SCALE: Semantic-Calibrated Guidance Enhancement for Prompt-Faithful Diffusion

Tianhang Lu[1⋆], Sudong Cai[2,1⋆✉], Bingzhi Chen[1], Shaodong Shen[3], Chunting Liu[4], Longguang Wang[5], and Bing Wang[2]

[1] Beijing Institute of Technology, Zhuhai [2] The Hong Kong Polytechnic University
[3] Southeast University [4] Kyoto University [5] Sun Yat-sen University, Shenzhen

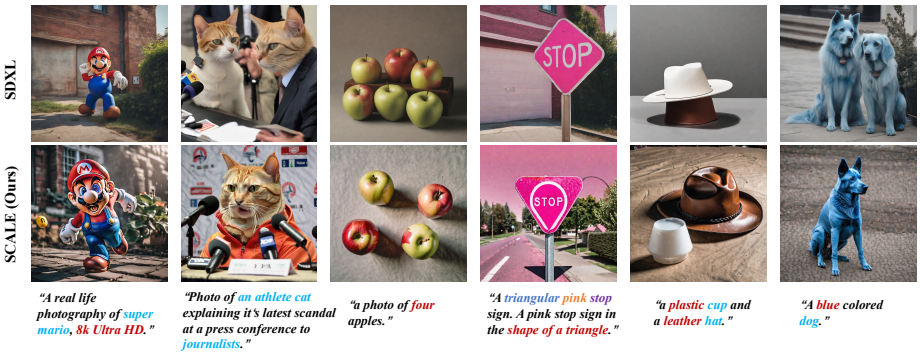

Fig. 1: Qualitative comparison: **SCALE** improves SDXL in prompt faithfulness across challenging cases involving fine details, counting, role assignment, and attribute binding.

Abstract. Ensuring prompt faithfulness remains a central challenge for text-to-image diffusion models. Classifier-Free Guidance (CFG) improves prompt adherence but exhibits an inherent quality–alignment tension: increasing the guidance scale to strengthen conditioning on the prompt often degrades visual quality and introduces artifacts. To probe the limit of greedy alignment maximization, we first introduce **SAP** (**S**emantic **A**lignment **P**rojection), a greedy update rule that projects each sampling update onto the guidance direction to maximize per-step alignment progress. We then show that SAP can fail due to the loss of orthogonal corrective freedom, discarding high-dimensional components that are crucial for rectifying accumulated trajectory drift. Based on this diagnosis, we propose **SCALE** (**S**emantic-**CAL**ibrated Guidance **E**nhancement), a drop-in, training-free guidance mechanism. SCALE selectively amplifies the semantic component along the guidance direction while preserving the orthogonal corrective component, improving semantic alignment without compromising structural fidelity. Across multiple text-to-image benchmarks, SCALE delivers substantial and consistent gains in prompt adherence and compositional alignment with negligible overhead over standard sampling. Project page: `https://github.com/SudongCAI/SCALE`.

Keywords: Diffusion Model · T2I Generation · Semantic Alignment

⋆ Tianhang Lu and Sudong Cai contributed equally to this work.
✉ Corresponding author: Sudong Cai (`sudong.cai@polyu.edu.hk`).

1 Introduction

Motivation: CFG's quality–alignment tension. Diffusion models have emerged as a leading paradigm for high-fidelity visual synthesis [11, 18, 37, 38, 43]. Yet in text-conditioned generation, *visual plausibility* does not guarantee *semantic faithfulness*: photorealistic images may still violate key attributes, relations, or compositional constraints specified by the prompt [14]. Ensuring that generated content faithfully follows the prompt rather than merely looking realistic remains a central challenge for controllable text-to-image diffusion models [16, 40].

Classifier-Free Guidance (CFG) [12] is widely leveraged as a foundational mechanism for conditional control in text-to-image synthesis. In essence, CFG performs a linear extrapolation between the unconditional and conditional noise predictions, thereby amplifying the influence of the text condition during sampling [17]. However, this linear extrapolation induces an inherent *quality–alignment tension*: increasing the guidance scale amplifies not only the semantic guidance signal but also structural perturbations that accumulate across denoising steps. As a result, aggressively pushing the guidance scale to maximize prompt adherence often destabilizes the denoising dynamics, driving the sample away from the data manifold and leading to structural distortions and visual artifacts [22, 36].

Existing efforts to navigate this quality–alignment tension broadly fall into two families. One line of work (*e.g.*, [35]) aims to support substantially higher guidance scales by explicitly mitigating over-guidance artifacts, seeking to preserve visual quality under strong conditional control. A second line (*e.g.*, [25]) instead operates within a conservative guidance range and improves alignment by reshaping the sampling trajectory, sometimes with added compute.

Motivated by this quality–alignment tension, we revisit the *linear extrapolation mechanism* of CFG [17]. Our central question is: *can we, with negligible inference overhead, optimally leverage the* ***semantic alignment signal****, while preserving the orthogonal corrective freedom essential for structural fidelity?* Concretely, under the standard noise-prediction formulation for text-to-image diffusion, we define the **conditional guidance vector** (*i.e.*, the container of the semantic alignment signal) at timestep t as

$$\boldsymbol{g}_t \triangleq \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \epsilon_\theta(\boldsymbol{x}_t, \varnothing)\,, \tag{1}$$

where $\epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c})$ and $\epsilon_\theta(\boldsymbol{x}_t, \varnothing)$ denote the conditional and unconditional noise predictions, respectively. Intuitively, $\boldsymbol{g}_t$ captures the prompt-induced change in the model's noise prediction and provides a text-specific steering direction for improving semantic alignment. To align with the sampler's denoising update convention (typically driven by a positive multiple of $-\epsilon_\theta$), we also introduce an update-induced reference direction

$$\hat{\boldsymbol{g}}_t \triangleq -\boldsymbol{g}_t, \tag{2}$$

noting that $\mathrm{span}(\hat{\boldsymbol{g}}_t) = \mathrm{span}(\boldsymbol{g}_t)$ and thus the semantic axis is invariant to this sign convention.

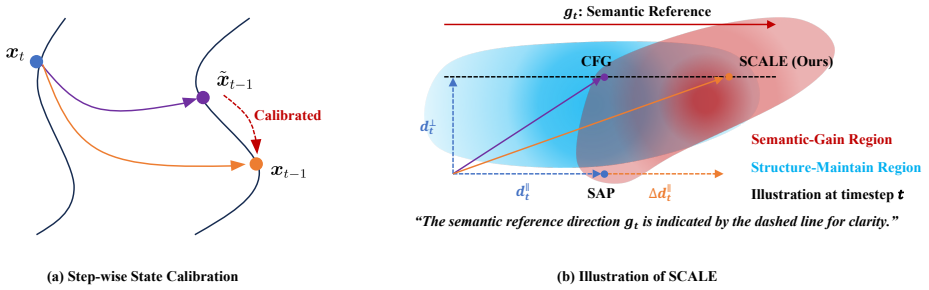

Fig. 2: Schematic illustration of the SCALE guidance mechanism. (a) illustrates the step-wise state calibration, where the trajectory is corrected from $\boldsymbol{x}_{t-1}$ to a calibrated state $\hat{\boldsymbol{x}}_{t-1}$ to maintain generative stability. **(b)** details the update dynamics at timestep t. We conceptualize the latent space into a *structure-preserving* region and a *semantic-gain* region. High-CFG tends to leave the structure-preserving region under coupled scaling, while SAP collapses by losing orthogonal corrective freedom. In contrast, SCALE follows the *Decoupled Satisfaction* paradigm: it selectively amplifies the semantic projection ($\boldsymbol{d}_{\parallel}$) while preserving the orthogonal restoring/corrective component ($\boldsymbol{d}_{\perp}$), improving alignment without compromising structural integrity.

