TIDE: TEXT INTERVENTION FOR DIVERSIFIED EXPLORATION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

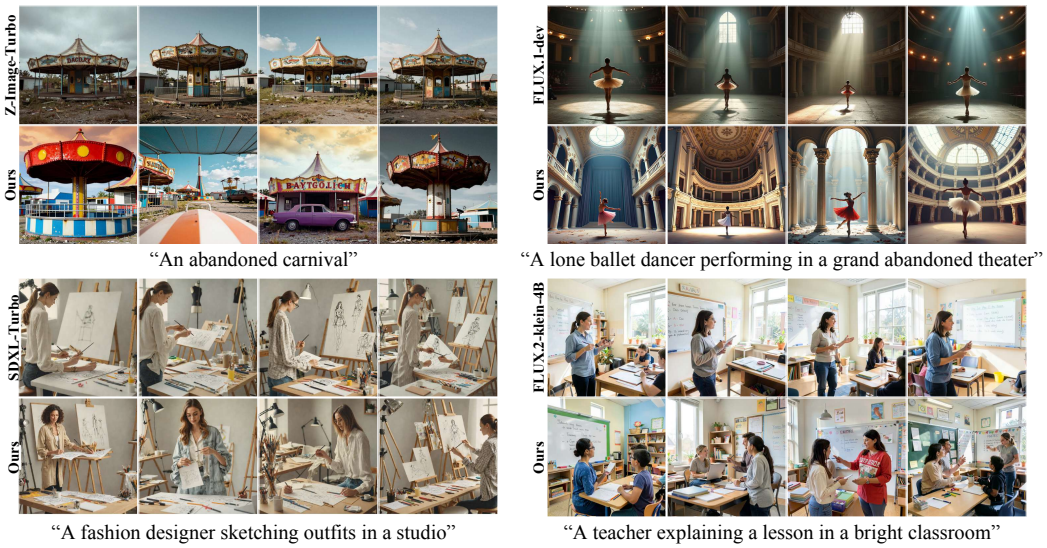


Figure 1: **Qualitative results across different text-to-image generators.** TIDE improves within-prompt diversity across Z-Image-Turbo, FLUX.1-dev, SDXL-Turbo, and FLUX.2-Klein-4B while preserving the semantic content of the input prompt.

ABSTRACT

Accelerated text-to-image diffusion models, including distilled generators such as Z-Image-Turbo and FLUX.2-Klein, can produce high-quality and semantically aligned images in only a few denoising steps. However, repeated sampling of the same prompt often yields highly similar compositions and styles despite different initial noise seeds. We refer to this reduced sensitivity to seed variation as *noise neglect*, which manifests as a near "one-to-one" prompt-to-image mapping and severe within-prompt diversity collapse. Existing training-free approaches improve diversity through various inference-time interventions, but often trade off generation quality, efficiency, or architectural generality. In this work, we propose to apply intervention in the text embedding space as a novel framework for achieving rich diversity in diffusion models. By explicitly decomposing the text representation with Singular Value Decomposition, we propose a new spherical linear interpolation method for the decomposed right-singular vectors of the text conditioning, which is enriched with emergent image structure. However, manipulating the prompt embedding alone does not sufficiently separate trajectories initialized from different random seeds. We further propose a latent-space repulsion method to coordinate trajectories sharing a prompt. By measuring similarities across multiple aspects, such as spatial and channel statistics, we derive a repulsive loss to explicitly push the latents apart and enhance diversity. We further demonstrate that our method extends beyond text-to-image generation to video generation and downstream 3D asset generation. Extensive experiments show that our method effectively enforces the "one-to-many" generative capability, significantly enhancing visual diversity while maintaining fidelity. Our code will be open-sourced.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Modern text-to-image generative models have achieved remarkable progress in visual fidelity and text–image alignment (Rombach et al., 2022; Podell et al., 2023; Esser et al., 2024). Conventional multi-step diffusion models can generate high-quality and diverse images, but require long iterative denoising trajectories, resulting in substantial inference cost (Ho et al., 2020; Rombach et al., 2022; Nguyen & Tran, 2023). To reduce this cost, recent diffusion distillation methods compress the sampling process into only a few model evaluations while maintaining high generation quality (Luo et al., 2023; Sauer et al., 2023; Yin et al., 2024). However, in accelerated few-step generators, it is desirable to preserve or even surpass the diversity of the original models while improving generation efficiency. Yet, compressed sampling trajectories limit the ability of different noise initializations to evolve toward distinct visual solutions, often yielding similar compositions, viewpoints, subject placements, and global appearances across random seeds (Dao et al., 2024; Dahary et al., 2026; Gandikota & Bau, 2026).

Existing approaches include both training-based and training-free methods (Han et al., 2025; Wu et al., 2026; Liu et al., 2026). Although training-based methods achieve promising results, they require computationally expensive training. Training-free methods instead intervene at different stages of the generation pipeline. CADS (Sadat et al., 2024) perturbs the conditioning signal with scheduled Gaussian noise, but its perturbations can degrade visual fidelity and introduce artifacts on prompts (Fig. 14 (3rd row)). SGI (Parmar et al., 2025) generates an enlarged candidate pool and selects a diverse, high-quality subset, incurring considerable computational overhead (Tabs. 7 to 10). Particle Guidance (Corso et al., 2024) and SPARKE (Jalali et al., 2025) separate samples through repulsion during denoising, but strong latent-space repulsion may cause semantic drift by pushing trajectories away from the learned data distribution (Dahary et al., 2026). More recently, Contextual Repulsion (Dahary et al., 2026) operates on joint text–visual contextual features. However, perturbing these semantically sensitive features may impair prompt consistency, while its dependence on architecture-specific activations limits direct transfer across generator families (Fig. 15 (2th row)).

In this paper, we introduce Text Intervention for Diversified Exploration in Diffusion Models (TIDE),an optimization-based, training-free inference-time method. Operating as a plug-and-play module, it requires no modifications to pretrained weights integrates into both Diffusion Transformer (DiT) and U-Net-based pipelines. Specifically, we introduce Residual Spectral Steering (*RSS*), a condition-space module that first applies Singular Value Decomposition (SVD) to the text embedding and then steers its right-singular vectors, which represent directions in the text-embedding space (Jolliffe & Cadima, 2016). We propose a method to automatically split the right singular vectors into *principal* and *residual subspaces*. Instead of indiscriminately perturbing the complete embedding, RSS leaves the principal directions untouched and changes the selected residual directions with a controlled spherical rotation method, which leads to structured flexibility without reducing both semantic information and image quality. As shown in Fig. 3, principal directions primarily stabilize the core semantics and structural constraints of the given prompt, whereas the residual directions contain underused degrees of freedom that can support alternative visual realizations. RSS introduces structured variation while preserving the dominant spectral structure of the prompt embedding. However, it does not explicitly separate trajectories initialized from different random seeds. We therefore introduce Fine-grained Spatio-channel Diversification (FSD), which operates in the latent space. FSD jointly coordinates a group of trajectories generated from the same prompt. FSD uses complementary descriptors of coarse spatial layout, global appearance, and channel statistics to measure inter-sample similarity. Based on these measurements, it applies bounded repulsive updates to reduce redundancy across trajectories while limiting latent displacement. RSS and FSD therefore play complementary roles: RSS identifies where structured variation can be introduced into the conditioning space, while FSD encourages different seeds to realize this variation along distinct latent trajectories.Component analyses further examine their individual contributions and the diversity–quality trade-off achieved by their combination.

We evaluate TIDE on four state-of-the-art models: Z-Image-Turbo (Team, 2025), SDXL-Turbo (Sauer et al., 2023), FLUX.1-dev (Black Forest Labs, 2024), and FLUX.2-Klein-4B (Black Forest Labs, 2026), using prompts from the COCO 2017 validation captions (Lin et al., 2014). Quantitative and qualitative evaluations show increased within-prompt diversity across multiple generators while preserving semantic alignment and perceptual quality. Beyond text-to-image generation,

we further demonstrate the generality of TIDE by extending it to video generation and downstream 3D asset generation.

In summary, our main contributions are:

- We identify two complementary factors for improving within-prompt diversity: underexploited residual directions in the prompt representation and insufficient separation among trajectories initialized from different random seeds. Based on this observation, we jointly address diversity at the text embedding and latent levels.
- We propose Text Intervention for Diversified Exploration in Diffusion Models (TIDE), an optimization-based, training-free inference-time method for improving within-prompt diversity. TIDE combines RSS, which automatically identifies and steers residual prompt directions, with FSD, which performs bounded group-wise separation of same-prompt latent trajectories. It requires no modification to pretrained model weights, supports both DiT- and U-Net-based pipelines, and incurs negligible additional inference-time overhead.
- We validate the proposed framework across four state-of-the-art models through main comparisons, component ablations, parameter analyses, and comparisons with representative training-free diversification methods, jointly evaluating diversity, semantic alignment, and perceptual quality.

2 Related Work

Text-to-image generation and few-step diffusion. Text-to-image diffusion models encompass Diffusion Transformer (DiT)-based architectures with multimodal attention (Esser et al., 2024; Black Forest Labs, 2024) and U-Net-based architectures with cross-attention conditioning (Rombach et al., 2022; Podell et al., 2023). While multi-step diffusion models can generate high-quality and diverse images, their iterative sampling incurs substantial inference cost. Recent consistency modeling and diffusion distillation techniques reduce generation to only a few model evaluations (Luo et al., 2023; Sauer et al., 2023; Yin et al., 2024), but accelerated sampling often suffers from reduced within-prompt diversity (Gandikota & Bau, 2026).

