Improving Diffusion Generalization with Weak-to-Strong Segmented Guidance

Liangyu Yuan[1,2*] Yufei Huang[1,3*] Mingkun Lei[1] Tong Zhao[1,3]
Ruoyu Wang[1] Changxi Chi[1,3] Yiwei Wang[4] Chi Zhang[1†]
[1]Westlake University [2]Tongji University
[3]Zhejiang University [4]University of California at Merced

Abstract

Diffusion models generate synthetic images through an iterative refinement process. However, the misalignment between the simulation-free objective and the iterative process often causes accumulated gradient error along the sampling trajectory, which leads to unsatisfactory results and a failure to generalize. Guidance techniques like Classifier Free Guidance (CFG) and AutoGuidance (AG) alleviate this by extrapolating between the main and inferior signal for stronger generalization. Despite empirical success, the effective operational regimes of prevalent guidance methods are still under-explored, leading to ambiguity when selecting the appropriate guidance method given a precondition. In this work, we first conduct synthetic comparisons to isolate and demonstrate the effective regime of guidance methods represented by CFG and AG from the perspective of weak-to-strong principle. Based on this, we propose a hybrid instantiation called SGG under the principle, taking the benefits of both. Furthermore, we demonstrate that the W2S principle along with SGG can be migrated into the training objective, improving the generalization ability of unguided diffusion models. We validate our approach with comprehensive experiments. At inference time, evaluations on SD3 and SD3.5 confirm that SGG outperforms existing training-free guidance variants. Training-time experiments on transformer architectures demonstrate the effective migration and performance gains in both conditional and unconditional settings. Code is available at https://github.com/Westlake-AGI-Lab/SGG

1. Introduction

Diffusion and flow matching models have become the defacto standard for modern image synthesis [2, 8, 16, 28, 30, 44, 46], prized for their ability to generate highly realistic images via iterative refinement. However, this multi-step process suffers from a misalignment between the local simulation-free training objective and the global, iterative sampling trajectory. This discrepancy, known as exposure bias [32], leads to the accumulation of network errors during sampling [5, 20]. Consequently, unguided models, particularly for complex conditional generation tasks like text-to-image generation, often fail to generalize properly, producing samples that are out of distribution and perceptually unacceptable [15].

To counteract this sampling drift, which is known to degrade generalization [19, 26, 45], inference-time guidance has become one of the standard practices, but the effective regimes of different prevalent guidance methods still present ambiguity. Classifier-Free Guidance (CFG) [15], for instance, is widely adopted due to its robustness. More recently, AutoGuidance (AG) [21] was proposed to address a flaw in CFG: the entanglement of condition-adherence and sample diversity. AG attempts to resolve this by guiding the generation with a condition-aligned, inferior model. However, despite its empirical success on specific scenarios [22], the idea of guiding with a condition-aligned weak model has not fully replaced CFG. In complex, large-scale tasks like text-to-image (T2I) generation, AG-inspired methods often serve as a complement to CFG [34] or are found to be less performant when used in isolation [29, 34].

To first give a better understanding of the operational regimes of two types of prevalent guidance methods analogous to CFG [15] and AG [21], we conceptualize them from the perspective of weak-to-strong principle, where we categorize these approaches into two classes: condition-dependent and condition-agnostic. Under this perspective, we conduct synthetic experiments to isolate and demonstrate the effective regimes and failure modes of each class. Our analysis reveals that appropriate guidance can be influenced by two key factors: the intrinsic **granularity of the condition** [57] and the **fitting capacity** [10, 26] of the model. Based on this insight, we propose **SeGmented Guidance** (SGG), a simple yet effective instantiation under the principle that synergizes

* Equal contribution. † Corresponding author. This work was done during Liangyu Yuan's visit at WestLake University. Code for 2d Toy example: https://github.com/851695e35/Leaves_Toy

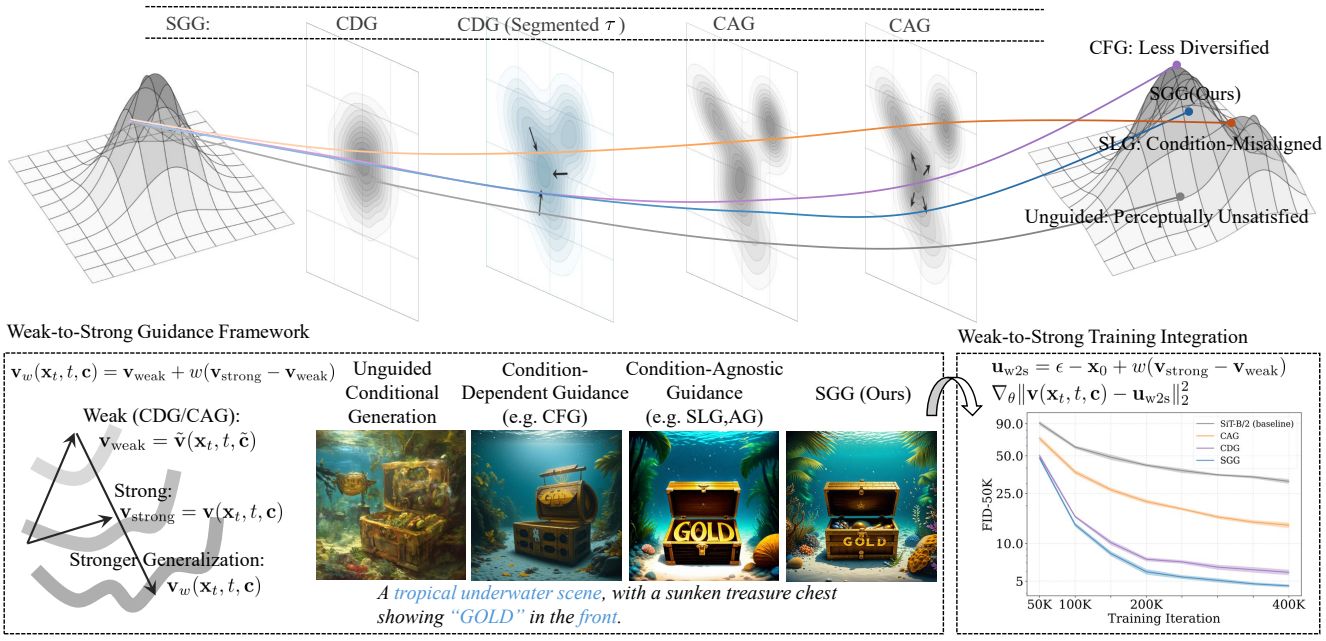

Figure 1. I: Weak-to-strong guidance principle: Guidance methods serve as tools for improving generalization capacity, we propose SGG to combine the benefits of condition-dependent (CDG) and condition-agnostic guidance (CAG). II: Integration to the training framework, improving the generalization ability of unguided diffusion models.