Limits of alignment maximization. Given the conditional guidance vector $\boldsymbol{g}_t$, a natural question arises: *can we maximize prompt adherence by constraining each update to lie* ***along*** *the guidance direction* $\boldsymbol{g}_t$*?* To investigate this, we formulate **Semantic Alignment Projection (SAP)** as a steepest-ascent–inspired greedy baseline. Specifically, let $\tilde{\boldsymbol{x}}_{t-1}$ denote the deterministic state transition produced at timestep t via one-step denoising, and define the corresponding update vector as $\tilde{\boldsymbol{d}}_t \triangleq \tilde{\boldsymbol{x}}_{t-1} - \boldsymbol{x}_t$. SAP projects this update onto the semantic axis $\mathrm{span}(\boldsymbol{g}_t)$ (equivalently $\mathrm{span}(\hat{\boldsymbol{g}}_t)$). It decomposes $\tilde{\boldsymbol{d}}_t$ into a component parallel to $\hat{\boldsymbol{g}}_t$ and an orthogonal remainder, and **retains only the parallel component**:

$$\boldsymbol{d}_t^{\parallel} = \mathrm{Proj}_{\hat{\boldsymbol{g}}_t}(\tilde{\boldsymbol{d}}_t) = \frac{\langle \tilde{\boldsymbol{d}}_t, \hat{\boldsymbol{g}}_t \rangle}{\|\hat{\boldsymbol{g}}_t\|^2} \hat{\boldsymbol{g}}_t, \quad \boldsymbol{d}_t^{\perp} = \tilde{\boldsymbol{d}}_t - \boldsymbol{d}_t^{\parallel}. \tag{3}$$

That is, SAP retains the update channel aligned with the semantic axis and removes its orthogonal complement. However, because $\hat{\boldsymbol{g}}_t$ is not designed to preserve the denoising dynamics, discarding $\boldsymbol{d}_t^{\perp}$ can eliminate crucial corrective degrees of freedom and lead to unstable trajectories.

Structural defect of SAP: Orthogonal correction collapse. SAP imposes a one-dimensional update constraint by projecting each sampling step onto the semantic axis $\mathrm{span}(\boldsymbol{g}_t)$, *i.e.*, restricting the update to $\mathrm{span}(\boldsymbol{g}_t)$ and retaining only the $\boldsymbol{g}_t$-axis component. Under this constraint, the best-approximation property of orthogonal projection implies that the minimum-perturbation way to enforce $\tilde{\boldsymbol{d}}_t \in \mathrm{span}(\boldsymbol{g}_t)$ is precisely to replace $\tilde{\boldsymbol{d}}_t$ with its projection onto $\mathrm{span}(\boldsymbol{g}_t)$ (Lemma 1). Consequently, SAP explicitly discards the orthogonal component $\boldsymbol{d}_t^{\perp}$ at every step. This elimination removes high-dimensional orthogonal corrective degrees of freedom that are crucial for stable denoising, allowing inter-step biases to accumulate.

As a result, the sampling trajectory can become increasingly unstable, leading to severe degradation in visual fidelity or, in extreme cases, noise-like collapse. *Discussion is detailed in §3.1.*

From SAP to SCALE: Decoupling alignment and stability. The failure of SAP reveals a structural incompatibility: strengthening conditional alignment tends to amplify the guidance vector $\boldsymbol{g}_t$, whereas stable denoising relies on high-dimensional orthogonal corrective freedom. To address this problem, we propose **S**emantic-**Cal**ibrated Guidance **E**nhancement (**SCALE**), a training-free guidance mechanism that explicitly decouples these two requirements. (i) ***Controlled gain modulation along the semantic axis.*** Rather than collapsing the update into a one-dimensional subspace as SAP does, SCALE treats $\boldsymbol{g}_t$ as a semantic reference axis and applies a controlled amplification only along this axis. This yields a minimum-perturbation semantic enhancement relative to the baseline dynamics (Theorem 2 in §4.2). (ii) ***Preservation of Orthogonal Corrective Freedom.*** Crucially, structural stability is largely supported by the update channel orthogonal to $\boldsymbol{g}_t$, which corrects accumulated trajectory errors. SCALE therefore preserves this orthogonal component via an invariance constraint, ensuring stable denoising (Theorem 3 in §3.1). Formally, SCALE can be viewed as an axis-aligned scaling operator with orthogonal invariance, enabling stronger alignment without sacrificing structural fidelity.

Our contributions:

- **Diagnosing the stability–alignment incompatibility of CFG.** By dissecting SAP as an exploratory strategy to probe the limits of alignment, we theoretically attribute its failure to structural degeneracy induced by one-dimensional constraints, which eliminates orthogonal corrective degrees of freedom and triggers sampling–trajectory collapse.
- **Introducing SCALE as a mechanistically transparent remedy.** We propose SCALE as a drop-in, training-free guidance mechanism that departs from the conventional coupled trade-off. With an *orthogonally invariant* semantic scaling operator, SCALE maximizes semantic-alignment gains while fully preserving the structural stability required for denoising. We further provide theoretical guarantees for SCALE, including its minimum-perturbation property and stability bounds.
- **SOTA performance.** Extensive experiments demonstrate that SCALE significantly improves text-to-image alignment across diverse prompts and achieves SOTA results.

2 Related Work

Classifier-Free Guidance (CFG). CFG [12, 28] is the standard text-to-image alignment baseline, operating via score extrapolation: $\tilde{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}) = \epsilon_\theta(\boldsymbol{x}_t, \varnothing) + \omega(\epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \epsilon_\theta(\boldsymbol{x}_t, \varnothing))$ $(\omega > 1)$. However, its alignment capacity is bounded by the base model. Increasing ω to resolve complex prompts inevitably induces off-manifold artifacts (*e.g.*, oversaturation and distortion) instead of true semantic comprehension [26, 31, 36].

Training-Based Alignment. To break this semantic ceiling and align models with complex intent, many rely on training-based paradigms. Early efforts employ supervised fine-tuning on curated datasets [6, 33]. To capture nuanced human preferences, recent methods utilize preference learning—such as RLHF or DPO—to fine-tune generators via reward models or pairwise ranking [5, 39]. While effective at embedding strong alignment priors, these approaches suffer from massive computational overhead and strict reliance on large-scale annotated datasets, limiting their downstream flexibility.

Inference-time alignment. To bypass the inflexibility and prohibitive costs of training-based methods, training-free inference-time guidance emerges as a plug-and-play alternative. Rather than updating parameters, these approaches dynamically steer the generative trajectory. Existing methods can be broadly categorized into three main directions: (i) **Attention intervention** [1, 13] refines cross-attention maps to prevent concept bleeding; (ii) **Trajectory optimization** [9, 25, 44] modulates the noise prior to enforce consistent generative paths; (iii) **Reward guidance** [4, 20] leverages auxiliary gradients to steer sampling toward target objectives;and (iv) **Projection-based guidance** [7, 21, 35] decomposes guidance or solver updates into projected geometric components, enabling direction-selective modulation of the denoising trajectory.This operator-level formulation formally distinguishes SCALE from existing projection-based decomposition methods; a detailed comparison is provided in Appendix B.6.

Our position. Different from methods that modify model internals (*e.g.*, attention), directly optimize an auxiliary reward, or perform general projection-based modulation of sampling updates, our work focuses on a lightweight, step-level control of guidance-induced updates. Specifically, SCALE selectively amplifies the semantic reference-axis component while preserving the orthogonal corrective component of the original sampling dynamics, aiming to improve prompt alignment without sacrificing generative stability.

