

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 ADOR: ATTENTION DILUTION AND OVERLAP RESOLVER FOR COMPLEX PROMPTS IN TEXT-TO- IMAGE DIFFUSION MODELS

Anonymous authors

Paper under double-blind review

## ABSTRACT

Text-to-image diffusion models have achieved remarkable progress, producing high-quality and realistic images. Nevertheless, these models still encounter challenges with semantic misalignment, particularly when required to understand complex prompts involving multiple objects and diverse attributes. Although several approaches have been proposed to address these issues, investigation into the causes of semantic misalignment has remained limited. In this work, we examine the behavior of cross-attention in text-to-image diffusion models and identify two key factors contributing to semantic misalignment: cross-attention overlap and cross-attention dilution. Building on these findings, we propose ADOR, a training-free framework that mitigates semantic misalignment in a single forward pass, without requiring external guidance. ADOR consists of two complementary modules: the Attention Overlap Disentangler (AO-Disentangler) and the Attention Dilution Reviver (AD-Reviver). The AO-Disentangler reduces cross-attention overlap between noun phrases via distance-based masking, thereby enhancing separation between object–attribute pairs. The AD-Reviver tackles the issue of reduced average cross-attention intensity that arises with longer prompts by applying L2-normalization or selective amplification. It ensures that semantic concepts remain represented during generation. We evaluate ADOR on standard benchmarks and demonstrate that it achieves state-of-the-art performance while preserving efficiency through its training-free, single-pass design.

## 1 INTRODUCTION

Recent advances in text-to-image diffusion models (Rombach et al., 2022; Podell et al., 2023; Peebles & Xie, 2023; Esser et al., 2024) have achieved remarkable progress in photorealistic generation from natural language. These models demonstrate exceptional capability in generating intricate details of objects and nuanced stylistic variations, enabling diverse creative and practical applications. For instance, they have been facilitating realistic scene rendering for virtual environments (Poole et al., 2022), detailed illustrations for storytelling (Liu et al., 2024), and personalized content generation (Ruiz et al., 2023).

Despite remarkable progress, existing text-to-image models struggle to accurately render complex prompts that specify multiple objects with diverse attributes (Feng et al., 2024; Meral et al., 2023; Yang et al., 2024; Wang et al., 2025). This semantic misalignment arises when the generated visual content fails to faithfully reflect the input semantics. It appears in various forms: *object entanglement*, where distinct subjects are erroneously merged into a single entity (Rassin et al., 2024; Zhuang et al., 2024); *improper attribute binding*, where characteristics such as color or texture are incorrectly assigned to objects (Li et al., 2024; Rassin et al., 2024; Zhuang et al., 2024; Meral et al., 2023); and *semantic neglect*, where entities or their specified properties are entirely omitted from the generated image (Marioriyad et al., 2025; Chefer et al., 2023; Meral et al., 2023; Rassin et al., 2024).

Previous strategies to mitigate semantic misalignment include finetuning with additional datasets (Jiang et al., 2024; Hu et al., 2024; Feng et al., 2024), optimizing latent representations during inference (Chefer et al., 2023; Li et al., 2024; Meral et al., 2023), and incorporating spatial guidance generated by large language models (Lian et al., 2024; Yang et al., 2024; Wang et al., 2025). While

054 effective to some extent, these methods often introduce considerable overhead, requiring additional  
 055 training, dependence on external modules, or significantly increased inference times. This raises the  
 056 critical challenge of how to resolve semantic misalignment effectively without external guidance or  
 057 prohibitive computational cost.

058 To address this challenge, we propose **ADOR** (Attention Dilution and Overlap Resolver), a frame-  
 059 work designed to mitigate the key causes of semantic misalignment in text-to-image diffusion mod-  
 060 els. ADOR is training-free, requires no external guidance, and avoids costly test-time optimization,  
 061 making it both efficient and widely accessible. It comprises two complementary modules. The AO-  
 062 Disentangler alleviates object entanglement and improper attribute binding by identifying attention  
 063 overlap regions and applying a locality-based masking strategy, which uses unambiguous regions  
 064 as anchors to ensure that only the correct object-attribute pairs contribute to the attention operation  
 065 within ambiguous areas. The AD-Reviver addresses semantic neglect arising from attention dilution  
 066 by rebalancing cross-attention maps, selectively amplifying the attention score of the corresponding  
 067 object-attribute pair. This independence from additional training and optimization leads to markedly  
 068 faster inference and improved usability compared to prior methods. Extensive experiments and abla-  
 069 tion studies demonstrate that ADOR achieves superior performance over existing approaches while  
 070 maintaining efficiency.

071 In summary, our contributions are as follows:

- 073 • We introduce **ADOR**, a training-free framework that mitigates semantic misalignment  
 074 without requiring external guidance or test-time optimization.
- 075 • We are the first to identify and empirically validate the phenomenon of *attention dilution*  
 076 in text-to-image diffusion models, establishing it as a key cause of semantic misalignment.
- 077 • We design two novel modules, the **AO-Disentangler** and **AD-Reviver**, which effectively  
 078 address cross-attention overlap and cross-attention dilution, thereby removing the primary  
 079 causes of semantic misalignment.
- 080 • We demonstrate that our method achieves state-of-the-art performance in T2I-CompBench  
 081 (Huang et al., 2023; 2025).

## 084 2 RELATED WORKS

085 **Finetuning methods** (Feng et al., 2024; Jiang et al., 2024; Hu et al., 2024) optimize either the  
 086 parameters of the pretrained model or those of auxiliary modules. ELLA (Hu et al., 2024) intro-  
 087 duces a timestep-aware semantic connector module that bridges LLMs and pre-trained diffusion  
 088 models, thereby leveraging the comprehensive language understanding capabilities of LLMs. Ranni  
 089 (Feng et al., 2024) finetunes a text-to-image diffusion model on the LLMs-augmented semantic-  
 090 panel dataset, which includes dense descriptions for each semantic object within an image. CoMat  
 091 (Jiang et al., 2024) proposes an end-to-end finetune methodology for a text-to-image diffusion model  
 092 that integrates a pretrained image-to-text model to enhance concept appearance consistency and a  
 093 pretrained segmentation model to enforce proper attribute binding. While these methods enhance  
 094 semantic alignment, they incur substantial computational and data costs due to the reliance on addi-  
 095 tional datasets and extensive retraining.

096 **Inference-time optimization methods** (Chefer et al., 2023; Li et al., 2024; Zhuang et al., 2024;  
 097 Meral et al., 2023; Zhang et al., 2025; Wang et al., 2025) refine the latent feature space by applying  
 098 their own task-specific loss functions during inference. Attend-and-Excite (Chefer et al., 2023) in-  
 099 troduces a loss that constrains the maximum values of the cross-attention maps for each object to be  
 100 one, ensuring that all target objects are properly attended to in the generated image. Divide & Bind  
 101 (Li et al., 2024) extends this idea by incorporating the Jensen-Shannon Divergence between object  
 102 attention maps and their corresponding attribute attention maps to achieve proper attribute bind-  
 103 ing. CONFORM (Meral et al., 2023) employs contrastive learning during the generation process,  
 104 encouraging attention maps of matching object-attribute pairs to be closer together while pushing  
 105 apart those of mismatched pairs. While these approaches sidestep the computational expense of  
 106 model retraining, they suffer from the drawback of slow inference speed and substantial memory  
 107 overhead.

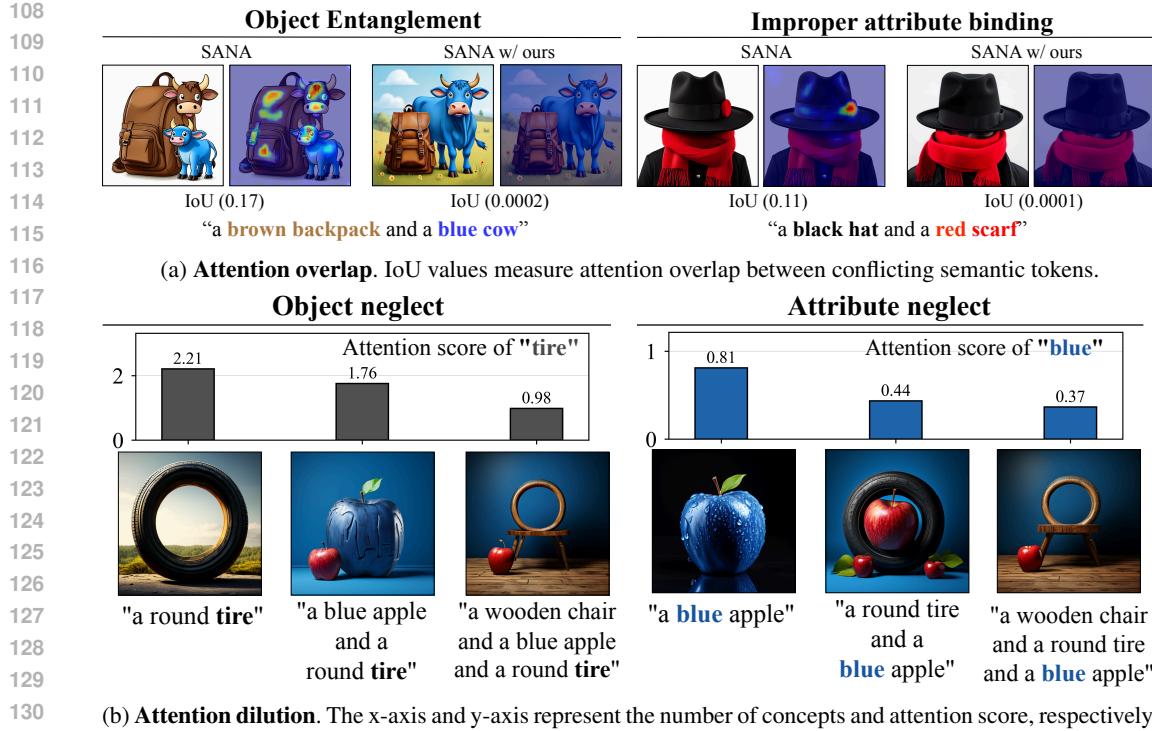


Figure 1: Analysis of semantic misalignment. (a) Comparison between base model and our method, with cross-attention heatmaps and IoU metrics highlighting *object entanglement* and *incorrect attribute binding*. (b) Example of *semantic neglect*, where key concepts are omitted; attention scores for “tire” (left) and “blue” (right) decline as prompt complexity increases.