the benefits of both classes to better handle practical, realistic generation scenarios. Specifically, SGG operates by first leveraging condition-dependent guidance to seek the correct manifold, then switching to condition-agnostic guidance to refine intra-condition details.

We take a step further by migrating the Weak-to-Strong (W2S) guidance principle and SGG from inference directly into the training objective. This approach enhances the generalization capacity of the unguided diffusion model, thereby reducing the reliance on extra guidance costs during sampling. We also explore various weak-model construction methods, providing a suite of practical choices tailored for transformer architectures. The overall pipeline is illustrated in Fig. 1. We validate our methods in both inference and training settings. For inference, SGG outperforms competing guidance variants on SD3 and SD3.5 [8]. For training, we verify the effectiveness of W2S principle and SGG on SiT models in both conditional and unconditional settings, elevating the generalization capacity of unguided diffusion models. Our contribution can be summarized as follows:

- We categorizes and analyze the operational regimes of condition-dependent and condition-agnostic guidance under W2S perspective.
- Based on this analysis, we introduce a hybrid instantiation called SGG, a simple yet effective technique that synergizes the benefits of both guidance paradigms.
- We migrate W2S principle and SGG from inference-time mechanism into the training objective, directly improving the generation ability of unguided diffusion models.

2. Related work

Condition-dependent guidance. Guidance techniques are crucial for controlling the synthesis process in diffusion models. An early approach, Classifier Guidance (CG) [7], leverages the gradients of a separately trained classifier to steer generation. The now-ubiquitous Classifier-Free Guidance (CFG) [15] eliminated this need for an external classifier by reformulating the guidance term using Bayes' rule, which requires the model to be jointly trained on conditional and unconditional outputs. Various variants have since been proposed to refine the application of CFG [9, 11, 24, 37–40, 52]. For instance, Guidance Interval [24] suggests skipping guidance during specific time intervals to mitigate observed negative effects. APG [38] alleviates the oversaturation problem in high guidance scale through decomposition of the guidance term. CFG-Zero-Star [9] proposes omitting guidance during the initial sampling steps to enhance performance.

Condition-agnostic guidance. Recently, the idea of using a condition-aligned inferior model for guidance has emerged in several methods [1, 3, 4, 17, 21, 34, 49], serving as either a complement or an alternative to CFG under certain conditions. These methods operate by constructing an inferior prediction to guide the expert output. The inferior

signal can be generated in several ways: by training a separate, inferior model, as in AutoGuidance (AG) [21]. Through self-perturbation, such as skipping residual or attention layers [1, 17], by using a stochastic subnetwork, as proposed in S^2-Guidance [4], or by perturbing the input tokens, as in TPG [34]. However, despite their practical success, these weak-model-based approaches have been reported to be less effective or robust than CFG when used in isolation [34], or often function only as a complement to CFG rather than a complete replacement [4].

Training acceleration in diffusion models. Recent works accelerate diffusion model training convergence via two main strategies: improving representation capacities or modifying the regression objective [6, 47, 48, 50, 54–56]. For representations, REPA [56] aligns the intermediate features of a Diffusion Transformer (DiT) with those from a base model like DINOv2 [33], VA-VAE [55] applies a similar principle to the features of the tokenizer. SRA [18] propose to align the features of a former block to a latter transformer blocks. For modification on regression objective, contrastive Flow Matching [47] introduces a contrastive term for separation of paths, and STF [54] replace the high-variance single-sample target with a more stable, lower-variance batch-level expectation. GFT [6] and MG [48] propose to modify the training target by adding the unconditional guidance term from CFG [15] to enhance generation without guidance.

3. Preliminaries

Diffusion models. Diffusion Models [16, 44, 46] are generative models that learn to reverse a process that gradually maps data to noise. Given a predefined data distribution p_{data}, the general diffusion forward process can be defined by the following perturbation kernel:

$$p(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \quad \mathbf{x}_0 \sim p_{\text{data}} \tag{1}$$

Where α_t, σ_t defines the noise schedule. Under the mathematical equivalence of various noise schedules [23, 25], we choose the parameterizations of flow matching (*i.e.*stochastic interpolants, $\alpha_t = 1 - t, \sigma_t = t$, along with velocity prediction model) [2, 28, 30] for brevity. The following reverse time ordinary differential equation is conducted to generate samples:

$$\mathrm{d}\mathbf{x}_t = \mathbf{v}_\theta(\mathbf{x}_t, t)\mathrm{d}t, \mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I}), \quad t : T \to 0 \tag{2}$$

To obtain the network approximate $\mathbf{v}_\theta(\mathbf{x}_t, t)$ given the state. One has to conduct the following simulation-free conditional flow matching training:

$$\mathbb{E}_{t, \mathbf{x}_t, \mathbf{x}_0, \epsilon} \left[\| v_\theta(\mathbf{x}_t, t) - (\epsilon - \mathbf{x}_0) \|_2^2 \right] \tag{3}$$

Where $\mathbf{x}_0, \epsilon$ are sampled from the data distribution p_{data} and isotropic gaussian $\mathcal{N}(0, \mathbf{I})$ respectively. The state $\mathbf{x}_t$ is sampled from the conditional probability path $p_t(\cdot \mid \mathbf{x}_0, \epsilon)$, which is $\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\epsilon$.