3 Problem Diagnosis

Probing boundaries via steepest ascent: SAP. To probe the theoretical boundaries of alignment maximization, SAP is inspired by a steepest-ascent upper bound within diffusion guidance under a locally linearized view of the dynamics. Let $\boldsymbol{d}_t$ denote a reference (baseline) sampler update at timestep t. Following §2, we use $\hat{\boldsymbol{g}}_t$ as the update-induced semantic reference direction (with $\mathrm{span}(\hat{\boldsymbol{g}}_t) = \mathrm{span}(\boldsymbol{g}_t)$). For a candidate update $\boldsymbol{z}_t$, we define the instantaneous semantic gain proxy

$$\mathcal{J}_t(\boldsymbol{z}_t) \triangleq \langle \boldsymbol{z}_t, \hat{\boldsymbol{g}}_t \rangle, \tag{4}$$

and consider a fixed step budget (*e.g.*, $\|\boldsymbol{z}_t\| \leqslant \|\boldsymbol{d}_t\|$). By Cauchy–Schwarz, $\mathcal{J}_t(\boldsymbol{z}_t) \leqslant \|\boldsymbol{z}_t\|\|\hat{\boldsymbol{g}}_t\|$, with equality if and only if $\boldsymbol{z}_t = \alpha\hat{\boldsymbol{g}}_t$ for some $\alpha \geqslant 0$ (*i.e.*, $\boldsymbol{z}_t$ is aligned with $\hat{\boldsymbol{g}}_t$). Motivated by this directional upper bound, SAP enforces an extreme one-dimensional subspace constraint $\boldsymbol{z}_t \in \mathrm{span}(\hat{\boldsymbol{g}}_t)$, collapsing each update into the semantic subspace. However, this greedy boundary probe deterministically

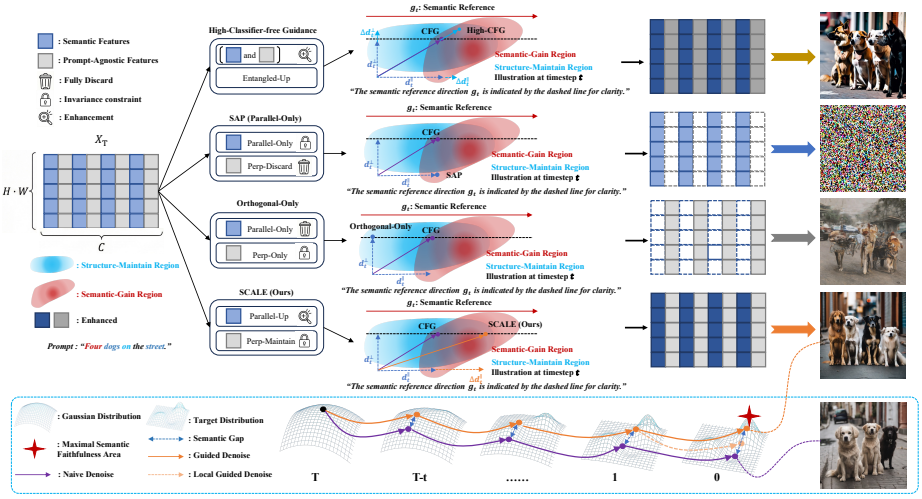

Fig. 3: Operational framework of SCALE. ***Top:*** At each timestep, we decompose the guidance update into a semantic component ($\boldsymbol{d}_{\parallel}$) and a prompt-agnostic (structural) component ($\boldsymbol{d}_{\perp}$) (orange vs. gray). Compared to **High-CFG** (coupled scaling) and **SAP** (discarding $\boldsymbol{d}_{\perp}$), SCALE follows a "Parallel-Up, Perp-Maintain" rule: it amplifies $\boldsymbol{d}_{\parallel}$ while preserving $\boldsymbol{d}_{\perp}$. **Orthogonal-Only** is included as a control to isolate the role of $\boldsymbol{d}_{\perp}$. ***Bottom:*** Over the denoising trajectory, this decoupled guidance improves semantic alignment while maintaining visual quality.

removes the orthogonal update channel, which carries the corrective freedom essential to stable denoising; consequently, cross-step deviations cannot be rectified and may accumulate into structural degradation.

What we analyze next. SAP is an intentionally extreme probe: by enforcing $\boldsymbol{z}_t \in \mathrm{span}(\hat{\boldsymbol{g}}_t)$, it saturates instantaneous semantic progress and exposes what comes at a cost. We first diagnose its failure from two complementary angles: geometrically (§3.1), the one-dimensional collapse eliminates the orthogonal correction channel; statistically (§3.2), $\boldsymbol{g}_t$ (hence $\hat{\boldsymbol{g}}_t$) corresponds to a likelihood-driven class-posterior force. We then cast this tension in a manifold-feasibility view (§3.3), which consolidates why semantic improvement and prior-feasible denoising impose decoupled requirements. *Together, these insights motivate SCALE: amplify semantic progress* without *sacrificing orthogonal corrective freedom.*

3.1 Geometric Analysis of SAP

Orthogonal correction collapse. SAP enforces $\boldsymbol{z}_t \in \mathrm{span}(\hat{\boldsymbol{g}}_t)$, which deterministically eliminates the orthogonal component of any reference update. We formalize this consequence via the best-approximation property of orthogonal projection.

Lemma 1 (Orthogonal projection as the best approximation in a subspace). *Let $\hat{\boldsymbol{g}}_t \neq 0$, and let $\boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\cdot)$ denote the orthogonal projection onto* $\mathrm{span}(\hat{\boldsymbol{g}}_t)$.

For any vector $\boldsymbol{u}$*, the constrained least-squares problem*

$$\min_{\boldsymbol{z}\in\mathrm{span}(\hat{\boldsymbol{g}}_t)} \|\boldsymbol{u}-\boldsymbol{z}\|^2 \tag{5}$$

admits a unique minimizer $\boldsymbol{z}^* = \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{u})$*, with minimum residual norm* $\|\boldsymbol{u} - \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{u})\|$*.*

Corollary 1 (Deterministic elimination of orthogonal correction under SAP). *Consider a reference update* $\boldsymbol{d}_t$ *at timestep* t *and decompose it into a parallel and an orthogonal component w.r.t.* $\hat{\boldsymbol{g}}_t$*:*

$$\boldsymbol{d}_t = \boldsymbol{d}_t^{\parallel} + \boldsymbol{d}_t^{\perp}, \tag{6}$$

$$\boldsymbol{d}_t^{\parallel} = \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{d}_t) \in \mathrm{span}(\hat{\boldsymbol{g}}_t), \qquad \boldsymbol{d}_t^{\perp} \perp \hat{\boldsymbol{g}}_t. \tag{7}$$

Any rule enforcing $\boldsymbol{z}_t \in \mathrm{span}(\hat{\boldsymbol{g}}_t)$ *necessarily satisfies* $\boldsymbol{P}^{\perp}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{z}_t) = 0$*, where* $\boldsymbol{P}^{\perp}_{\hat{\boldsymbol{g}}_t} \triangleq \boldsymbol{I} - \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}$*. Hence, whenever* $\boldsymbol{d}_t^{\perp} \neq 0$ *exists in the reference dynamics, SAP deterministically discards this orthogonal corrective direction.*

By Lemma 1, the projection $\boldsymbol{z}_t = \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{d}_t)$ is the *minimum-perturbation* update among all vectors constrained to $\mathrm{span}(\hat{\boldsymbol{g}}_t)$, yet it fully removes $\boldsymbol{d}_t^{\perp}$. Across timesteps, this hard one-dimensional constraint systematically suppresses the orthogonal correction channel, limiting the dynamics' ability to absorb modeling errors and accumulated drift. Empirically, such suppression can amplify trajectory deviation and manifest as structural artifacts and loss of fine details.

