

SAND: Stage-Aware Noise Decomposition for Training-Free Diffusion Guidance

DaeHyun Kim¹, Hyo-Jun Lee², Hanul Kim³, and Yeong Jun Koh^{†1}

¹ Chungnam National University

² Kangwon National University

³ Seoul National University of Science and Technology

Abstract. Diffusion models have achieved state-of-the-art performance across 2D, video, and 3D generation, but their sampling quality remains heavily dependent on guidance methods. Existing training-free guidance techniques construct a weakened auxiliary branch through static perturbations, yet these methods fundamentally disregard the stage-wise structure of the denoising process: early timesteps govern magnitude-driven global formation, whereas later timesteps refine directional geometry. We propose a principled training-free diffusion guidance with Stage-Aware Noise Decomposition (SAND) that considers this temporal asymmetry by decomposing Gaussian noise into orthogonal magnitude- and direction-oriented components and modulating them with a time-adaptive weighting schedule aligned to diffusion dynamics. The decomposed perturbations are injected directly into the self-attention queries, keys, and values, forming a dynamically degraded branch that requires no retraining while faithfully mirroring the evolving geometry of the sample. This representation-level perturbation provides a temporally coherent guidance signal that better balances global structure and fine detail. Extensive experiments on Stable Diffusion XL (SDXL) and Stable Diffusion 3 (SD3) demonstrate that our method consistently improves global coherence, sharpens fine-grained texture, and yields superior sample quality in both unconditional and conditional image generation. We further show that SAND can be extended to the guidance-distilled image generation model FLUX.1 and the text-to-video model Wan2.2, highlighting its broad applicability.

Keywords: Diffusion Models · Training-Free Guidance · Stage-Aware Noise Decomposition

1 Introduction

Diffusion models [10, 16, 34, 35] have emerged as one of the most powerful tools of generative models, enabling high-quality generation across a broad range of domains, including 2D image [6, 18, 25, 26, 29], video [2, 9, 12, 27], and 3D generation [5, 23, 28, 38]. Sampling guidance techniques [1, 11, 13, 17] have significantly

[†] : Corresponding author.

contributed to this success. These methods commonly contain a weaker auxiliary generation branch that produces undesirable samples and guide the denoising trajectory away from these degraded results, effectively steering the sampling process toward high-quality regions of the data manifold.

Among these guidance methods, Classifier-Free Guidance (CFG) has been widely adopted, enhancing conditional generation performance by intentionally utilizing the unconditional generation path as a weak model and guiding the sampling trajectory away from unconditional predictions toward condition-consistent results. However, CFG requires an additional training strategy and is applicable only to conditional generation in the inference phase, as its guidance signal relies on the difference between conditional and unconditional denoising results, which can lead to degraded results when the guidance scale is large. As alternatives, some methods [1, 13, 14, 17] employ self-perturbation to mimic a weak generation, thereby enabling guidance without relying on conditional information. For example, Autoguidance [17] uses an earlier-trained or lower-capacity version of the diffusion model as the weaker generator. However, Autoguidance still requires retraining a network to obtain the weaker variant, which makes it inefficient and impractical for large-scale or publicly released models [6, 26, 29]. Meanwhile, attention-based self-perturbation methods that do not require additional training have also been explored. For instance, PAG [1] substitutes the self-attention maps of the diffusion model with an identity matrix during inference to produce undesirable samples. SEG [13] applies Gaussian smoothing to the self-attention maps to construct an implicit weak generation.

However, PAG and SEG impose a fixed degradation throughout sampling, without considering the stage-wise dynamics of diffusion. Because early steps mainly establish the global semantic layout whereas later steps refine high-frequency details, this fixed strategy can disrupt the sampling trajectory and limit the model’s ability to adaptively balance global structure and fine details.

To address these limitations, we introduce Stage-Aware Noise Decomposition (SAND), a novel training-free diffusion guidance designed to adaptively align the perturbation strategy with the intrinsic temporal dynamics of the denoising process. Unlike prior methods that rely on static and indiscriminate degradation, our work is motivated by the key insight that diffusion timesteps exhibit semantically distinct behaviors: early iterations shape global structure through magnitude-driven modulation, whereas later iterations refine high-frequency geometry through directional adjustments. To this end, SAND decomposes Gaussian noise into orthogonal magnitude- and direction-oriented components, and modulates them via a time-adaptive weighting schedule aligned with diffusion dynamics. These decomposed perturbations are injected directly into the self-attention queries, keys, and values, forming a dynamically weakened branch that guides the diffusion model toward a more desirable trajectory, while faithfully mirroring the geometric evolution of the sample without any retraining. Extensive experiments on both unconditional and conditional image generation using SDXL [26] and SD3 [6] demonstrate that our method yields more coherent global layouts, sharper local details, and consistently higher sample quality. Further-

more, SAND can be seamlessly extended to CFG distillation-based models like FLUX.1 [22] and text-to-video generation models such as Wan2.2 [37], highlighting its broad applicability. To summarize, the main contributions of this work are as follows:

- We identify the stage-wise semantic roles of diffusion timesteps and show that early and late iterations require fundamentally different forms of perturbation for effective guidance.
- We propose a training-free guidance method that decomposes Gaussian noise into magnitude- and direction-oriented components and modulates them with a time-adaptive weighting schedule aligned to diffusion dynamics.
- We introduce a representation-level perturbation mechanism that injects the decomposed noise into self-attention projections, yielding consistent improvements in global structure, local detail, and overall sample quality across multiple diffusion models.