Classifier and Classifier-Free Guidance. To control the generation process, guidance techniques modify the score or velocity field at inference time. Classifier Guidance [7] was first introduced to steer generation by sampling from a distribution $p_w(\mathbf{x}_t \mid \mathbf{c}) \propto p(\mathbf{x}_t) p(\mathbf{c} \mid \mathbf{x}_t)^w$, which in practice approximate the score function as:

$$\nabla_{\mathbf{x}_t} \log p_w(\mathbf{x}_t \mid \mathbf{c}) = \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t) + w \cdot \nabla_{\mathbf{x}_t} \log p(\mathbf{c} \mid \mathbf{x}_t) \tag{4}$$

Classifier-Free Guidance (CFG) [15] avoids the need for a separate classifier by instead training a single diffusion model to learn both conditional and unconditional distributions. This is achieved by randomly dropping the condition $\mathbf{c}$ during training (i.e., replacing it with a null token $\emptyset$). The guided velocity is then formed by extrapolating from the unconditional prediction to the conditional one:

$$\mathbf{v}_w(\mathbf{x}_t, t, \mathbf{c}) = w \cdot \mathbf{v}(\mathbf{x}_t, t, \mathbf{c}) + (1 - w) \cdot \mathbf{v}(\mathbf{x}_t, t, \emptyset) \tag{5}$$

where w is the guidance scale.

Inferior Model Guidance/Condition-Agnostic Guidance. Inferior Model Guidance [1, 4, 17, 21, 34] bears a parameterization-level similarity to CFG but arises from a different motivation. Instead of using an unconditional estimate, it leverages an inferior but condition-aligned model, $\tilde{\mathbf{v}}_\theta$, to guide the primary strong model, $\mathbf{v}_\theta$. The guided velocity is computed via a similar extrapolation:

$$\mathbf{v}_w(\mathbf{x}_t, t, \mathbf{c}) = w \cdot \mathbf{v}(\mathbf{x}_t, t, \mathbf{c}) + (1 - w) \cdot \tilde{\mathbf{v}}(\mathbf{x}_t, t, \mathbf{c}) \tag{6}$$

Here, the weak model $\tilde{\mathbf{v}}$ is conditioned on the same inputs and is designed to be less accurate than $\mathbf{v}_\theta$. This is typically achieved by using a smaller network [21] or by perturbing the architecture of the strong model at inference time [1, 17, 34].

4. Method

In this section, from the perspective of weak-to-strong (W2S) principle, we first categorize existing guidance methods into two major groups: condition-dependent and condition-agnostic approaches. We analyze the operation regimes of these two categories under various preconditions and, based on our findings, propose SGG that combines their respective benefits. Finally, we extend W2S with SGG beyond inference by migrating it into the training phase, offering a suite of choices to directly improve the unguided diffusion models' generalization capacity.

4.1. Weak-to-strong guidance principle

A general extrapolation formula for weak-to-strong guidance can be expressed as:

$$\begin{aligned} \mathbf{v}_w(\mathbf{x}_t, t, \mathbf{c}) &= \mathbf{v}_{\text{weak}} + w(\mathbf{v}_{\text{strong}} - \mathbf{v}_{\text{weak}}) \\ \mathbf{v}_{\text{strong}} &= \mathbf{v}(\mathbf{x}_t, t, \mathbf{c}), \quad \mathbf{v}_{\text{weak}} = \tilde{\mathbf{v}}(\mathbf{x}_t, t, \tilde{\mathbf{c}}) \end{aligned} \tag{7}$$

