical failure modes of CFG on compositional prompts. CFG struggles with text rendering (producing misspelled words), spatial-attribute binding (confusing positions or attributes), and complex interactions (generating semantically ambiguous compositions). In contrast, CDG consistently achieves accurate rendering, precise spatial-semantic alignment, and correct action semantics—improvements consistent with quantitative gains on compositional benchmarks (Tab. 2). Additional results are in Appendix C.9.

Table 1. Quantitative comparison on the MS-COCO 2017 validation set. Results are shown for SD3, SD3.5, FLUX.1, and Qwen-Image. Best results in **bold**, second-best underlined. (↓) indicates lower is better, (↑) higher is better.

Method		FID↓	CLIP Score↑	Aesthetic Score↑	VQA Score↑
SD3	CFG	35.69	31.73	5.66	91.44
	CADS	36.16	31.72	5.65	91.44
	ICG	39.09	31.41	**5.79**	90.89
	SEG	41.90	30.28	5.56	84.15
	PAG	50.60	30.15	5.52	81.27
	SFG	38.92	31.72	5.52	90.52
	DNP	34.68	31.30	5.51	89.65
	CDG	**34.05**	**32.00**	5.70	**92.40**
SD3.5	CFG	34.56	31.85	6.21	91.94
	CADS	34.58	31.86	6.21	91.83
	ICG	35.41	31.70	**6.32**	91.80
	SEG	38.90	30.71	6.16	88.49
	PAG	39.70	30.60	6.25	87.96
	CDG	**33.07**	**31.96**	6.26	**92.61**
FLUX.1	CFG	38.55	31.20	6.06	90.31
	CADS	38.73	**31.21**	6.01	90.05
	ICG	37.44	31.15	6.11	90.21
	CDG	**37.11**	**31.21**	**6.15**	**90.62**
Qwen	CFG	42.45	32.11	**2.57**	93.66
	CDG	**39.02**	**32.31**	2.54	**93.93**

Benchmark Results. We evaluate CDG on GenAI-Bench. Tab. 2 shows CDG consistently outperforms all baselines on SD3.5, with particularly strong gains on *Differentiation* (+3.64) and *Comparison* (+2.36)—tasks requiring subtle semantic contrasts where CDG's "good vs. almost good" paradigm excels. CDG also significantly outperforms structure-level methods (PAG, SEG) that lack semantic awareness. Full results are in Appendix C.7.

6.3. Ablation Study

Our ablation study examines four aspects: (1) component necessity (Sec. 6.3.1), (2) hyperparameter configuration (Sec. 6.3.2), (3) mechanism validation (Sec. 6.3.3), and (4) computational efficiency (Sec. 6.3.4). In Sec. 6.3.1, we use a fixed degradation budget ($R_{\text{deg}} = 1.1$) to isolate the contribution of each component; Sec. 6.2 reports results at

Table 2. GenAI-Bench compositional reasoning results: Spatial Relation (Spatial), Comparison (Comp), Differentiation (Differ), and Universal (Univ).

Method		Spatial↑	Comp↑	Differ↑	Univ↑
SD3	CFG	78.66	74.08	77.22	70.74
	CADS	78.55	74.05	76.98	70.11
	ICG	78.09	73.10	75.89	69.39
	SEG	72.26	69.51	70.46	65.52
	PAG	69.65	66.40	67.09	64.80
	CDG	**79.84**	**74.86**	**78.26**	**71.53**
SD3.5	CFG	79.66	73.70	75.10	72.21
	CADS	79.58	73.54	75.08	71.94
	ICG	78.90	74.13	75.63	70.23
	SEG	76.34	71.43	72.80	67.73
	PAG	75.64	71.02	72.38	66.17
	CDG	**80.69**	**76.06**	**78.74**	**73.13**
FLUX.1	CFG	77.47	72.97	75.39	71.13
	CADS	77.42	72.58	74.87	70.81
	ICG	77.07	72.83	74.50	71.47
	CDG	**77.56**	**73.47**	**76.17**	**71.55**
Qwen	CFG	83.26	80.19	83.41	77.36
	CDG	**83.79**	**80.24**	**83.54**	**77.56**

the default $R_{\text{deg}} = 1.0$ identified in Sec. 6.3.2.