3.2 Statistical Analysis of SAP

Bayesian lens on the CFG guidance signal. We next interpret the semantic reference direction via score decomposition. Let $\boldsymbol{x}_t$ denote the diffusion state with marginal density $p_t(\boldsymbol{x}_t)$, and let $\boldsymbol{c}$ denote the condition with prior $p(\boldsymbol{c}) > 0$.

Lemma 2 (Bayesian decomposition of conditional score). *Assume* $p_t(\boldsymbol{x}_t)$ *and* $p_t(\boldsymbol{x}_t|\boldsymbol{c})$ *are differentiable w.r.t.* $\boldsymbol{x}_t$ *and* $p(\boldsymbol{c}) > 0$*. Then the conditional score decomposes as*

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t|\boldsymbol{c}) = \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) + \nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{c}|\boldsymbol{x}_t). \tag{8}$$

The decomposition in Eq. 8 isolates the *class-posterior* term beyond prior-consistent denoising. Under standard diffusion parameterizations (*e.g.*, noise prediction), the denoiser is proportional to the score, $\epsilon_\theta(\boldsymbol{x}_t) \simeq -\sigma_t \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$, where σ_t is the noise scale. Thus the CFG increment $\boldsymbol{g}_t$ aligns with the class-posterior score direction (up to a positive scaling), and the update-induced reference direction $\hat{\boldsymbol{g}}_t$ aligns with the *descent* direction of the conditional potential $\Psi_t(\boldsymbol{x}) \triangleq -\log p(\boldsymbol{c}|\boldsymbol{x})$:

Corollary 2 (Statistical interpretation of the guidance signal). *Under standard diffusion parameterizations (e.g., noise prediction), define the CFG increment* $\boldsymbol{g}_t \triangleq \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \epsilon_\theta(\boldsymbol{x}_t, \varnothing)$*. Then*

$$\boldsymbol{g}_t \propto -\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{c}|\boldsymbol{x}_t). \tag{9}$$

Let $\Psi_t(\boldsymbol{x}) \triangleq -\log p(\boldsymbol{c}|\boldsymbol{x})$ *and define the update-induced semantic reference direction* $\hat{\boldsymbol{g}}_t$ *by* $\hat{\boldsymbol{g}}_t = -\boldsymbol{g}_t$ *(so* $\mathrm{span}(\hat{\boldsymbol{g}}_t) = \mathrm{span}(\boldsymbol{g}_t)$*). Then*

$$\hat{\boldsymbol{g}}_t \propto -\nabla_{\boldsymbol{x}_t} \Psi_t(\boldsymbol{x}_t). \tag{10}$$

Remark 1. Consequently. Increasing $\mathcal{J}_t(\boldsymbol{z}_t) = \langle \boldsymbol{z}_t, \hat{\boldsymbol{g}}_t \rangle$ prioritizes the first-order decrease of the conditional potential Ψ_t under update $\boldsymbol{z}_t$ (up to a positive scaling factor). *However.* Overly concentrating updates onto this posterior-driven direction (as SAP does by enforcing $\boldsymbol{z}_t \in \mathrm{span}(\hat{\boldsymbol{g}}_t)$) can disturb the balance with the prior-consistent component, potentially pulling trajectories away from high-density regions and degrading distributional fidelity—consistent with the geometric loss of orthogonal correction in Corollary 1.

3.3 Manifold View

Manifold-feasibility. According to the manifold hypothesis, natural images concentrate near a low-dimensional manifold embedded in the ambient space. In a local neighborhood, small perturbations can be decomposed into components that vary along intrinsic degrees of freedom and components that deviate from the underlying structure.

Definition 1 (Potential functions and update sets). *Let* $\boldsymbol{x}_t \in \mathbb{R}^n$ *denote the diffusion state at timestep* t*, and for any update* $\boldsymbol{z} \in \mathbb{R}^n$ *define* $\boldsymbol{x}_t^+ \triangleq \boldsymbol{x}_t + \boldsymbol{z}$*. Define the prior potential* $\Phi_t(\boldsymbol{x}) \triangleq -\log p_t(\boldsymbol{x})$ *and the conditional potential* $\Psi_t(\boldsymbol{x}) \triangleq -\log p(\boldsymbol{c}|\boldsymbol{x})$*. Given thresholds* $\beta > 0$ *and* $\alpha > 0$*, define the prior-feasible set*

$$\mathcal{S}_t(\beta) \triangleq \{\boldsymbol{z} : \ \Phi_t(\boldsymbol{x}_t^+) - \Phi_t(\boldsymbol{x}_t) \leqslant \beta\}, \tag{11}$$

and the conditional-improvement set

$$\mathcal{A}_t(\alpha) \triangleq \{\boldsymbol{z} : \ \Psi_t(\boldsymbol{x}_t^+) - \Psi_t(\boldsymbol{x}_t) \leqslant -\alpha\}. \tag{12}$$

Proposition 1 (Local anisotropy of the prior-feasible set). *Under the manifold hypothesis, let* $\boldsymbol{x}_t$ *lie in a local neighborhood of the data manifold* $\mathcal{M}$*, with tangent space* $T \triangleq T_{\Pi_{\mathcal{M}}(\boldsymbol{x}_t)}\mathcal{M}$ *and normal complement* $N \triangleq T^\perp$*. For an infinitesimal update* $\boldsymbol{z} = \boldsymbol{z}_T + \boldsymbol{z}_N$*, assume* Φ_t *is twice differentiable and exhibits significantly larger curvature along* N *than along* T *(formalized by the subspace spectral regularity condition in the Appendix). Then the feasible set* $\mathcal{S}_t(\beta)$ *is locally anisotropic: compared to* $\boldsymbol{z}_T$*, the allowable magnitude of* $\boldsymbol{z}_N$ *is strictly more constrained, yielding a thin-shell region in directions deviating from the manifold.*

Theorem 1 (Local decoupled structure). *Under Proposition 1, there exists* $\delta > 0$ *such that for any* $\|\boldsymbol{z}\| \leqslant \delta$*:*

1. ***Feasibility constraint.*** *For any* $\boldsymbol{z} \in \mathcal{S}_t(\beta)$*, the normal deviation is tightly bounded by the local curvature/noise scale,* i.e.*,* $\|\boldsymbol{z}_N\|$ *must remain small.*
2. ***Improvement orientation.*** *For any* $\boldsymbol{z} \in \mathcal{A}_t(\alpha)$*, the first-order improvement is dominated by the projection onto the conditional score direction* $\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{c}|\boldsymbol{x}_t)$*, up to second-order terms controlled by the smoothness of* Ψ_t*.*

Remark 2. Consequently, feasible and improving updates exhibit a **decoupled structure**: feasibility primarily constrains structure-deviating components, while semantic improvement is driven by progress along the conditional score direction within the feasible region.

Note. This diagnosis suggests that improving alignment should mainly amplify semantic progress along the semantic axis while preserving the orthogonal corrective channel of the original dynamics, motivating our methodology in §4.

