Anchoring and Steering Diffusion: Enhancing the Faithfulness of Text-to-Image Generation at Inference Time

Xinyi Wang, Yuyang Huang, Yalin Su, Pengcheng Luan,
Tao Zhang(✉), Feiming Wei, and Wenxian Yu

Shanghai Jiao Tong University, Shanghai, China
{xinyi_wang, huangyuyang, syl_move, pengcheng_luan,
sjtu--zt, weifeiming, wxyu}@sjtu.edu.cn

Fig. 1: Visual comparison of text-to-image results. Our **AnchorSteer** demonstrates better text–image alignment across various aspects.

Abstract. While text-to-image diffusion models achieve impressive visual quality, they frequently struggle to maintain precise alignment with complex compositional prompts. An effective strategy is to improve the inference process of diffusion models, thereby better leveraging their pretrained priors to address misalignment. Existing training-free methods can be divided into two categories. The first category focuses on improving the randomly sampled initial noise to obtain an initialization that encapsulates semantics relevant to the target text. However, existing approaches either perform costly search over noise pools, with no guarantee of finding a truly prompt-compatible noise within the limited candidates, or manipulate sampled noise without simultaneously ensuring reliable semantic injection and preservation of the Gaussian distribution. The second category focuses on improving the denoising trajectory. However, these methods lack explicit mechanisms to timely diagnose and correct

Corresponding author: Tao Zhang.

semantic errors, and therefore fail to prevent error propagation during generation. To address these limitations, we propose **AnchorSteer**, a training-free framework that exerts fine-grained control over **both initialization** and **the denoising trajectory**. AnchorSteer consists of two synergistic components: **Semantic Anchoring** replaces uninformative Gaussian noise with text-aligned initializations via CLIP-based prior extraction and a novel Latent-Prior Score Distillation Sampling (LP-SDS) objective. Specifically, LP-SDS distills CLIP visual priors into the knowledge distribution of diffusion models, mitigating the domain gap between CLIP-based priors and diffusion-based priors. **Reflective Steering** transforms passive denoising with an active **Think–Erase–Retouch loop** that enables mid-generation self-correction. *Think* phase employs VLM-based dual diagnosis to detect semantic deviations. *Erase* phase performs a targeted latent rollback with negative-guided inversion to suppress erroneous content, and *Retouch* phase subsequently applies positive-guided refinement to recover the missing attributes. Therefore, **Reflective Steering** enables timely correction of semantic deviations along the denoising trajectory, preventing error propagation. Extensive experiments on GenEval and T2I-CompBench++ demonstrate that AnchorSteer consistently outperforms existing baselines in text–image alignment while preserving high visual quality.

Keywords: Text–Image Alignment · Latent Diffusion Models · Latent-Prior Score Distillation Sampling · Reflective Denoising

1 Introduction

Text-to-image (T2I) diffusion models [19,48,51,56] have achieved remarkable success in generating high-quality images. However, they still frequently encounter challenges in maintaining precise semantic alignment [8, 12, 14, 21, 47, 62], particularly **under complex compositional prompts** (e.g., involving multiple objects, attributes, and spatial constraints). As is shown in Figure 1, under such scenarios, key semantic details are often lost or distorted, leading to missing objects, attribute leakage, or spatial inconsistencies [6, 7, 27, 38, 46, 70].

To address these issues, existing works can be broadly categorized into two groups: *training-based* and *training-free* methods. Training-based methods improve text–image alignment by fine-tuning diffusion models on customized image–text datasets involving complex semantics [4, 32, 49]. However, the diversity of possible compositional scenarios is effectively unbounded, making it difficult for the training distribution to cover all cases. As a result, these approaches often exhibit limited generalization to unseen combinations, which restricts their overall effectiveness.

Recent studies observe that pretrained diffusion models already demonstrate strong capabilities in handling individual sub-prompts [8, 18, 35, 37], suggesting that improving the inference process can serve as an effective complement to training-based methods for enhancing alignment with complex textual prompts.

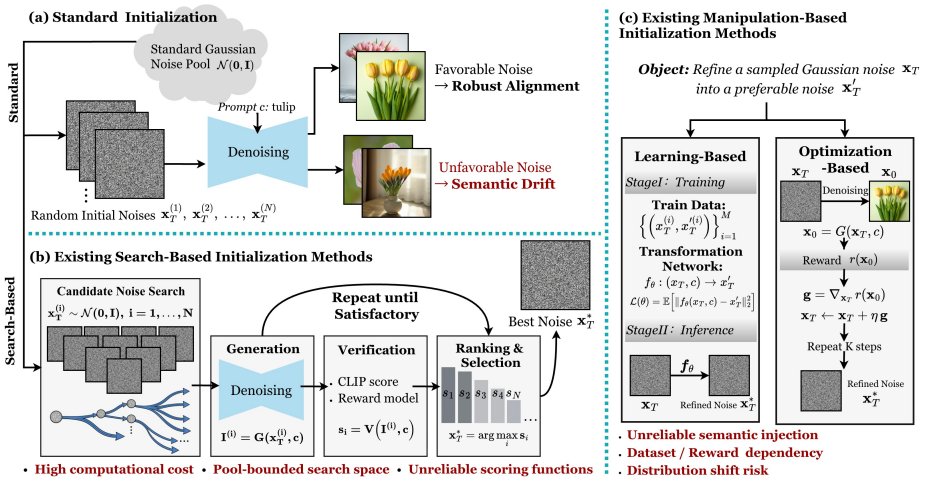

Fig. 2: Existing methods for noise initialization and their limitations. (a) Standard Gaussian initialization treats all noises as semantically equivalent, potentially assigning an unfavorable noise to the prompt and causing semantic drift during denoising. (b) Search-based methods rely on candidate selection from a predefined noise pool, leading to high computational cost and unreliable alignment scoring. (c) Manipulation-based approaches refine sampled noise via learning or optimization, yet suffer from unstable semantic injection and potential distribution shift.

This insight motivates training-free methods that improve text–image alignment by enhancing the inference process and better utilizing pretrained models.

To begin with, misalignment during inference mainly stems from two components. First, at ***noise initialization***, randomly sampling from standard Gaussian sampling often produces starting points unfavorable for complex compositional prompts(Figure 2a). Second, along the ***denoising trajectory***, the inference process lacks mechanisms to detect or correct semantic deviations, hindering the satisfaction of multiple prompt constraints. Therefore, existing training-free methods mainly focus on improving these two components.