2 Related Works

2.1 Guidance with Trained Weak Model

Classifier-Free Guidance (CFG) [11] is a widely adopted technique for boosting conditional generation in diffusion models, primarily by leveraging an unconditional model to serve as the weak branch. However, this approach introduces challenges: the discrepancy in objectives between the conditional and unconditional models often results in diminished sample diversity [17].

Autoguidance [17] addresses these limitations by replacing the weak model with a deliberately degraded version of the original model, trained with reduced capacity to ensure alignment. This inherent alignment enables the guidance algorithm to more reliably identify and correct errors by comparing the prediction outputs of the two models. Although effective in improving sample quality while preserving diversity, this method requires an additional training phase, which poses a substantial practical challenge, particularly for large-scale diffusion models that are already publicly deployed.

2.2 Guidance with Training-Free Weak Model

Alternative approaches leverage the model’s own architecture to create a contrastive signal, specifically by degrading the self-attention blocks during inference. Rather than relying on auxiliary training, these techniques induce a weak model on-the-fly. For example, SAG [14] applies localized blurring to high-attention zones, PAG [1] replaces attention matrices with identity matrices, and SEG [13] smooths the attention weights with Gaussian blur. Although effective, these methods rely on static perturbations to the attention mechanism. In contrast, our method introduces noise directly to the attention inputs. By injecting decomposed, timestep-dependent noise into the query, key, and value representations, we construct a more principled and stage-aligned weak trajectory.

Another line of training-free guidance methods modifies the guidance direction in the prediction or sampling-update space using geometric decomposition. APG [30] decomposes the CFG update into components parallel and orthogonal to the conditional prediction, and down-weights the parallel component to mitigate oversaturation and artifacts at high guidance scales. TCFG [21] analyzes the misalignment between conditional and unconditional scores and damps the tangential component of the unconditional score to keep the sampling trajectory closer to the target manifold. TAG [3] decomposes the sampling increment into radial and tangential components, amplifying the tangential component to improve semantic consistency. These methods operate after the denoising prediction or sampling update has been obtained. In contrast, SAND applies decomposition before the final prediction is formed, directly in the feature space of self-attention. Rather than reshaping an already computed guidance direction, SAND decomposes Gaussian noise with respect to the query, key, and value features and uses the resulting magnitude- and direction-oriented perturbations to generate the weak prediction itself.

3 Preliminaries

3.1 Diffusion models

Diffusion models generate samples through an iterative denoising process that reverses a forward process, which gradually corrupts data into Gaussian noise. The forward process that adds noise to an image $\mathbf{x}$ over time $t \in [0, T]$ can be represented by a stochastic differential equation (SDE) [35]:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \quad (1)$$

where $\mathbf{f}$ and g are predefined functions representing the drift and diffusion coefficients, respectively, and $d\mathbf{w}$ denotes the standard Wiener process. Correspondingly, the denoising process can be formulated by the reverse SDE:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla_{\mathbf{x}} \log p_t(\mathbf{x})] dt + g(t)d\bar{\mathbf{w}} \quad (2)$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ denotes the score of the noisy data distribution, and $d\bar{\mathbf{w}}$ represents the standard Wiener process in the reverse direction. The score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is approximated by a neural network $\mathbf{s}_\theta(\mathbf{x}, t)$ [4, 11, 34, 35] parameterized by weights θ , trained using denoising score matching [36]. Therefore, the denoising process of the diffusion model can be rewritten as:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \mathbf{s}_\theta(\mathbf{x}, t)] dt + g(t)d\bar{\mathbf{w}}. \quad (3)$$

3.2 Sampling Guidance

While diffusion models trained with denoising score matching are built on solid theoretical principles, the estimated score functions are frequently imprecise because of constrained model capacity. Consequently, sampling solely based on

the learned score often produces structurally incomplete or distorted images. To address the problem and enhance sampling quality, sampling guidance methods [1, 11, 13, 17] incorporate weaker generation samples into the denoising process, guiding it away from these degraded samples toward high-quality regions of the data manifold. In particular, the denoising process in Equation (3) can be extended as follows:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \tilde{\mathbf{s}}_\theta(\mathbf{x}, t)] dt + g(t) d\bar{\mathbf{w}} \quad (4)$$

$$\tilde{\mathbf{s}}_\theta(\mathbf{x}, t) = \gamma \mathbf{s}_\theta(\mathbf{x}, t) - (\gamma - 1) \mathbf{s}_\theta^{\text{weak}}(\mathbf{x}, t) \quad (5)$$

where $\mathbf{s}_\theta^{\text{weak}}(\mathbf{x}, t)$ is the weaker model that generates degraded samples, and γ is a hyperparameter that controls the strength of the guidance.

4 Method

In this study, we introduce a novel training-free guidance technique, called SAND (Stage-Aware Noise Decomposition), which decomposes Gaussian noise into directional and magnitude components and injects them into the self-attention module to promote more stochastic and diverse guidance behaviors during sampling. In the denoising process of diffusion models, the early steps primarily establish the global semantic layout, whereas the later steps refine the structural details. The displacement of a data vector can be interpreted as moving the corresponding point in the vector space along its direction and magnitude. From this perspective, the early denoising steps mainly modulate the magnitude of pixel vectors, defining the overall intensity and coarse structure. In contrast, the later steps adjust the direction of pixel vectors to refine fine-grained appearance and local geometry. To provide guidance that aligns with these distinct denoising stages, we first decompose Gaussian noise into directional and magnitude components (Section 4.1) and then adaptively regulate the contribution of each component according to the denoising stage (Section 4.2). Finally, the adjusted components are directly injected into the query, key, and value of the self-attention mechanism (Section 4.3), enabling more dynamic and stage-aware guidance generation throughout the sampling process.