**LLMs-based methods** (Lian et al., 2024; Yang et al., 2024; Chen et al., 2024) leverage large language models (LLMs) to extract additional conditional information from complex text prompts, thereby enabling more effective conditional image generation. LMD (Lian et al., 2024) proposes a two-stage pipeline which LLM generates an explicit layout including object locations and attributes and layout-to-image generation by controlling the diffusion model’s attention maps. Self-Coherence Guidance (Wang et al., 2025) dynamically controls the cross-attention map through a mask obtained in the previous step, using ratios determined by machine learning. However, the additional priors generated by LLMs without consideration of initial noise characteristics can lead to conflicts between generated elements, resulting in degraded image quality (Ban et al., 2024; Xu et al., 2025; Dahary et al., 2025; Battash et al., 2024).

Despite the promising advances made by these prior methods, they each introduce distinct limitations, such as extensive retraining, slow inference speed, or reliance on external guidance that disregards the inherent characteristics of latent representations. To overcome these challenges, we propose a framework that effectively addresses semantic misalignment by capturing and mitigating attention conflicts at risk of semantic misalignment on the fly.

### 3 METHOD

#### 3.1 OBSERVATION

**Attention overlap** Our analysis identifies two critical failure modes in cross-attention: object entanglement and improper attribute binding. To investigate these, we visualize attention heatmaps in Figure 1a, constructed by aggregating the intersections of the top 5% high-attention regions for semantically conflicting text tokens across all denoising steps and layers. For quantitative evaluation, we compute the average Intersection-over-Union (IoU) of the top 5% high-attention regions for these conflicting tokens across all denoising steps and layers. In both failure modes, we observe consis-

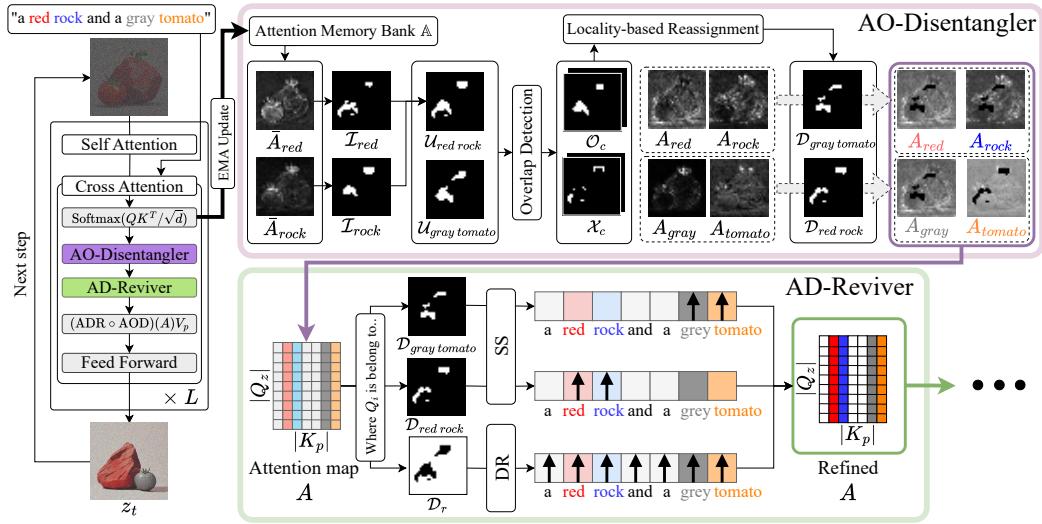


Figure 2: **Overview of the proposed ADOR framework.** ADOR modifies  $L$  cross-attention layers using two components: the Attention-Overlap Disentangler (AO-Disentangler) and the Attention-Dilution Reviver (AD-Reviver). Initially, attention maps for each object and attribute token are extracted from each cross attention and do an EMA update to an attention memory bank  $\mathbb{A}$ . The **AO-Disentangler** then leverages  $\mathbb{A}$  to detect overlapping regions between attribute-object pairs (e.g., “red rock” and “gray tomato”) and performs a locality-based reassignment to create disentangled masks for each attribute-object pair. Subsequently, the **AD-Reviver** reinforces diluted attention scores for visual queries  $Q_i$ . For queries within the disentangled regions ( $D_{gray\ tomato}$  and  $D_{red\ rock}$ ), it applies selective strengthening (SS) to enhance attribute-object binding. For all remaining regions  $D_r$ , it performs dilution-aware rescaling (DR) to uniformly enhance attention across all text tokens, thereby mitigating semantic neglect.

tently high IoU values, indicating that a single visual token often exhibits strong correlations with unrelated text tokens. For instance, object entanglement arises when attention for distinct objects like a “brown backpack” and a “blue cow” incorrectly overlaps, while improper attribute binding occurs when attention from an attribute-object pair like “red scarf” spills onto “black hat.”

**Attention dilution** We also observe a complementary issue, which we term *semantic neglect*. When multiple concepts are appended in a prompt, certain text tokens are overlooked during the denoising process, as shown in Figure 1b. This issue frequently occurs when a single visual token is forced to aggregate information from multiple semantic tokens, leading to weak or flattened correlations. The phenomenon mirrors attention dilution, a well-known problem in natural language processing, where increasing text length leads to progressively flattened attention scores (Zhang et al., 2024; Xu et al., 2024; Liu et al., 2023a). The root cause lies in the softmax operation, which enforces that attention weights sum to one. To quantify this phenomenon, we compute the sum of the top 5% of the highest attention scores for a given token within its attention map and then average across all denoising steps and cross-attention layers. As the number of key tokens increases, the attention weight multiplied by each value token tends to decrease, as illustrated in the graphs in Figure 1b.

### 3.2 OVERVIEW

As illustrated in Figure 2, our proposed method, ADOR, is a training-free, single-pass procedure that modifies cross-attention within each denoising step. The standard cross-attention mechanism, with attention map  $A \in \mathbb{R}^{|Q_z| \times |K_p|}$ , is defined as

$$CA(Q_z, K_p, V_p) = A \cdot V_p, \quad A = \text{softmax} \left( Q_z K_p^\top / \sqrt{d} \right), \quad (1)$$

where  $z$  denotes the latent representation being iteratively denoised, and  $p$  denotes the textual prompt that guides the generation process. The query matrix  $Q_z \in \mathbb{R}^{|Q_z| \times d}$  is derived from the latent feature, while the key and value matrices,  $K_p, V_p \in \mathbb{R}^{|K_p| \times d}$ , are obtained from the prompt  $p$ .

To enhance the semantic alignment with a given text prompt  $p$ , we first perform syntactic parsing with an NLP library, such as spaCy (Honnibal & Montani, 2017). This syntactic information is then used to categorize the sequence of text tokens  $\mathbb{E}_p = \{p_1, \dots, p_S\}$  into three groups: a set of  $N$  object tokens  $\mathbb{E}_o = \{o_1, \dots, o_N\}$ , their corresponding attribute tokens  $\mathbb{E}_a = \bigcup_{i=1}^N \{a_1^i, \dots, a_{M_i}^i\}$ , where  $M_i$  is the number of attributes associated with an object  $o_i$ ; and a set of  $R$  remaining tokens  $\mathbb{E}_r = \{r_1, \dots, r_R\}$ .

ADOR modifies the attention map  $A$  through the integration of AO-Disentangler (AOD) in Section 3.3 and AD-Reviver (ADR) in Section 3.4, defined as

$$\text{ADOR}(Q_z, K_p, V_p) = (\text{ADR} \circ \text{AOD})(A)V_p. \quad (2)$$

The AO-Disentangler addresses issues of *object entanglement* and *improper attribute binding*, while the AD-Reviver specifically focuses on *semantic neglect*. In the following sections, we provide a detailed description of two components, explaining how each component is designed to resolve these specific semantic failures.

### 3.3 AO-DISENTANGLER

The AO-Disentangler is designed to resolve attention overlap that arises when multiple object-attribute concepts compete for the same visual regions. Its pipeline consists of three main stages: (i) constructing an attention memory bank to stabilize identification of attention overlap region, (ii) performing attention overlap detection to identify ambiguous and exclusive regions for each concept, and (iii) applying a locality-based reassignment strategy to ensure that the visual token is uniquely assigned to the most relevant object-attribute concept.