Training-free diversity enhancement. Existing training-free methods improve diversity through different inference-time interventions. CADS perturbs the conditioning signal with scheduled Gaussian noise (Sadat et al., 2024), while CNO optimizes initial noise latents (Kim et al., 2025) and MinorityPrompt searches for low-density conditioning signals (Um & Ye, 2025). Particle Guidance and SPARKE introduce repulsion among samples during denoising (Corso et al., 2024; Jalali et al., 2025), whereas SGI generates an enlarged candidate pool and selects a diverse subset (Parmar et al., 2025). More recently, Contextual Repulsion operates on multimodal representations inside Diffusion Transformer blocks (Dahary et al., 2026). Despite these advances, existing methods compromise semantic alignment, increase inference cost, or rely on architecture-specific internal representations.

In this work, we systematically analyze the spectral structure of prompt embeddings and observe that different spectral subspaces play distinct roles in generation. Principal directions are critical for preserving dominant prompt semantics, whereas residual directions provide greater flexibility for alternative visual realizations. Based on this observation, we introduce RSS to selectively steer residual prompt directions while preserving the principal structure. We further introduce FSD to explicitly diversify same-prompt latent trajectories. Together, they improve within-prompt diversity without additional training.

3 Method

In this section, we first conduct a comprehensive analysis that shows the limited diversity could be relaxed by decomposing the text embeddings with SVD (see Sec. 3.1). This observation motivates us to formalize our approach to generative diversity, namely Residual Spectral Steering (RSS), by shifting the intervention focus to the text-embedding space (see Sec. 3.2). Furthermore, we propose Fine-grained Spatio-channel Diversification (FSD) which promotes separation among same-prompt latent trajectories during early denoising (Sec. 3.2). Fig. 4 summarizes the overall generation pipeline.

3.1 EMPIRICAL MOTIVATION

We motivate our method through two observations in the text-embedding space. First, lower-energy residual components provide degrees of freedom for increasing diversity while largely preserving the dominant prompt semantics. Second, the corresponding right-singular vectors have unit norm and lie on a unit hypersphere, naturally supporting spherical interpolation (Golub & Van Loan, 2013). More experimental details are provided in Appendix F.

Observation 1: Diversity is governed by the residual space. Let $\mathbf{E}_p \in \mathbb{R}^{N\times D}$ denote the prompt embedding, where N is the number of text tokens and D is the embedding dimension. Its thin Singular Value Decomposition (SVD) is

$$\mathbf{E}_p = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \tag{1}$$

where the rows of $\mathbf{V}^\top$ form orthonormal directions in the embedding space (Golub & Van Loan, 2013). We divide these directions into principal and residual subspaces using the Maximum Distance to Chord (MDC) criterion on the normalized singular-value spectrum (Douglas & Peucker, 1973):

$$\mathbf{V}^\top = \begin{bmatrix} \mathbf{V}_{\text{pr}}^\top \\ \mathbf{V}_{\text{res}}^\top \end{bmatrix}. \tag{2}$$

The MDC-selected elbow point partitions the right-singular directions: $\mathbf{V}_{\text{pr}}^\top$ contains the principal directions before the elbow point, while $\mathbf{V}_{\text{res}}^\top$ contains the residual directions from the elbow point onward.

To examine their roles, we independently replace the principal or residual directions with null-text reference directions during the first two denoising steps, while retaining $\mathbf{U}$ and $\mathbf{\Sigma}$. As shown in Fig. 2, principal-space replacement causes substantial semantic drift and structural degradation. For example, the generated subject for the "balloon elephant" prompt loses its characteristic balloon-like appearance. Correspondingly, the average CLIP score decreases from 28.50 to 23.01. In contrast, residual-space replacement preserves the core subject semantics while producing visible variation in backgrounds and textures. Its average CLIP score is 28.55, close to the vanilla score of 28.50, over 1,000 prompts. These results indicate that the principal subspace carries the dominant information required for prompt alignment, whereas the residual subspace offers greater flexibility for diversification with substantially less semantic disruption.

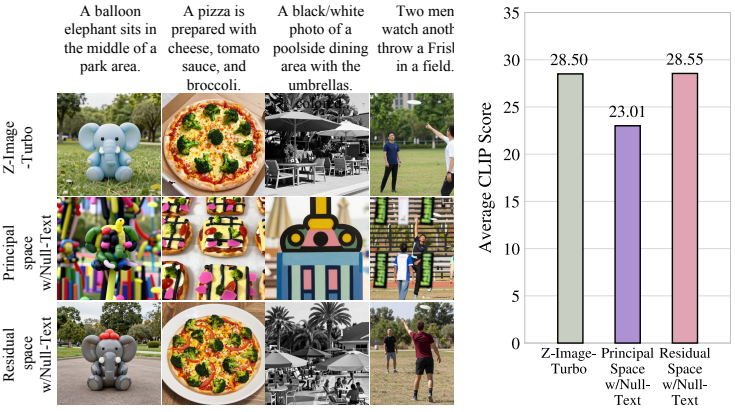


Figure 2: **Principal–residual replacement analysis.** Effects of null-text replacement on generation and prompt alignment.

Is the null-text direction meaningful? We compare Null-SLERP (RSS) with residual zeroing, random orthogonal steering, and residual shuffling to determine whether its benefit arises from the null-text direction rather than arbitrary residual perturbation. As shown in Tab. 1, both zeroing and shuffling increase diversity but substantially degrade semantic alignment and quality, while random orthogonal steering offers only limited gains. Null-SLERP instead improves DINO-div. from 0.165 to 0.175 and LPIPS from 0.526 to 0.582 while remaining close to the baseline in CLIPScore and ImageReward, demonstrating a more favorable diversity–quality trade-off.

Table 1: **Residual-direction intervention comparison.** Diversity–Quality trade-offs of alternative residual interventions.

Method	DINO↑	LPIPS↑	CLIP Score↑	Image Reward↑
Baseline	0.165	0.526	0.246	0.922
Null-SLERP	0.175	**0.582**	**0.244**	**0.917**
Random orthogonal	0.175	0.540	0.243	0.695
Shuffled residual	0.250	0.549	0.230	0.307
Residual zeroing	0.763	0.773	0.144	−1.933

Table 2: **Quantitative comparison of interpolation strategies.** LERP improves the diversity metrics but substantially degrades image quality and prompt alignment. In contrast, SLERP achieves a better diversity–quality trade-off.

Method	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑
Baseline	1.638	1.412	0.534	0.258	1.167
SLERP	**1.699**	**1.478**	**0.588**	0.255	**1.201**
Residual-LERP	1.690	1.422	0.583	0.232	0.068



Figure 3: **Effect of interpolation geometry on semantic steering.** SLERP produces a smoother and more coherent semantic transition, whereas LERP introduces visible artifacts.

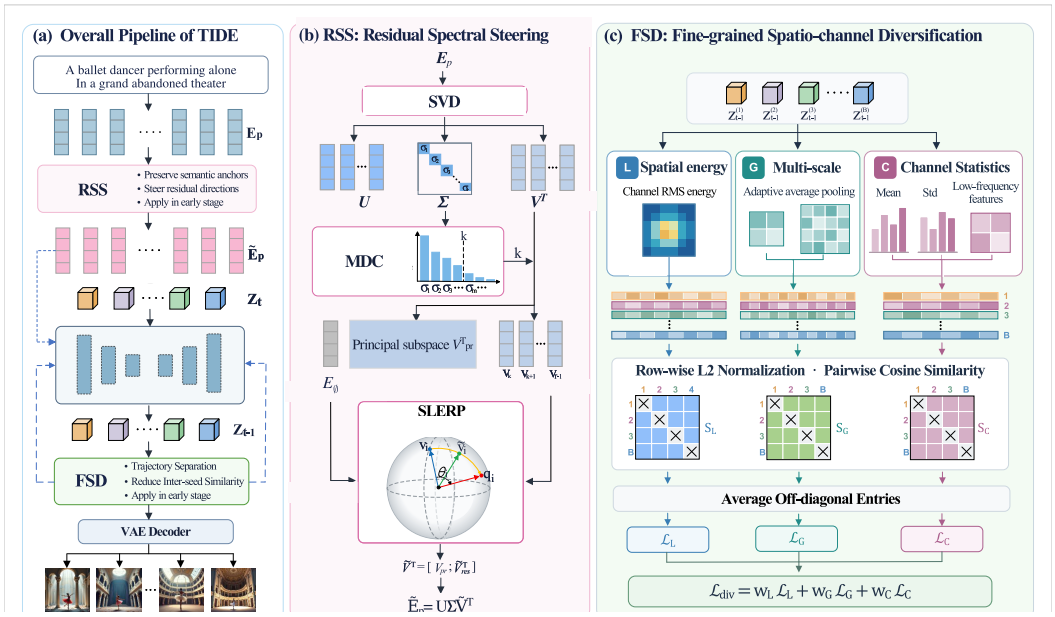


Figure 4: **Overview of TIDE. (a)** RSS intervenes in the prompt embedding before denoising, while FSD separates latent trajectories during early denoising. **(b)** RSS uses SVD and MDC to preserve principal right-singular directions and steer residual directions toward the null-text reference via SLERP. **(c)** FSD combines spatial-energy (L), multi-scale (G), and channel-statistics (C) objectives for latent diversification.

Observation 2: Spherical Interpolation Better Preserves Semantics. The rows of $\mathbf{V}^\top$ are orthonormal and therefore lie on the unit hypersphere (Golub & Van Loan, 2013), motivating the use of Spherical Linear Interpolation (SLERP) for steering residual right-singular directions. To examine the effect of interpolation geometry, we compare SLERP with linear interpolation (LERP) between the embeddings of "a photo of a smiling man" and "a photo of an angry man." As shown in Fig. 3, SLERP produces a smooth and visually coherent semantic transition, whereas LERP introduces visible artifacts.