For *noise initialization*, recent studies [1,17,30,38,39,52,57,66,69] has focused on constructing initial noise that encapsulates semantics aligned with the given text prompt, thereby improving the generated results. Existing methods can be broadly grouped into two categories.

The first category consists of **search-based methods**(Figure 2b) [30, 38, 39, 52], which aim to identify a prompt-compatible initial noise by selecting the best candidate from a predefined noise pool. However, search-based methods are inefficient and fundamentally limited by the noise pool and imperfect alignment scoring employed for searching. Another category is **manipulation-based methods** (Figure 2c), which refine a randomly sampled initial noise into a more prompt-compatible one [1, 17, 57, 66, 69]. However, manipulation-based methods

cannot simultaneously guarantee reliable semantic injection and preservation of the Gaussian noise distribution after manipulation.

For *denoising trajectory*, existing methods can be broadly categorized into two groups: unidirectional and bidirectional. **Unidirectional methods** attempt to optimize the trajectory within a single forward pass through various strategies, such as increasing denoising steps [15] or scaling up classifier-free guidance [9, 50, 54]. However, they still cannot recover from intermediate semantic deviations. **Bidirectional methods** [3, 11, 28, 55] introduce re-noising and re-denoising transitions, allowing them to revisit earlier states. However, they still lack explicit mechanisms to diagnose and correct semantic errors. Therefore, semantic errors may propagate and accumulate along the denoising trajectory.

To tackle the aforementioned limitations in *noise initialization* and *denoising trajectory*, we propose **AnchorSteer**, a training-free framework that exerts fine-grained control over both noise initialization and denoising trajectory. To replace unreliable noise initialization, we introduce **Semantic Anchoring**, which constructs initial noise that is both text-aligned and distributionally consistent. To overcome the lack of reliable error rectification during the denoising trajectory, we further develop **Reflective Steering**, a strategy that enables faithful diagnosis and targeted correction of semantic deviations throughout the diffusion process. Extensive experiments show that AnchorSteer achieves superior alignment across different diffusion models without training.

The main contributions in our paper are summarized as follows:

- We propose **AnchorSteer**, a training-free framework that explicitly enhances initial noise with text-aligned semantics and performs timely rectification of the denoising trajectory to prevent the propagation of semantic deviations.
- We propose **Semantic Anchoring**, a principled initialization strategy that first extracts a deterministic semantic prior via CLIP, then bridges the domain gap between CLIP outputs and the diffusion latent manifold through **Latent-Prior Score Distillation Sampling (LP-SDS)**, and finally applies DDPM forward diffusion to yield initial noise that is both text-aligned and distributionally consistent.
- We introduce **Reflective Steering**, a bidirectional self-corrective mechanism that diagnoses and corrects semantic deviations via a *Think–Erase–Retouch* loop, enabling explicit suppression of inconsistent content and reinforcement of missing semantics.

2 Related Work

Text-to-Image Diffusion Models. Image diffusion models have achieved remarkable success in text-to-image generation [19, 56] and a wide range of other visual tasks [24, 29, 59]. Latent Diffusion Models [48] such as SDXL [42] improve efficiency via latent-space generation, while recent Transformer-based architectures [10, 33, 41] further enhance scalability and visual quality. Despite these

advances, models still struggle with complex compositional prompts, often producing missing objects, attribute leakage, or spatial inconsistencies [27,38,46,67].

Noise Initialization Strategies. Recent studies [16,44] show that different initial noises lead to different generation outcomes. *Search-based methods* [38,39,52] select favorable samples from noise pools but are computationally expensive and limited by pool diversity. *Manipulation-based methods* modify noise via learned mappings [1, 69] or gradient optimization [57, 66], yet often lack reliable semantic injection or introduce distribution shift. Our LP-SDS addresses both issues by ensuring semantic alignment while remaining consistent with the diffusion model's training distribution.

Inference-Time Trajectory Optimization. Prior work improves alignment by refining the denoising trajectory at inference time. *Unidirectional methods* adjust guidance strength or introduce gradient-based updates [9, 50, 65], but cannot recover once the trajectory deviates from the desired semantics. *Bidirectional methods* [3, 23, 25] introduce re-noising steps to revisit earlier states, yet lack explicit mechanisms for diagnosing and correcting semantic errors. In contrast, our Reflective Steering introduces a VLM-guided Think–Erase–Retouch loop that actively detects and rectifies semantic inconsistencies.

3 Preliminaries

3.1 Latent Diffusion Models

Latent Diffusion Models (LDMs) [48] operate in the latent space of a pretrained autoencoder $(\mathcal{E}, \mathcal{D})$. A clean latent $z_0 = \mathcal{E}(x)$ is corrupted by the forward process $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$, where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. A text-conditional denoising network ϵ_ϕ is trained to predict the added noise by minimizing [56]

$$\mathcal{L}_{\text{Diff}} = \mathbb{E}_{t,\epsilon}\big[w(t)\, \big\|\epsilon_\phi(z_t; t, c) - \epsilon\big\|_2^2\big], \tag{1}$$

with text condition c and time-dependent weight $w(t)$. Starting from $z_T \sim \mathcal{N}(0, \mathbf{I})$, the reverse process iteratively denoises with ϵ_ϕ. Classifier-Free Guidance (CFG) [20] enhances text alignment by extrapolating between conditional and unconditional estimates with guidance scale s:

$$\hat{\epsilon}_\phi(z_t; t, c) = \epsilon_\phi(z_t; t, \varnothing) + s\big(\epsilon_\phi(z_t; t, c) - \epsilon_\phi(z_t; t, \varnothing)\big), \quad s > 1. \tag{2}$$

3.2 Score Distillation Sampling

Score Distillation Sampling (SDS) [43] optimizes a differentiable representation $x = g(\theta)$ using a frozen diffusion model as a probabilistic critic. After encoding and perturbing $z_0 = \mathcal{E}(x)$ to z_t, the parameter gradient is

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\Big[w(t)\big(\hat{\epsilon}_\phi(z_t; t, c) - \epsilon\big)\, \tfrac{\partial x}{\partial \theta}\Big], \tag{3}$$

which drives x toward the natural-image manifold consistent with c by distilling the generative prior of the frozen model, where constant scaling factors are absorbed into $w(t)$.