**Attention memory bank** To stabilize identification and minimize noise-induced overfitting, we build an attention memory bank  $\mathbb{A} := \{\bar{A}_e\}$ , where  $e \in \mathbb{E}_o \cup \mathbb{E}_a$ . The accumulated attention map  $\bar{A}_e \in \mathbb{R}^{H \times W}$  is updated at every cross attention layer via an exponential moving average (EMA):

$$\bar{A}_e \leftarrow (1 - \alpha)\bar{A}_e + \alpha A_e, \quad (3)$$

where  $A_e \in \mathbb{R}^{H \times W}$  is the current layer’s attention map for that token and  $\alpha \in (0, 1]$  is the EMA rate. Here,  $H$  and  $W$  denote the height and width of the latent representation  $z$ , respectively.

**Attention overlap detection** To identify attention overlap, we define object-attribute concepts  $\mathcal{C} = \{c_i\}_{i=1}^N$ , where  $c_i = \{o_i, a_1^i, \dots, a_{M_i}^i\}$  and its high-correlated region  $\mathcal{U}_{c_i}$  as follows:

$$\mathcal{U}_{c_i} = \mathcal{I}_{o_i} \cup \bigcup_{j=1}^{M_i} \mathcal{I}_{a_j^i}, \quad \mathcal{I}_e := \left\{ (h, w) \mid [\bar{A}_e]_{h,w} \geq \text{Percentile}(\bar{A}_e, \beta) \right\} \quad (4)$$

where  $\mathcal{I}_e$  is the set of spatial indices from the attention map  $\bar{A}_e$  corresponding to a score above the  $\beta$  percentile threshold. Given the object-attribute concept region  $\mathcal{U}_c$ , we can compute the exclusive attention region  $\mathcal{X}_c$  and the attention overlap region  $\mathcal{O}_c$  as follows:

$$\mathcal{X}_{c_i} = \mathcal{U}_{c_i} \setminus \bigcup_{j \neq i} \mathcal{U}_{c_j}, \quad \mathcal{O}_{c_i} = \mathcal{U}_{c_i} \setminus \mathcal{X}_{c_i}, \quad (5)$$

where  $\mathcal{X}_c$  represents the unambiguous anchor regions exclusively associated with a single object-attribute concept, whereas  $\mathcal{O}_c$  represents the semantically conflicted regions that must be reassigned.

**Locality-based reassignment** To resolve ambiguity in the overlap regions  $\mathcal{O}_c$ , we construct the disentangled region  $\mathcal{D}_c$  by reassigning each ambiguous point to a single concept among the set of competing concepts based on spatial proximity:

$$\mathcal{D}_{c_i} := \mathcal{X}_{c_i} \cup \left\{ \mathbf{x} \mid \mathbf{x} \in \mathcal{O}_{c_i} \text{ and } c_i = \arg \min_{c_k \in C(\mathbf{x})} D(\mathbf{x}, \mathcal{X}_{c_k}) \right\}, \quad (6)$$

where  $C(\mathbf{x}) = \{c_k \mid \mathbf{x} \in \mathcal{O}_{c_k}\}$  represents the collection of all concepts whose ambiguous overlap regions contain the spatial point  $\mathbf{x} = (h, w)$ . Proximity is measured by the point-to-set distance,

$$D(\mathbf{x}, \mathcal{X}_{c_k}) = \min_{\mathbf{y} \in \mathcal{X}_{c_k}} \|\mathbf{x} - \mathbf{y}\|_2, \quad (7)$$

270 which computes the minimum Euclidean distance from a point to any location within an exclusive  
 271 region. Given  $\mathcal{D}_c$ , we apply a mask  $M \in \mathbb{B}^{|Q_z| \times |K_p|}$  that filters attention from tokens which belong  
 272 to the other disentangled regions to the attention map  $A$ :

$$274 \quad \text{AOD}(A) = M \odot A, \quad M_{i,j} = \begin{cases} 0 & \text{if } p_j \in \mathbb{E}_o \cup \mathbb{E}_a \text{ and } \mathbf{x}_i \in \bigcup_{c \in \mathcal{C} \setminus \kappa(p_j)} \mathcal{D}_c \\ 1 & \text{otherwise} \end{cases}, \quad (8)$$

277 where  $\mathbf{x}_i = (h_i, w_i)$  are the 2D latent coordinates for the query index  $i$ , where  $h_i$  and  $w_i$  correspond  
 278 to its position in the latent feature grid.  $\odot$  represents the elementwise product and the function  $\kappa(p_j)$   
 279 denotes the object-attribute concept which contains the token  $p_j$ . By doing so, the AO-Disentangler  
 280 ensures that each ambiguous visual token is uniquely assigned to the most semantically and spatially  
 281 relevant object-attribute concept, effectively resolving the semantic conflict.

### 282 3.4 AD-REVIVER

284 Building on the observation discussed in Section 3.1, we propose the AD-Reviver, an adaptive attention  
 285 rescaling strategy to resolve attention dilution. It consists of two complementary components:  
 286 dilution-aware rescaling, which globally stabilizes attention distributions, and selective strengthening,  
 287 which locally reinforces semantically relevant tokens. Formally, we define

$$289 \quad \text{ADR}(A)_{i,k} = \begin{cases} \text{SS}(A)_{i,k} & \text{if } p_j \in \mathbb{E}_o \cup \mathbb{E}_a \text{ and } \mathbf{x}_i \in \mathcal{D}_{\kappa(p_j)} \\ \text{DR}(A)_{i,k} & \text{otherwise} \end{cases}, \quad (9)$$

291 where SS denotes selective strengthening and DR denotes dilution-aware rescaling. In this manner,  
 292 AD-Reviver enhances semantically relevant regions while mitigating attention dilution across dis-  
 293 persed distributions, thereby ensuring that the contribution of each semantic token is not diminished.

294 **Dilution-aware rescaling** To mitigate attention dilution, we rescale the cross-attention map  $A$  on  
 295 a per-query basis. Let  $A'$  denote the query-wise normalized attention map, obtained by dividing each  
 296 row by  $\sum_k A_{i,k}$ . The rescaled map is then defined as

$$298 \quad \text{DR}(A)_{i,k} = \frac{A'_{i,k}}{\|A'_{i,:}\|_2}, \quad (10)$$

300 This normalization prevents divergence of the inverse of  $\ell_2$  norm, regardless of whether attention  
 301 masking is applied. The vector  $A'_{i,:} \in \mathbb{R}^{|K_p|}$  represents the normalized attention scores associated  
 302 with the  $i$ -th query token. We adopt the inverse  $\ell_2$  norm as the scaling factor due to its adaptive be-  
 303 havior. Specifically, for a concentrated attention vector, the factor remains close to one. This minimal  
 304 adjustment preserves the original scores, ensuring stable generation for these well-defined concepts.  
 305 In contrast, for a dispersed (diluted) vector, the factor is close to  $\sqrt{|K_p|}$ , substantially amplifying  
 306 the scores and alleviating semantic neglect.

308 **Selective strengthening** For queries corresponding to the disentangled region  $\mathcal{D}_c$ , we selectively  
 309 enhance the concept tokens associated with the most relevant concept. The rescaling is defined as

$$311 \quad \text{SS}(A)_{i,k} = \begin{cases} \lambda_{i,j} A_{i,j} & \text{if } p_j \in \mathbb{E}_o \cup \mathbb{E}_a \text{ and } \mathbf{x}_i \in \mathcal{D}_{\kappa(p_j)} \\ A_{i,j} & \text{otherwise} \end{cases},$$

$$313 \quad \lambda_{i,j} = \left( 1 + \frac{\sum_j^{|K_t|} \text{DR}(A)_{i,j} - \sum_j^{|K_t|} A_{i,j}}{\sum_{p_k \in \kappa(p_j)} A_{i,k}} \right).$$

316 This formulation adaptively strengthens the concept tokens in proportion to their original attention  
 317 scores.

## 318 4 EXPERIMENT

### 321 4.1 EXPERIMENTAL SETTINGS

323 **Implementation details.** Our experiments utilize Sana (Xie et al., 2024) and PixArt- $\alpha$  (Chen et al.,  
 324 2023) as base models. We generate images with 20 diffusion steps for Sana (Xie et al., 2024) and

324 Table 1: Quantitative comparison of models on T2I-CompBench benchmark, evaluating performance  
 325 on color, shape, texture, 2D/3D spatial reasoning, non-spatial, and numeracy, in addition to  
 326 inference speed, and resolution. The best scores in T2I-CompBench are highlighted in **bold**. Results  
 327 marked with  $\dagger$  are from (Huang et al., 2025) and  $\ddagger$  are from (Feng et al., 2024). All other results  
 328 are measured using the official codebases.