4 SCALE

From SAP to Decoupled Satisfaction. §3 shows that SAP fails for a structural reason: enforcing $\boldsymbol{z}_t \in \mathrm{span}(\hat{\boldsymbol{g}}_t)$ collapses the update to a one-dimensional semantic axis and inevitably discards the orthogonal correction channel, which is essential for absorbing modeling errors and cross-step drift. Meanwhile, $\hat{\boldsymbol{g}}_t$ encodes a likelihood-driven (class-posterior) force that improves alignment but may destabilize denoising when overly prioritized. These two views suggest a simple remedy: *amplify semantic progress only along the semantic axis while preserving the orthogonal corrective freedom.* We formalize this principle as *Decoupled Satisfaction* and instantiate it as **S**emantic-**Cal**ibrated Guidance **E**nhancement (**SCALE**). In this section, we work in the sampler *update space* and take the semantic reference direction as $\hat{\boldsymbol{g}}_t$ (Eq. 2), where $\hat{\boldsymbol{g}}_t = -\boldsymbol{g}_t$ and $\boldsymbol{g}_t = \epsilon_\theta(\boldsymbol{x}_t, \boldsymbol{c}) - \epsilon_\theta(\boldsymbol{x}_t, \varnothing)$ (Eq. 1). Let $\boldsymbol{d}_t$ denote the reference update produced by a baseline sampler at timestep t, and let $\boldsymbol{z}_t$ denote the SCALE-corrected update. We measure instantaneous semantic gain in the update space by the inner product $\langle \boldsymbol{z}_t, \hat{\boldsymbol{g}}_t \rangle$.

4.1 Minimal-Perturbation Optimality

Gain-constrained minimal perturbation. SCALE modifies the baseline update as little as possible while enforcing a controlled increase of semantic progress along $\hat{\boldsymbol{g}}_t$. Concretely, we impose the gain constraint

$$\langle \boldsymbol{z}_t, \hat{\boldsymbol{g}}_t \rangle = \lambda \langle \boldsymbol{d}_t, \hat{\boldsymbol{g}}_t \rangle, \qquad \lambda \geqslant 1. \tag{13}$$

Among all updates satisfying Eq. 13, SCALE chooses the closest one to $\boldsymbol{d}_t$ in Euclidean distance. In practice, SCALE is activated only when the baseline update exhibits non-negative semantic progress along $\hat{\boldsymbol{g}}_t$ (e.g., via step gating in Appendix C.4); otherwise we set $\lambda = 1$ and keep the baseline update unchanged.

Theorem 2 (SCALE as minimal-perturbation extrapolation). *Let $\boldsymbol{d}_t \in \mathbb{R}^n$ be a reference update, $\hat{\boldsymbol{g}}_t \in \mathbb{R}^n$ with $\hat{\boldsymbol{g}}_t \neq 0$ be the semantic reference direction, and $\lambda \geqslant 1$ be a semantic scaling factor. Consider the convex program*

$$\min_{\boldsymbol{z}\in\mathbb{R}^n} \ \|\boldsymbol{z} - \boldsymbol{d}_t\|^2 \quad s.t. \quad \langle \boldsymbol{z}, \hat{\boldsymbol{g}}_t \rangle = \lambda \langle \boldsymbol{d}_t, \hat{\boldsymbol{g}}_t \rangle. \tag{14}$$

It admits a unique optimizer $\boldsymbol{z}$ given by

$$\boldsymbol{z} = \boldsymbol{d}_t + (\lambda - 1)\boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{d}_t) = \lambda \boldsymbol{d}_t^{\parallel} + \boldsymbol{d}_t^{\perp}, \tag{15}$$

where $\boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\cdot)$ is the orthogonal projector onto $\mathrm{span}(\hat{\boldsymbol{g}}_t)$*, $\boldsymbol{d}_t^{\parallel} = \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{d}_t)$, and $\boldsymbol{d}_t^{\perp} = \boldsymbol{d}_t - \boldsymbol{d}_t^{\parallel}$.*

The closed form in Eq. 15 makes the decoupling explicit: SCALE scales only the semantic-axis component and leaves the orthogonal component intact.

Corollary 3 (Orthogonal-channel invariance). *Let $\boldsymbol{z}$ be the minimizer in Theorem 2. Then the orthogonal component is preserved exactly:*

$$\boldsymbol{P}_{\hat{\boldsymbol{g}}_t}^{\perp}(\boldsymbol{z}) = \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}^{\perp}(\boldsymbol{d}_t) = \boldsymbol{d}_t^{\perp}, \qquad \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}^{\perp} \triangleq \boldsymbol{I} - \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}. \tag{16}$$

Corollary 3 states a precise guarantee: SCALE removes SAP's failure mechanism (discarding orthogonal correction) while still allowing controllable semantic amplification along $\hat{\boldsymbol{g}}_t$.

4.2 Operator View and Stability

Key observation: Orthogonal-invariant semantic scaling. Theorem 2 implies that SCALE can be written as a simple axis-aligned linear operator applied to the baseline update.

Theorem 3 (Orthogonal-invariant semantic scaling operator). *Let $\hat{\boldsymbol{g}}_t \neq 0$ and $\lambda \geqslant 1$. Define*

$$\boldsymbol{A}_\lambda \triangleq \boldsymbol{P}_{\hat{\boldsymbol{g}}_t}^{\perp} + \lambda \boldsymbol{P}_{\hat{\boldsymbol{g}}_t} \ = \ \boldsymbol{I} + (\lambda - 1)\boldsymbol{P}_{\hat{\boldsymbol{g}}_t}, \qquad \textit{so that} \qquad \boldsymbol{z} = \boldsymbol{A}_\lambda \boldsymbol{d}_t. \tag{17}$$

Then:

1. ***Invariant subspaces.*** $\mathrm{span}(\hat{\boldsymbol{g}}_t)$ *and* $\mathrm{span}(\hat{\boldsymbol{g}}_t)^{\perp}$ *are invariant under $\boldsymbol{A}_\lambda$.*
2. ***Spectrum.*** *$\boldsymbol{A}_\lambda$ has eigenvalue λ on* $\mathrm{span}(\hat{\boldsymbol{g}}_t)$ *and eigenvalue* 1 *on* $\mathrm{span}(\hat{\boldsymbol{g}}_t)^{\perp}$*.*
3. ***No distortion on the orthogonal channel.*** *For any $\boldsymbol{v} \perp \hat{\boldsymbol{g}}_t$, $\boldsymbol{A}_\lambda \boldsymbol{v} = \boldsymbol{v}$; hence SCALE acts as the identity on the orthogonal complement.*

This operator view isolates where SCALE can expand: only along the semantic axis. The orthogonal correction channel is strictly undistorted.

Corollary 4 (Exact perturbation bound and degenerate case). *Let $\bar{\boldsymbol{g}}_t \triangleq \hat{\boldsymbol{g}}_t / \|\hat{\boldsymbol{g}}_t\|$. Under Theorem 3,*

$$\|\boldsymbol{z} - \boldsymbol{d}_t\| = |\lambda - 1| \, \|\boldsymbol{P}_{\hat{\boldsymbol{g}}_t}(\boldsymbol{d}_t)\| = |\lambda - 1| \, |\langle \boldsymbol{d}_t, \bar{\boldsymbol{g}}_t \rangle|. \tag{18}$$

In particular, if $\langle \boldsymbol{d}_t, \hat{\boldsymbol{g}}_t \rangle = 0$, then $\boldsymbol{z} = \boldsymbol{d}_t$ (SCALE degenerates to the identity update).