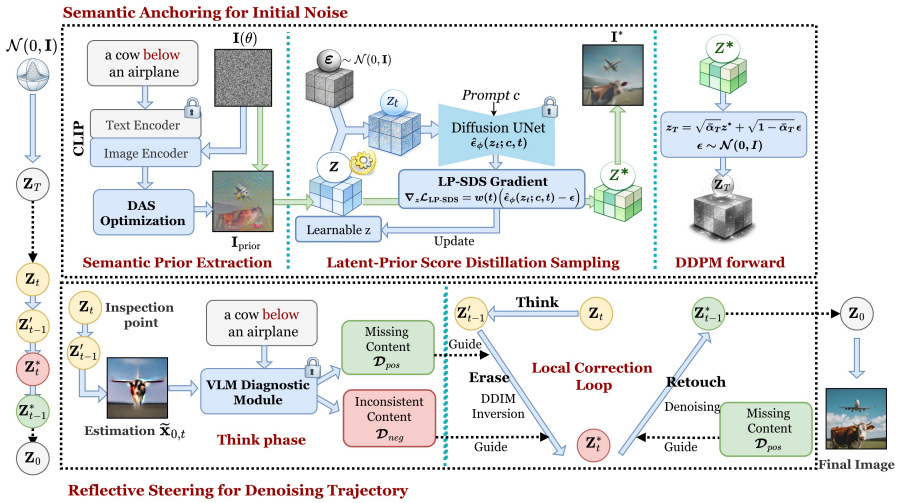

Fig. 3: Overview of AnchorSteer. (Top) Semantic Anchoring for Initial Noise (Section 4.1). Given a text prompt c, a semantic prior I_{prior} is first extracted from the CLIP model via Direct Ascent Synthesis (DAS), then encoded into the latent space and refined through Latent-Prior Score Distillation Sampling (LP-SDS) to obtain an optimized latent z^* that bridges the domain gap. The standard DDPM forward noising process is finally applied to produce an initial noise z_T that is both semantically aligned and distributionally consistent. **(Bottom) Reflective Steering for Denoising Trajectory (Section 4.2).** At selected timesteps, a Think–Erase–Retouch loop is inserted into the denoising trajectory $z_t \to z_{t-1}$. The *Think* module employs a VLM to diagnose the current estimate z'_{t-1}, producing a negative prompt $\mathcal{D}_{\text{neg}}$ (inconsistent content) and a positive prompt $\mathcal{D}_{\text{pos}}$ (missing content). The *Erase* module performs negative-guided DDIM inversion ($z'_{t-1} \to z^*_t$) to suppress erroneous semantics. The *Retouch* module applies positive guidance during re-noising and re-denoising ($z'_{t-1} \to z^*_t \to z^*_{t-1}$) to reinforce missing content, actively correcting semantic drift.

4 Methodology

Our framework **AnchorSteer anchors the initial noise** for **a good start** and reflectively **steers the denoising trajectory** for **precise guidance**. The overall framework is illustrated in Figure 3.

4.1 Semantic Anchoring for Initial Noise

To ensure effective semantic injection, we **extract text-aligned semantic prior** from **the CLIP** [45], thereby providing a deterministic and reliable semantic enhancement to the manipulated noise. To eliminate the domain gap of the noise distribution, we further propose **Latent-Prior Score Distillation Sampling (LP-SDS)** to refine this prior and apply the DDPM forward process to inject the refined semantic prior into the initial noise.

Fig. 4: CLIP-based semantic priors and their refinement via LP-SDS. The first row shows CLIP semantic priors I_{prior}, which are semantically aligned but OOD for diffusion models. The second row shows refined images I^* after LP-SDS, which become in-distribution while preserving the same semantics.

CLIP-based Semantic Prior Extraction. CLIP encodes rich semantic knowledge about objects, attributes, and relational constraints for nearly arbitrary prompts. We therefore seek to extract these semantics for complex prompts and leverage them to construct the initial noise. Inspired by Direct Ascent Synthesis (DAS) [13] to extract the semantic prior embedded within CLIP. Given a target text prompt c, we parameterize a learnable image $I(\theta)$ and directly optimize it to maximize its alignment with c in the CLIP embedding space using its image encoder $\mathcal{E}_I(\cdot)$ and text encoders $\mathcal{E}_T(\cdot)$ by $\mathcal{L}_{\text{DAS}}(\theta) = -[\mathcal{E}_I(I(\theta)) \cdot \mathcal{E}_T(c)]/[\|\mathcal{E}_I(I(\theta))\| \, \|\mathcal{E}_T(c)\|]$. After N_{das} steps, the resulting pseudo-image $I_{\text{prior}} = I(\theta_{N_{\text{das}}})$ captures spatial structure and color distribution that are deterministically anchored to the semantics of c.

Bridging Domain Gap via LP-SDS. As illustrated in Fig. 4, the CLIP-based pseudo-images I_{prior} exhibits severe visual degradation, including high-frequency artifacts and unnatural color distributions, despite preserving text-aligned semantic structures. Such images deviate significantly from the natural image distribution learned by pre-trained diffusion models. To mitigate the domain gap, we propose **Latent-Prior Score Distillation Sampling (LP-SDS)**, illustrated in Fig. 5, which extends Score Distillation Sampling (SDS) [43] to the noise optimization setting.

We initialize the optimization variable as $z_0 = \mathcal{E}_{\text{VAE}}(I_{\text{prior}})$. We add noise to z_0 and input it to the diffusion model to reconstruct z_0. Then we compute the gradient of the diffusion loss $\mathcal{L}_{\text{Diff}}$ with respect to z. Following SDS [43], we omit the U-Net Jacobian term, then the gradient simplifies to:

$$\nabla_z \mathcal{L}_{\text{LP-SDS}}(\phi, z) \triangleq \mathbb{E}_{t,\epsilon} \left[w(t) \, (\hat{\epsilon}_\phi(z_t; \, c, t) - \epsilon) \right]. \tag{4}$$

From a theoretical perspective, it can be proved that the gradient implicitly minimizes a Kullback–Leibler (KL) divergence weighted across time steps [43]:

$$\nabla_z \mathcal{L}_{\text{LP-SDS}}(\phi, z) = \nabla_z \, \mathbb{E}_t \left[\frac{\sigma_t}{\alpha_t} \, w(t) \, \text{KL}(q(z_t \mid z) \, \| \, p_\phi(z_t; \, c, t)) \right], \tag{5}$$