6.3.1. Core Component Analysis

We demonstrate that WPR-based ranking effectively captures semantic structure. Details of experimental setup are in Appendix C.3.

Tab. 3 reveals that Stratified Degradation is the primary driver of CDG's effectiveness. Both stratified variants (rows 1–2) dramatically outperform all non-stratified variants (rows 3–5), with VQA improvements of +5.9–12.2 points and FID reductions of 0.9–16.8 points, confirming that treating content and context-aggregating tokens as separate degradation pools is crucial for precise semantic control.

Rows 1 and 2 show comparable performance between WPR-based and random ranking within the stratified framework (FID: 33.89 vs. 34.17), confirming that WPR serves as an analysis tool providing determinism and theoretical grounding for the $R_{\text{deg}} = 1.0$ boundary, rather than a necessary component.

Within non-stratified variants, WPR-based ranking (row 3) significantly outperforms reverse ranking (row 4, FID 50.73) and random ranking (row 5, FID 47.02), validating that WPR correctly identifies semantically important tokens when ranking is applied.

6.3.2. Optimal Hyperparameter Analysis

Having established that WPR effectively captures semantic structure (Sec. 6.3.1), we now analyze hyperparameter selection. This section demonstrates our hyperparameter selection process on the SD3 model, which guided the configurations reported in Tab. 1.

Table 3. Ablation study on CDG core components at fixed degradation budget ($R_{\text{deg}} = 1.1$). '✓' and '✗' denote active and inactive components. Asterisk (*) indicates reverse ranking (least important tokens).

Components		Metrics			
Importance	Stratified	FID↓	CLIP Score↑	Aesthetic Score↑	VQA Score↑
✓	✓	**33.89**	31.98	**5.68**	92.21
✗	✓	34.17	**32.02**	**5.68**	**92.27**
✓	✗	35.06	30.93	5.48	86.31
✓*	✗	50.73	29.86	5.22	80.10
✗	✗	47.02	30.39	5.21	83.55

We analyze two key hyperparameters: the intervention block λ_{block} and the unified Degradation Ratio $R_{\text{deg}} \in [0, 2]$ introduced in Sec. 5, which maps to per-type degradation ratios $(r_{\text{content}}, r_{\text{CtxAgg}})$ via Eq. (9).

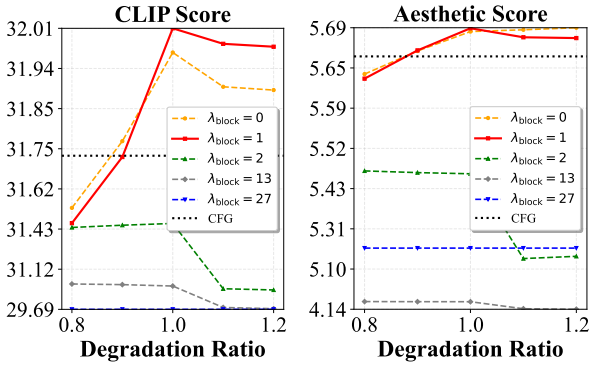

Figure 5. **Hyperparameter analysis:** joint effect of intervention block (λ_{block}) and Degradation Ratio (R_{deg}) on SD3.

As shown in Fig. 5, metrics exhibit an asymmetric response around $R_{\text{deg}} = 1.0$, consistent with the content/context-aggregating dichotomy. The steep slope in $[0, 1.0]$ (removing content tokens) transitions to a gentler slope in $[1.0, 2.0]$ (removing context-aggregating tokens), with the latter region exhibiting relative stability for fine-grained style control. Experiments confirm $\lambda_{\text{block}} = 1$ provides the most robust performance. We adopt $R_{\text{deg}} = 1.0$ as the default due to: (1) multi-metric balance, (2) cross-model applicability (detailed in Appendix C.4), and (3) computational efficiency (all content tokens are degraded at this boundary, bypassing WPR entirely).

To further illustrate the impact of degradation ratio adjustments, Appendix C.5 demonstrates qualitative results of adjusting degradation ratio near optimal parameters.