4.3 Algorithmic Application and Implementation

SCALE is a lightweight inference-time intervention. SCALE requires no fine-tuning or architectural changes. At each step, given a baseline update $\boldsymbol{d}_t$ and semantic direction $\hat{\boldsymbol{g}}_t$, SCALE computes $\boldsymbol{d}_t^{\parallel}$ and $\boldsymbol{d}_t^{\perp}$ and outputs

$$\boldsymbol{z}_t = \lambda_t \boldsymbol{d}_t^{\parallel} + \boldsymbol{d}_t^{\perp}, \tag{19}$$

where $\lambda_t \geqslant 1$ may be chosen adaptively (e.g., as a function of the alignment between $\boldsymbol{d}_t$ and $\hat{\boldsymbol{g}}_t$). All theoretical statements above hold *step-wise*: for each timestep t, conditioning on the chosen λ_t, SCALE equals the unique solution of Eq. 14. In our implementation, we adopt the orthogonal-invariant form (*i.e.*, the orthogonal component is kept unchanged, corresponding to $\beta \equiv 1$), with optional step gating and λ_t schedules as described in Appendix C.4.

5 Experiments

We evaluate the effectiveness and generalizability of SCALE through extensive comparisons with state-of-the-art training-free diffusion methods.

5.1 Settings

Base models and datasets. We primarily evaluate **SCALE** on SDXL [29] at 1024×1024 resolution. To validate generalizability, we also report results on SD1.5 [32] and FLUX.1-Dev [23]. Extensive evaluations are conducted on three benchmarks: DrawBench [36], GenEval [10], and a subset of T2I-CompBench [15].

Metrics and baselines. To evaluate semantic fidelity, we apply four distinct metrics: *CLIPScore* [30], *BLIPScore* [24], *VQAScore* [27], and *DSGScore* [8]. We also report *ImageReward* [42] as an image-quality preference metric. For comparison, we include the base model and a suite of strong inference-time baselines: PAG [1], SEG [13], AYS [34], Z-Sampling [25], CFG++ [9], APG [35], Golden Noise [44], GA-Eval [41], TAG [7], DAS [20], and Diffusion-DPO [39].

5.2 Main Experiments

Implementation details. All pre-trained models are initialized from their official public checkpoints. For fair comparison, we reproduce all baselines using their *canonical open-source implementations* within consistent experimental settings. For the calculation of DSGScore, we employ *Qwen2.5-VL-7B-Instruct* [3] as the question-answering backbone. Further details are provided in the Appendix.

Main results. Table 1 summarizes quantitative results of **SCALE** against standard sampling and recent baselines across multiple datasets, with qualitative comparisons in Fig. 4. We follow the official GenEval evaluation protocol and report the corresponding accuracies in Table 2. Overall, SCALE achieves consistent gains across datasets and metrics. Table 3 further compares **SCALE** with DAS,

Table 1: Quantitative comparisons on DrawBench, GenEval, and T2I-CompBench. We report the Text-to-Image Alignment performance of SCALE and baseline methods.

Prompt Set	Method	CLIPScore↑	BLIPScore↑	VQAScore↑	DSGScore↑	ImageReward↑
DrawBench [36]	SDXL [29]	0.2775	0.5087	0.7417	0.7068	0.6402
	PAG [1](ECCV'24)	0.2696	0.4896	0.7239	0.6727	0.5509
	SEG [13](NeurIPS'24)	0.2723	0.4998	0.7155	0.6662	0.6255
	AYS [34](ICML'24)	0.2703	0.4983	0.7239	0.7003	0.3235
	Z-Sampling [25](ICLR'25)	0.2811	0.5129	0.7438	0.7064	0.7983
	CFG++ [9](ICLR'25)	0.2804	0.5067	0.7487	0.7133	0.6022
	APG [35](ICLR'25)	0.2863	0.5176	0.7800	0.6912	0.7857
	Golden Noise [44](ICCV'25)	0.2814	0.5043	0.7439	0.7084	0.6694
	GA-Eval [41](ICLR'26)	**0.2889**	0.5198	0.7867	0.7034	**0.8522**
	TAG [7](ICML'26)	0.2717	0.4957	0.7259	0.6697	0.4028
	SCALE (Ours)	0.2885	**0.5244**	**0.7889**	**0.7305**	0.8305
GenEval [10]	SDXL [29]	0.2802	0.5120	0.7998	0.8439	0.5746
	PAG [1](ECCV'24)	0.2744	0.4973	0.7455	0.7854	0.4867
	SEG [13](NeurIPS'24)	0.2751	0.5038	0.7568	0.7946	0.5509
	AYS [34](ICML'24)	0.2784	0.5105	0.7647	0.8051	0.3521
	Z-Sampling [25](ICLR'25)	0.2860	0.5180	0.8165	0.8455	0.7624
	CFG++ [9](ICLR'25)	0.2814	0.5128	0.8170	0.8331	0.6537
	APG [35](ICLR'25)	0.2873	0.5173	0.8176	0.8295	0.7599
	Golden Noise [44](ICCV'25)	0.2810	0.5119	0.8065	0.8369	0.6277
	GA-Eval [41](ICLR'26)	0.2872	0.5189	0.8182	0.8394	0.8104
	TAG [7](ICML'26)	0.2763	0.5044	0.7699	0.7876	0.4870
	SCALE (Ours)	**0.2887**	**0.5190**	**0.8345**	**0.8556**	**0.8168**
T2I-CompBench [15]	SDXL [29]	0.2771	0.5194	0.7192	0.7306	0.4852
	PAG [1](ECCV'24)	0.2657	0.4995	0.6889	0.6989	0.3053
	SEG [13](NeurIPS'24)	0.2687	0.5064	0.6974	0.7078	0.4293
	AYS [34](ICML'24)	0.2710	0.5114	0.7141	0.7140	0.2448
	Z-Sampling [25](ICLR'25)	0.2811	0.5261	0.7347	0.7490	0.7125
	CFG++ [9](ICLR'25)	0.2760	0.5194	0.7305	0.7418	0.5260
	APG [35](ICLR'25)	0.2815	**0.5282**	0.7382	0.7462	0.7424
	Golden Noise [44](ICCV'25)	0.2752	0.5148	0.7311	0.7232	0.4638
	GA-Eval [41](ICLR'26)	0.2811	0.5257	0.7289	0.7293	0.7125
	TAG [7](ICML'26)	0.2710	0.4993	0.6941	0.6763	0.2264
	SCALE (Ours)	**0.2823**	**0.5282**	**0.7593**	**0.7593**	**0.7725**

* *Note.* All baselines are reproduced using their official open-source implementations under a unified experimental setting. We adopt evaluation code from official codebases to ensure fair and standardized comparisons. GA-Eval [41] is applied on top of Z-Sampling [25].

Fig. 4: Qualitative comparison of SCALE against baseline methods. SCALE demonstrates robust prompt faithfulness, particularly when conditioned on complex and challenging prompts. To ensure a strictly fair comparison, all baselines were reproduced in an identical environment using their official open-source implementations.

highlighting their differences in performance across benchmarks. In addition, we include further comparisons with PLADIS [19] and W2SD [2] on the DrawBench

Table 2: Accuracy evaluation on the GenEval benchmark. We reproduce all baselines using their official open-source implementations for fair comparison.

Method	Single object	Two object↑	Counting↑	Colors↑	Position↑	Color attribution↑	Overall↑
SDXL [29]	**100.00%**	68.69%	40.00%	81.91%	9.00%	18.00%	53.44%
Z-Sampling [25](ICLR'25)	**100.00%**	**76.77%**	38.75%	86.17%	12.00%	**20.00%**	55.61%
SCALE (Ours)	**100.00%**	71.72%	**42.50%**	**88.30%**	**17.00%**	19.00%	**56.41%**

* We follow the official GenEval evaluation protocol to ensure standardized evaluation.

Table 3: Quantitative comparisons with DAS. DAS is reproduced using its official implementation with its default CLIPScore reward.