330 Model	331 Base	332 Color $\uparrow$	333 Shape $\uparrow$	334 Texture $\uparrow$	335 T2I-CompBench				336 Speed (sec/image) $\downarrow$	337 Resolution
338	339	340	341	342	2D-spatial $\uparrow$	3D-spatial $\uparrow$	343 Non-spatial $\uparrow$	344 Numeracy $\uparrow$	345	346
347 SD2 $\dagger$	348 none	349 0.5065	350 0.4221	351 0.4922	352 0.1342	353 0.3300	354 0.3127	355 0.4582	356 2.36	357 512 $\times$ 512
348 Composable $\dagger$	349 SD2	350 0.4063	351 0.3299	352 0.3645	353 0.0800	354 0.2847	355 0.2980	356 0.4272	357 11.88	358 512 $\times$ 512
348 A&E $\dagger$	349 SD2	350 0.6400	351 0.4517	352 0.5963	353 0.1455	354 0.3222	355 0.3109	356 0.4773	357 10.77	358 512 $\times$ 512
348 Ranni $\ddagger$	349 SD2	350 0.6893	351 0.4934	352 0.6325	353 0.3167	354 –	355 –	356 –	357 20.24	358 768 $\times$ 768
347 PixArt- $\alpha$	348 none	349 0.3964	350 0.4062	351 0.4696	352 0.1994	353 0.3421	354 0.3081	355 0.4971	356 2.74	357 512 $\times$ 512
347 SCG	348 PixArt- $\alpha$	349 0.5538	350 0.4115	351 0.4633	352 0.1921	353 0.3444	354 0.3094	355 0.5021	356 8.08	357 512 $\times$ 512
347 Ours	348 PixArt- $\alpha$	349 0.6817	350 0.5425	351 0.6339	352 0.2190	353 0.3706	354 0.3104	355 0.5451	356 8.59	357 512 $\times$ 512
347 Sana	348 none	349 0.7703	350 0.5405	351 0.6744	352 0.3794	353 0.4128	354 0.3137	355 0.6096	356 9.57	357 1024 $\times$ 1024
347 Ours	348 Sana	349 <b>0.8240</b>	350 <b>0.6143</b>	351 <b>0.7425</b>	352 <b>0.3862</b>	353 <b>0.4180</b>	354 <b>0.3149</b>	355 <b>0.6398</b>	356 14.71	357 1024 $\times$ 1024

339 50 diffusion steps for PixArt- $\alpha$  (Chen et al., 2023). In both configurations, our framework employs  
 340 an exponential moving average (EMA) rate of 0.5, a percentile- $\beta$  rate of 0.05, and a classifier-free  
 341 guidance scale of 4.5 (Ho & Salimans, 2022). All experiments are conducted using a single NVIDIA  
 342 GeForce RTX 3090 GPU.

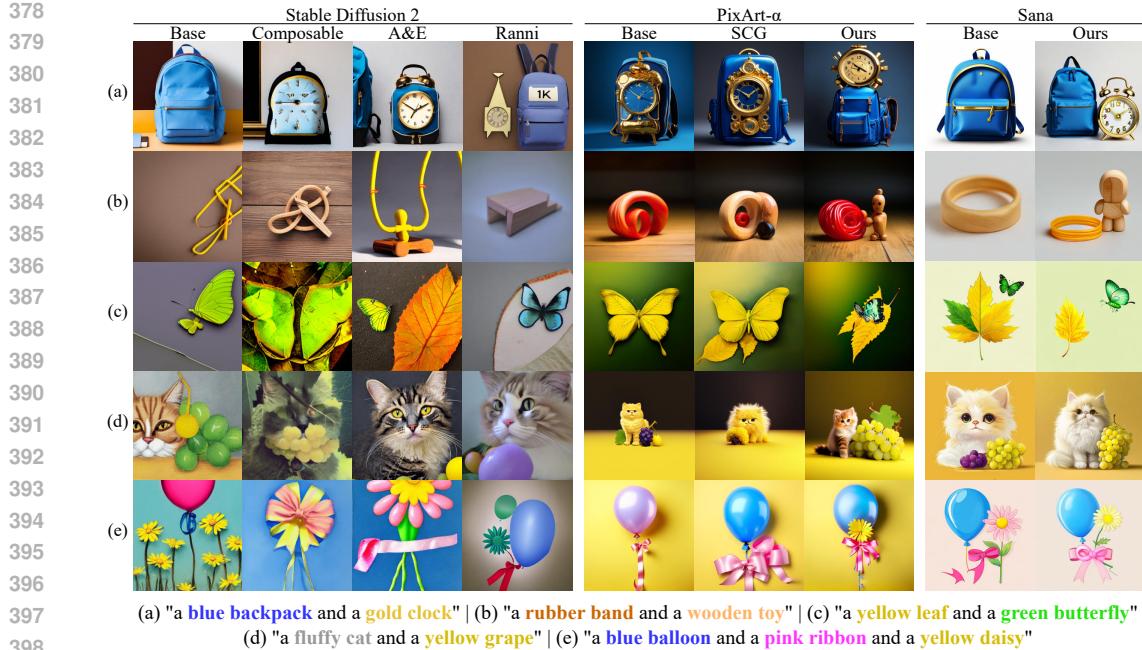
343 **Evaluation metrics.** The T2I-CompBench (Huang et al., 2023; 2025) evaluates three main categories: attribute binding (color, shape, texture), object relationships (2D-spatial, 3D-spatial, non-spatial), and numeracy. For attribute binding, the BLIP model (Li et al., 2022) is employed to assess whether attributes such as color, shape, and texture are correctly associated with their respective objects in the generated image. The evaluation of 2D/3D spatial relationships and numeracy employs the UniDet model (Zhou et al., 2022). This model detects objects to compare their positions—using bounding box coordinates for 2D relationships and depth estimation for 3D relationships—and to verify that the number of generated objects matches the prompt. Non-spatial relationships are evaluated using CLIPScore (Radford et al., 2021; Hessel et al., 2021), which measures the alignment between the generated image and the provided text prompt by calculating the cosine similarity between their feature representations.

354 **Baselines.** We evaluate our method against a comprehensive set of baseline models that represent  
 355 different approaches to compositional generation. Our comparison includes inference-time optimi-  
 356 zation methods such as Attend-and-Excite (A&E) (Rombach et al., 2022) and Composable Diffusion  
 357 (Composable) (Liu et al., 2023b). Additionally, we include Ranni (Feng et al., 2024) and self-coherence  
 358 guidance (SCG) (Wang et al., 2025) for comparison with a fine-tuning and LLM-based approach,  
 359 respectively.

## 360 361 4.2 COMPARISON WITH OTHER MODELS ON T2I-COMPBENCH

362 **Quantitative results.** As shown in Table 1, we evaluate baseline performance on T2I-CompBench  
 363 (Huang et al., 2023; 2025). When applied to Sana (Xie et al., 2024), our method achieves state-  
 364 of-the-art results with consistent improvements across all categories. Notably, the performance  
 365 gains are most pronounced for attribute types such as color (+7.0%), shape (+13.6%), and texture  
 366 (+10.1%). We attribute this significant improvement to our method’s targeted mitigation of a  
 367 critical vulnerability in cross-attention: its tendency to incorrectly bind or dilute semantic signals  
 368 across multiple concepts in complex prompts. On 512 $\times$ 512 generation, our PixArt- $\alpha$  implemen-  
 369 tation demonstrates strong efficiency (8.59 sec/image) while outperforming SD2-based Composable  
 370 Diffusion (11.88 sec/image) and A&E (10.77 sec/image) in both speed and accuracy. Although SCG  
 371 achieves marginally faster inference (8.08 sec/image), our model achieves substantially higher accu-  
 372 racy, with improvements of +23.1% for color and +36.8% for texture, indicating that the modest  
 373 computational overhead yields significant performance gains. Collectively, these results demon-  
 374 strate robust improvements across semantic alignment benchmarks, particularly in attribute binding, while  
 375 maintaining competitive computational efficiency.

376 **Qualitative results.** Figure 3 presents a qualitative comparison between baseline models and our  
 377 proposed method. Across all prompts, the generated images reveal that semantic misalignment con-  
 378 stitutes a persistent challenge for existing approaches. For prompts (a) and (b), the base models



378  
379 (a) "a **blue backpack** and a **gold clock**" | (b) "a **rubber band** and a **wooden toy**" | (c) "a **yellow leaf** and a **green butterfly**"  
380 (d) "a **fluffy cat** and a **yellow grape**" | (e) "a **blue balloon** and a **pink ribbon** and a **yellow daisy**"  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396

397 Figure 3: Qualitative comparison with other models.  
398  
399  
400  
401

402 exhibit object neglect or entanglement, and even previous methods fail to resolve these issues fully.  
403 In contrast, our method successfully separates "backpack" and "clock" as well as "band" and "toy",  
404 while preserving their semantic independence. For prompts (c) and (d), the base models suffer from  
405 either improper attribute binding or attribute neglect. Alternative approaches still fall short of fully  
406 addressing these challenges. In contrast, our extensions correctly match attributes and objects across  
407 pairs, ensuring that none of these concepts are neglected and that their semantics are faithfully pre-  
408 served in the generated results. For prompt (e), which includes more concepts, baseline models  
409 show compounded semantic misalignment. In contrast, our method separates objects and correctly  
410 binds attributes, representing all concepts. Collectively, these results demonstrate that our approach  
411 effectively mitigates semantic misalignment, yielding outputs that are more faithfully aligned with  
412 compositional prompts.  
413  
414