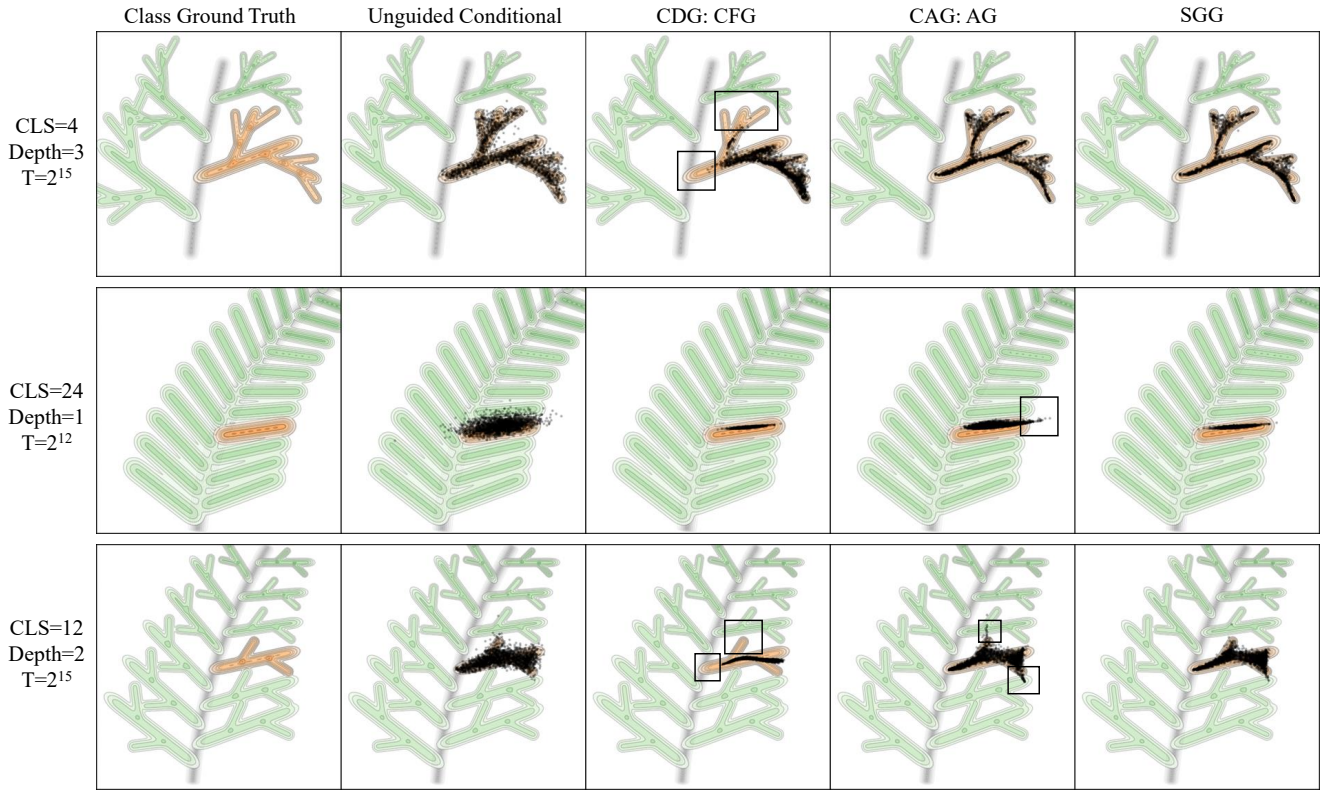

Figure 2. Recursive toy example with varying class complexity and in-class distribution (granular of the condition). 1st row: In a well fitted model and the conditional information is blurry, CFG [15] exhibits mode-seeking capacity while lack diversity. 2nd row: In a less fitted model and the conditional information is sharp, AG [21] improves diversity while leads to outliers. 3rd row: In practice, SGG incorporates the mode-seeking capacity of CFG in high noise levels while applying AG in low noise levels to preserve the in-class distribution.

where $\mathbf{v}, \mathbf{c}$ and $\tilde{\mathbf{v}}, \tilde{\mathbf{c}}$ is the strong and weak velocity output and their corresponding condition input. w is the guidance scale. The primary distinction between guidance methods lies in how the weak signal $\tilde{\mathbf{v}}(\mathbf{x}_t, t, \tilde{\mathbf{c}})$ is constructed.

In **Condition-Dependent Guidance (CDG)**, exemplified by CFG [15], creates a weak signal by manipulating the *condition*: the model architecture is identical, but the condition is dropped ($\tilde{\mathbf{v}} = \mathbf{v}, \tilde{\mathbf{c}} = \emptyset$). On the other hand, **Condition-Agnostic Guidance (CAG)** [1, 17, 21] creates a weak signal by manipulating the *model*: the condition is preserved (either with or without), but the model itself is made inferior ($\tilde{\mathbf{v}} = \mathbf{v}_{\text{inferior}}, \tilde{\mathbf{c}} = \mathbf{c}$) by either using a separate smaller network [21] or by perturbing the main model [4, 17].

4.2. Effective regimes of CAG and CDG

The choice between CAG and CDG is not absolute. On one hand, CAG, such as AG [21] with EDM2 [22] models, has been shown to outperform CDG (*e.g.*CFG) on class-conditional benchmarks like ImageNet [21, 22]. On the other hand, CDG remains the dominant and more robust method for large-scale text-to-image [1, 34] and audio generation [49] tasks, where CAG-based methods like AG [21] and PAG [1] fall short.

While varying factors such as data distribution, training iterations, weak model construction, and sampling initial noise can all contribute to the performance gap, this work investigates the following two perspectives. We interpret that the effectiveness of each guidance type is not absolute, but can be influenced by two key factors: **granularity of the condition** and the **model's fitting capacity.** To visually substantiate this hypothesis, we conduct synthetic experiments across settings to isolate the effective operational regimes of CAG and CDG. For dataset construction, we follow the principle of [21] by creating a toy dataset based on a recursive mixture of Gaussians. This setup allows us to precisely control the class number (granularity of the condition) and the recursive depth (in-class complexity). We choose CFG and AG as instances of CDG and CAG respectively. Detail of the configuration can be referred in the appendix.

Failure mode of CDG. In our first experiment, we simulate a task with conditional ambiguity (*i.e.*, fewer classes) but high in-class complexity: CLS $= 4$, Depth $= 3$. We train the model for $T = 2^{15}$ iterations to ensure a relatively strong fit to the in-class distributions. Illustrated in

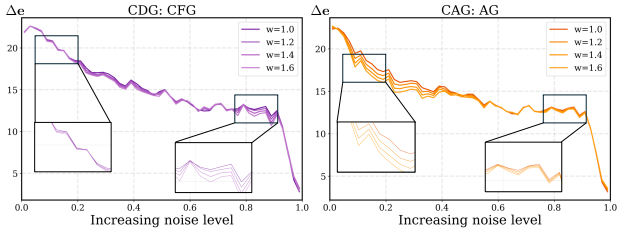

Figure 3. Applying guidance reduces the gap to optimal velocity $\dot{\mathbf{v}}$. The error-correction of CFG is prominent at high noise levels, while the effect of AG is prominent at low noise levels.

1st row of Fig. 2, we observe that once the model has captured the overall shape of each cluster, applying CDG causes mode-seeking behavior [7]: it pushes samples toward high-density regions, failing to cover the lower-density parts of the class manifold. In contrast, CAG avoids mode collapse, sharpening the class distribution while preserving intra-class coverage. This finding is analogous to results on well-fitted models on ImageNet-1K [36, 58], where CFG is inferior to AG [22] or even under the performance of unguided generation [58]. However, this trend is inverted in large-scale text-to-image generation [8]. Given the complexity of the task along with the poor unguided generation results, the robustness of CFG consistently surpasses CAG variants like TPG [34] and PAG [1].

Failure mode of CAG. To provide a counter-example where CAG loses its effectiveness in synthetic settings, we now increase the task's conditional complexity while keeping the in-class distribution simple: CLS $= 24$, Depth $= 1$. We use $T = 2^{12}$ training iterations, which is insufficient to fit the data. As shown in 2nd row of Fig. 2, the unguided conditional generation from this model produces outliers. In this underfitted regime, CAG struggles to generate plausible samples and produces artifacts that lie off-manifold or belong to incorrect classes. CDG, in contrast, successfully mitigates this failure by strongly enforcing the condition, it steers the errant samples back toward their classes, removing outliers. We therefore infer that CDG excels at **inter-class separation** and class manifold seeking. And CAG is better suited for **intra-class refinement** once the model is already well-fitted to the condition manifolds.