Prompt Set	Method	CLIPScore↑	BLIPScore↑	VQAScore↑	DSGScore↑	ImageReward↑
DrawBench [36]	SDXL [29]	0.2771	0.5188	0.7184	0.6945	0.5921
	DAS [20](ICLR'25)	**0.2912**	0.5187	0.7483	0.7215	0.7793
	SCALE (Ours)	0.2885	**0.5244**	**0.7889**	**0.7305**	**0.8305**
GenEval [10]	SDXL [29]	0.2802	0.5120	0.7900	0.8373	0.5487
	DAS [20](ICLR'25)	**0.2912**	0.5187	0.8245	0.8510	0.7793
	SCALE (Ours)	0.2887	**0.5190**	**0.8345**	**0.8556**	**0.8168**
T2I-CompBench [15]	SDXL [29]	0.2771	0.5194	0.7192	0.7306	0.4852
	DAS [20](ICLR'25)	**0.2867**	0.5271	0.7265	0.7401	0.6802
	SCALE (Ours)	0.2823	**0.5282**	**0.7593**	**0.7593**	**0.7725**

* While DAS leads on CLIPScore, **SCALE** achieves robust improvements across various metrics.

dataset. We also evaluate SCALE under the long-prompt setting by constructing a GPT-expanded variant of DrawBench. Quantitative results are discussed in Appendix C.5 (Table 2-3).

Compatibility with training-based methods. To further examine whether **SCALE** is complementary to training-based approaches, we apply SCALE at inference time on top of a Diffusion-DPO–fine-tuned model to evaluate potential improvements. Quantitative results are reported in Table 4, showing that **SCALE** yields clear additional gains over Diffusion-DPO.

Efficiency. As a lightweight drop-in module, **SCALE** introduces negligible computational overhead (see §4.3). As shown in Table 5, SCALE achieves an inference latency nearly identical to standard SDXL sampling (Ratio $\approx$ **1.0038**), indicating minimal overhead without compromising performance.

5.3 Ablation Study

We conduct ablation studies on DrawBench to analyze the effect of key design choices in **SCALE**. We report two representative ablations here, with additional results provided in the Appendix.

Ablation of feasibility and improvement. To validate the decoupled paradigm, we isolate the roles of the parallel and orthogonal-to-$\hat{\boldsymbol{g}}_t$ components ($\boldsymbol{d}_{\parallel}$ and $\boldsymbol{d}_{\perp}$). We compare **SCALE** against four baselines: (1) Standard CFG; (2) Simultaneous Enhancement (scaling both $\boldsymbol{d}_{\parallel}$ and $\boldsymbol{d}_{\perp}$ by the same factor); (3) Parallel-Only (discarding $\boldsymbol{d}_{\perp}$); and (4) Perp-Only (discarding $\boldsymbol{d}_{\parallel}$). Comparative results are reported in Table 6, where SCALE improves the baseline and variants significantly.

Table 4: Compatibility evaluation with preference-tuned models. We utilize the official open-source weights and inference code of Diffusion-DPO for batch testing.

Prompt Set	Method	CLIPScore↑	BLIPScore↑	VQAScore↑	DSGScore↑	ImageReward↑
DrawBench [36]	SDXL [29]	0.2771	0.5118	0.7184	0.6945	0.5921
	SDXL-DPO [39](CVPR'24)	0.2853	0.5159	0.7685	0.7172	0.8178
	SDXL-DPO+SCALE (Ours)	**0.2911**	**0.5301**	**0.7955**	**0.7638**	**0.9167**
GenEval [10]	SDXL [29]	0.2802	0.5120	0.7900	0.8373	0.5487
	SDXL-DPO [39](CVPR'24)	0.2912	0.5241	0.8340	0.8342	0.9856
	SDXL-DPO+SCALE (Ours)	**0.2943**	**0.5279**	**0.8501**	**0.8642**	**1.0278**
T2I-CompBench [15]	SDXL [29]	0.2771	0.5194	0.7192	0.7306	0.7306
	SDXL-DPO [39](CVPR'24)	0.2839	0.5296	0.7542	0.7707	0.8607
	SDXL-DPO+SCALE (Ours)	**0.2863**	**0.5358**	**0.7707**	**0.7834**	**0.9212**

* Concurrently, we evaluate the efficacy of applying SCALE to Diffusion-DPO. Both scenarios are executed within an identical experimental environment to ensure strict fairness.

Table 5: Comparison of inference time relative to SDXL. To guarantee strict fairness in efficiency benchmarking, we employ the official implementation of each method.

Method	Ratio	Method	Ratio	Method	Ratio
SDXL	1.0000×	CFG++ (ICLR'25)	0.9741×	TAG (ICML'26)	1.0114×
AYS (ICML'24)	0.2315×	Golden Noise (ICCV'25)	1.0395×	GA-Eval (ICLR'26)	3.5506×
Z-Sampling (ICLR'25)	2.8159×	DAS (ICLR'25)	55.6001×	**SCALE (Ours)**	**1.0038**×

* *Note.* For each baseline, we measure runtime independently and normalize it to standard sampling measured in the same session. In our implementation, standard sampling takes 5.3486 s per image, while **SCALE** takes 5.3687 s, indicating negligible latency overhead.

Ablation of the hybrid gain coefficient. SCALE adopts a hybrid gain schedule that combines a tunable lower bound with an adaptive modulation. Let $\hat{\boldsymbol{g}}_t$ be the update-induced semantic reference direction (Eq. 2), and let $\tilde{\boldsymbol{d}}_t$ be the baseline trial update in Algorithm(See Appendix B.5). We define the alignment score in the update space as

$$\cos(\Theta_t) \triangleq \frac{\langle \tilde{\boldsymbol{d}}_t, \hat{\boldsymbol{g}}_t \rangle}{\|\tilde{\boldsymbol{d}}_t\| \, \|\hat{\boldsymbol{g}}_t\|}. \tag{20}$$

We set the semantic scaling factor as

$$\lambda_t \triangleq 1 + \lambda_{\text{base}} + \cos(\Theta_t), \tag{21}$$

where λ_{base} is a hyperparameter controlling the minimum semantic gain. When SCALE is activated (late steps and $\cos(\Theta_t) > 0$), larger $\cos(\Theta_t)$ yields larger λ_t, producing stronger semantic amplification. Since $\cos(\Theta_t) \in (0, 1]$ in the activated regime, we have

$$\lambda_t \in \big(1 + \lambda_{\text{base}},\, 2 + \lambda_{\text{base}}\big]. \tag{22}$$

Fig. 5 reports performance under different choices of λ_{base} and with/without the adaptive term. In our main experiments, we set $\lambda_{\text{base}} = 1$ and enable the adaptive modulation.

Table 6: Ablation of guidance components on DrawBench. **SCALE** is compared with component-wise variants and improves both semantic alignment and visual quality.