#### 4.3 ABLATION ON AO-DISENTANGLER AND AD-REVIVER

415 **Quantitative results.** As summarized in Table 2, we assess the AO-  
416 Disentangler (AOD) and AD-Reviver (ADR) via selective ablations while  
417 holding all other conditions fixed. For  
418 BLIP-VQA, AOD provides noticeable improvements, which become even larger when ADR is added on top (e.g., PixArt- $\alpha$ : +9.1% vs. +22.4% in shape). This trend is even stronger in PixArt- $\alpha$ , highlighting the role of ADR in reducing dilution effects caused by diverse text tokens in cross-attention. In contrast, AOD demonstrates a dominant contribution to numeracy performance, with substantially larger improvements compared to ADR. This suggests that attention overlaps become more probable when the semantic number of objects increases through numerical expression (e.g., changing "one bear and one horse" to "two bears and three horses"), making AOD especially effective. Overall, ADR and AOD serve distinct but complementary roles, and their combination consistently yields the strongest performance across all categories.  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431

432 Table 2: Ablation studies on the proposed method. AOD and ADR denote AO-Disentangler and AD-Reviver, respectively. **Bold** denotes the best performance.  
433  
434

Baseline	AOD	ADR	BLIP-VQA			Numeracy $\uparrow$
			Color $\uparrow$	Shape $\uparrow$	Texture $\uparrow$	
Sana	$\times$	$\times$	0.7703	0.5405	0.6744	0.6096
	$\checkmark$	$\times$	0.7843	0.5550	0.7076	0.6397
	$\checkmark$	$\checkmark$	<b>0.8240</b>	<b>0.6143</b>	<b>0.7425</b>	<b>0.6398</b>
PixArt- $\alpha$	$\times$	$\times$	0.3964	0.4062	0.4696	0.4971
	$\checkmark$	$\times$	0.4952	0.4432	0.5296	0.5289
	$\checkmark$	$\checkmark$	<b>0.6817</b>	<b>0.5425</b>	<b>0.6339</b>	<b>0.5451</b>

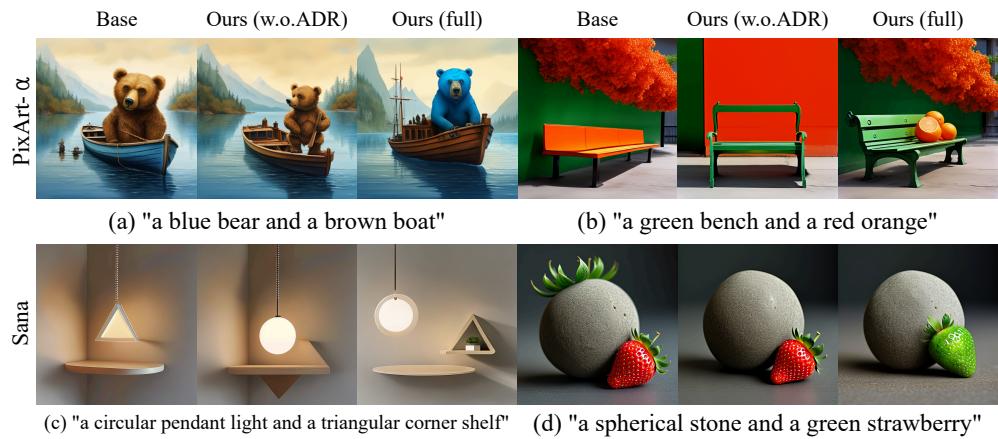


Figure 4: Ablation studies on our method.

**Qualitative results.** Figure 4 presents a qualitative ablation study highlighting the contributions of our key modules. Images generated by the base models consistently exhibit semantic misalignment. For example, in prompt (a), the attribute “blue” is incorrectly assigned to “boat” rather than “bear”, while “bear” itself appears in “brown”, reflecting both attribute neglect and improper binding. In prompt (b), the object “orange” is omitted, and “bench” is rendered with an incorrect color. Prompt (c) demonstrates improper attribute binding of “triangular” to “pendant light” instead of “corner shelf”. Finally, in prompt (d), “stone” and “strawberry” are entangled into a single incoherent object, and “green” is insufficient from the strawberry. When the AOD module is added to the base, these issues are partially alleviated. Improper attribute binding is corrected across prompts (a), (b), and (c), while object entanglement is resolved in prompt (d). This indicates that AOD effectively disentangles mixed concepts and prevents attributes from being incorrectly matched. Nevertheless, certain deficiencies remain: “bear” and “strawberry” still lack their intended colors, and the object “orange” continues to be neglected. With the subsequent inclusion of ADR, forming our full model, these shortcomings are largely addressed. Prompts (a), (c), and (d) show attributes correctly emphasized on their respective objects. For instance, “blue” is properly bound to “bear”, “triangular” to “corner shelf”, and “green” to “strawberry”. Moreover, ADR revives previously neglected elements, ensuring that missing objects and attributes are faithfully generated. Together, these results demonstrate that AOD and ADR complement one another, each targeting distinct sources of semantic misalignment, and that their integration is critical for achieving faithful compositional alignment.

## 5 CONCLUSION

We presented ADOR, a training-free framework that tackles the root causes of semantic misalignment in text-to-image diffusion models. Building on analysis and prior studies, we identified cross-attention overlap and cross-attention dilution as two key factors responsible for object entanglement, improper attribute binding, and neglect of visual concepts. To mitigate these issues, we designed two complementary modules. AO-Disentangler separates overlapped cross-attention signals via distance-based masking, while AD-Reviver restores balanced attention strength through normalization and selective amplification. Extensive experiments demonstrated that ADOR consistently improves semantic alignment, delivering more faithful object–attribute correspondences while preserving efficiency through a single forward pass and avoiding additional training or external guidance. These results highlight the importance of understanding and controlling cross-attention behavior as a pathway to more reliable generative modeling. Looking forward, our work opens promising avenues for attention-aware inference strategies and the extension of training-free alignment techniques to broader multimodal generation tasks, including video synthesis and text-conditioned 3D generation. By addressing the mechanisms behind semantic misalignment, ADOR offers a practical and effective step toward more semantically faithful image generation.

486 REFERENCES  
487

488 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-  
489 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical  
490 report. *arXiv preprint arXiv:2303.08774*, 2023.

491 Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. The  
492 crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise, 2024.  
493 URL <https://arxiv.org/abs/2406.01970>.

494 Barak Battash, Amit Rozner, Lior Wolf, and Ofir Lindenbaum. Obtaining favorable layouts for  
495 multiple object generation, 2024. URL <https://arxiv.org/abs/2405.00791>.

496 Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite:  
497 Attention-based semantic guidance for text-to-image diffusion models, 2023. URL <https://arxiv.org/abs/2301.13826>.

498 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James  
499 Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer  
500 for photorealistic text-to-image synthesis, 2023. URL <https://arxiv.org/abs/2310.00426>.

501 Zhennan Chen, Yajie Li, Haofan Wang, Zhibo Chen, Zhengkai Jiang, Jun Li, Qian Wang, Jian Yang,  
502 and Ying Tai. Region-aware text-to-image generation via hard binding and soft refinement, 2024.  
503 URL <https://arxiv.org/abs/2411.06558>.

504 Omer Dahary, Yehonathan Cohen, Or Patashnik, Kfir Aberman, and Daniel Cohen-Or. Be decisive:  
505 Noise-induced layouts for multi-subject generation, 2025. URL <https://arxiv.org/abs/2505.21488>.

506 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam  
507 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion En-  
508 glish, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow  
509 transformers for high-resolution image synthesis, 2024. URL <https://arxiv.org/abs/2403.03206>.

510 Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-  
511 image diffusion for accurate instruction following, 2024. URL <https://arxiv.org/abs/2311.17002>.

512 Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A  
513 reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

514 Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint  
515 arXiv:2207.12598*, 2022.

516 Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embed-  
517 dings, convolutional neural networks and incremental parsing. *To appear*, 2017.

518 Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models  
519 with llm for enhanced semantic alignment, 2024. URL <https://arxiv.org/abs/2403.05135>.

520 Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive  
521 benchmark for open-world compositional text-to-image generation. *Advances in Neural Infor-  
522 mation Processing Systems*, 36:78723–78747, 2023.

523 Kaiyi Huang, Chengqi Duan, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench++:  
524 An enhanced and comprehensive benchmark for compositional text-to-image generation. *IEEE  
525 Transactions on Pattern Analysis and Machine Intelligence*, 2025.

526 Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu,  
527 and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept  
528 matching, 2024. URL <https://arxiv.org/abs/2404.03653>.

540 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-  
 541 training for unified vision-language understanding and generation. In *International conference on*  
 542 *machine learning*, pp. 12888–12900. PMLR, 2022.

543

544 Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide bind your attention for im-  
 545 proved generative semantic nursing, 2024. URL <https://arxiv.org/abs/2307.10864>.

546 Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt  
 547 understanding of text-to-image diffusion models with large language models, 2024. URL  
 548 <https://arxiv.org/abs/2305.13655>.

549

550 Bingbin Liu, Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Exposing attention  
 551 glitches with flip-flop language modeling. *Advances in Neural Information Processing Systems*,  
 552 36:25549–25583, 2023a.