4.1 Noise Decomposition

Let $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ denote an input feature map within a self-attention layer, where H and W denote spatial dimensions and C is the channel dimension. Each spatial position (h, w) corresponds to a feature vector $\mathbf{F}(h, w) \in \mathbb{R}^C$. Directly adding Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\epsilon \in \mathbb{R}^{H \times W \times C}$, to $\mathbf{F}$ induces a stochastic displacement of each feature vector $\mathbf{F}(h, w)$ in the feature space. This noise perturbation, when applied to the self-attention layers within the weaker model $\mathbf{s}_\theta^{\text{weak}}(\mathbf{x}, t)$, can yield degraded generations. However, such indiscriminate perturbation fails to reflect the characteristics of each denoising step, where the early stages primarily reconstruct the global semantic layout, while the later stages refine high-frequency and fine-grained details.

To provide more stage-aware guidance for the denoising process, the perturbation should be aligned with the distinct roles of each denoising stage. Specifically, the early stages focus on magnitude modulation to define the overall intensity and coarse structure, whereas the later stages perform directional adjustment to refine fine-grained appearance and local geometry. Therefore, to align the sampling guidance with these stage-specific roles, we decompose Gaussian noise ϵ into two components:

$$\epsilon_{\parallel}(h, w) = \frac{\mathbf{F}(h, w)^{\top} \epsilon(h, w)}{\|\mathbf{F}(h, w)\|_2^2} \mathbf{F}(h, w) \quad (6)$$

$$\epsilon_{\perp}(h, w) = \epsilon(h, w) - \epsilon_{\parallel}(h, w). \quad (7)$$

The parallel component $\epsilon_{\parallel}(h, w)$ corresponds to the portion of Gaussian noise that is aligned with the direction of the feature vector $\mathbf{F}(h, w)$. Therefore, adding $\epsilon_{\parallel}(h, w)$ to $\mathbf{F}(h, w)$ primarily affects the magnitude of the feature vector, while its orientation remains unchanged. Consequently, the orthogonal component $\epsilon_{\perp}(h, w)$ can be interpreted as directional noise that adjusts the geometric structure of $\mathbf{F}(h, w)$, as it is derived as the residual of the parallel component that affects magnitude.

4.2 Stage-wise Weight Strategy of Decomposed Noise

To align the perturbation of the weaker model $\mathbf{s}_{\theta}^{\text{weak}}(\mathbf{x}, t)$ with the denoising objective of $\mathbf{s}_{\theta}(\mathbf{x}, t)$, we introduce a stage-wise weight strategy that balances the contributions of the parallel and orthogonal noise components, $\epsilon_{\parallel}(h, w)$ and $\epsilon_{\perp}(h, w)$. The parallel component $\epsilon_{\parallel}(h, w)$ dominates the early denoising stages by governing magnitude variations, whereas the orthogonal component $\epsilon_{\perp}(h, w)$ gradually becomes influential in directional refinement at later stages. To this end, we formulate the perturbation applied to each self-attention layer of $\mathbf{s}_{\theta}^{\text{weak}}(\mathbf{x}, t)$ as:

$$\mathbf{F}_{\epsilon}^{(t)}(h, w) = \mathbf{F}^{(t)}(h, w) + \lambda_{\parallel}^{(t)} \epsilon_{\parallel}^{(t)}(h, w) + \lambda_{\perp}^{(t)} \epsilon_{\perp}^{(t)}(h, w) \quad (8)$$

where $t \in [0, 1]$ denotes the normalized diffusion timestep and the time-dependent scaling scalars $\lambda_{\parallel}^{(t)}$ and $\lambda_{\perp}^{(t)}$ are defined as:

$$\lambda_{\parallel}^{(t)} = \alpha(1 - t), \quad \lambda_{\perp}^{(t)} = \alpha t. \quad (9)$$

Here, α is a global coefficient balancing the overall perturbation. This formulation ensures that the parallel component $\epsilon_{\parallel}^{(t)}$, which modulates the magnitude of the feature vectors, dominates during the early denoising stages $t \rightarrow 0$, while the orthogonal component $\epsilon_{\perp}^{(t)}$, which adjusts their directional variation, gradually becomes more influential in later stages $t \rightarrow 1$.