6.3.3. CFG* Validation Experiment

To validate the semantic properties of our constructed $\boldsymbol{c}_{\text{deg}}$, we design a CFG* experiment where we replace the positive prompt $\boldsymbol{c}$ in Eq. (3) with $\boldsymbol{c}_{\text{deg}}$, effectively using the degraded condition to guide generation. This allows us to directly probe what semantic information remains in $\boldsymbol{c}_{\text{deg}}$ at different degradation levels. Detailed formulation and additional qualitative results are provided in Appendix C.6.

Qualitative Analysis. As shown in Fig. 6(a), under a fixed prompt, the generated results progressively lose semantic information as R_{deg} increases from 0.00 to 2.00. In the range [0, 1], specific details such as "sleeping" and "sofa" are lost, while in [1, 2], the main subject "cat" disappears.

Quantitative Analysis. As shown in Fig. 6(b), the CLIP Score exhibits monotonic decline with a noticeable slope change near $R_{\text{deg}} \approx 1.0$. This inflection point aligns with the content/context-aggregating boundary revealed by WPR (Fig. 4) and corresponds with the qualitative analysis, supporting our hypothesis that these two token types encode different semantic granularities. Content token removal ($R_{\text{deg}} \in [0, 1.0]$) causes steep decline due to loss of specific semantics, while context-aggregating token removal ($R_{\text{deg}} > 1.0$) shows gentler decline as only global context degrades. Together with the asymmetric pattern in Fig. 5, these observations provide converging evidence for the dichotomy-driven design.

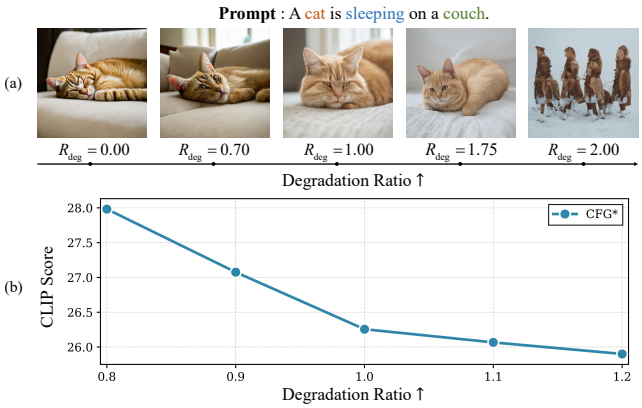

Figure 6. (a) Semantic degradation sequence under CFG* as R_{deg} increases from 0.00 to 2.00. (b) Quantitative analysis.

6.3.4. Computational Efficiency Analysis

CDG's efficiency is critical. As shown in Tab. 4, a naive per-step recomputation incurs substantial overhead (+47.2%) with negligible performance difference compared to one-time computation. In contrast, our one-time computation strategy introduces only minimal overhead (+3.6%). Note that we use $R_{\text{deg}} = 1.1$ for this analysis to evaluate WPR's computational cost; at $R_{\text{deg}} = 1.0$, WPR computation is bypassed entirely (all content tokens are degraded), making overhead assessment infeasible. Crucially, at our default $R_{\text{deg}} = 1.0$ setting, the strategy becomes a simple "replace all content tokens" operation, achieving near-zero overhead.

Table 4. Efficiency comparison of CDG implementation variants (one-time vs. per-step WPR computation) on SD3 with $R_{\text{deg}} = 1.1$, $\lambda_{\text{block}} = 1$.

Method	Time (s)	Aesthetic	VQA
CFG (Baseline)	5.456	5.66	91.44
CDG (Per-Step)	8.031 (+47.2%)	5.68	92.33
CDG	**5.655 (+3.6%)**	5.68	92.21

6.4. Modularity and Applications

CDG is a plug-and-play module that operates directly on text embeddings, making it naturally compatible with existing methods and downstream tasks. We demonstrate its extensibility across three scenarios: (1) combination with orthogonal methods like PAG [1], (2) image-to-image translation, and (3) ControlNet-based controllable generation, with comprehensive results in Appendix C.9.

Appendix B illustrates the CDG pipeline and key code, demonstrating the high extensibility of our method.