Simulating realistic scenarios. Practical applications, such as large-scale text-to-image models, are characterized by complex condition and detailed in-class distributions. Usually needs large guidance scale (*e.g.*7.5 for Stable Diffusion [8, 35]) for better generalization. We now increase the recursive depth to 2 with 12 classes. In this setting, the model ($T = 2^{15}$) captures the approximate in-class shape but still produces significant outliers. Illustrated in 3rd row of Fig. 2, both standard guidance methods fail in distinct ways: CDG, as before, exhibits mode-seeking behavior [7, 15] and collapses the in-class structure, while CAG preserves the general shape but fails to correct the outliers, leaving them far from the data manifold. Given the trade-offs of CAG and CDG, it is natural to take a step further by devising a practical implementation based on their operational regimes.

Introducing Segmented Guidance (SGG). To bridge the gap between our 2D synthetic analysis and high-dimensional images, we now quantify the error-correction capacities of CDG and CAG on ImageNet [36]. This allows us to investigate their operational regimes in a realistic, high-dimensional setting.

Theoretically, a perfectly fitted model could reconstruct the entire training set (memorization), but in practice, network inductive biases and inevitable approximation error lead to generalization [13]. In large-scale tasks [8, 36], this approximation error accumulates, often causing the unguided model's trajectory to drift far from the data manifold and resulting in perceptually unsatisfying samples. To understand how CDG and CAG alleviate this, we pretrain a SiT-B/2 model on ImageNet. We then compute the guided velocity $\mathbf{v}_w(\mathbf{x}_t, t, \mathbf{c})$ for both CFG (as CDG) and AG (as CAG) across all timesteps during generation and measure its distance to the theoretical optimal velocity, $\dot{\mathbf{v}}(\mathbf{x}_t, t, \mathbf{c})$.

The optimal conditional velocity $\dot{\mathbf{v}}$, derived from the dataset (please refer to appendix for derivation and configuration), is:

$$\dot{\mathbf{v}}(\mathbf{x}_t, t, \mathbf{c}) = \mathbb{E}_{\mathbf{x}_0 \sim p(\cdot|\mathbf{c}), \epsilon \sim \mathcal{N}(0,\mathbf{I})}[\mathbf{u} = \boldsymbol{\epsilon} - \mathbf{x}_0 \mid \mathbf{x}_t, t] \quad (8)$$

$$= \frac{\sum_{i=1}^{N} (\mathbf{x}_t - \mathbf{x}_0^i) \mathcal{N}(\mathbf{x}_t; (1-t)\mathbf{x}_0^i, t^2\mathbf{I})}{t \sum_{j=1}^{N} \mathcal{N}(\mathbf{x}_t; (1-t)\mathbf{x}_0^j, t^2\mathbf{I})} \quad (9)$$

We measure the guidance error as the Inception distance [14] between the guided and optimal velocities, $\Delta\mathbf{e} = \mathbb{E}_{\mathbf{x}_t}[\mathrm{d}(\dot{\mathbf{v}}, \mathbf{v}_w)]$, capturing the perceptual alignment on high dimension images. As observed in Fig. 3, the error-correction properties of the two classes are temporally separated: CDG (CFG) is most effective at high noise levels, while CAG (AG) is more effective at low noise levels. This corroborates the finding that semantic, high-level information (inter-class) is resolved in early sampling steps [53, 57], while fine-grained perceptual details (intra-class) are resolved in late steps close to data [5].

Inspired by these distinct operational regimes, we propose a simple yet effective hybrid mechanism called **SGG** (Segmented Guidance). Formally, the guided velocity $\mathbf{v}_w$ is:

$$\mathbf{v}_w(\mathbf{x}_t, t, \mathbf{c}) = \mathbf{v}(\mathbf{x}_t, t, \mathbf{c}) + (w-1) \cdot \mathbf{g}(\mathbf{x}_t, t, \mathbf{c}) \quad (10)$$

where $\mathbf{v}$ is the strong model, w is the guidance scale, and the guidance direction $\mathbf{g}$ is segmented by time τ:

$$\mathbf{g}(\mathbf{x}_t, t, \mathbf{c}) = \begin{cases} \mathbf{v}(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}(\mathbf{x}_t, t, \emptyset) & \text{if } t > \tau \\ \mathbf{v}(\mathbf{x}_t, t, \mathbf{c}) - \tilde{\mathbf{v}}(\mathbf{x}_t, t, \mathbf{c}) & \text{if } t \leq \tau \end{cases} \quad (11)$$

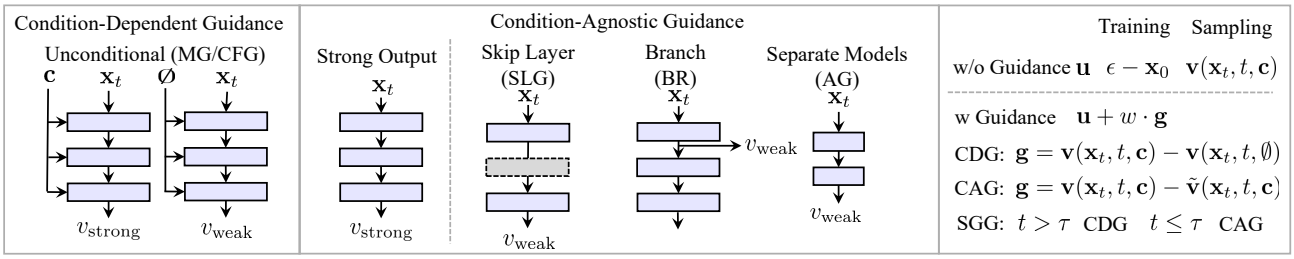

Figure 4. I: Two groups of construction of the weak models, condition-dependent and condition-agnostic. II: Segmented Guidance applied in training and sampling

The core idea is to first leverage CDG for **condition manifold seeking** at high noise levels ($t \geq \tau$) and subsequently apply CAG for **in-condition refinement** at low noise levels ($t < \tau$).

4.3. Training integration

While both the regression target and guidance mechanisms are critical for generalization [21, 45], they remain fundamentally decoupled, applied separately during training and sampling. We take a step forward by integrating the Weak-to-Strong (W2S) principle with SGG directly into the training phase. This approach aims to improve the generalization of unguided diffusion model, thereby boosting inference efficiency by reducing the need for an extra guidance call.