Quantitatively, both strategies improve diversity over the baseline (Tab. 2). SLERP outperforms LERP across all three diversity metrics while maintaining CS close to the baseline and improving IR. LERP instead substantially degrades both alignment and image quality. We therefore adopt SLERP for residual-direction steering in RSS.

3.2 TIDE: Text Intervention for Diversified Exploration in Diffusion Models

Fig. 4(a) summarizes the TIDE pipeline. RSS spectrally steers the prompt embedding during the early denoising steps, while FSD separates the latent trajectories associated with different random seeds using spatial-energy, multi-scale, and channel-statistics descriptors. The final latents are independently decoded into a diverse image group.

RSS: Residual Spectral Steering. As shown in Fig. 4(b), RSS decomposes the text embedding as $\mathbf{E}_p = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and uses the MDC boundary to separate the principal and residual

right-singular directions in $\mathbf{V}^\top$. We preserve $\mathbf{U}$, $\mathbf{\Sigma}$, and the principal right-singular subspace $\mathbf{V}_{\text{pr}}^\top = [\mathbf{v}_0, \ldots, \mathbf{v}_{k-1}]^\top$, and steer only the residual directions $\mathbf{v}_k, \ldots, \mathbf{v}_{r-1}$ based on Observation 1 (see Fig. 2), indexed by $\mathcal{I}_{\text{res}} = \{k, k+1, \ldots, r-1\}$, where $r = \min(N, D)$ denotes the number of singular directions in the thin SVD.

Based on Observation 1(Tab. 1), we use the null-text embedding $\mathbf{E}_\emptyset \in \mathbb{R}^D$ as the reference. Following the shortest-arc convention of spherical linear interpolation (SLERP) (Shoemake, 1985), we choose the orientation of the null-text reference separately for each residual direction. Specifically, we select between $\mathbf{E}_\emptyset$ and $-\mathbf{E}_\emptyset$ such that the selected reference lies in the same hemisphere as $\boldsymbol{v}_i$:

$$\boldsymbol{q}_i = \begin{cases} \mathbf{E}_\emptyset, & \langle \boldsymbol{v}_i, \mathbf{E}_\emptyset \rangle \geq 0, \\ -\mathbf{E}_\emptyset, & \langle \boldsymbol{v}_i, \mathbf{E}_\emptyset \rangle < 0. \end{cases} \tag{3}$$

This ensures $\langle \boldsymbol{v}_i, \boldsymbol{q}_i \rangle \geq 0$ and therefore selects the shorter SLERP path, limiting the perturbation applied to the residual direction. To adapt the steering strength across the residual spectrum, we define

$$b_i = \frac{\sigma_i - \sigma_{r-1}}{\sigma_k - \sigma_{r-1} + \epsilon}, \qquad a_i = \max(a_{\min}, a \exp[\gamma(b_i - 1)]), \tag{4}$$

where a is the global steering strength, γ controls the decay, and $a_{\min}$ prevents negligible rotations. Unless otherwise specified, $\gamma = 2.3$ and $a_{\min} = 0.01$.

We then apply SLERP independently to each residual direction:

$$\widetilde{\boldsymbol{v}}_i = \frac{\sin((1-a_i)\theta_i)}{\sin\theta_i} \frac{\boldsymbol{v}_i}{\|\boldsymbol{v}_i\|_2} + \frac{\sin(a_i\theta_i)}{\sin\theta_i} \frac{\boldsymbol{q}_i}{\|\boldsymbol{q}_i\|_2}, \qquad \theta_i = \arccos \frac{\langle \boldsymbol{v}_i, \boldsymbol{q}_i \rangle}{\|\boldsymbol{v}_i\|_2 \|\boldsymbol{q}_i\|_2}. \tag{5}$$

We then combine the unchanged principal directions with the steered residual directions as $\widetilde{\mathbf{V}}^\top = [\mathbf{V}_{\text{pr}}^\top; \widetilde{\mathbf{V}}_{\text{res}}^\top]$ and reconstruct the steered prompt embedding as $\widetilde{\mathbf{E}}_p = \mathbf{U}\mathbf{\Sigma}\widetilde{\mathbf{V}}^\top$.

Thus, RSS preserves $\boldsymbol{U}$, $\mathbf{\Sigma}$, and all principal directions in $\mathbf{V}^\top$ and modifies only the residual directions. It is applied only during an early steps, since extending the intervention to later steps provides limited additional diversity (Appendix A).

FSD: Fine-grained Spatio-channel Diversification. RSS modifies the generator's conditioning but does not explicitly separate trajectories initialized from different random seeds. FSD complements it by reducing similarity among same-prompt latent trajectories during early denoising. As shown in Fig. 4(c), FSD constructs complementary descriptors from the latent group produced by an enabled denoising step and applies a bounded update before the next step.

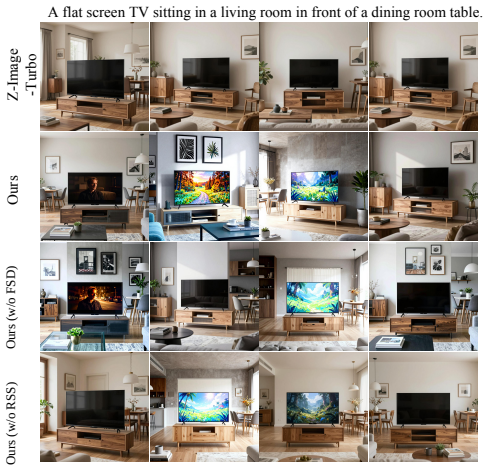


Figure 5: **Qualitative ablation on Z-Image-Turbo.** Rows denote different configurations.

Let $\mathbf{Z}_t = \{z_t^{(b)}\}_{b=1}^B \in \mathbb{R}^{B \times C_z \times H \times W}$ denote the latent group, where B is the number of different random seeds. For each latent, we construct three complementary descriptor families: a spatial-energy descriptor L from the downsampled channel-energy map; a multi-scale descriptor G from concatenated 2×2 and 4×4 average-pooled features; and a channel-statistics descriptor C from channel means, standard deviations, and 2×2 low-frequency features. Together, they capture spatial organization, multi-scale structure, and location-insensitive channel statistics. Let $\mathcal{M} = \{\text{L}, \text{G}, \text{C}\}$ denote the set of descriptor families. For each latent $z_t^{(b)}$, we compute all three descriptor vectors $\{f_m(z_t^{(b)})\}_{m \in \mathcal{M}}$, where f_m denotes the feature-extraction operation associated with descriptor family m. Specifically, f_{L} extracts the spatial-energy descriptor, f_{G} extracts the multi-scale descriptor, and f_{C} extracts the channel-statistics descriptor. For each descriptor family,

Table 4: **Quantitative comparison across four generators on COCO 2017.** Results are averaged over 1,000 four-image groups. VI, VC, and LPIPS measure diversity, while CS, IR, and VQA measure alignment and perceptual quality.

(a) Z-Image-Turbo (DiT)

Method	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑	VQA ↑
Baseline	1.524	1.340	0.535	0.258	1.011	0.906
CADS	2.645	2.099	0.799	0.189	−1.518	0.486
PG	1.524	1.340	0.535	0.258	1.012	0.906
SPARKE	1.554	1.376	0.581	0.259	1.051	0.908
SGI	1.544	1.346	0.548	0.259	0.995	0.906
CR	1.525	1.340	0.538	0.258	1.010	0.906
Ours	1.647	1.482	0.671	0.258	1.017	0.906

(b) SDXL-Turbo (U-Net)

Method	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑	VQA ↑
Baseline	1.675	1.369	0.524	0.259	0.539	0.859
CADS	1.961	1.543	0.558	0.260	0.447	0.855
PG	1.684	1.370	0.531	0.259	0.524	0.860
SPARKE	1.691	1.383	0.538	0.259	0.539	0.861
SGI	1.716	1.391	0.559	0.259	0.520	0.856
CR	1.682	1.370	0.537	0.259	0.534	0.860
Ours	1.690	1.376	0.538	0.259	0.541	0.859

(c) FLUX.1-dev (DiT)

Method	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑	VQA ↑
Baseline	1.761	1.509	0.618	0.639	0.955	0.882
CADS	2.566	2.544	0.711	0.424	−1.355	0.409
PG	1.760	1.510	0.618	0.639	0.955	0.882
SPARKE	1.772	1.520	0.655	0.639	0.995	0.882
SGI	1.762	1.518	0.619	0.639	0.952	0.884
CR	1.761	1.509	0.618	0.639	0.956	0.882
Ours	1.855	1.638	0.741	0.641	0.983	0.873

(d) FLUX.2-Klein-4B (DiT)

Method	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑	VQA ↑
Baseline	1.595	1.439	0.616	0.260	1.174	0.921
CADS	1.951	1.832	0.699	0.249	0.616	0.890
PG	1.596	1.440	0.616	0.260	1.172	0.920
SPARKE	1.623	1.486	0.664	0.259	1.175	0.921
SGI	1.618	1.467	0.640	0.260	1.157	0.920
CR	1.595	1.440	0.615	0.259	1.163	0.919
Ours	1.667	1.573	0.706	0.260	1.175	0.920

we measure group similarity using the average off-diagonal cosine similarity:

$$\mathcal{L}_m(\mathbf{Z}_t) = \frac{1}{B(B-1)} \sum_{b \neq b'} \frac{\left\langle f_m(z_t^{(b)}), f_m(z_t^{(b')}) \right\rangle}{\left\| f_m(z_t^{(b)}) \right\|_2 \left\| f_m(z_t^{(b')}) \right\|_2}, \qquad m \in \mathcal{M}. \tag{6}$$

The descriptor losses are combined as

$$\mathcal{L}_{\text{div}} = w_{\text{L}} \mathcal{L}_{\text{L}} + w_{\text{G}} \mathcal{L}_{\text{G}} + w_{\text{C}} \mathcal{L}_{\text{C}}, \tag{7}$$

where $\mathcal{L}_{\text{L}}$, $\mathcal{L}_{\text{G}}$, and $\mathcal{L}_{\text{C}}$ denote the average cosine-similarity losses computed using the spatial-energy, multi-scale, and channel-statistics descriptors, respectively, as defined in Equation (6). The weights w_{L}, w_{G}, and w_{C} control their respective contributions. Minimizing this objective separates the latents under multiple complementary views of the evolving latent representation.