7. Conclusion

CFG's reliance on the semantically vacuous null prompt $\emptyset$ produces geometrically entangled guidance signals, limiting compositional accuracy. Our Condition-Degradation Guidance (CDG) addresses this by constructing a semantically degraded condition $\boldsymbol{c}_{\text{deg}}$ through attention-based analysis, reframing guidance as "good vs. almost good" discrimination. This design enables common-mode rejection, synthesizing guidance signals with superior orthogonality to the denoising manifold. Validated on state-of-the-art models, CDG consistently improves compositional reasoning and text-image alignment with negligible overhead. This work establishes a principle: adaptive, semantically-aware negative synthesis is essential for precise semantic control in conditional diffusion models.

Acknowledgements

We sincerely thank four anonymous reviewers for their valuable and constructive feedback, which has greatly improved our work. This work was supported by the following grants: the National Natural Science Foundation of China (Grant No. 12471401, 12401419).

References

[1] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, Seonhwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXIII*, pages 1–17. Springer, 2024. 1, 2, 6, 8

[2] Kambiz Azarian, Debasmit Das, Qiqi Hou, and Fatih Porikli. Segmentation-free guidance for text-to-image diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7520–7529, 2024. 2, 6

[3] Lichen Bai, Masashi Sugiyama, and Zeke Xie. Weak-to-strong diffusion with reflection. In *The Fourteenth International Conference on Learning Representations*, 2026. 1, 2

[4] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. 3

[5] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. Accessed: 2025-11-11. 6

[6] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 18392–18402. IEEE, 2023. 1

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[8] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack, 2023. 1

[9] Alakh Desai and Nuno Vasconcelos. Improving image synthesis with diffusion-negative sampling. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LIII*, pages 199–214. Springer, 2024. 1, 2, 6

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, pages 12606–12633. PMLR / OpenReview.net, 2024. 6

[11] Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. 3

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 1

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017. 6

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 1, 2, 3, 6

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[16] Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 2, 6

[17] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13753–13773. PMLR, 2023. 1

[18] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, pages 26565–26577. Curran Associates, Inc., 2022. 3

[19] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. 1, 2

[20] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6007–6017. IEEE, 2023. 1

[21] Mingi Kwon, Shin seong Kim, Jaeseok Jeong, Yi Ting Hsiao, and Youngjung Uh. TCFG: tangential damping classifier-free guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 2620–2629. Computer Vision Foundation / IEEE, 2025. 1, 2, 3

[22] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Gra-

ham Neubig, and Deva Ramanan. Genai-bench: Evaluating and improving compositional text-to-visual generation, 2024. 6

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 3, 6

[24] Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part IX*, pages 366–384. Springer, 2024. 6

[25] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 4296–4304. AAAI Press, 2024. 1

[26] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 4172–4182. IEEE, 2023. 1

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, pages 8748–8763. PMLR, 2021. 6

[28] Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 1

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE, 2022. 1

[30] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 1, 2, 6

[31] Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 1, 2, 4

[32] Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M. Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. 1, 2, 6

[33] Seyedmorteza Sadat, Tobias Vontobel, Farnood Salehi, and Romann M. Weber. Guidance in the frequency domain enables high-fidelity sampling at low cfg scales, 2025. 1

[34] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 8871–8879. IEEE, 2024. 1

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 3

[36] Kaiyu Song and Hanjiang Lai. Rethinking oversaturation in classifier-free guidance via low frequency, 2025. 1

[37] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11895–11907, 2019. 1

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1, 3

[39] Jan Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, pages 46412–46440. PMLR / OpenReview.net, 2024. 3

[40] Hongjie Wang, Bhishma Dedhia, and Niraj K. Jha. Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pre-trained transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 16070–16079. IEEE, 2024. 4

[41] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 6

[42] Wenpu Xing and Ali A. Ghorbani. Weighted pagerank algorithm. In *2nd Annual Conference on Communication Networks and Services Research (CNSR 2004), 19-21 May 2004, Fredericton, N.B., Canada*, pages 305–314. IEEE Computer Society, 2004. 4

[43] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952, 2023. 6

[44] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3813–3824. IEEE, 2023. 1

[45] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. 1