Algorithm 1 Sampling Guidance with Stage-Aware Noise Decomposition

Require: Initial noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, timesteps t down to 1, guidance scale γ , global coefficient α
Ensure: Generated image $\mathbf{x}_0$

```

1: for  $t = 0, \dots, 1$  do
2:   Compute original score  $\mathbf{s}_\theta(\mathbf{x}_t, t)$ 
3:   // Construct Weaker Model  $\mathbf{s}_\theta^{\text{weak}}$  via Perturbation
4:   for each self-attention layer in  $\mathbf{s}_\theta^{\text{weak}}$  do
5:     for  $\mathbf{F} \in \{\mathbf{Q}, \mathbf{K}, \mathbf{V}\}$  do
6:       1. Noise Decomposition:
7:       Sample  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$  and decompose:  $\epsilon_{\parallel} \leftarrow \frac{\mathbf{F}^\top \epsilon}{\|\mathbf{F}\|_2} \mathbf{F}$ ,  $\epsilon_{\perp} \leftarrow \epsilon - \epsilon_{\parallel}$ 
8:       2. Stage-wise Scheduling:
9:        $\lambda_{\parallel}^{(t)} \leftarrow \alpha(1-t)$ ,  $\lambda_{\perp}^{(t)} \leftarrow \alpha t$ 
10:      3. Integrate Perturbation:
11:       $\mathbf{F}_{\epsilon}^{(t)} \leftarrow \mathbf{F}^{(t)} + \lambda_{\parallel}^{(t)} \epsilon_{\parallel} + \lambda_{\perp}^{(t)} \epsilon_{\perp}$ 
12:    end for
13:    Compute perturbed attention:  $\mathbf{O}_{\epsilon} \leftarrow \text{Softmax}(\mathbf{Q}_{\epsilon} \mathbf{K}_{\epsilon}^\top / \sqrt{C}) \mathbf{V}_{\epsilon}$ 
14:  end for
15:  Let  $\mathbf{s}_\theta^{\text{weak}}(\mathbf{x}_t, t)$  be the output using the perturbed attention layers
16:  4. Apply Extrapolated Guidance:
17:   $\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t) \leftarrow \gamma \mathbf{s}_\theta(\mathbf{x}_t, t) - (\gamma - 1) \mathbf{s}_\theta^{\text{weak}}(\mathbf{x}_t, t)$ 
18:  Update  $\mathbf{x}_t \rightarrow \mathbf{x}_{t-1}$  using  $\tilde{\mathbf{s}}_\theta(\mathbf{x}_t, t)$  via SDE/ODE solver
19: end for
20: return  $\mathbf{x}_0$ 

```

4.3 Integrating Perturbations into Self-Attention

The perturbation in Equation 8 is applied to the self-attention operation in $\mathbf{s}_\theta^{\text{weak}}(\mathbf{x}, t)$. Specifically, in each self-attention module, the query $\mathbf{Q}^{(t)} \in \mathbb{R}^{H \times W \times C}$, key $\mathbf{K}^{(t)} \in \mathbb{R}^{H \times W \times C}$, and value $\mathbf{V}^{(t)} \in \mathbb{R}^{H \times W \times C}$ are first projected from the input features. We then perturb them to obtain $\mathbf{Q}_{\epsilon}^{(t)}$, $\mathbf{K}_{\epsilon}^{(t)}$, and $\mathbf{V}_{\epsilon}^{(t)}$ using Equation 8, where each perturbation is computed with its corresponding noise $\epsilon^{(t)}$. The perturbed attention map at timestep t is computed as:

$$\mathbf{A}_{\epsilon}^{(t)} = \text{Softmax} \left(\frac{\mathbf{Q}_{\epsilon}^{(t)} \mathbf{K}_{\epsilon}^{(t)\top}}{\sqrt{C}} \right). \quad (10)$$

The final degraded self-attention output is then given by $\mathbf{A}_{\epsilon}^{(t)} \mathbf{V}_{\epsilon}^{(t)}$. By perturbing the query, key, and value rather than the attention weights, our method injects stochasticity directly into the feature space. Consequently, the resulting attention map $\mathbf{A}_{\epsilon}^{(t)}$ enables more dynamic and diverse guidance signals by modulating both magnitude and directional variations over timesteps. Moreover, $\mathbf{A}_{\epsilon}^{(t)}$ provides a time-adaptive guidance mechanism aligned with the denoising objective of each timestep t , effectively coupling the guidance with the diffusion dynamics for more reliable sampling. Our method is summarized in Algorithm 1.

5 Experiments**5.1 Experimental Setup**

Implementation Details. We evaluate the proposed guidance mechanism using two widely adopted models: Stable Diffusion XL (SDXL) [26] and Stable

Table 1: Quantitative comparison of guidance methods on unconditional generation. The best results are shown in **bold**, and the second-best results are underlined.

Model	Metric	Vanilla	PAG [1]	SEG [13]	SAND
SDXL	FID↓	126.51	100.58	<u>83.46</u>	75.25
	IS↑	8.61	<u>12.94</u>	12.52	13.90
SD3	FID↓	101.66	102.74	<u>83.30</u>	75.64
	IS↑	11.09	12.10	11.69	13.37

Table 2: Quantitative comparison of guidance methods on conditional generation. The best results are shown in **bold**, and the second-best results are underlined.

Model	Metric	Vanilla	PAG [1]	SEG [13]	SAND
SDXL	FID↓	51.11	23.20	24.49	17.56
	IS↑	23.2	31.46	28.61	32.37
SD3	FID↓	22.88	18.13	<u>14.49</u>	11.06
	IS↑	31.09	35.83	38.77	<u>36.87</u>

Table 3: Quantitative comparison of text-image alignment and perceptual aesthetics on SDXL [26]. The best results are shown in **bold**, and the second-best results are underlined.

Dataset	Metric	Vanilla	PAG [1]	SEG [13]	SAND
DrawBench	CLIPScore↑	19.73	<u>22.41</u>	22.04	23.46
	PickScore↑	20.07	<u>20.95</u>	20.81	21.15
	HPSv2↑	0.1694	0.2022	<u>0.2037</u>	0.2253
	AES↑	5.35	5.55	<u>5.61</u>	5.71
	ImageReward↑	0.1023	<u>0.3069</u>	0.2196	0.6256

Diffusion 3 (SD3) [6]. All experiments use official implementations and publicly available weights from Hugging Face. For SDXL, we employ the Euler Discrete Scheduler [16] with 25 denoising steps. For SD3, we follow the default settings. The guidance scale γ was set to 3 specifically for the unconditional generation of the SDXL model, while $\gamma = 2$ was used for all other experimental settings. The same random seed sets are used across various comparison methods and baseline generative models to enable fair comparisons. For more details, please refer to Section B in the supplementary material.