Table 3: **Comparison on complex prompts.** Results on PartiPrompts using Z-Image-Turbo.

Method	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑	VQA ↑
Baseline	1.512	1.258	0.414	0.323	0.913	0.862
CADS	2.522	2.074	0.719	0.275	−0.738	0.775
PG	1.517	1.242	0.414	0.323	0.906	0.865
SPARKE	1.590	1.268	0.454	0.322	0.905	0.844
SGI	1.531	1.262	0.432	0.321	0.940	0.865
CR	1.517	1.250	0.414	0.322	0.903	0.862
Ours	1.618	1.347	0.509	0.328	0.920	0.861

Each runtime batch contains latents generated from the same prompt using different random seeds and is treated as a single group. Therefore, the objective directly measures and reduces within-prompt latent similarity. Additional implementation details are provided in Appendix F.4.

4 Experiments

4.1 Experimental Setups

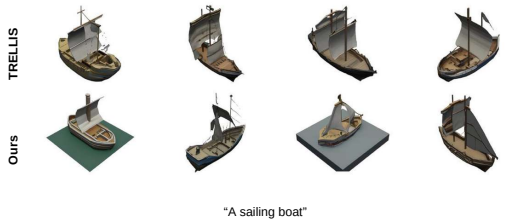
Methods and benchmarks. We evaluate TIDE on four U-Net- and Transformer-based generators: Z-Image-Turbo, SDXL-Turbo, FLUX.1-dev, and FLUX.2-Klein-4B. Baselines include the frozen

generator, CADS (Sadat et al., 2024), Particle Guidance (PG) (Corso et al., 2024), SPARKE (Jalali et al., 2025), SGI (Parmar et al., 2025), and Contextual Repulsion (CR) (Dahary et al., 2026).

We evaluate 1,000 prompts sampled from the COCO 2017 validation captions, generating four images per prompt and method. To verify that TIDE remains effective on compositionally complex prompts, we further evaluate it on PartiPrompts (Yu et al., 2022) using Z-Image-Turbo. The results are reported in Tab. 3 and Fig. 16. Further experimental details are provided in Appendix E.



(a) Text-to-video comparison on LTX-2.3.



(b) Text-to-3D comparison on TRELLIS.

Figure 6: **Extension to video and 3D generation.**

Metrics. Within-prompt diversity is measured using Vendi-Inception (VI) and Vendi-CLIP (VC) (Friedman & Dieng, 2023), and LPIPS (Zhang et al., 2018). Semantic alignment and perceptual quality are evaluated using CLIPScore (CS) (Hessel et al., 2021), ImageReward (IR) (Xu et al., 2023), and VQAScore (VQA) (Lin et al., 2024). Higher values indicate better performance for all metrics.

4.2 Experimental Results

Qualitative Results. As shown in Fig. 7, frozen generators often produce similar compositions, viewpoints, subject placements, and global appearances across seeds. CADS introduces stronger variation but can reduce prompt consistency and visual fidelity, whereas groupwise guidance methods achieve varying degrees of trajectory separation. TIDE yields visible differences in poses, layouts, backgrounds, and appearance while generally preserving the core subject and prompt semantics. These qualitative trends are observed across the evaluated U-Net- and Transformer-based generators and are consistent with the quantitative diversity–quality trade-off. Additional comparisons are provided in the Appendix (Figs. 9 and 10).

Quantitative comparison. As shown in Tab. 4, TIDE consistently improves VI, VC, and LPIPS over the frozen baselines across all four generators, with larger gains on Z-Image-Turbo, FLUX.1-dev, and FLUX.2-Klein-4B. Unlike high-diversity CADS settings, these improvements preserve alignment and perceptual quality. For example, on Z-Image-Turbo, CADS raises VI to 2.645 but reduces CS, IR, and VQA to 0.189, -1.518, and 0.486, respectively. TIDE instead achieves VI, VC, and LPIPS of 1.647, 1.482, and 0.671 while retaining near-baseline quality, yielding a more balanced diversity–quality trade-off.

Ablation study. As shown in Tab. 5, RSS and FSD independently improve within-prompt diversity over the baseline, confirming the contributions of both condition-space steering and latent-trajectory separation. Their combination achieves the highest VI, VC, and LPIPS while keeping CS and IR close to the baseline and slightly improving VQA. The qualitative results in Fig. 5 further illustrate their complementary roles: RSS introduces condition-side variation (fourth row), FSD promotes variation across same-prompt trajectories (third row), and their combination produces the strongest overall diversification (second row).

Table 5: **Ablation of RSS and FSD on Z-Image-Turbo.** VI, VC, and LPIPS measure diversity; CS, IR, and VQA measure alignment and perceptual quality.

Configuration	VI ↑	VC ↑	LPIPS ↑	CS ↑	IR ↑	VQA ↑
Baseline	1.524	1.340	0.535	0.258	1.011	0.906
RSS only	1.630	1.478	0.637	0.257	1.007	0.905
FSD only	1.587	1.438	0.600	0.258	1.000	0.901
Combined	**1.647**	**1.482**	**0.671**	0.258	**1.017**	**0.906**

We further ablate the spatial-energy (L), multi-scale (G), and channel-statistics (C) descriptors used by FSD, with the corresponding quantitative and qualitative results reported in Appendix C. In addition, a user study across four state-of-the-art models shows that participants consistently prefer our



Figure 7: **Qualitative comparison across prompts and text-to-image generators.** Each row presents four samples generated by the corresponding method under the same prompt.

results over the corresponding baseline generations; complete protocols and results are provided in Tab. 6.

Extensions to video and 3D generation. We apply our method to LTX-2.3 (HaCohen et al., 2026) and use images sampled from different seeds to condition TRELLIS (Xiang et al., 2025), while keeping both pipelines unchanged. As shown in Figs. 6a and 6b, our method produces more varied video appearances and 3D shapes, structures, and viewpoints, demonstrating transfer beyond image generation.Additional comparisons are provided in the Appendix Fig. 12 and Fig. 13

5 Conclusion

We introduce TIDE, an optimization-based, training-free inference-time method that improves diversity across four text-to-image generators while largely preserving generation quality. Future work will examine distributional fidelity and computational efficiency.

AI use statement

Generative AI tools were used to assist with manuscript polishing and code debugging. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work.

Ethics statement

This work does not raise significant ethical concerns.

References

Black Forest Labs. FLUX. https://github.com/black-forest-labs/flux, 2024. Official inference repository.

Black Forest Labs. FLUX.2 [klein]: Towards interactive visual intelligence. https://bfl.ai/blog/flux2-klein-towards-interactive-visual-intelligence, 2026. Accessed: 2026-09-23.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Gabriele Corso, Yilun Xu, Valentin de Bortoli, Regina Barzilay, and Tommi S. Jaakkola. Particle guidance: non-i.i.d. diverse sampling with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KqbCvIFBY7.

Omer Dahary, Benaya Koren, Daniel Garibi, and Daniel Cohen-Or. On-the-fly repulsion in the contextual space for rich diversity in diffusion transformers. *arXiv preprint arXiv:2603.28762*, 2026.

Trung Dao, Thuan Hoang Nguyen, Thanh Le, Duc Vu, Khoi Nguyen, Cuong Pham, and Anh Tran. Swiftbrush v2: Make your one-step diffusion model better than its teacher. *arXiv preprint arXiv:2408.14176*, 2024.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

David H. Douglas and Thomas K. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica*, 10(2):112–122, 1973.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 4 edition, 2013.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026. doi: 10.48550/arXiv.2601.03233. URL https://arxiv.org/abs/2601.03233.

Jiyeon Han, Dahee Kwon, Gayoung Lee, Junho Kim, and Jaesik Choi. Enhancing creative generation on stable diffusion-based models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28609–28618, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020.

Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. SPARKE: Scalable prompt-aware diversity and novelty guidance in diffusion models via RKE score. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.10173.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Byungjun Kim, Soobin Um, and Jong Chul Ye. Diverse text-to-image generation via contrastive noise optimization. *arXiv preprint arXiv:2510.03813*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755, 2014.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, 2024.

Yuxi Liu, Haoyu Li, Yixiang Cai, Tengxu Sun, Zekun Zhang, Baole Ai, Ang Wang, Jiamang Wang, Lin Qu, Kun Yuan, et al. Crossdistill: Balancing quality and diversity via trajectory-level hybrid few-step distillation. *arXiv preprint arXiv:2609.14725*, 2026.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Thuan Hoang Nguyen and Anh Tran. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. *arXiv preprint arXiv:2312.05239*, 2023.