Method	ClipScore↑	BlipScore↑	VQAScore↑	DSGScore↑	ImageReward↑
SDXL [29]	0.2771	0.5188	0.7184	0.6945	0.5921
Simultaneous Enhancement	0.2624	0.4784	0.6519	0.5881	-0.1666
Parallel-Only	0.1418	0.2950	0.1446	0.0092	-2.2422
Perp-Only	0.1202	0.2864	0.2900	0.1097	-2.0469
SCALE (Original)	**0.2885**	**0.5244**	**0.7889**	**0.7305**	**0.8305**

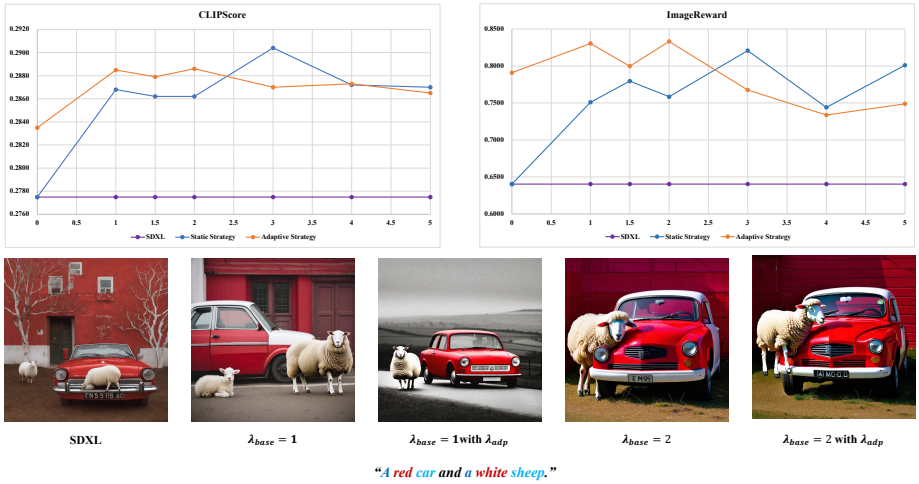

Fig. 5: Alignment performance under different hybrid gain settings.

6 Conclusion

To probe the limits of alignment maximization, we introduce SAP and show that collapsing updates to a one-dimensional semantic axis inevitably discards orthogonal correction, which is crucial for stable denoising. Motivated by this diagnosis, we propose SCALE, a training-free guidance mechanism that follows the Decoupled Satisfaction principle: it amplifies semantic progress along the semantic axis while preserving the orthogonal corrective channel. Our theoretical analysis establishes SCALE's minimal-perturbation optimality and orthogonal-channel invariance, and comprehensive experiments across standard CFG-based diffusion baselines validate its ability to improve text–image alignment with negligible runtime overhead while preserving visual fidelity.

Acknowledgments

This work was jointly supported by the the Research Grants Council of Hong Kong (25206524, 15212925), National Natural Science Foundation of China (42301520, 62302172), and the Innovation and Technology Fund (PRP/068/23FX).

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: European Conference on Computer Vision (ECCV). pp. 1–17. Springer (2024)
2. Bai, L., Sugiyama, M., Xie, Z.: Weak-to-strong diffusion with reflection. In: International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=tg19FVh3p1
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923
4. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, R., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=pzpWBbnwiJ
5. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. In: The Twelfth International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=YCWjhGrJFD
6. 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 (CVPR). pp. 18392–18402 (June 2023)
7. Cho, H., Ahn, D., Hong, S., Kim, J.E., Kim, S., Jin, K.H.: Tag: Tangential amplifying guidance for hallucination-resistant sampling. arXiv preprint arXiv:2510.04533 (2025)
8. Cho, J., Hu, Y., Baldridge, J.M., Garg, R., Anderson, P., Krishna, R., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. In: International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=ITq4ZRUT4a
9. Chung, H., Kim, J., Park, G.Y., Nam, H., Ye, J.C.: CFG++: Manifold-constrained classifier free guidance for diffusion models. In: International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=E77uvbOTtp
10. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 52132–52152. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/a3bf71c7c63f0c3bcb7ff67c67b1e7b1-Paper-Datasets_and_Benchmarks.pdf
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
12. 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
13. Hong, S.: Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing

Systems (NeurIPS). vol. 37, pp. 66743–66772. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-2132, https://proceedings.neurips.cc/paper_files/paper/2024/file/7b3f7b6670fdab2933411b5b922cdcc3-Paper-Conference.pdf
14. Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A.: Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 20406–20417 (October 2023)
15. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 78723–78747. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/f8ad010cdd9143dbb0e9308c093aff24-Paper-Datasets_and_Benchmarks.pdf
16. 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. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 76177–76209. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-2425, https://proceedings.neurips.cc/paper_files/paper/2024/file/8b54ecd9823fff6d37e61ece8f87e534-Paper-Conference.pdf
17. Jin, C., Xiao, Z., Liu, C., Gu, Y.: Angle domain guidance: Latent diffusion requires rotation rather than extrapolation. In: International Conference on Machine Learning (ICML) (2025), https://openreview.net/forum?id=DidTLeezyp
18. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 26565–26577. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/a98846e9d9cc01cfb87eb694d946ce6b-Paper-Conference.pdf
19. Kim, K., Sim, B.: Pladis: Pushing the limits of attention in diffusion models at inference time by leveraging sparsity. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 16238–16248 (October 2025)
20. Kim, S., Kim, M., Park, D.: Test-time alignment of diffusion models without reward over-optimization. In: International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=vi3DjUhFVm
21. Kwon, M., Kim, S.s., Jeong, J., Hsiao, Y.T., Uh, Y.: Tcfg: Tangential damping classifier-free guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 2620–2629 (June 2025)
22. Kynkäänniemi, T., Aittala, M., Karras, T., Laine, S., Aila, T., Lehtinen, J.: Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 122458–122483. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-3892, https://proceedings.neurips.cc/paper_files/paper/2024/file/dd540e1c8d26687d56d296e64d35949f-Paper-Conference.pdf
23. Labs, B.F.: Flux. https://github.com/black-forest-labs/flux (2024)
24. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning (ICML). pp. 12888–12900. PMLR (2022)

25. LiChen, B., Shao, S., zikai zhou, Qi, Z., zhiqiang xu, Xiong, H., Xie, Z.: Zigzag diffusion sampling: Diffusion models can self-improve via self-reflection. In: International Conference on Learning Representations (ICLR) (2025), https://openreview.net/forum?id=MKvQH1ekeY
26. Lin, S., Liu, B., Li, J., Yang, X.: Common diffusion noise schedules and sample steps are flawed. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5404–5411 (January 2024)
27. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision (ECCV). pp. 366–384. Springer (2024)
28. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
29. 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: International Conference on Learning Representations (ICLR) (2024), https://openreview.net/forum?id=di52zR8xgf
30. 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 (ICML). pp. 8748–8763. PmLR (2021)
31. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 3536–3559. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/0b08d733a5d45a547344c4e9d88bb8bc-Paper-Conference.pdf
32. 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 (CVPR). pp. 10684–10695 (June 2022)
33. 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 (CVPR). pp. 22500–22510 (June 2023)
34. Sabour, A., Fidler, S., Kreis, K.: Align your steps: Optimizing sampling schedules in diffusion models (2024)
35. Sadat, S., Hilliges, O., Weber, R.M.: Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In: International Conference on Learning Representations (ICLR) (2025)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 36479–36494. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=St1giarCHLP

38. 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 (ICLR) (2021), https://openreview.net/forum?id=PxTIG12RRHS
39. 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 (CVPR). pp. 8228–8238 (June 2024)
40. Wang, Z., Sha, Z., Ding, Z., Wang, Y., Tu, Z.: Tokencompose: Text-to-image diffusion with token-level supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8553–8564 (June 2024)
41. Xie, D., Shao, S., Bai, L., zikai zhou, Cheng, B., Yang, S., JUN, W., Xie, Z.: Guidance matters: Rethinking the evaluation pitfall for text-to-image generation. In: International Conference on Learning Representations (ICLR) (2026), https://openreview.net/forum?id=T9xcbgFD3k
42. 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: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (NeurIPS). vol. 36, pp. 15903–15935. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf
43. Zhang, Z., Tu, Z., Yuan, J., Soh, D.W., Du, B.: Leveraging text-to-image diffusion models for unsupervised visual object tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–19 (2026). https://doi.org/10.1109/TPAMI.2026.3697145
44. Zhou, Z., Shao, S., Bai, L., Zhang, S., Xu, Z., Han, B., Xie, Z.: Golden noise for diffusion models: A learning framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17688–17697 (October 2025)