Metrics. To evaluate both unconditional and conditional image generation, we employ multiple quantitative metrics. For visual quality, we calculate the Fréchet Inception Distance (FID) [8] and Inception Score (IS) [32] using 30k reference images and 30k text prompts randomly sampled from the MS-COCO 2014 validation set [24]. To assess text-image alignment and perceptual aesthetics, we measure CLIP Score [7], PickScore [20], Human Preference Score (HPS v2.1) [39], LAION Aesthetic Score (AES) [33], and ImageReward [40] using text prompts from the DrawBench benchmark [31].

5.2 Experiment Results

Unconditional Generation. Table 1 compares our method with PAG [1] and SEG [13] on the two models (SDXL [26] and SD3 [6]). Quantitative results in Table 1 show that FID is substantially reduced relative to unguided sampling and prior training-free approaches such as PAG and SEG. On SDXL, compared to unguided sampling, FID decreases from 126.51 to 75.25, while IS increases from 8.61 to 13.90, the highest among the compared methods. SD3 also exhibits



Fig. 1: Qualitative comparisons of our method with vanilla SDXL [26], PAG [1], and SEG [13] on unconditional generation.

similar improvements, achieving the best FID and IS across all methods. These results demonstrate that our approach provides superior fidelity and semantic quality compared to baseline training-free guidance methods.

Figure 1 illustrates generated samples from different guidance methods. Our guidance method consistently produces images with sharper details, clearer semantic structures, and richer textures compared to prior methods. While the unguided SDXL and PAG results often exhibit desaturated colors, structural inconsistencies, and SEG tends to generate unrealistic scenes, our method preserves both local fidelity and global coherence across diverse categories. Notably, the proposed approach recovers fine-grained details such as facial features, object contours, and lighting variations, yielding visually realistic and semantically consistent results.

Conditional (Text-to-Image) Generation. Table 2 presents the quantitative results under the conditional setting. Our method achieves the best performance

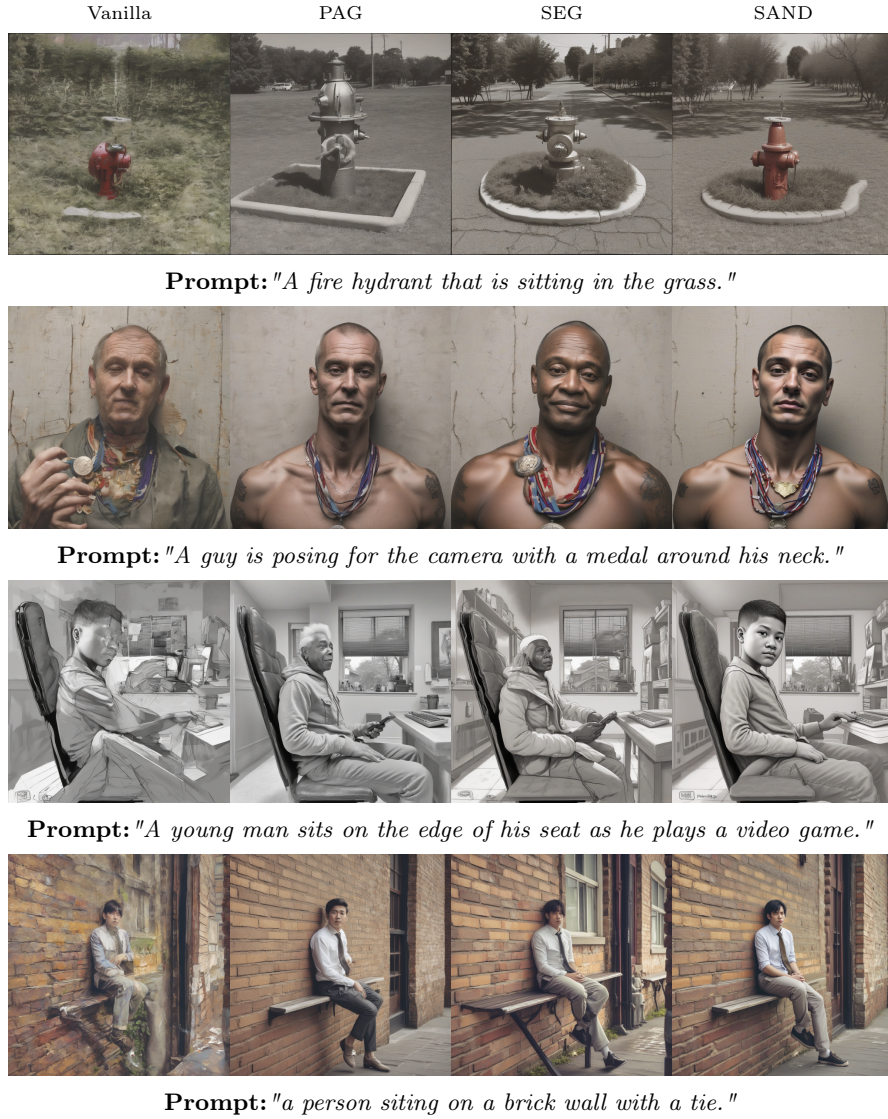


Fig. 2: Qualitative comparisons of our method with unguided SDXL [26], PAG [1], and SEG [13] on conditional generation.

across both SDXL and SD3. On SDXL, FID improves to 17.56, outperforming both PAG and SEG, while IS increases to 32.37. For SD3, our method achieves the lowest FID, outperforming the second-best score (FID=14.49) by a significant margin of +3. Additionally, Table 3 reports a quantitative evaluation of text-image alignment and human preference on the DrawBench dataset. Notably, our approach consistently outperforms all baselines. These results demonstrate

Table 4: Ablation study on noise decomposition. The best results are shown in **bold**, and the second-best results are underlined.