Gaurav Parmar, Or Patashnik, Daniil Ostashev, Kuan-Chieh Wang, Kfir Aberman, Srinivasa Narasimhan, and Jun-Yan Zhu. Scaling group inference for diverse and high-quality generation. *arXiv preprint arXiv:2508.15773*, 2025. URL https://arxiv.org/abs/2508.15773.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 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.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zMoNrajk2X.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Ken Shoemake. Animating rotation with quaternion curves. *ACM SIGGRAPH Computer Graphics*, 19(3):245–254, 1985. doi: 10.1145/325165.325242.

Z-Image Team. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Soobin Um and Jong Chul Ye. Minority-focused text-to-image generation via prompt optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20926–20936, 2025.

Tianhe Wu, Ruibin Li, Lei Zhang, and Kede Ma. Diversity-preserved distribution matching distillation for fast visual synthesis. *arXiv preprint arXiv:2602.03139*, 2026.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, June 2025. doi: 10.1109/CVPR52734.2025.02000.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research*, 2022.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

APPENDIX

A ANALYSIS OF THE EARLY INTERVENTION WINDOW

We further investigate when the text embedding has the strongest influence on the generation trajectory.We use Z-Image-Turbo as the representative accelerated generator. The original prompt is replaced by a null-text prompt at different denoising steps. All other generation settings remain unchanged.We first apply the null-text condition at a single denoising step. As shown in Fig. 8, interventions during the early steps produce clear changes in the generated structure. The effect becomes substantially weaker at later steps. We compare each modified generation with its vanilla counterpart using LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2020). Both metrics change rapidly during the early stage and then become stable. This indicates that the generation trajectory is more structurally responsive to conditioning changes during early denoising.

We also evaluate semantic preservation using CLIP (Hessel et al., 2021) and DINO (Caron et al., 2021). The two metrics exhibit a similar stabilization trend after the early steps. This suggests that later conditioning changes have only a limited effect on the final generation.

We next extend the null-text intervention over multiple denoising steps. For a selected step, null-text conditioning is applied at that step and all preceding steps. As the intervention window grows, the generated structure changes more substantially. However, prolonged null-text conditioning also leads to increasing semantic degradation. The quantitative results in Fig. 8 show the same trend.These results reveal a useful but fragile intervention window. The early denoising stage remains flexible enough to alter the emerging generation structure. However, directly modifying the complete text condition for an extended period can damage prompt semantics. This motivates us to restrict RSS to an early conditioning window.

B RUNTIME AND COMPUTATIONAL OVERHEAD

Evaluation protocol. We evaluate the runtime overhead of TIDE on Z-Image-Turbo, SDXL-Turbo, FLUX.1-dev, and FLUX.2-Klein-4B. For each generator, all methods use the same model checkpoint, image resolution, denoising budget, prompt, seed group, batch size, and hardware environment. Each invocation generates a group of four images. We report the average wall-clock time for generating the complete group, together with the amortized time per image. For SGI, the reported runtime includes both candidate generation and subset selection, while the final output group contains four images.

Runtime results. Detailed results are reported in Tabs. 7 to 10. Across all four generators, TIDE largely preserves the inference efficiency of the corresponding vanilla pipeline. The measured runtime overhead is 2.40% on Z-Image-Turbo, 2.78% on SDXL-Turbo, 0.09% on FLUX.1-dev, and 3.90% on FLUX.2-Klein-4B. Thus, the complete framework adds less than 4% runtime overhead on every evaluated generator.

The additional computation includes both components of our framework. RSS performs spectral decomposition and residual-direction steering. FSD performs bounded latent-space updates during selected early steps. Neither component introduces additional denoising-model evaluations. Therefore, the number of backbone model calls remains identical to the vanilla pipeline.

SGI follows a different inference strategy. It generates an enlarged candidate pool before selecting the final output group. This results in substantially higher inference cost under the evaluated settings. Contextual Repulsion also incurs higher runtime than our method in the configurations where it is evaluated. Overall, TIDE improves generation diversity while largely preserving the inference efficiency of the underlying accelerated generator.

C ABLATION OF FSD DESCRIPTOR FAMILIES

We further examine the contributions of the spatial-energy (L), multi-scale (G), and channel-statistics (C) descriptors in FSD. As shown in Fig. 18, each individual descriptor improves within-prompt diversity over the baseline. The L-only configuration provides moderate diversity gains and

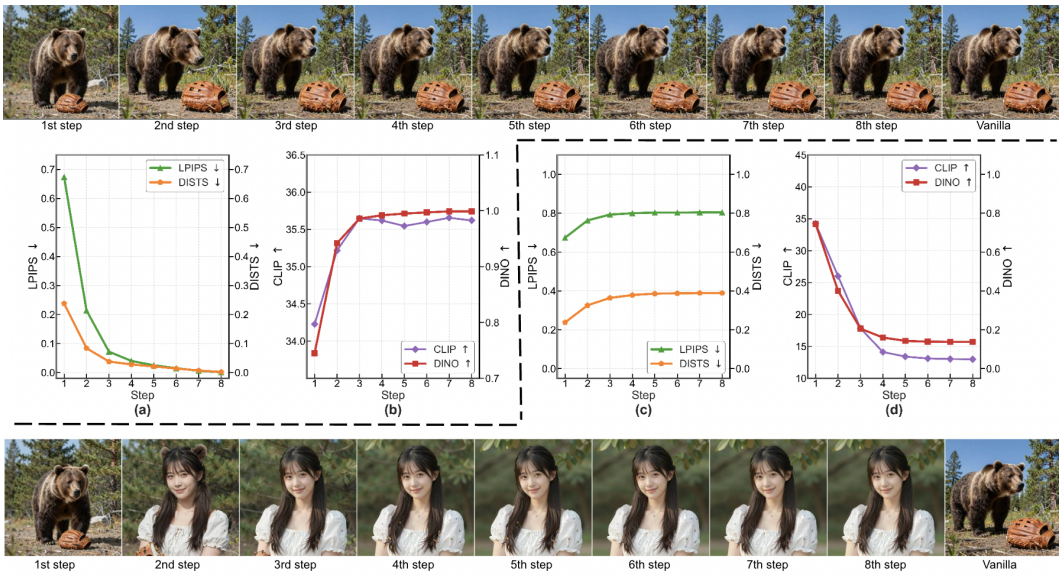


Figure 8: **Effect of the RSS intervention window.** We progressively extend spectral steering to later denoising steps and evaluate the resulting generations. The qualitative results and quantitative metrics show that the largest changes occur during the early denoising stage, while extending the intervention to later steps yields diminishing changes. These results motivate restricting RSS to an early conditioning window.

achieves the highest VQAScore. The G-only configuration substantially improves VI and LPIPS while achieving the highest ImageReward. The C-only configuration obtains the highest LPIPS and CLIPScore, but its ImageReward decreases markedly.

Combining the three descriptors achieves the highest VI and VC, while maintaining strong LPIPS, ImageReward, and VQAScore. The qualitative results in Fig. 18 further illustrate the variations produced by the individual descriptors and their combination. Overall, $L+G+C$ provides the most balanced performance across diversity, semantic alignment, and perceptual-quality metrics, and is therefore used as the default FSD configuration.

D Evaluation on Compositionally Complex Prompts

We further evaluate the compared methods on compositionally complex prompts selected from PartiPrompts. As shown in Tab. 3, CADS achieves the largest increases in VI, VC, and LPIPS, but these gains are accompanied by substantial decreases in CLIPScore, ImageReward, and VQAScore. Excluding this aggressive operating point, our method achieves the highest VI, VC, and LPIPS while also improving CLIPScore and ImageReward over the baseline. Although its VQAScore decreases, the overall results indicate a more balanced trade-off between diversity and generation quality. The qualitative comparison in Fig. 16 further illustrates the outputs produced by the different methods under the same compositionally complex prompt.

E Additional Experimental Details

Evaluation protocol. For the main evaluation, we use a fixed set of 1,000 prompts sampled from the COCO 2017 validation captions. Each method generates four images per prompt. Trajectory-based methods use identical prompt–seed pairs to ensure a controlled comparison. SGI instead generates an enlarged candidate pool and selects four images as its final output group. Each prompt and its four-image group constitute one evaluation unit, resulting in 1,000 groups and 4,000 retained images for each model–method configuration.

For the PartiPrompts evaluation, we select compositionally complex prompts containing multiple objects, attributes, and relations. All methods use identical prompts, random seeds, and generation settings.

Implementation settings. All experiments use the original pretrained model weights, schedulers, image resolutions, and numbers of denoising-model evaluations. Compared methods are implemented using their official configurations and evaluated under the same generation budget whenever applicable. TIDE introduces no additional denoising-model evaluations. Detailed method-specific operating points are reported alongside the corresponding experiments.

Metric aggregation. VI, VC, and LPIPS are computed within each four-image group to measure within-prompt diversity. CLIPScore, ImageReward, and VQAScore are first averaged within each group and then across all prompts. Since metric scales vary across generator families, comparisons are made within each model rather than directly across models.

F Additional Details for Empirical Motivation

F.1 Spectral Decomposition and Subspace Partition

Let $\mathbf{E}_p \in \mathbb{R}^{N\times D}$ denote the prompt embedding, where N is the number of text tokens and D is the embedding-channel dimension. Its thin singular value decomposition is

$$\mathbf{E}_p = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{8}$$

where

$$\mathbf{U} \in \mathbb{R}^{N\times r}, \qquad \mathbf{\Sigma} = \mathrm{diag}(\sigma_0, \ldots, \sigma_{r-1}) \in \mathbb{R}^{r\times r}, \qquad \mathbf{V}^\top \in \mathbb{R}^{r\times D}, \tag{9}$$

and $r = \min(N, D)$. The singular values are ordered as $\sigma_0 \geq \cdots \geq \sigma_{r-1} \geq 0$.