Training target modification. To integrate the extrapolation capacity of guidance explicitely into the training phase, thus reducing the extra forward call of guidance during inference, we modify the standard velocity-matching objective. The conventional training target is the coupling level [28] optimal transport $\mathbf{u} = \boldsymbol{\epsilon} - \mathbf{x}_0$. We augment this target with a guidance term derived from the difference between the strong and weak signal:

$$\mathbf{u}_{\text{w2s}} = \mathbf{u} + w \cdot \mathbf{g}(\mathbf{x}_t, t, \mathbf{c}) \tag{12}$$

This modification encourages the strong model to move beyond the conservative fit of standard MSE training and explcitely improve its extrapolative capacity. The training objective is:

$$\mathcal{L}_s = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\Big[\big\|\mathbf{v}_\theta(\mathbf{x}_t, t, \mathbf{c}) - \big(\mathbf{u} + w \cdot \text{sg}[\mathbf{g}(\mathbf{x}_t, t, \mathbf{c})]\big)\big\|_2^2\Big] \tag{13}$$

Stop-gradient (sg) is used to stablize training, following protocols of [6, 12, 48]. The primary model network serves as the strong model. The main design choice lies in constructing an effective and efficient weak signal $\mathbf{v}_{\text{weak}}$.

Construction of Weak Signals for Training. We adapt existing inference-time guidance methods for the training phase and introduce a novel, highly efficient Condition-Agnostic Guidance (CAG) variant:

- **CDG: CFG/MG.** Migrating the unconditional term ($\mathbf{v}(x_t, t, \emptyset)$) in CFG [15] into the training objective, an approach similar to MG [48].
- **CAG: AG.** Following AutoGuidance [21], we maintain a separate, smaller and less-trained network during training to function as the weak model.
- **CAG: BR.** Inspired by the sequential structure of transformer blocks, this approach generates the weak signal by supervising an auxiliary output **br**anching from an intermediate layer.

BR is condition-agnostic and requires no extra forward calls during training for guidance. We also explored layer-perturbation methods (*e.g.*, SLG [17]) but found they degraded performance when integrated into training, thus excluded them (Further discussion are provided in the appendix). Subsequently, we apply the idea of **Segmented Guidance (SGG)** directly to the training framework. The training-time version of SGG uses the condition-dependent guidance (CFG) signal for high noise levels ($t \geq \tau$) and switches to the condition-agnostic guidance (BR) signal for low noise levels ($t < \tau$). Illustration of the pipeline is provided in Fig. 4.

5. Experiments

We validate our methods in two settings. First, we demonstrate the effectiveness of our inference-time SGG on state-of-the-art text-to-image models (SD3, SD3.5 [8]). Second, to perform a controlled and computationally feasible analysis of our training-time integration, we follow standard practice [56] and use the SiT-B/2 model on ImageNet, which allows us to ablate the W2S training targets (MG, AG, BR, and SGG) and measure the impact on training convergence and generalization.

5.1. Implementation details

Inference-time guidance. For pre-trained model, we use the SD3-Medium and SD3.5-Medium as base models [8]. We use MS-COCO-1k [27] subset and LAION-1k [42] subset for prompt instantiation. We compare our method against several baselines, including standard conditional generation (no guidance), CFG [15], and Skip-Layer Guidance (SLG). We also include comparisons to recent advanced guidance variants, such as S^2-Guidance [4], Guidance Interval [24], CFG+SLG [17], CFG-Zero* [9] and Rectified-CFG++ [40].

Models		SD3-medium				SD3.5-medium			
Dataset		MS-COCO-1K		LAION-5B-1K		MS-COCO-1K		LAION-5B-1K	
Metrics	NFE/s	HPSv2.1	Aes.	HPSv2.1	Aes.	HPSv2.1	Aes.	HPSv2.1	Aes.
Conditional (w/o Guidance)	1	21.118	4.864	20.215	4.938	21.204	4.978	20.801	4.978
CFG [15]	2	29.199	5.267	28.174	5.193	29.199	5.279	28.333	5.203
SLG [17]	2	26.685	<u>5.581</u>	25.000	5.415	27.295	<u>5.714</u>	26.050	<u>5.512</u>
CFG+SLG [17]	3	27.759	5.468	26.147	5.285	28.331	5.678	27.246	5.421
CFG-zero-star [9]	2	28.296	5.250	28.139	5.259	27.954	5.249	28.272	5.244
S^2 Guidance [4]	3	**29.614**	5.333	<u>28.442</u>	5.244	<u>29.614</u>	5.342	<u>28.491</u>	5.250
Rectified CFG++ [40]	3	28.932	5.399	28.306	<u>5.443</u>	28.540	5.425	28.122	5.406
Guidance Interval [24]	2	29.126	5.321	28.174	5.257	29.077	5.326	28.125	5.254
SGG	2	<u>29.541</u>	**5.614**	**28.638**	**5.489**	**29.736**	**5.717**	**28.685**	**5.518**

Table 1. Quantitative comparison of guidance methods on MS-COCO-1K and LAION-5B-1K, evaluating both HPSv2.1 and aesthetic scores for SD3 and SD3.5 models. Best results are in **bold**, second-best are <u>underlined</u>.

We use the standard 28 inference steps throughout experiments. All methods are evaluated using HPSv2.1 Score [51] and Aesthetic Score [41]. We select standard CFG [15] as CDG and SLG [17] as CAG in SGG implementations.

Training-time guidance. We conduct training evaluation mainly on SiT-B/2 model [31] due to computational constraints. We use lognormal-timestep sampling throughout all experiments to boost convergence, following [43]. We perform experiments in both unconditional and conditional settings. CAG methods are applied in both settings, whereas CDG method is naturally applied only in conditional training. For the conditional setting. All models are trained for 400k iterations. The sampling configuration is SDE Euler-Maruyama sampler with steps=250. We report the FID, sFID and Inception Score for all methods.

NFE/s & time/it. We report the NFE per sampling step (NFE/s) during sampling. We also report wall-clock time per training iteration, normalized by the baseline configuration's time (time/it) to track the computation of guidance during training. Details of the implementation configurations across experiments could be referred in the appendix.