Metric	$\epsilon_{\parallel}$	$\epsilon_{\perp}$	ϵ	Ours
FID↓	84.31	76.31	<u>76.26</u>	75.25
IS↑	12.50	<u>13.94</u>	13.97	13.90

Table 5: Ablation study on different noise schedules. The best results are shown in **bold**, and the second-best results are underlined.

	Coupled increase	Coupled decrease	Swap	Ours
Metric				
FID↓	83.18	83.58	83.69	75.25
IS↑	12.48	13.10	13.13	13.90

that our method offers superior prompt fidelity and image quality compared to baseline guidance techniques.

Figure 2 presents a qualitative comparison of text-to-image generation across four prompts using different guidance methods. Our method yields high-quality images with sharp details, strong adherence to the textual descriptions, and stable structural composition. In contrast, unguided SDXL, PAG, and SEG frequently produce outputs with weakened semantic alignment, exhibiting issues such as desaturated colors, distorted geometry, inconsistent textures, and prompt mismatches. These artifacts highlight the limitations of existing baselines, whereas our method demonstrates improved robustness and more faithful rendering aligned with the given prompts. These results highlight the robustness and generalization ability of our degradation-based guidance mechanism across diffusion architectures of different scales.

5.3 Ablation study

To validate the effectiveness of our proposed perturbation mechanism, we conduct extensive ablation studies. All experiments use SDXL [26] with 25 sampling steps in the unconditional setting using 30K generated samples.

Effectiveness of Noise Decomposition. To demonstrate the necessity of decomposing noise, we perform an ablation study on the contributions of the parallel component $\epsilon_{\parallel}$ and the orthogonal component $\epsilon_{\perp}$. We compare our full method against three variants: (1) parallel-only noise $\epsilon_{\parallel}$, (2) orthogonal-only noise $\epsilon_{\perp}$, and (3) non-decomposed noise ϵ . In our method, the noise scales α are set to 12, 12, and 2 for the query, key, and value projections, respectively. For fairness, the fixed scales used in the three baseline variants are matched to the time-averaged magnitudes of our scheduling functions, resulting in strengths of 6, 6, and 1.

Figure 3 reveals a notable phenomenon during the sampling process. During the early denoising stages, all four methods produce a comparable global semantic layout. However, a clear divergence emerges in the transitional stage (third column of Figure 3). At this point, the parallel-only variant begins to deviate from the others, progressing toward a lower-quality trajectory characterized by desaturated appearance and diminished detail. In contrast, the remaining three methods continue to converge toward a high-fidelity solution.

This divergence offers strong evidence supporting our design. The parallel component $\epsilon_{\parallel}$ is sufficient for guiding early-stage magnitude modulation that es-



Fig. 3: Qualitative comparison between the proposed method and various noise component combinations.

Table 6: Component-wise scheduling analysis of the parallel $\epsilon_{\parallel}$ and orthogonal $\epsilon_{\perp}$ perturbations on SDXL [26] for unconditional generation. The best results are shown in **bold**, and the second-best results are underlined.

Metric	$\epsilon_{\parallel} \uparrow$	$\epsilon_{\parallel} \downarrow$	$\epsilon_{\perp} \uparrow$	$\epsilon_{\perp} \downarrow$	Ours
FID↓	94.98	84.31	<u>79.18</u>	83.47	75.25
IS↑	11.10	<u>13.60</u>	12.42	13.40	13.90

establishes the coarse global structure, but the orthogonal component $\epsilon_{\perp}$ becomes essential for late-stage refinement of geometry and high-frequency details. The parallel-only method fails precisely because it lacks this $\epsilon_{\perp}$ signal, which serves as a critical geometric correction pathway. Meanwhile, the non-decomposed variant retains both forms of perturbation, which allows it to outperform the parallel-only method and achieve results similar to the orthogonal-only method. However, compared to our decomposed approach, it still falls short in terms of fine details and overall realism.

The quantitative comparison in Table 4 reinforces this hierarchy. Our method achieves the best FID, followed by the non-decomposed variant ϵ and the orthogonal only variant $\epsilon_{\perp}$, while the parallel-only variant $\epsilon_{\parallel}$ performs the worst. These results demonstrate that the two noise components contribute fundamentally different—and non-interchangeable—roles in guiding the denoising trajectory, providing quantitative support for our decomposition strategy.

Effectiveness of Stage-Wise Weight Schedule. Table 5 compares the quantitative performance of our stage-wise weight schedule with three alternative schedules. The coupled schedules apply the same linear profile (either increasing or decreasing) to both components simultaneously, ignoring their distinct functional roles in the denoising process. The swapped schedule inverts our proposed profiles, increasing the parallel component while decreasing the orthogonal component over time. All three alternatives fail to align with the stage-dependent requirements of diffusion sampling, and consequently yield noticeably worse performance than our method.

Table 7: Quantitative comparison of our guidance applied to the down, mid, and up layers for unconditional generation on SDXL [26], evaluated using 30k generated samples. The best results are shown in **bold**.