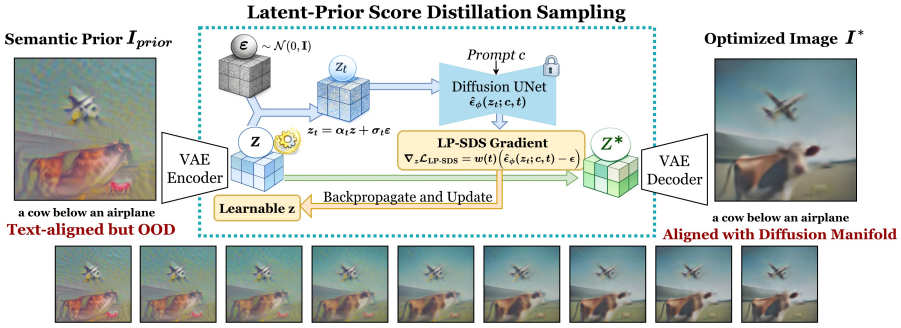

Fig. 5: Illustration of LP-SDS. The semantic prior I_{prior} produced by DAS is encoded into the latent space as $z_0 = \mathcal{E}_{\text{VAE}}(I_{\text{prior}})$ and iteratively refined via LP-SDS gradients from a frozen diffusion U-Net. The optimized latent z^* is projected onto the diffusion manifold, bridging the domain gap between CLIP-based synthesis and diffusion-based generation while preserving text-aligned semantics.

where $q(z_t \mid z) = \mathcal{N}(\alpha_t z, \sigma_t^2 I)$ is the Gaussian distribution induced by the forward diffusion process from the current latent variable z, and $p_\phi(z_t; c, t)$ is the marginal distribution conditioned on text c at time step t. Thus, this optimization can be viewed as a **probability density distillation process** , driving z towards high-density regions of the conditional distribution $p_\phi(z_t; c, t)$, thereby regularizing the optimized latent to remain consistent with the generative prior of the diffusion model.

After N_{lp} steps of optimization, we obtain a refined image I^*. As shown in Fig. 4, I^* exhibit significantly improved visual coherence compared with the degraded priors I_{prior}, while still preserving the underlying text-aligned semantic structures.

DDPM Forward Diffusion for Noise Initialization. We diffuse z^* to the terminal step T via the DDPM forward process to obtain the ultimate initial noise z_T by $z_T = \sqrt{\bar{\alpha}_T}\, z^* + \sqrt{1 - \bar{\alpha}_T}\, \epsilon$. Therefore, z_T still follows $\mathcal{N}(0, \mathbf{I})$.

4.2 Reflective Steering for Denoising Trajectory

For denoising trajectory, we also adopt a fine-grained semantic correction method via a **Think–Erase–Retouch** loop. We evenly select inspection timesteps $\mathcal{T}_{\text{insp}} \subset \{1, \ldots, T\}$ with stride τ. For each $t \in \mathcal{T}_{\text{insp}}$, we expand the standard denoising trajectory into $z_t \to z'_{t-1} \to z^*_t \to z^*_{t-1}$, where z'_{t-1} denotes the initial denoising estimate, and z^*_t, z^*_{t-1} denote the refined latents after correction loop.

Think: Semantic Diagnosis via Vision–Language Feedback. At each inspection timestep t, once the initial estimate z'_{t-1} is obtained via standard denoising, the *Think* phase performs an on-the-fly semantic diagnosis to assess its

alignment with the text prompt. This involves two pivotal steps: Intermediate State Visualization and Dual Semantic Gap Diagnosis.

1) Intermediate State Visualization. Following DDIM prediction, we first obtain the clean latent $z'_{0|t-1}$ of z'_{t-1} and decode it into pixel space via the VAE decoder to yield the visualized image $\tilde{x}'_{0|t-1} = \mathcal{D}(\tilde{z}'_{0|t-1})$, which is then used for subsequent VLM-based diagnosis.

2) Dual Semantic Gap Diagnosis. Based on the estimated $\tilde{x}'_{0|t-1}$, we leverage the robust cross-modal understanding capabilities of a pre-trained Vision-Language Model (VLM) to assess its alignment with the text prompt c. Departing from prior methods that rely on implicit scalar scoring, we structure semantic deviations into two distinct, actionable guidance signals:

- **Missing Content ($\mathcal{D}_{\text{pos}}$):** Identifies contents explicitly specified in c but insufficiently manifested in $\tilde{x}'_{0|t-1}$ (*e.g.*, specific objects, attributes, or spatial relationships). $\mathcal{D}_{\text{pos}}$ forms the target set for reinforcement in the subsequent *Retouch* phase.
- **Inconsistent Content ($\mathcal{D}_{\text{neg}}$):** Detects hallucinatory or extraneous content in $\tilde{x}'_{0|t-1}$ that contradicts or is irrelevant to c. $\mathcal{D}_{\text{neg}}$ defines the target set for suppression in the subsequent *Erase* phase.

Through this structured diagnosis, vague image-text misalignments are transformed into precise, navigational correction signals that enable targeted optimization of the denoising trajectory.

Erase and Retouch: Local Correction Loop for Targeted Semantic Refinement. Upon acquiring the diagnostic sets $\mathcal{D}_{\text{pos}}$ and $\mathcal{D}_{\text{neg}}$ from the Think stage, we perform a local bidirectional correction at the inspection timestep t: $z'_{t-1} \xrightarrow{\text{Erase}} z^*_t \xrightarrow{\text{Retouch}} z^*_{t-1}$. This loop exploits the local reversibility of diffusion forward–reverse processes, applying negative guidance during noising to suppress errors and positive guidance during both noising and denoising to reinforce missing attributes.