5.2. Inference time comparison

We first conduct experiments to validate the effectiveness of our Segmented Guidance (SGG) principle against standard CFG and other prevalent guidance variants. As shown in Table 1, a clear compromise between prompt-adherence (correlated to HPSv2.1) and aesthetic quality is evident in using CFG or SLG alone. For example, on the SD3.5/MS-COCO benchmark, SLG achieves a high aesthetic score (5.714), but at a significant cost to its HPSv2.1 score (27.295). Conversely, standard CFG achieves a competitive HPSv2.1 score (29.199) but produces a comparatively low aesthetic score (5.279). As a hybrid approach to take the benefits of two, our Segmented Guidance (SGG) achieves the competitive scores in both categories (HPSv2.1: 29.736 and Aesthetic: 5.717). This pattern holds across models and datasets in our evaluation, where SGG reach comparable results to other guidance variants. We also provide qualitative comparison of our methods, As illustrated in Fig. 5.

5.3. Training convergence acceleration

We subsequently evaluate the effectiveness of the migration of the weak-to-strong principle to boost training convergence, in both conditional and unconditional settings (Table 2). In the conditional setting, our experiments demonstrate that all weak-to-strong guidance integrations consistently outperform the baseline. with the hybrid SGG approach yields the best result. In unconditional setting, where CDG is not applicable, CAG methods (*e.g.*BR, AG) still provide a notable

Model	NFE/s	time/it	FID ↓	sFID ↓	IS ↑
Conditional Generation					
SiT-B/2	1	1.00	31.22	6.41	49.59
+ CFG	2	1.00	6.02	5.47	183.83
AG	1	1.27	13.96	88.36	4.68
BR	1	1.02	16.02	5.13	76.21
MG	1	1.23	5.88	6.19	253.74
SGG	1	1.22	**4.58**	**4.95**	**264.06**
SiT-B/2+REPA	1	1.00	20.46	6.31	73.00
SGG+REPA	1	1.19	**3.07**	**4.88**	**242.15**
Unconditional Generation					
SiT-B/2	1	1.00	61.27	7.00	17.33
AG	1	1.26	45.97	**4.94**	20.32
BR	1	1.02	**43.25**	5.11	**20.66**

Table 2. Training-time integration results on ImageNet 256×256 with SiT-B/2, in conditional and unconditional settings.

Figure 5. Qualitative comparison between Conditional (w/o guidance), CFG [15], SLG [17], SGG (Ours).

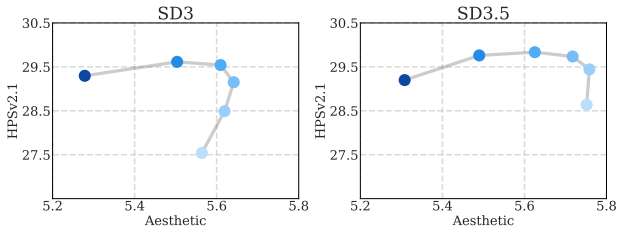

Figure 6. Ablation study on the segmentation timestep τ. We vary τ from 4 (lightest) to 24 (darkest) in increments of 4, out of 28 total sampling steps. The results indicate that a mid-range segmentation point yields the best performance.

performance boost over the baseline. We also observe that SGG in training time can be complemented with REPA [56], providing further performance gains.

Training integration of W2S guidance introduces a an extra forward call per iteration, *e.g.*, an additional 22% for the full SGG method. The cost could be offset by the resulting model's inference efficiency. The trained model's unguided output (NFE/s=1) achieves an FID of 4.58, which is superior to the guided (NFE/s=2) output of the baseline model (FID 6.02). Furthermore, BR variant in CAG incurs only a 2% training overhead. This minimal cost still yields a reasonable FID improvement over the baseline, from 31.22 to 16.02 in conditional setting and from 61.27 to 43.25 in unconditional setting.

Segmented τ	0.0	0.1	0.2	0.3	0.4
FID ↓	5.88	5.26	4.58	4.99	5.79
sFID ↓	6.19	5.44	4.95	5.37	6.37
Inception Score ↑	253.74	267.34	264.06	249.52	236.51

Table 3. Ablation study on segmentation timestamp τ conditional training with SGG on ImageNet 256x256. We choose $\tau = 0.2$ as our default setting.

5.4. Ablation study

We ablate two critical components: (1) The segmented timestep τ between CDG and CAG. (2) The guidance weight w. As illustrated in Fig. 6, our ablation on the guidance segmentation point reveals a Pareto frontier. This frontier traces the trade-off between HPSv2.1 (prompt adherence) and aesthetic score as the segmented step transitions from high noise level to low noise level. We also conducted ablation on τ in conditional training configuration, shown in Table 3.

6. Conclusion

In this work, we first clarify the generalization issues in common diffusion models and the alleviation by guidance. We then systematically analyze the operational regimes of condition-dependent and condition-agnostic approaches under the perspective of weak-to-strong principle. Based on this analysis, we proposed Segmented Guidance (SGG), a simple and effective approach that synergizes the benefits of both guidance types. We subsequently migrate W2S principle along with SGG into the training objective, thereby reducing the need for guidance during inference. Comprehensive qualitative and quantitative comparisons validate the effectiveness of both Segmented Guidance and training-time integration of weak-to-strong principle.

Limitations and future work. Our approach is limited to continuous diffusion, future work could benefit from migrating the segmentation idea of SGG to other modalities (*e.g.*discrete diffusion) and further explore the combination of different guidance instances under W2S principle.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 6250070674) and the Zhejiang Leading Innovative and Entrepreneur Team Introduction Program (2024R01007).