Setting	Metric	Vanilla	Down	Mid	Up
Uncond.	FID↓	126.51	119.46	75.25	122.96
	IS↑	8.61	6.69	13.90	7.70

Analysis on Component-wise Scheduling Dynamics. To more directly examine the temporal roles of the parallel and orthogonal noise components, we evaluate four variants that apply either a linear increase ($\uparrow$) or linear decrease ($\downarrow$) schedule to the magnitude-oriented parallel component $\epsilon_{\parallel}$ or the direction-oriented orthogonal component $\epsilon_{\perp}$. The results are summarized in Table 6.

As reported in Table 6, the two variants that rely solely on $\epsilon_{\parallel}$ yield the weakest performance (FID= 94.98 for $\epsilon_{\parallel} \uparrow$ and FID= 84.31 for $\epsilon_{\parallel} \downarrow$). This result indicates that magnitude-only perturbations are insufficient for the detail refinement required in the later stages of the diffusion process. In contrast, the orthogonal-only variants achieve better results, and $\epsilon_{\perp} \uparrow$ performs particularly well (FID= 79.18 and IS= 12.42). This outcome demonstrates that increasing directional influence toward the end of the sampling process is important for reconstructing high-frequency details.

The $\epsilon_{\perp} \downarrow$ variant (FID= 83.47) further reveals that applying directional perturbation too early, when the features are still close to random noise, provides little meaningful guidance. This aligns with the intuition that early stages are dominated by global magnitude modulation rather than geometric refinement. Taken together, these results validate the underlying principle of our method: early diffusion steps benefit more from parallel noise, which modulates feature magnitude and shapes coarse structure, whereas later steps rely increasingly on orthogonal noise to adjust feature direction and recover fine details. Our full stage-aware formulation, which couples $\epsilon_{\parallel}$ with early timesteps and $\epsilon_{\perp}$ with later ones, achieves the best overall performance (FID= 75.25 and IS= 13.90), confirming the effectiveness of the proposed noise decomposition and temporally aligned scheduling strategy.

Layer Selection. We conducted an independent analysis of perturbing the downsampling, mid, and upsampling layers in the U-Net backbone of SDXL [26]. Table 7 shows that our guidance achieves the best performance when perturbations are applied exclusively to the mid layers throughout the entire diffusion sampling process. Perturbing the other two layer groups provides only limited improvements in FID and even degrades performance in IS compared with the unguided baseline. These results highlight the importance of selecting the appropriate layers for perturbation. Layer selection experiments for SD3 are provided in Section B.2 of the supplementary material.

Table 8: Quantitative comparison of text-image alignment and perceptual aesthetics on FLUX.1-dev [22] using text prompts from DrawBench [31]. The best results are **bold**, and the second-best results are underlined.

Metric	Vanilla	PLADIS	SAND	CFG	CFG+SAND
CLIPScore $\uparrow$	26.61	26.93	22.65	27.37	<u>27.29</u>
PickScore $\uparrow$	22.76	22.78	22.83	<u>22.82</u>	22.83
HPSv2 $\uparrow$	0.2869	0.2876	0.2890	<u>0.3006</u>	0.3008
AES $\uparrow$	5.72	5.70	5.68	5.83	<u>5.81</u>
ImageReward $\uparrow$	0.8357	0.8741	0.8884	<u>0.9935</u>	1.004



Prompt: "A sign that says 'Google Brain Toronto'."

Fig. 4: Qualitative comparisons of our method with unguided FLUX.1-dev [22], PLADIS [19], CFG, and CFG+SAND on text-to-image generation.

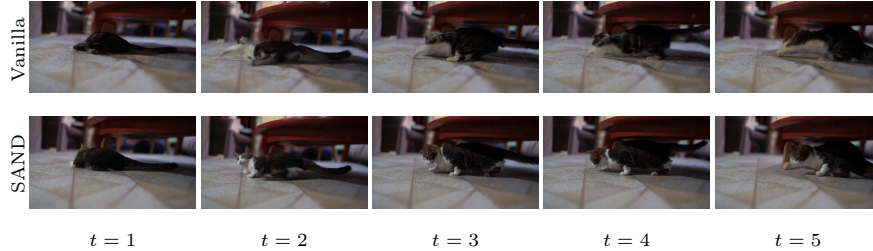
5.4 Generalization to Other Models

In this section, we evaluate the generalizability of SAND on the image generation model FLUX and the video generation model Wan2.2. Detailed implementations and experiment settings are provided in Section B of the supplementary material. **FLUX with CFG Distillation.** We further extend SAND to a CFG-distilled FLUX model using a one-forward inference scheme. Instead of performing two separate forward passes, we compute the original and degraded attention outputs within the selected MMDiT blocks and add their weighted difference to the original attention output. As shown in Table 8, SAND improves several metrics when applied to both vanilla and CFG-distilled FLUX, with a particularly large gain in ImageReward. It also outperforms PLADIS [19] on several metrics, such as PickScore, HPSv2, and ImageReward. This is consistent with the qualitative results in Figure 4, where SAND produces images with better prompt alignment and visual coherence.

Text-to-Video Model. To further verify the extensibility of SAND beyond image generation, we apply it to the Wan2.2 text-to-video model. We compare the vanilla Wan2.2 model with its SAND-guided variant using six VBench quality metrics [15]: subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. As shown in Table 9, SAND consistently improves all six metrics over the vanilla model, demonstrating that the proposed guidance is also effective for video generation. The qualitative comparisons in Fig. 5 further show that SAND produces videos with

Table 9: Quantitative comparisons on Wan2.2 [37]. The best results are shown in bold.