1) Erase: Targeted Removal via Negative-Guided Inversion. This step executes the re-noising trajectory $z'_{t-1} \to z^*_t$. We replace the unconditional estimate $\epsilon_\phi(z_t; t, \varnothing)$ in Eq. (2) with $\epsilon_\phi(z_t; t, \mathcal{D}_{\text{neg}})$, steering the latent away from the identified erroneous semantics. The guided noise estimate $\hat{\boldsymbol{\epsilon}}_{\text{e}}$ is computed as:

$$\hat{\epsilon}_{\text{e}}(z'_{t-1}; t{-}1) \;=\; (1+\gamma_{\text{e}})\,\epsilon_\phi(z'_{t-1}; t{-}1,\, c) \;-\; \gamma_{\text{e}}\,\epsilon_\phi(z'_{t-1}; t{-}1,\, \mathcal{D}_{\text{neg}}) \tag{6}$$

We then compute the predicted clean latent $\hat{z}'_{0|t-1}$ via Tweedie's formula and apply the DDIM inversion rule to obtain the re-noised state z^*_t:

$$\hat{z}'_{0|t-1} \;=\; \frac{z'_{t-1} - \sqrt{1-\bar{\alpha}_{t-1}}\,\hat{\epsilon}_{\text{e}}(z'_{t-1}; t{-}1)}{\sqrt{\bar{\alpha}_{t-1}}} \tag{7}$$

$$z^*_t \;=\; \sqrt{\bar{\alpha}_t}\,\hat{z}'_{0|t-1} \;+\; \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_{\text{e}}(z'_{t-1}; t{-}1) \tag{8}$$

By substituting $\mathcal{D}_{\text{neg}}$ into the negative-prompt position, the guidance gradient actively steers the inversion trajectory away from the undesirable semantic manifold, effectively suppressing the corresponding latent feature activations and achieving targeted semantic erasure.

2) Retouch: Semantic Completion via Positive-Guided Refinement. The *Retouch* phase restores structural coherence while injecting the enhanced concepts $\mathcal{D}_{\text{pos}}$, spanning the re-noising and re-denoising trajectory $z'_{t-1} \to z^*_t \to z^*_{t-1}$. We construct an enhanced prompt $c^+ = c \cup \mathcal{D}_{\text{pos}}$ and adopt it as the conditional input throughout both sub-steps, enabling progressive semantic accumulation following [3]. For re-noising, we follow the same inversion procedure as *Erase* (Eqs. (7)–(8)), but replace the original prompt c with c^+ in the guided noise estimate:

$$\hat{\epsilon}_{\text{e}}(z'_{t-1}; t{-}1) \;=\; (1+\gamma_{\text{e}})\,\epsilon_\phi(z'_{t-1}; t{-}1,\, c^+) \;-\; \gamma_{\text{e}}\,\epsilon_\phi(z'_{t-1}; t{-}1,\, \mathcal{D}_{\text{neg}}) \tag{9}$$

The re-noised state z^*_t is then obtained via Eqs. (7) and (8). For re-denoising, starting from z^*_t, we apply standard CFG (Eq. (2)) with c^+ and a retouch guidance scale γ_{r}:

$$\hat{\epsilon}_{\text{r}}(z^*_t; t) \;=\; (1+\gamma_{\text{r}})\,\epsilon_\phi(z^*_t; t,\, c^+) \;-\; \gamma_{\text{r}}\,\epsilon_\phi(z^*_t; t,\, \varnothing) \tag{10}$$

where $\varnothing$ denotes the unconditional input. The refined state z^*_{t-1} is obtained via the DDIM denoising rule:

$$z^*_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{z^*_t - \sqrt{1-\bar{\alpha}_t}\,\hat{\epsilon}_{\text{r}}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\hat{\epsilon}_{\text{r}} \tag{11}$$

By applying c^+ in both sub-steps, the positive semantic signal is progressively accumulated into the latent representation, completing the missing content within the erased regions.

After iterating over all inspection timesteps, the final image is obtained by decoding z^*_0 as $x^* = \mathcal{D}_{\text{vae}}(z^*_0)$.

5 Experiments

5.1 Experimental Setting

Implementation Details. We adopt two representative diffusion models in our main experiments: the UNet-based SDXL [42] and the Transformer-based HunyuanDiT-v1.2 [34]. Unless otherwise specified, we use the DDIM sampler with T=50 denoising steps and CFG scale γ=5 for both models, following their default configurations. For *Semantic Anchoring*, we leverage OpenCLIP ViT-B/32 [26] to extract semantic priors via DAS, followed by LP-SDS refinement with N_{lp}=400. For *Reflective Steering*, we employ Qwen-VL-Chat [2] as the representative VLM, with inspection stride τ=1 and guidance scales γ_{e}=1.0, γ_{r}=5.0. All experiments are conducted on NVIDIA H100 GPUs.

Table 1: Cross-architecture evaluation on GenEval. AnchorSteer consistently improves faithfulness across different architectures.

Model	Method	Single↑	Two↑	Count↑	Color↑	Pos.↑	Attr.↑	Overall↑
SDXL	Standard	99.69	81.06	30.31	92.29	11.00	21.25	55.933
	+ **AnchorSteer**	**100.00**	**91.16**	**52.50**	**94.95**	**17.75**	**24.00**	**63.393**
	Δ	+0.31	+10.10	+22.19	+2.66	+6.75	+2.75	+7.460
HunyuanDiT	Standard	99.69	93.18	64.69	95.74	17.75	39.00	68.342
	+ **AnchorSteer**	**100.00**	**95.71**	**72.50**	**97.61**	**20.75**	**48.25**	**72.470**
	Δ	+0.31	+2.53	+7.81	+1.87	+3.00	+9.25	+4.128

Datasets and Metrics. We conduct experiments on two benchmarks: GenEval [14] and T2I-CompBench++ [21, 22], to evaluate text-to-image alignment. GenEval specializes in fine-grained object-centric metrics and T2I-CompBench++ targets more challenging compositional scenarios. Besides alignment, we also evaluate image quality and sample diversity, including HPSv2 [63], PickScore [31], ImageReward [64], and Aesthetic score [53] for image quality, and LPIPS [68] and DINO [5, 40] similarity for diversity. For all experiments, we follow the official benchmark protocols. Additional details of the benchmarks and evaluation metrics are provided in the supplementary material.

5.2 Main Results

Cross-Architecture Effectiveness. We evaluate AnchorSteer on two representative diffusion models with distinct architectures: SDXL [42] and HunyuanDiT [34]. As shown in Table 1, AnchorSteer consistently improves the standard inference baseline across all six GenEval categories on both models.