553 Chang Liu, Haoning Wu, Yujie Zhong, Xiaoyun Zhang, Yanfeng Wang, and Weidi Xie. Intelligent  
 554 grimm - open-ended visual storytelling via latent diffusion models. In *Proceedings of the*  
 555 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6190–6200,  
 556 June 2024.

557

558 Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional vi-  
 559 sual generation with composable diffusion models, 2023b. URL <https://arxiv.org/abs/2206.01714>.

560

561 Arash Marioriyad, Mohammadali Banayeeanzade, Reza Abbasi, Mohammad Hossein Rohban, and  
 562 Mahdieh Soleymani Baghshah. Attention overlap is responsible for the entity missing problem in  
 563 text-to-image diffusion models!, 2025. URL <https://arxiv.org/abs/2410.20972>.

564

565 Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is  
 566 all you need for high-fidelity text-to-image diffusion models, 2023. URL <https://arxiv.org/abs/2312.06059>.

567

568 William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023. URL <https://arxiv.org/abs/2212.09748>.

569

570 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe  
 571 Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image  
 572 synthesis, 2023. URL <https://arxiv.org/abs/2307.01952>.

573

574 Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d  
 575 diffusion, 2022. URL <https://arxiv.org/abs/2209.14988>.

576

577 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
 578 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
 579 models from natural language supervision. In *International conference on machine learning*, pp.  
 580 8748–8763. PMLR, 2021.

581

582 Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik.  
 583 Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map  
 584 alignment, 2024. URL <https://arxiv.org/abs/2306.08877>.

585

586 Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-  
 587 resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF confer-  
 588 ence on computer vision and pattern recognition*, pp. 10684–10695, 2022.

589

590 Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman.  
 591 Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.  
 592 URL <https://arxiv.org/abs/2208.12242>.

593

594 Shulei Wang, Wang Lin, Hai Huang, Hanting Wang, Sihang Cai, WenKang Han, Tao Jin, Jingyuan  
 595 Chen, Jiacheng Sun, Jieming Zhu, and Zhou Zhao. Towards transformer-based aligned generation  
 596 with self-coherence guidance, 2025. URL <https://arxiv.org/abs/2503.17675>.

594 Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang  
 595 Li, Ligeng Zhu, Yao Lu, and Song Han. Sana: Efficient high-resolution image synthesis with  
 596 linear diffusion transformers, 2024. URL <https://arxiv.org/abs/2410.10629>.

597 Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering se-  
 598 cret seeds in text-to-image diffusion models, 2025. URL <https://arxiv.org/abs/2405.14828>.

600 Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of  
 601 large language models. *arXiv preprint arXiv:2401.11817*, 2024.

602 Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-  
 603 to-image diffusion: Recaptioning, planning, and generating with multimodal llms, 2024. URL  
 604 <https://arxiv.org/abs/2401.11708>.

605 Xuechen Zhang, Xiangyu Chang, Mingchen Li, Amit Roy-Chowdhury, Jiasi Chen, and Samet Oy-  
 606 mak. Selective attention: Enhancing transformer through principled context control. *Advances in  
 607 Neural Information Processing Systems*, 37:11061–11086, 2024.

608 Yang Zhang, Rui Zhang, Xuecheng Nie, Haochen Li, Jikun Chen, Yifan Hao, Xin Zhang, Luoqi Liu,  
 609 and Ling Li. Spdiffusion: Semantic protection diffusion models for multi-concept text-to-image  
 610 generation, 2025. URL <https://arxiv.org/abs/2409.01327>.

611 Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *Pro-  
 612 ceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7571–7580,  
 613 2022.

614 Chenyi Zhuang, Ying Hu, and Pan Gao. Magnet: We never know how text-to-image diffusion models  
 615 work, until we learn how vision-language models function, 2024. URL <https://arxiv.org/abs/2409.19967>.

620

621

622

623

624

625

626

627

628

629

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

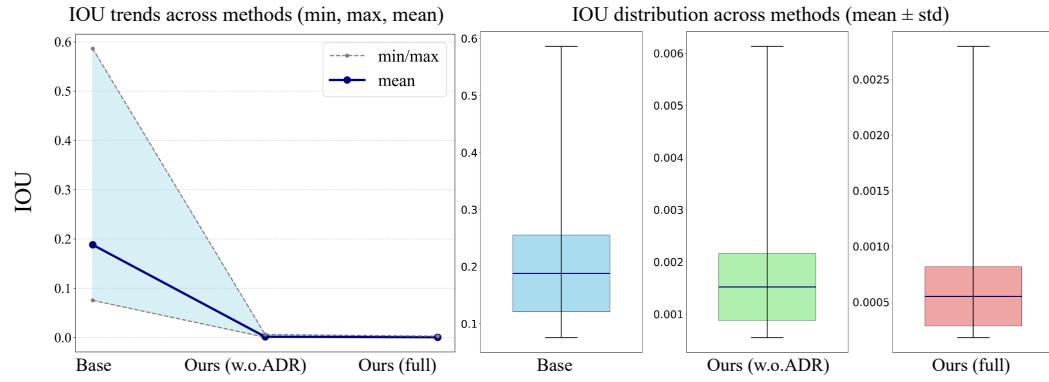
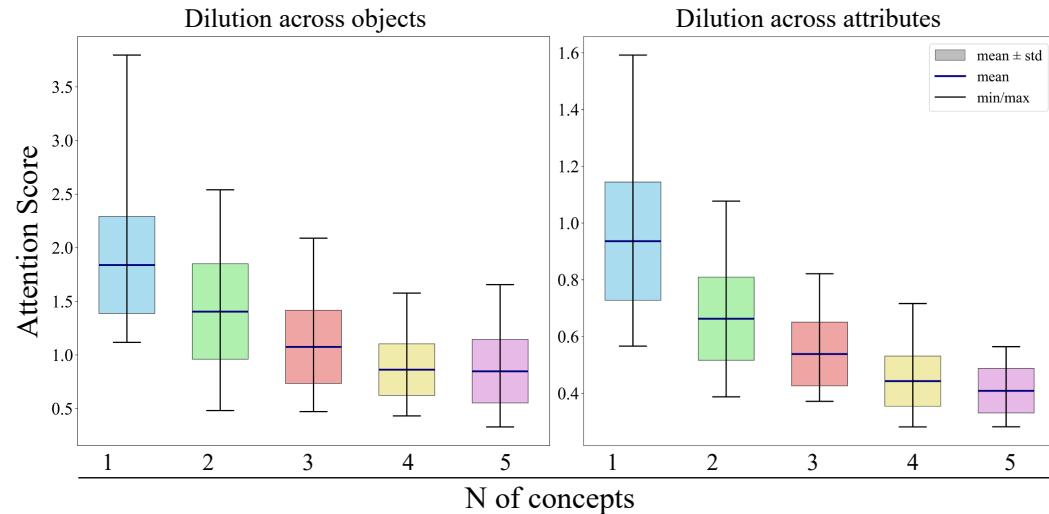
648 A STATISTICAL ANALYSIS ON OBSERVATION  
649650 We extend the experiments presented in Section 3.1 with quantitative analysis.  
651652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701

Figure 5: IOU comparison between base model and our variants. Left: a line plot summarizing IOU per method, with the solid line indicating the mean, dashed lines marking the minimum and maximum, and the shaded band spanning the min–max range. Right: per-method distributions shown as box-style, where the box reflects the mean  $\pm 1$  standard deviation and whiskers denote the full range.

**Attention overlap** In Section 3.1, comparisons are restricted to the base model (Sana (Xie et al., 2024)) and our Sana variants. Here, we explicitly compare three methods: (1) the base Sana, (2) Sana without ADR (Sana + AOD), and (3) Sana with full modules (Sana + AOD + ADR). We evaluate on 800 prompts from T2I-CompBench (Huang et al., 2025) of the form “a/an ⟨attribute1⟩⟨object1⟩ and a/an ⟨attribute2⟩⟨object2⟩”. Figure Figure 5 shows that adding AOD alone leads to a sharp reduction in overlap: the average IoU across the 800 prompts drops from 0.18818 to 0.00152, a 99.2% decrease. Building on this, ADR—by selectively enhancing features within the separated regions—further reduces overlap: relative to Sana + AOD, the average IoU decreases from 0.00152 to 0.00055, a 63.8% reduction. These results indicate that AOD effectively separates the attention regions corresponding to each concept, and ADR further reinforces this effect.