Each row $\boldsymbol{v}_i^\top \in \mathbb{R}^{1\times D}$ of $\mathbf{V}^\top$ defines an orthonormal direction in the embedding-channel space (Golub & Van Loan, 2013). Modifying $\boldsymbol{v}_i$ changes the direction of the corresponding spectral component, whereas retaining $\mathbf{U}$ and $\mathbf{\Sigma}$ preserves its token-side mixing coefficients and spectral magnitude.

To separate dominant and lower-energy components, we min–max normalize the singular-value spectrum and apply the Maximum Distance to Chord (MDC) criterion (Douglas & Peucker, 1973). MDC selects the zero-based index k whose point has the largest perpendicular distance to the chord connecting the first and last points of the normalized spectrum. We treat k as the index of the first residual direction:

$$\mathbf{V}^\top = \begin{bmatrix} \mathbf{V}_{\mathrm{pr}}^\top \\ \mathbf{V}_{\mathrm{res}}^\top \end{bmatrix}, \qquad \mathbf{V}_{\mathrm{pr}}^\top = [\boldsymbol{v}_0, \ldots, \boldsymbol{v}_{k-1}]^\top, \qquad \mathbf{V}_{\mathrm{res}}^\top = [\boldsymbol{v}_k, \ldots, \boldsymbol{v}_{r-1}]^\top. \tag{10}$$

F.2 Principal- and Residual-Space Replacement

Let $\mathbf{R}_\emptyset^{\mathrm{pr}}$ and $\mathbf{R}_\emptyset^{\mathrm{res}}$ denote null-text reference directions matched to the numbers of principal and residual rows, respectively. To isolate the role of each subspace, we apply the replacement only during the first two denoising calls and retain $\mathbf{U}$ and $\mathbf{\Sigma}$ from the original prompt embedding.

Principal-space replacement is defined as

$$(\mathbf{V}^{\mathrm{rpr}})^\top = \begin{bmatrix} \mathbf{R}_\emptyset^{\mathrm{pr}} \\ \mathbf{V}_{\mathrm{res}}^\top \end{bmatrix}, \qquad \mathbf{E}_p^{\mathrm{rpr}} = \mathbf{U}\mathbf{\Sigma}\,(\mathbf{V}^{\mathrm{rpr}})^\top. \tag{11}$$

Residual-space replacement is defined as

$$(\mathbf{V}^{\mathrm{rres}})^\top = \begin{bmatrix} \mathbf{V}_{\mathrm{pr}}^\top \\ \mathbf{R}_\emptyset^{\mathrm{res}} \end{bmatrix}, \qquad \mathbf{E}_p^{\mathrm{rres}} = \mathbf{U}\mathbf{\Sigma}\,(\mathbf{V}^{\mathrm{rres}})^\top. \tag{12}$$

Because both the prompt right-singular vectors and the null-text reference directions lie in the same D-dimensional embedding-channel space, the latter can be used to replace or steer individual spectral directions. We additionally repeat the replacement experiment using Gaussian reference directions; the results are shown in Fig. 17.

F.3 Residual-Direction Controls

We compare the following residual-space interventions:

- **Baseline:** no modification is applied.
- **Null-SLERP (RSS):** residual directions are spherically steered toward the null-text reference directions.
- **Residual zeroing:** residual directions are set directly to zero.
- **Random orthogonal:** residual directions are steered toward randomly sampled orthogonal directions.
- **Shuffled residual:** each residual direction is steered toward another randomly shuffled residual direction.

DINO-div. and LPIPS measure within-prompt visual diversity, whereas CLIPScore and ImageReward evaluate semantic alignment and perceptual quality. All configurations use the same prompts, random seeds, and generation settings.

F.4 FSD Update and Moment Restoration

Given the diversity objective in Equation (7), we compute an RMS-normalized, element-wise bounded latent update:

$$\Delta \mathbf{Z}_t = -\eta \, \mathrm{clip}\left(\frac{\nabla_{\mathbf{Z}_t} \mathcal{L}_{\mathrm{div}}(\mathbf{Z}_t)}{\mathrm{RMS}(\nabla_{\mathbf{Z}_t} \mathcal{L}_{\mathrm{div}}(\mathbf{Z}_t))}, -c, c \right), \qquad \widehat{\mathbf{Z}}_t = \mathbf{Z}_t + \Delta \mathbf{Z}_t. \tag{13}$$

Here, η controls the update magnitude and c bounds each normalized gradient element.

To preserve the global latent statistics, we restore the original mean and standard deviation of each latent:

$$\widetilde{\mathbf{Z}}_t^{(b)} = \frac{\widehat{\mathbf{Z}}_t^{(b)} - \mu\left(\widehat{\mathbf{Z}}_t^{(b)}\right)}{\sigma\left(\widehat{\mathbf{Z}}_t^{(b)}\right)} \sigma\left(\mathbf{Z}_t^{(b)}\right) + \mu\left(\mathbf{Z}_t^{(b)}\right), \tag{14}$$

where $\mu(\cdot)$ and $\sigma(\cdot)$ denote the global mean and standard deviation, respectively.

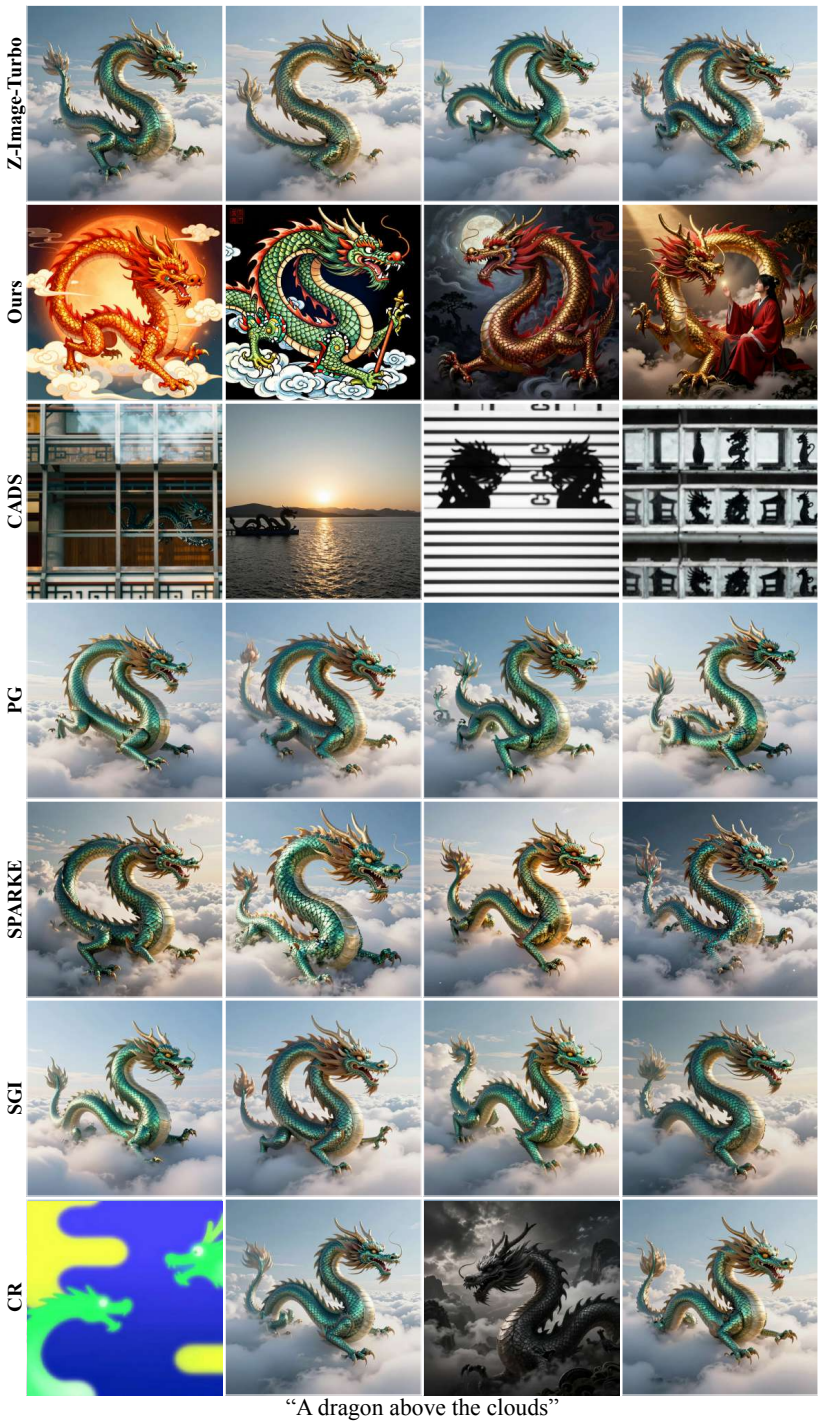
G Additional Experimental Results

G.1 User Study

We recruited 23 participants to compare our method with the corresponding baseline across four text-to-image generators. We used 24 prompts in total, with six prompts assigned to each generator. Every participant evaluated all 24 comparisons, resulting in 552 preference judgments Tab. 6. For each prompt, participants were shown two anonymized sets of generated images, one produced by the baseline and the other by our method. The left–right order of the two sets was randomized independently to reduce positional bias. Participants were asked to select the set they preferred overall, jointly considering within-prompt diversity and generation quality. The method identities were not revealed during evaluation.

Table 6: **User preference study across four text-to-image generators.** Preference rates for our method and the corresponding baselines.