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 *Proc. ECCV*, 2024. 2, 3, 4, 5

[2] Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. In *Proc. ICLR*, 2023. 1, 3

[3] Lichen Bai, Masashi Sugiyama, and Zeke Xie. Weak-to-strong diffusion with reflection. In *Proc. ICLR workshop*, 2025. 2

[4] Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. S^2-guidance: Stochastic self guidance for training-free enhancement of diffusion models, 2025. 2, 3, 4, 6, 7

[5] Defang Chen, Zhenyu Zhou, Can Wang, Chunhua Shen, and Siwei Lyu. On the trajectory regularity of ODE-based diffusion sampling. In *Proc. ICML*, 2024. 1, 5

[6] Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. In *Proc. ICML*, 2025. 3, 6

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Proc. NeurIPS*, 2021. 2, 3, 5

[8] 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 *Proc. ICML*, 2024. 1, 2, 5, 6

[9] Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. Cfg-zero*: Improved classifier-free guidance for flow matching models, 2025. 2, 6, 7

[10] Alexandre Galashov, Ashwini Pokle, Arnaud Doucet, Arthur Gretton, Mauricio Delbracio, and Valentin De Bortoli. Learn to guide your diffusion model. *arXiv preprint arXiv:2510.00815*, 2025. 1

[11] Zhengqi Gao, Kaiwen Zha, Tianyuan Zhang, Zihui Xue, and Duane S Boning. Reg: Rectified gradient guidance for conditional diffusion models. In *Proc. ICML*, 2025. 2

[12] Zhengyang Geng, Mingyang Deng, Xingjian Bai, J Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Proc. NeurIPS*, 2025. 6

[13] Xiangming Gu, Chao Du, Tianyu Pang, Chongxuan Li, Min Lin, and Ye Wang. On memorization in diffusion models. *TMLR*, 2025. 5

[14] 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 *Proc. NeurIPS*, 2017. 5

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *Proc. NeurIPS Workshop*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020. 1, 3

[17] Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *Proc. CVPR*, 2025. 2, 3, 4, 6, 7, 8

[18] Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025. 3

[19] Zahra Kadkhodaie, Florentin Guth, Eero P. Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *Proc. ICLR*, 2024. 1

[20] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 1

[21] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Proc. NeurIPS*, 2024. 1, 2, 3, 4, 6

[22] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *Proc. CVPR*, pages 24174–24184, 2024. 1, 4, 5

[23] Diederik Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation. In *Proc. NeurIPS*, 2023. 3

[24] 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 *Proc. NeurIPS*, 2024. 2, 6, 7

[25] Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. In *Proc. NeurIPS*, 2024. 3

[26] Puheng Li, Zhong Li, Huishuai Zhang, and Jiang Bian. On the generalization properties of diffusion models. In *Proc. NeurIPS*, 2023. 1

[27] 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 *Proc. ECCV*, 2014. 6

[28] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proc. ICLR*, 2023. 1, 3, 6

[29] Qihao Liu, Xi Yin, Alan Yuille, Andrew Brown, and Mannat Singh. Flowing from words to pixels: A noise-free framework for cross-modality evolution. In *Proc. CVPR*, 2025. 1

[30] Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. ICLR*, 2023. 1, 3

[31] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring

flow and diffusion-based generative models with scalable interpolant transformers. In *Proc. ECCV*, 2024. 7

[32] Mang Ning, Mingxiao Li, Jianlin Su, Albert Ali Salah, and Itir Onal Ertugrul. Elucidating the exposure bias in diffusion models. In *Proc. ICLR*, 2024. 1

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. In *TMLR*, 2024. 3

[34] Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token perturbation guidance for diffusion models. In *Proc. NeurIPS*, 2025. 1, 2, 3, 4, 5

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022. 5

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 5

[37] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. In *Proc. ICLR*, 2024. 2

[38] Seyedmorteza Sadat, Otmar Hilliges, and Romann M Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *Proc. ICLR*, 2024. 2

[39] Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *Proc. ICLR*, 2025.

[40] Shreshth Saini, Shashank Gupta, and Alan C. Bovik. Rectified-cfg++ for flow based models. In *Proc. NeurIPS*, 2025. 2, 6, 7

[41] Christoph Schuhmann. Laion-aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. 7

[42] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Proc. NeurIPS*, 2022. 6

[43] Inkyu Shin, Chenglin Yang, and Liang-Chieh Chen. Deeply supervised flow-based generative models. In *Proc. ICCV*, 2025. 7

[44] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. ICML*, 2015. 1, 3

[45] Kiwhan Song, Jaeyeon Kim, Sitan Chen, Yilun Du, Sham Kakade, and Vincent Sitzmann. Selective underfitting in diffusion models. *arXiv preprint arXiv:2510.01378*, 2025. 1, 6

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. ICLR*, 2022. 1, 3

[47] George Stoica, Vivek Ramanujan, Xiang Fan, Ali Farhadi, Ranjay Krishna, and Judy Hoffman. Contrastive flow matching. In *Proc. ICCV*, 2025. 3

[48] Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025. 3, 6

[49] Junyou Wang, Zehua Chen, Binjie Yuan, Kaiwen Zheng, Chang Li, Yuxuan Jiang, and Jun Zhu. Audiomog: Guiding audio generation with mixture-of-guidance. *arXiv preprint arXiv:2509.23727*, 2025. 2, 4

[50] Runqian Wang and Kaiming He. Diffuse and disperse: Image generation with representation regularization. *arXiv preprint arXiv:2506.09027*, 2025. 3

[51] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. In *Proc. ICCV*, 2023. 7

[52] Mengfei Xia, Nan Xue, Yujun Shen, Ran Yi, Tieliang Gong, and Yong-Jin Liu. Rectified diffusion guidance for conditional generation. In *Proc. CVPR*, 2025. 2

[53] Fu Xiaomeng and Li Jia. Tcfg: Truncated classifier-free guidance for efficient and scalable text-to-image acceleration. In *Proc. ICCV*, 2025. 5

[54] Yilun Xu, Shangyuan Tong, and Tommi S Jaakkola. Stable target field for reduced variance score estimation in diffusion models. In *Proc. ICLR*, 2023. 3

[55] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proc. CVPR*, 2025. 3

[56] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *Proc. ICLR*, 2025. 3, 6, 8

[57] Zheyuan Zhan, Defang Chen, Jian-Ping Mei, Zhenghe Zhao, Jiawei Chen, Chun Chen, Siwei Lyu, and Can Wang. Conditional image synthesis with diffusion models: A survey. *TMLR*, 2024. 1, 5

[58] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 5