Method	Subj.Cons. $\uparrow$	BG Cons. $\uparrow$	Mot. Smooth. $\uparrow$	Dyn. Deg. $\uparrow$	Aes. Qual. $\uparrow$	Img. Qual. $\uparrow$
Vanilla	0.9216	0.9453	0.9912	0.64	0.4530	0.5975
SAND	0.9236	0.9478	0.9913	0.67	0.4713	0.6134

**Fig. 5:** Qualitative comparison between Vanilla Wan2.2 and SAND. Each column shows frames extracted at different timestamps. The text prompt is "A cat is hiding under the table, then it suddenly jumps out and starts scratching the furniture."

more coherent subjects, stable backgrounds, and improved visual quality. These results confirm the general applicability of SAND to text-to-video models.

6 Conclusion

We presented a principled training-free guidance framework that revisits diffusion sampling through a stage-aware lens. By recognizing that early and late timesteps fulfill fundamentally different semantic functions, we decomposed Gaussian noise into magnitude- and direction-oriented components and introduced a time-adaptive weighting schedule that aligns these perturbations with the diffusion dynamics. Injecting the decomposed noise directly into the self-attention projections yields a dynamically weakened branch that requires no retraining yet faithfully tracks the geometric evolution of the sample. Extensive experiments on both unconditional and conditional settings demonstrate consistent improvements in global coherence, fine-grained detail, and overall sample quality over existing training-free guidance techniques. Beyond its practical effectiveness, our work highlights the value of temporally structured perturbations in diffusion models and opens up new opportunities for stage-aware guidance, representation-level regularization, and geometry-informed generative modeling.

Acknowledgements

This work was partly supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (No. RS-2024-00397293, No. RS-2024-00352566, No. RS-2025-00559165, and No. RS-2025-25429004)

References

1. Ahn, D., Cho, H., Min, J., Jang, W., Kim, J., Kim, S., Park, H.H., Jin, K.H., Kim, S.: Self-rectifying diffusion sampling with perturbed-attention guidance. In: ECCV. pp. 1–17. Springer (2024) [1](#), [2](#), [3](#), [5](#), [8](#), [9](#), [10](#)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendeleevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) [1](#)
3. Cho, H., Ahn, D., Hong, S., Kim, J.E., Kim, S., Jin, K.H.: TAG: Tangential amplifying guidance for hallucination-resistant sampling. In: ICML (2026), <https://openreview.net/forum?id=1XJqPy2LhA> [4](#)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. NeurIPS **34**, 8780–8794 (2021) [4](#)
5. Ding, L., Dong, S., Huang, Z., Wang, Z., Zhang, Y., Gong, K., Xu, D., Xue, T.: Text-to-3d generation with bidirectional diffusion using both 2d and 3d priors. In: CVPR. pp. 5115–5124 (2024) [1](#)
6. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024) [1](#), [2](#), [8](#)
7. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: EMNLP. pp. 7514–7528 (2021) [8](#)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017) [8](#)
9. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) [1](#)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020) [1](#)
11. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021) [1](#), [3](#), [4](#), [5](#)
12. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. NeurIPS **35**, 8633–8646 (2022) [1](#)
13. Hong, S.: Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. NeurIPS **37**, 66743–66772 (2024) [1](#), [2](#), [3](#), [5](#), [8](#), [9](#), [10](#)
14. Hong, S., Lee, G., Jang, W., Kim, S.: Improving sample quality of diffusion models using self-attention guidance. In: ICCV. pp. 7462–7471 (2023) [2](#), [3](#)
15. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024) [14](#)
16. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. NeurIPS **35**, 26565–26577 (2022) [1](#), [8](#)
17. Karras, T., Aittala, M., Kynkäänniemi, T., Lehtinen, J., Aila, T., Laine, S.: Guiding a diffusion model with a bad version of itself. NeurIPS **37**, 52996–53021 (2024) [1](#), [2](#), [3](#), [5](#)
18. Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., Laine, S.: Analyzing and improving the training dynamics of diffusion models. In: CVPR. pp. 24174–24184 (2024) [1](#)

19. Kim, K., Sim, B.: Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. In: ICCV. pp. 16238–16248 (2025) 14
20. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-pic: An open dataset of user preferences for text-to-image generation. *NeurIPS* **36**, 36652–36663 (2023) 8
21. Kwon, M., Jeong, J., Hsiao, Y.T., Uh, Y., et al.: Tcfg: Tangential damping classifier-free guidance. In: CVPR. pp. 2620–2629 (2025) 4
22. Labs, B.F.: Flux. <https://github.com/black-forest-labs/flux> (2024) 3, 14
23. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR. pp. 300–309 (2023) 1
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014) 8
25. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4195–4205 (2023) 1
26. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: ICLR. vol. 2024, pp. 1862–1874 (2024) 1, 2, 7, 8, 9, 10, 11, 12, 13
27. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024) 1
28. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: ICLR (2023) 1
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 1, 2
30. Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: ICLR (2024) 4
31. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* **35**, 36479–36494 (2022) 8, 14
32. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. *NeurIPS* **29** (2016) 8
33. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS* **35**, 25278–25294 (2022) 8
34. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. *NeurIPS* **32** (2019) 1, 4
35. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021) 1, 4
36. Vincent, P.: A connection between score matching and denoising autoencoders. *Neural computation* **23**(7), 1661–1674 (2011) 4
37. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W.,

- Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) [3](#), [15](#)
38. Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., Zhu, J.: Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *NeurIPS* **36**, 8406–8441 (2023) [1](#)
39. Wu, X., Hao, Y., Sun, K., Chen, Y., Zhu, F., Zhao, R., Li, H.: Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. arXiv preprint arXiv:2306.09341 (2023) [8](#)
40. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS* **36**, 15903–15935 (2023) [8](#)