Model	Baseline (%)	Ours (%)
Z-Image-Turbo	5.80	**94.20**
FLUX.2-Klein-4B	15.94	**84.06**
FLUX.1-dev	5.80	**94.20**
SDXL-Turbo	7.25	**92.75**
Overall	**8.70**	**91.30**

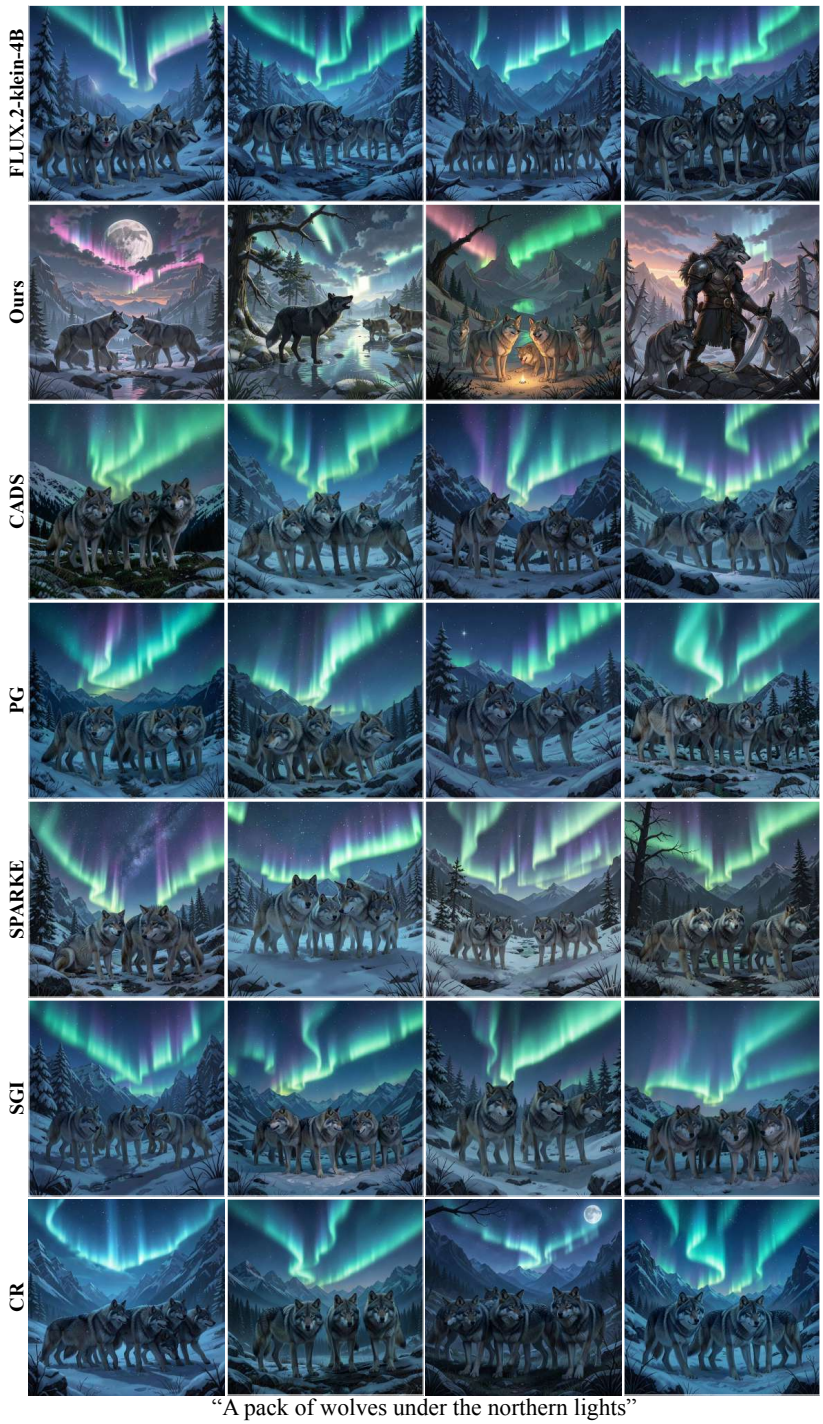


Qualitative comparison on Z-Image-Turbo. Each row shows four samples generated from the same prompt using the indicated method.



"A scientist conducting an experiment in a high-tech research laboratory"

Qualitative comparison on FLUX.1-dev. Each row shows four samples generated from the same prompt using the indicated method.



"A pack of wolves under the northern lights"

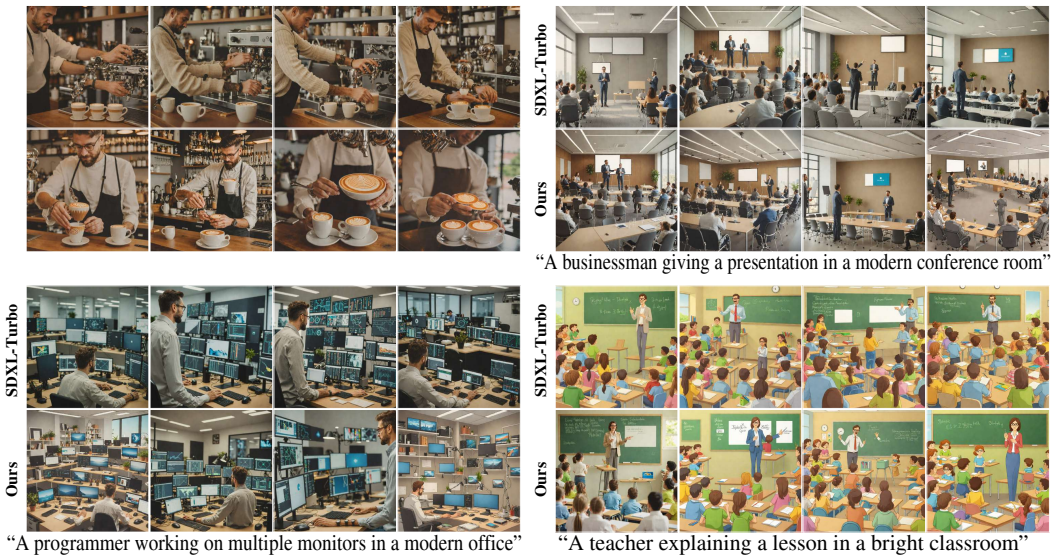
Qualitative comparison on FLUX.2-Klein. Each row shows four samples generated from the same prompt using the indicated method.



Figure 9: **Qualitative comparisons across four text-to-image generators.** Each row shows four samples generated from the same prompt using the indicated method. The complete comparison includes results on Z-Image-Turbo, SDXL-Turbo, FLUX.1-dev, and FLUX.2-Klein. CR denotes Contextual Repulsion.



Additional qualitative comparisons on Z-Image-Turbo.



Additional qualitative comparisons on SDXL-Turbo.



Additional qualitative comparisons on FLUX.1-dev.



Figure 11: **Additional qualitative comparisons across four text-to-image generators.** Results are presented for Z-Image-Turbo, SDXL-Turbo, FLUX.1-dev, and FLUX.2-Klein. Prompts and compared configurations are indicated within the panels.



Figure 12: **Extension to text-to-3D generation.** Qualitative comparisons on TRELLIS for the prompts "A small robot" and "A ceramic teapot." Each row contains four generated 3D assets.

Table 7: **Computational overhead comparison on Z-Image-Turbo.** Runtime is measured for generating a group of four images. Relative overhead is computed with respect to the vanilla baseline.CR denotes Contextual Repulsion

Method	Time / 4 images (s) ↓	Std. (s) ↓	Time / image (s) ↓
Baseline	9.5232	0.0108	2.3808
CADS	9.6349	0.0133	2.4087
PG	9.5375	0.0054	2.3844
SPARKE	9.6462	0.0140	2.4116
Ours	9.7521	0.0184	2.4380
CR	10.1309	0.0230	2.5327
SGI	11.0849	0.0032	2.7712