698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
10010  
10011  
10012  
10013  
10014  
10015  
10016  
10017  
10018  
10019  
10020  
10021  
10022  
10023  
10024  
10025  
10026  
10027  
10028  
10029  
10030  
10031  
10032  
10033  
10034  
10035  
10036  
10037  
10038  
10039  
10040  
10041  
10042  
10043  
10044  
10045  
10046  
10047  
10048  
10049  
10050  
10051  
10052  
10053  
10054  
10055  
10056  
10057  
10058  
10059  
10060  
10061  
10062  
10063  
10064  
10065  
10066  
10067  
10068  
10069  
10070  
10071  
10072  
10073  
10074  
10075  
10076  
10077  
10078  
10079  
10080  
10081  
10082  
10083  
10084  
10085  
10086  
10087  
10088  
10089  
10090  
10091  
10092  
10093  
10094  
10095  
10096  
10097  
10098  
10099  
100100  
100101  
100102  
100103  
100104  
100105  
100106  
100107  
100108  
100109  
100110  
100111  
100112  
100113  
100114  
100115  
100116  
100117  
100118  
100119  
100120  
100121  
100122  
100123  
100124  
100125  
100126  
100127  
100128  
100129  
100130  
100131  
100132  
100133  
100134  
100135  
100136  
100137  
100138  
100139  
100140  
100141  
100142  
100143  
100144  
100145  
100146  
100147  
100148  
100149  
100150  
100151  
100152  
100153  
100154  
100155  
100156  
100157  
100158  
100159  
100160  
100161  
100162  
100163  
100164  
100165  
100166  
100167  
100168  
100169  
100170  
100171  
100172  
100173  
100174  
100175  
100176  
100177  
100178  
100179  
100180  
100181  
100182  
100183  
100184  
100185  
100186  
100187  
100188  
100189  
100190  
100191  
100192  
100193  
100194  
100195  
100196  
100197  
100198  
100199  
100200  
100201  
100202  
100203  
100204  
100205  
100206  
100207  
100208  
100209  
100210  
100211  
100212  
100213  
100214  
100215  
100216  
100217  
100218  
100219  
100220  
100221  
100222  
100223  
100224  
100225  
100226  
100227  
100228  
100229  
100230  
100231  
100232  
100233  
100234  
100235  
100236  
100237  
100238  
100239  
100240  
100241  
100242  
100243  
100244  
100245  
100246  
100247  
100248  
100249  
100250  
100251  
100252  
100253  
100254  
100255  
100256  
100257  
100258  
100259  
100260  
100261  
100262  
100263  
100264  
100265  
100266  
100267  
100268  
100269  
100270  
100271  
100272  
100273  
100274  
100275  
100276  
100277  
100278  
100279  
100280  
100281  
100282  
100283  
100284  
100285  
100286  
100287  
100288  
100289  
100290  
100291  
100292  
100293  
100294  
100295  
100296  
100297  
100298  
100299  
100300  
100301  
100302  
100303  
100304  
100305  
100306  
100307  
100308  
100309  
100310  
100311  
100312  
100313  
100314  
100315  
100316  
100317  
100318  
100319  
100320  
100321  
100322  
100323  
100324  
100325  
100326  
100327  
100328  
100329  
100330  
100331  
100332  
100333  
100334  
100335  
100336  
100337  
100338  
100339  
100340  
100341  
100342  
100343  
100344  
100345  
100346  
100347  
100348  
100349  
100350  
100351  
100352  
100353  
100354  
100355  
100356  
100357  
100358  
100359  
100360  
100361  
100362  
100363  
100364  
100365  
100366  
100367  
100368  
100369  
100370  
100371  
100372  
100373  
100374  
100375  
100376  
100377  
100378  
100379  
100380  
100381  
100382  
100383  
100384  
100385  
100386  
100387  
100388  
100389  
100390  
100391  
100392  
100393  
100394  
100395  
100396  
100397  
100398  
100399  
100400  
100401  
100402  
100403  
100404  
100405  
100406  
100407  
100408  
100409  
100410  
100411  
100412  
100413  
100414  
100415  
100416  
100417  
100418  
100419  
100420  
100421  
100422  
100423  
100424  
100425  
100426  
100427  
100428  
100429  
100430  
100431  
100432  
100433  
100434  
100435  
100436  
100437  
100438  
100439  
100440  
100441  
100442  
100443  
100444  
100445  
100446  
100447  
100448  
100449  
100450  
100451  
100452  
100453  
100454  
100455  
100456  
100457  
100458  
100459  
100460  
100461  
100462  
100463  
100464  
100465  
100466  
100467  
100468  
100469  
100470  
100471  
100472  
100473  
100474  
100475  
100476  
100477  
100478  
100479  
100480  
100481  
100482  
100483  
100484  
100485  
100486  
100487  
100488  
100489  
100490  
100491  
100492  
100493  
100494  
100495  
100496  
100497  
100498  
100499  
100500  
100501  
100502  
100503  
100504  
100505  
100506  
100507  
100508  
100509  
100510  
100511  
100512  
100513  
100514  
100515  
100516  
100517  
100518  
100519  
100520  
100521  
100522  
100523  
100524  
100525  
100526  
100527  
100528  
100529  
100530  
100531  
100532  
100533  
100534  
100535  
100536  
100537  
100538  
100539  
100540  
100541  
100542  
100543  
100544  
100545  
100546  
100547  
100548  
100549  
100550  
100551  
100552  
100553  
100554  
100555  
100556  
100557  
100558  
100559  
100560  
100561  
100562  
100563  
100564  
100565  
100566  
100567  
100568  
100569  
100570  
100571  
100572  
100573  
100574  
100575  
100576  
100577  
100578  
100579  
100580  
100581  
100582  
100583  
100584  
100585  
100586  
100587  
100588  
100589  
100590  
100591  
100592  
100593  
100594  
100595  
100596  
100597  
100598  
100599  
100600  
100601  
100602  
100603  
100604  
100605  
100606  
100607  
100608  
100609  
100610  
100611  
100612  
100613  
100614  
100615  
100616  
100617  
100618  
100619  
100620  
100621  
100622  
100623  
100624  
100625  
100626  
100627  
100628  
100629  
100630  
100631  
100632  
100633  
100634  
100635  
100636  
100637  
100638  
100639  
100640  
100641  
100642  
100643  
100644  
100645  
100646  
100647  
100648  
100649  
100650  
100651  
100652  
100653  
100654  
100655  
100656  
100657  
100658  
100659  
100660  
100661  
100662  
100663  
100664  
100665  
100666  
100667  
100668  
100669  
100670  
100671  
100672  
100673  
100674  
100675  
100676  
100677  
100678  
100679  
100680  
100681  
100682  
100683  
100684  
100685  
100686  
100687  
100688  
100689  
100690  
100691  
100692  
100693  
100694  
100695  
100696  
100697  
100698  
100699  
100700  
100701  
100702  
100703  
100704  
100705  
100706  
100707  
100708  
100709  
100710  
100711  
100712  
100713  
100714  
100715  
100716  
100717  
100718  
100719  
100720  
100721  
100722  
100723  
100724  
100725  
100726  
100727  
100728  
100729  
100730  
100731  
100732  
100733  
100734  
100735  
100736  
100737  
100738  
100739  
100740  
100741  
100742  
100743  
100744  
100745  
100746  
100747  
100748  
100749  
100750  
100751  
100752  
100753  
100754  
100755  
100756  
100757  
100758  
100759  
100760  
100761  
100762  
100763  
100764  
100765  
100766  
100767  
100768  
100769  
100770  
100771  
100772  
100773  
100774  
100775  
100776  
100777  
100778  
100779  
100780  
100781  
100782  
100783  
100784  
100785  
100786  
100787  
100788  
100789  
100790  
100791  
100792  
100793  
100794  
100795  
100796  
100797  
100798  
100799  
100800  
100801  
100802  
100803  
100804  
100805  
100806  
100807  
100808  
100809  
100810  
100811  
100812  
100813  
100814  
100815  
100816  
100817  
100818  
100819  
100820  
100821  
100822  
100823  
100824  
100825  
100826  
100827  
100828  
100829  
100830  
100831  
100832  
100833  
100834  
100835  
100836  
100837  
100838  
100839  
100840  
100841  
100842  
100843  
100844  
100845  
100846  
100847  
100848  
100849  
100850  
100851  
100852  
100853  
100854  
100855  
100856  
100857  
100858  
100859  
100860  
100861  
100862  
100863  
100864  
100865  
100866  
100867  
100868  
100869  
100870  
100871  
100872  
100873  
100874  
100875  
100876  
100877  
100878  
100879  
100880  
100881  
100882  
100883  
100884  
100885  
100886  
100887  
100888  
100889  
100890  
100891  
100892  
100893  
100894  
100895  
100896  
100897  
100898  
100899  
100900  
100901  
100902  
100903  
100904  
100905  
100906  
100907  
100908  
100909  
100910  
100911  
100912  
100913  
100914  
100915  
100916  
100917  
100918  
100919  
100920  
100921  
100922  
100923  
100924  
100925  
100926  
100927  
100928  
100929  
100930  
100931  
100932  
100933  
100934  
100935  
100936  
100937  
100938  
100939  
100940  
100941  
100942  
100943  
100944  
100945  
100946  
100947  
100948  
100949  
100950  
100951  
100952  
100953  
100954  
100955  
100956  
100957  
100958  
100959  
100960  
100961  
100962  
100963  
100964  
100965  
100966  
100967  
100968  
100969  
100970  
100971  
100972  
100973  
100974  
100975  
100976  
100977  
100978  
100979  
100980  
100981  
100982  
100983  
100984  
100985  
100986  
100987  
100988  
100989  
100990  
100991  
100992  
100993  
100994  
100995  
100996  
100997  
100998  
100999  
1001000  
1001001  
1001002  
1001003  
1001004  
1001005  
1001006  
1001007  
1001008  
1001009  
1001010  
1001011  
1001012  
1001013  
1001014  
1001015  
1001016  
1001017  
1001018  
1001019  
1001020  
1001021  
1001022  
1001023  
1001024  
1001025  
1001026  
1001027  
1001028  
1001029  
1001030  
1001031  
1001032  
1001033  
1001034  
1001035  
1001036  
1001037  
1001038  
1001039  
1001040<br

**Attention dilution** In Section 3.1, we examine attention dilution with up to three attribute–object pairs (i.e., up to three concepts) and a limited set of prompts. We expand this to up to five concepts and construct 100 prompts for each concept cardinality. For example, a single-concept prompt is “a/an  $\langle$ attribute $\rangle$  $\langle$ object $\rangle$ ”, and a three-concept prompt is “a/an  $\langle$ attribute1 $\rangle$  $\langle$ object1 $\rangle$  and a/an  $\langle$ attribute2 $\rangle$  $\langle$ object2 $\rangle$  and a/an  $\langle$ attribute3 $\rangle$  $\langle$ object3 $\rangle$ ”. Following this template, we manually create prompts with assistance from a large language model such as GPT (Achiam et al., 2023). For each prompt, we measure the attention scores associated with the first attribute and the first object, and summarize the mean, minimum, maximum, and standard deviation in Figure 6. For both objects and attributes, the mean attention score decreases as the number of concepts increases, following an approximately logarithmic trend. Specifically, increasing the number of concepts from 1 to 2 in attributes reduces the mean from 0.93301 to 0.66255, a 29% reduction. Overall, the results indicate that within the cross-attention mechanism of text-to-image diffusion models, increasing the number of non-padding text tokens dilutes the information allocated to each token.