On SDXL, AnchorSteer increases the overall score from 55.93 to 63.39, with large improvements in *Counting* and *Two Objects*. On HunyuanDiT, which already achieves a stronger baseline of 68.34, AnchorSteer raises the overall score to 72.47, with notable gains in *Attribute Binding* and *Counting*. These results demonstrate that AnchorSteer is architecture-agnostic and consistently enhances compositional faithfulness across both U-Net and Transformer backbones.

Comparison with Existing Methods. Table 2 presents quantitative comparisons with several representative training-based and training-free alignment methods on GenEval benchmark. AnchorSteer achieves the best overall performance with a score of 63.39, outperforming all baselines by a clear margin. It shows notable gains on challenging categories such as *Counting*, *Spatial Relation*, and *Attribute Binding*, indicating stronger compositional accuracy.

Table 3 further evaluates AnchorSteer on T2I-CompBench++, where it achieves the best performance on the majority of semantic categories. These results demonstrate that AnchorSteer consistently improves compositional faithfulness and text-image alignment compared with existing methods.

Table 2: Quantitative comparison on GenEval (SDXL). AnchorSteer achieves the best overall score and outperforms baselines on most compositional categories.

Method	Single↑	Two↑	Count↑	Color↑	Pos.↑	Attr.↑	Overall↑
Standard	99.69	81.06	30.31	92.29	11.00	21.25	55.933
AaE [8]	99.06	83.59	31.56	90.96	15.00	24.00	57.361
InitNO [16]	99.69	85.10	37.81	89.89	16.00	22.75	58.540
DNO [57]	100.00	80.81	44.06	90.96	12.75	22.00	58.430
Zigzag [3]	99.69	85.35	40.94	92.55	12.25	20.50	58.547
Diffusion DPO [60]	100.00	**92.42**	46.25	92.29	15.00	18.50	60.744
SPO [36]	100.00	85.86	33.44	90.96	13.00	18.50	56.959
NPNet [69]	99.38	84.09	36.88	92.82	13.50	22.00	58.110
AnchorSteer (Ours)	**100.00**	91.16	**52.50**	**94.95**	**17.75**	**24.00**	**63.393**

Table 3: Quantitative comparison on T2I-CompBench++ (SDXL). AnchorSteer improves compositional alignment across diverse semantic categories.

Method	Color↑	Shape↑	Texture↑	2D-S↑	3D-S↑	Non-S↑	Num↑	Complex↑
Standard	0.5766	0.4830	0.5353	0.2015	0.3824	0.3139	0.5029	0.3382
AaE [8]	0.6084	0.4991	0.5511	0.1743	0.3673	0.3126	0.5017	0.3400
InitNO [16]	0.5636	0.4967	0.5430	0.1968	0.3699	0.3113	0.5037	0.3388
DNO [57]	0.5982	0.4986	0.5402	0.2050	0.4024	0.3144	0.5185	0.3446
Zigzag [3]	0.6172	0.5187	0.5759	0.2011	0.4024	**0.3206**	0.5349	0.3504
Diffusion DPO [60]	0.6351	0.5291	0.6110	0.2153	0.4017	0.3156	0.5223	0.3536
SPO [36]	0.6194	0.4955	0.5553	0.1930	0.3928	0.3084	0.5061	0.3556
NPNet [69]	0.5849	0.4824	0.5348	0.1905	0.3876	0.3132	0.5126	0.3397
AnchorSteer (Ours)	**0.6712**	**0.5345**	**0.6369**	**0.2158**	**0.4135**	0.3171	**0.5796**	**0.3759**

Qualitative Analysis. Figure 6 presents qualitative comparisons on a set of challenging compositional prompts constructed for evaluation. The baseline SDXL model and several representative alignment methods often exhibit semantic inconsistencies, such as missing objects, incorrect counts, attribute leakage, or violated spatial relations. In contrast, **AnchorSteer** produces results that more faithfully follow the prompt semantics, demonstrating improved compositional reasoning and text-image alignment.

Figure 7 further illustrates how AnchorSteer corrects semantic errors during generation. We visualize the predicted clean images x_0 from the first eight denoising steps. While SDXL consistently produces two giraffes, AnchorSteer gradually erases the incorrect giraffe instance and introduces the missing horse, leading to a semantically consistent result. Additional qualitative results are provided in the supplementary material.

Image Quality and Diversity Analysis. We further evaluate the visual quality and sample diversity of the standard baseline and AnchorSteer. As shown in Table 4, AnchorSteer improves all quality and diversity metrics over the base-

Fig. 6: Qualitative comparison on challenging compositional prompts. AnchorSteer produces more prompt-faithful results than SDXL and several representative alignment methods, which often exhibit semantic inconsistencies.

Table 4: Quality and diversity analysis on GenEval(SDXL). AnchorSteer further enhances visual quality and sample diversity over the standard baseline.

Method	Image Quality				Sample Diversity	
	HPSv2↑	PickScore↑	ImageReward↑	Aesthetic↑	1-LPIPS↓	DINOSim↓
Standard	0.2769	22.55	0.4891	5.465	0.3860	0.6638
AnchorSteer (Ours)	**0.2963**	**22.83**	**0.8781**	**5.561**	**0.3794**	**0.6314**

line, with a particularly notable gain on ImageReward [64] from 0.4891 to 0.8781. These results indicate that the improved text-image alignment is accompanied by better perceptual quality and sample diversity, rather than at their expense. Supplementary Human preference results further support this observation.

5.3 Ablation Studies

Component-wise Analysis. We progressively integrate Semantic Anchoring and Reflective Steering into the standard diffusion baseline to evaluate their individual and joint effects (Table 5). Semantic Anchoring consistently improves structural and attribute-related metrics, particularly *Counting* and *Attribute Binding*, suggesting that semantically aligned initialization reduces early-stage semantic drift. Reflective Steering mainly enhances relational reasoning during denoising, leading to clear improvements on the *Two Objects* category and further strengthening overall compositional consistency. When combined, AnchorSteer achieves the best overall performance on both SDXL and Hunyuan-DiT, demonstrating that jointly controlling the initialization and the denoising trajectory provides complementary benefits.