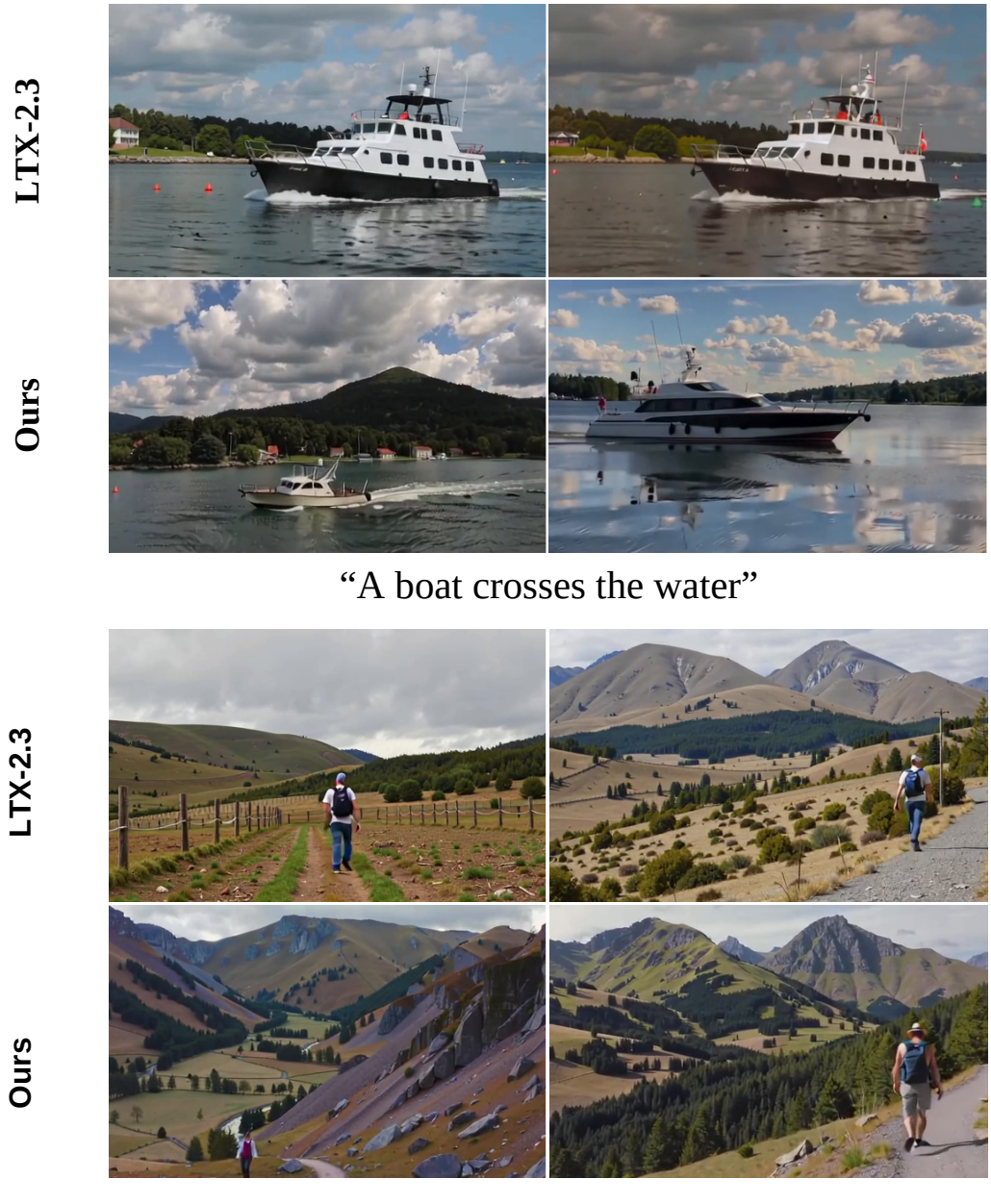


Figure 13: **Extension to text-to-video generation.** Qualitative comparison on LTX-2.3 under the prompt "A boat crosses the water.","A traveler walks through a quiet landscape." The top row shows results from the original model, while the bottom row shows results from our method.

Table 8: **Computational overhead comparison on SDXL-Turbo.** Runtime is measured for generating a group of four images. Results are measured at 512×512 resolution with 4 denoising steps.

Method	Time / 4 images (s) ↓	Std. (s) ↓	Time / image (s) ↓
Baseline	0.9728	0.0022	0.2432
CADS	0.9799	0.0032	0.2450
PG	0.9780	0.0014	0.2445
SPARKE	0.9924	0.0040	0.2481
Ours	0.9998	0.0022	0.2499
CR	1.4603	0.0181	0.3651
SGI	8.0319	0.0170	2.0080

Table 9: **Computational overhead comparison on FLUX.1-dev.** Runtime is measured for generating a group of four images. Results are measured at 512×512 resolution with 20 denoising steps.

Method	Time / 4 images (s) ↓	Std. (s) ↓	Time / image (s) ↓
Baseline	118.8695	0.1061	29.7174
CADS	118.8628	0.0366	29.7157
PG	118.8563	0.0428	29.7141
SPARKE	118.8343	0.0627	29.7086
Ours	118.9816	0.1126	29.7454
CR	656.8852	0.7500	164.2213
SGI	184.7037	0.2222	46.1759

Table 10: **Computational overhead comparison on FLUX.2-Klein-4B.** Runtime is measured for generating a group of four images. Results are measured at 512×512 resolution with 4 denoising steps.

Method	Time / 4 images (s) ↓	Std. (s) ↓	Time / image (s) ↓
Baseline	3.2371	0.0084	0.8093
CADS	3.2241	0.0022	0.8060
PG	3.2830	0.0036	0.8208
SPARKE	3.3642	0.0019	0.8410
Ours	3.3632	0.0048	0.8408
CR	8.4257	0.0386	2.1064
SGI	10.6326	0.3415	2.6582

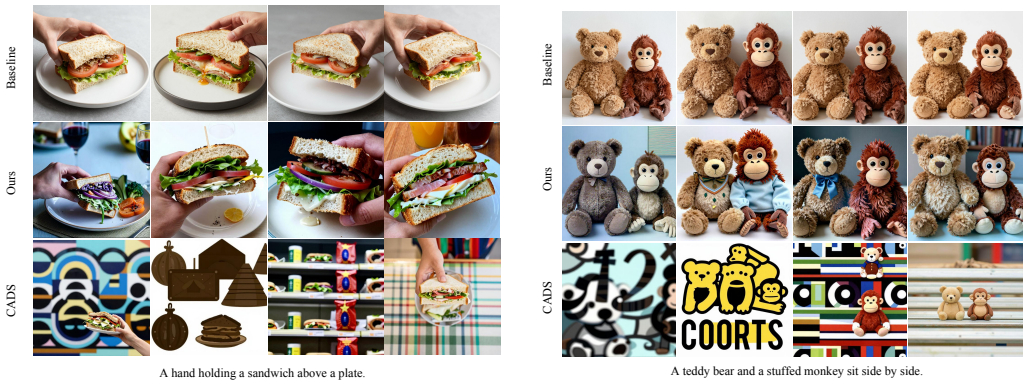


Figure 14: **Failure cases of condition-level perturbation.** Representative generations produced by CADS when condition perturbation is applied to text-prompt-conditioned image generators. Strong perturbations may weaken prompt consistency and introduce visible visual artifacts.



Figure 15: **Architecture dependence of contextual-space intervention.** Representative results illustrating the limited direct applicability of architecture-specific contextual intervention to U-Net-based generators.

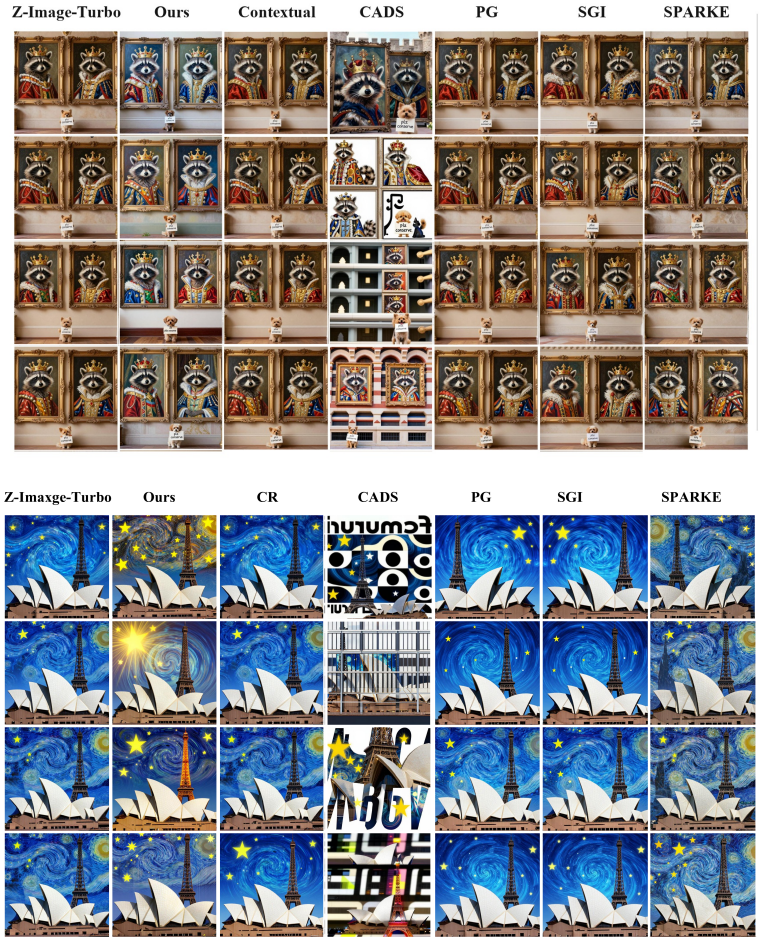


Figure 16: **Qualitative comparisons on compositionally complex prompts selected from PartiPrompts using Z-Image-Turbo.** Each visualization corresponds to a different prompt, and each column presents four samples generated by the indicated diversity-enhancement method. All methods use identical random seeds and generation settings within each comparison.



Figure 17: **Principal–residual replacement study using Gaussian noise on Z-Image-Turbo.** We repeat the replacement experiment using Gaussian noise instead of the null-text embedding. The first row shows vanilla samples, the second row shows samples generated after replacing the principal subspace with Gaussian noise, and the third row shows samples generated after replacing the residual subspace with Gaussian noise.



(a) Qualitative descriptor ablation

Configuration	VI ↑	VC ↑	LPIPS ↑	CLIPScore ↑	ImageReward ↑	VQAScore ↑
Baseline	1.537	1.346	0.526	0.246	0.952	0.886
L-only	1.572	1.410	0.546	0.247	0.951	**0.892**
G-only	1.658	1.477	0.664	0.247	**0.997**	0.882
C-only	1.647	1.486	0.668	**0.248**	0.814	0.880
$L + G + C$	**1.663**	**1.495**	**0.674**	0.247	0.989	0.890

(b) Quantitative descriptor ablation

Figure 18: **Qualitative and quantitative ablation of the spatial-energy (L), multi-scale (G), and channel-statistics (C) descriptors on Z-Image-Turbo. (a)** Four samples generated for the same prompt using the baseline, each individual descriptor, and their combined configuration. **(b)** Quantitative results, where VI, VC, and LPIPS measure within-prompt diversity, while CLIPScore, ImageReward, and VQAScore measure semantic alignment or perceptual quality. Combining all three descriptors achieves the highest diversity while maintaining semantic alignment and generation quality comparable to the baseline.