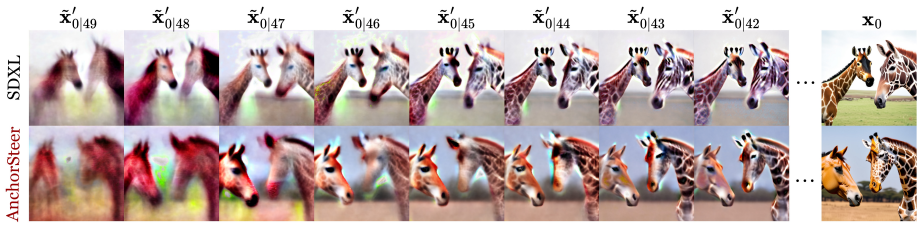

Fig. 7: Visualization of semantic correction during denoising. Predicted x_0 from the first eight steps for *A horse and a giraffe*. SDXL confuses the two concepts and generates two giraffes, while AnchorSteer progressively removes the incorrect giraffe and restores the missing horse.

Table 5: Component-wise ablation on GenEval. Each row progressively adds one module to the vanilla baseline.

Model	Method	Single↑	Two↑	Count↑	Color↑	Pos.↑	Attr.↑	Overall↑
SDXL	Standard	99.69	81.06	30.31	92.29	11.00	21.25	55.933
	+ Anchor	**100.00**	88.64	44.38	93.62	13.75	**24.00**	60.731
	+ Steer	**100.00**	**91.67**	42.50	92.29	12.50	20.25	59.868
	+ AnchorSteer	**100.00**	91.16	**52.50**	**94.95**	**17.75**	**24.00**	**63.393**
HunyuanDiT	Standard	99.69	93.18	64.69	95.74	17.75	39.00	68.342
	+ Anchor	**100.00**	93.43	66.88	96.81	18.50	46.75	70.395
	+ Steer	**100.00**	**96.72**	71.56	97.34	19.25	45.50	71.728
	+ AnchorSteer	**100.00**	95.71	**72.50**	**97.61**	**20.75**	**48.25**	**72.470**

Effect of VLM Scale. We further examine the effect of VLM scale in Reflective Steering by comparing Qwen2-VL-2B-Instruct [61], Qwen2.5-VL-3B-Instruct [58], and Qwen-VL-Chat [2]. As shown in Table 6, smaller VLMs still substantially improve over the standard SDXL baseline and achieve stronger performance than the compared baselines in Table 2, indicating that AnchorSteer does not rely on an oversized VLM for semantic diagnosis.

5.4 Efficiency Analysis

Table 7 reports the average per-image runtime of the methods evaluated in Table 2. AnchorSteer takes a runtime of 68.5 seconds per image, which is slower than standard sampling and lightweight inference-time methods, yet remains more efficient than DNO. The runtime ablation reveals that the primary computational overhead stems from the Reflective Steering module: under the 50-step SDXL setting adopted for fair comparison, the steering loop is executed at every denoising step. To improve efficiency and practicality, we further apply AnchorSteer to few-step distilled models, which provides a promising direction for improving efficiency without sacrificing alignment quality. More details are provided in the supplementary material.

Table 6: Effect of VLM scale on GenEval (SDXL). We replace the VLM used for semantic diagnosis in Reflective Steering with models of different scales.

VLM	Standard	Qwen2-VL-2B	Qwen2.5-VL-3B	Qwen-VL-Chat
Scale	–	2B	3B	9.6B
Overall	55.933	62.073	62.843	**63.393**

Table 7: Runtime analysis. Average per-image runtime and ablation of AnchorSteer. Diffusion DPO, SPO, and NPNet share the same inference runtime as Standard.

	Runtime Comparison						Runtime Ablation			
	Standard	Zigzag	AaE	InitNO	Ours	DNO	Anchor	Steer	Other	Total
Time (s)	3.7	8.3	14.9	34.7	68.5	155.5	18.0	46.7	3.8	68.5

6 Conclusion, Limitations and Future Work

Limitations. The main limitation of AnchorSteer is its additional inference cost. Both Semantic Anchoring and Reflective Steering introduce extra computation compared with standard sampling, making AnchorSteer more suitable for scenarios where alignment faithfulness is prioritized over generation speed. Improving its efficiency while maintaining strong alignment remains an important direction for future work.

Conclusion. In this work, we present AnchorSteer, a training-free framework for improving T2I alignment under complex compositional prompts. By jointly refining the initial noise and denoising trajectory, AnchorSteer consistently enhances compositional faithfulness across different diffusion backbones while preserving visual quality. Overall, AnchorSteer offers a practical inference-time solution for more faithful text-to-image generation.

Acknowledgements

This work was supported in part by the United Fund of National Key Laboratory of Automatic Target Recognition (Shanghai) under Grant ATR(S)2025-006.

References

1. Ahn, D., Kang, J., Lee, S., Min, J., Kim, M., Jang, W., Cho, H., Paul, S., Kim, S., Cha, E., Jin, K.H., Kim, S.: A noise is worth diffusion guidance. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=xEWooSOgaz
2. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)

3. Bai, L., Shao, S., Zhou, Z., Qi, Z., Xu, Z., Xiong, H., Xie, Z.: Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In: The Thirteenth International Conference on Learning Representations (2024), https://openreview.net/forum?id=MKvQH1ekeY
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18392–18402 (2023)
5. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the International Conference on Computer Vision (ICCV) (2021)
6. Chang, Z., Li, M., Wang, J., Liu, Y., Wang, Q., Liu, Y.: Repairing catastrophic-neglect in text-to-image diffusion models via attention-guided feature enhancement. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 11379–11390 (2024)
7. Chatterjee, A., Stan, G.B.M., Aflalo, E., Paul, S., Ghosh, D., Gokhale, T., Schmidt, L., Hajishirzi, H., Lal, V., Baral, C., et al.: Getting it right: Improving spatial consistency in text-to-image models. In: European Conference on Computer Vision. pp. 204–222. Springer (2024)
8. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
9. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: CFG++: Manifold-constrained classifier free guidance for diffusion models. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=E77uvbOTtp
10. 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: Forty-first international conference on machine learning (2024)
11. Feng, H., Ding, Z., Xia, Z., Niklaus, S., Abrevaya, V., Black, M.J., Zhang, X.: Explorative inbetweening of time and space. In: European Conference on Computer Vision. pp. 378–395. Springer (2024)
12. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=PUIqjT4rzq7
13. Fort, S., Whitaker, J.: Direct ascent synthesis: Revealing hidden generative capabilities in discriminative models (2025), https://arxiv.org/abs/2502.07753
14. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
15. Grimal, P., Borgne, H.L., Ferret, O.: Text-to-image alignment in denoising-based models through step selection (2025), https://arxiv.org/abs/2504.17525
16. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9380–9389 (2024)
17. Harrington, A., Koepke, A., Karthik, S., Darrell, T., Efros, A.A.: It's never too late: Noise optimization for collapse recovery in trained diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 43124–43134 (2026)

18. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross-attention control. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=_CDixzkzeyb
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
20. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications (2021), https://openreview.net/forum?id=qw8AKxfYbI
21. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-Image Generation . IEEE Transactions on Pattern Analysis Machine Intelligence (01), 1–17 (Jan 5555), https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3531907
22. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
23. Huang, Y., Chen, Y., Ding, L., Zhang, X., Dai, W., Zou, J., Xiong, H., Tian, Q.: Im-zero: Instance-level motion controllable video generation in a zero-shot manner. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7265–7275 (2025)
24. Huang, Y., Chen, Y., Liu, Y., Zhang, X., Dai, W., Xiong, H., Tian, Q.: Domainfusion: Generalizing to unseen domains with latent diffusion models. In: European Conference on Computer Vision. pp. 480–498. Springer (2024)
25. Huang, Y., Chen, Y., Zhou, J., Dai, W., Zhang, X., Zou, J., Xiong, H., Tian, Q.: Diffusion-driven progressive target manipulation for source-free domain adaptation. Advances in Neural Information Processing Systems **38**, 108999–109019 (2026)
26. Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., et al.: Openclip. Zenodo (2021)
27. Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., Li, H.: Comat: Aligning text-to-image diffusion model with image-to-text concept matching. Advances in Neural Information Processing Systems **37**, 76177–76209 (2024)
28. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)
29. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
30. Kim, K., Kim, S.: Model already knows the best noise: Bayesian active noise selection via attention in video diffusion model. In: The Fourteenth International Conference on Learning Representations (2026), https://openreview.net/forum?id=11dzFZ2UM1
31. 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. Advances in neural information processing systems **36**, 36652–36663 (2023)
32. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)

33. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
34. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., Chen, D., He, J., Li, J., Li, W., Zhang, C., Quan, R., Lu, J., Huang, J., Yuan, X., Zheng, X., Li, Y., Zhang, J., Zhang, C., Chen, M., Liu, J., Fang, Z., Wang, W., Xue, J., Tao, Y., Zhu, J., Liu, K., Lin, S., Sun, Y., Li, Y., Wang, D., Chen, M., Hu, Z., Xiao, X., Chen, Y., Liu, Y., Liu, W., Wang, D., Yang, Y., Jiang, J., Lu, Q.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding (2024)
35. Lian, L., Li, B., Yala, A., Darrell, T.: LLM-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=hFALpTb4fR, featured Certification
36. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13199–13208 (2025)
37. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European conference on computer vision. pp. 423–439. Springer (2022)
38. Liu, Y., Zhang, Y., Jaakkola, T., Chang, S.: Correcting diffusion generation through resampling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8713–8723 (2024)
39. Ma, N., Tong, S., Jia, H., Hu, H., Su, Y.C., Zhang, M., Yang, X., Li, Y., Jaakkola, T., Jia, X., et al.: Inference-time scaling for diffusion models beyond scaling denoising steps. arXiv preprint arXiv:2501.09732 (2025)
40. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), https://openreview.net/forum?id=a68SUt6zFt, featured Certification
41. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
42. 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: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=di52zR8xgf
43. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=FjNys5c7VyY
44. Qi, Z., Bai, L., Xiong, H., Xie, Z.: Not all noises are created equally:diffusion noise selection and optimization (2024), https://arxiv.org/abs/2407.14041
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
46. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems **36**, 3536–3559 (2023)

47. Rassin, R., Ravfogel, S., Goldberg, Y.: Dalle-2 is seeing double: Flaws in word-to-concept mapping in text2image models. In: Proceedings of the Fifth BlackboxNLP workshop on analyzing and interpreting neural networks for NLP. pp. 335–345 (2022)
48. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
49. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
50. Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: The Thirteenth International Conference on Learning Representations (2024)
51. 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. Advances in neural information processing systems **35**, 36479–36494 (2022)
52. Samuel, D., Ben-Ari, R., Raviv, S., Darshan, N., Chechik, G.: Generating images of rare concepts using pre-trained diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4695–4703 (2024)
53. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.W., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.R., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: LAION-5b: An open large-scale dataset for training next generation image-text models. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022), https://openreview.net/forum?id=M3Y74vmsMcY
54. Shen, D., Song, G., Xue, Z., Wang, F.Y., Liu, Y.: Rethinking the spatial inconsistency in classifier-free diffusion guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9370–9379 (2024)
55. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=St1giarCHLP
56. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=PxTIG12RRHS
57. Tang, Z., Peng, J., Tang, J., Hong, M., Wang, F., Chang, T.H.: Inference-time alignment of diffusion models with direct noise optimization. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=JpbqiD7n9r
58. Team, Q.: Qwen2.5-vl (January 2025), https://qwenlm.github.io/blog/qwen2.5-vl/
59. Tian, J., Aggarwal, L., Colaco, A., Kira, Z., Gonzalez-Franco, M.: Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3554–3563 (2024)
60. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)

61. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
62. Wang, R., Chen, Z., Chen, C., Ma, J., Lu, H., Lin, X.: Compositional text-to-image synthesis with attention map control of diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5544–5552 (2024)
63. 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)
64. 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. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp. 15903–15935 (2023)
65. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184 (2023)
66. Zeng, Q., Song, J., Zheng, H., Jiang, H., Song, M.: D^2-dpm: Dual denoising for quantized diffusion probabilistic models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9806–9814 (2025). https://doi.org/10.1609/aaai.v39i9.33063
67. Zhang, G., Fu, B., Fan, Q., Zhang, Q., Liu, R., Gu, H., Zhang, H., Liu, X.: Compass: Enhancing spatial understanding in text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15253–15265 (2025)
68. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018). https://doi.org/10.1109/CVPR.2018.00068
69. Zhou, Z., Shao, S., Bai, L., Zhang, S., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. In: International Conference on Computer Vision (2025)
70. Zhuang, C., Hu, Y., Gao, P.: Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. Advances in Neural Information Processing Systems **37**, 57115–57149 (2024)