## B ABLATION ON HYPERPARAMETERS

We investigate the influence of two key hyperparameters in our framework: the EMA rate and the Percentile- $\beta$  rate. As shown in Figure 7, we conduct systematic variations of each hyperparameter while holding all other conditions fixed. This analysis highlights how different choices affect performance beyond the default setting and provides practical guidance for selecting stable operating ranges.

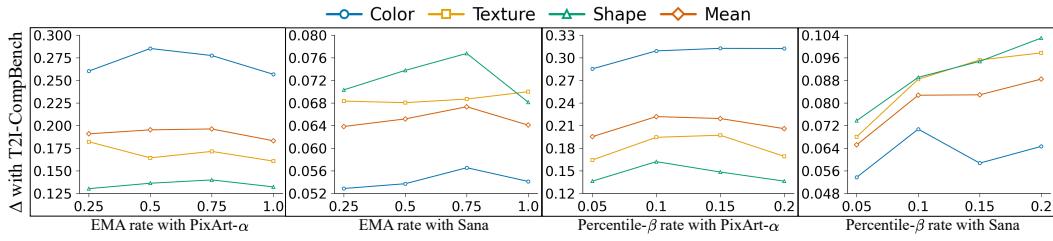
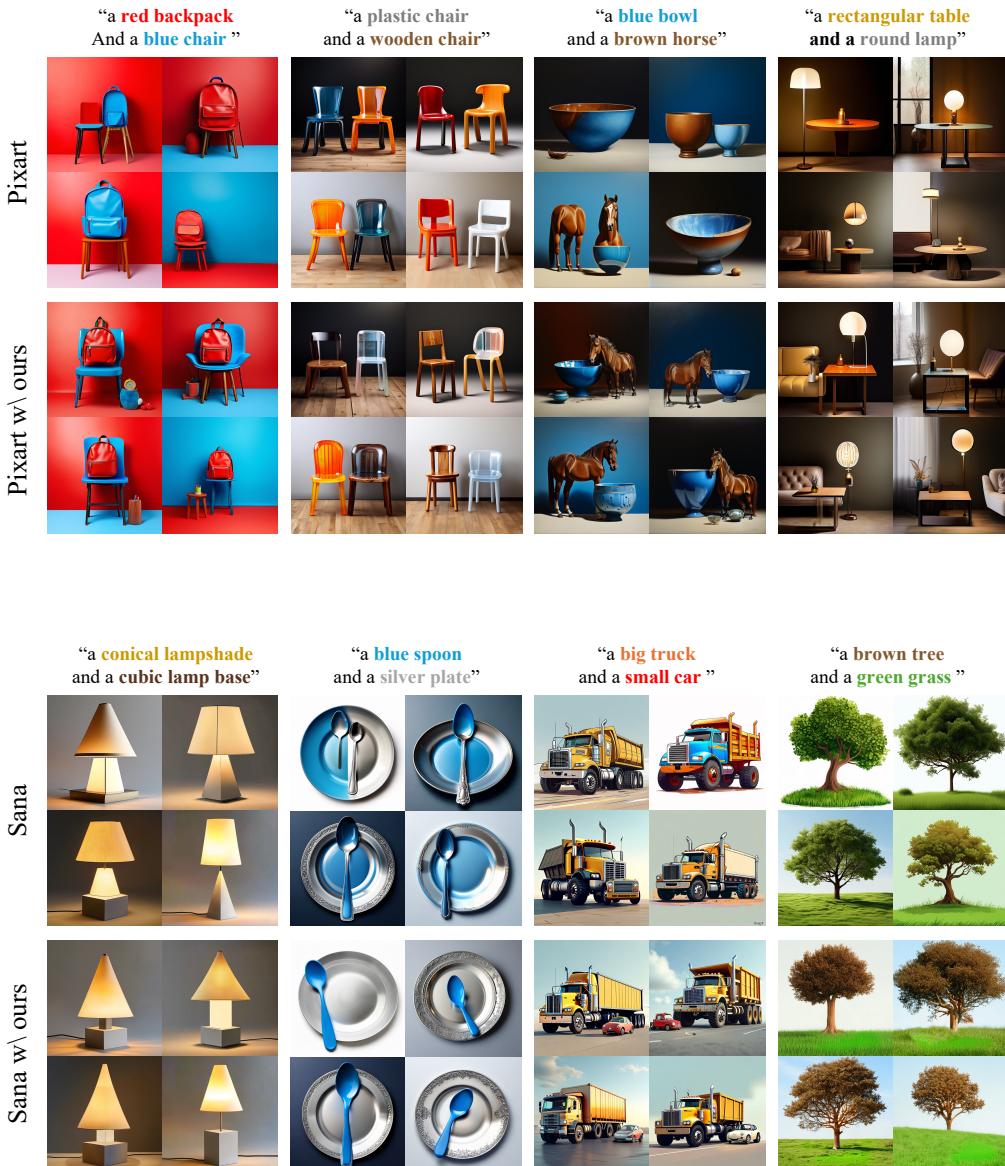


Figure 7: Effect of EMA rate and Percentile- $\beta$  rate on T2I-CompBench performance. The y-axis label indicates improvement over the corresponding base models. Results are shown for PixArt- $\alpha$  and Sana across Color, Texture, Shape, and their Mean performance.

**EMA rate.** We investigate the effect of the EMA rate while fixing the percentile- $\beta$  parameter at 0.05, varying the EMA rate over 0.25, 0.5, 0.75, 1.0. The EMA rate controls the sensitivity of the moving average: lower values dampen responsiveness to recent cross-attention signals, whereas higher values place greater weight on them. An EMA rate of 1.0 corresponds to using only the current cross-attention value (i.e., no averaging). Except for Sana in shape, both models are robust to the choice of EMA rate, showing minimal difference across categories. Overall, these findings indicate that moderate EMA rates provide the most effective balance, supporting better semantic alignment while avoiding the limitations of excessively small or overly large values.

**Percentile- $\beta$  rate.** We investigate the effect of the percentile- $\beta$  rate by fixing the EMA rate to 0.5 and varying the percentile- $\beta$  rate across 0.05, 0.1, 0.15, 0.2. The percentile- $\beta$  rate specifies a threshold for selecting indices from the EMA cross-attention weights of each text token. Based on this threshold, indices with the highest attention values are chosen in descending order. Higher percentile- $\beta$  rates result in more indices being extracted per token, while lower rates restrict the selection to fewer indices. PixArt- $\alpha$  achieves its strongest results on average at 0.1 and 0.15, with a slight decline at 0.2, while Sana continues to improve consistently except for the color category. Since larger percentile- $\beta$  rates entail extracting more indices per token and thus increase computational cost, we adopt 0.05 as the default setting.

Overall, the ablation results demonstrate that both hyperparameters yield consistent benefits and that our framework maintains stable performance without demanding fine-grained tuning. The integration of EMA and percentile- $\beta$ , therefore, proves essential for mitigating semantic misalignment.

756 C ADDITIONAL QUALITATIVE RESULTS  
757758 We provide additional samples below. Figure 8 demonstrates improved performance over the base  
759 model, and Figure 9 shows that our method mitigates semantic misalignment more effectively than  
760 competing models.  
761800 Figure 8: Additional semantically aligned images generated by our method, applied to the PixArt-  
801  $\alpha$  (top two rows) (Chen et al., 2023) and SANA (bottom two rows) (Xie et al., 2024) base models.  
802 Results are shown for two random seeds for each baseline. Zooming in is recommended for a detailed  
803 view.  
804  
805  
806  
807  
808  
809

810

811

812

813

814

815

816

817

818

819

820

821

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837

838

839

840

841

842

843

844

845

846

847

848

849

850

851

852

853

854

855

856

857

858

859

860

861

862

863



Figure 9: Additional quantitative comparison of our method against competing approaches. Results are shown for two random seeds per prompt. Zooming in is recommended for a detailed view.

864 **D ETHICS STATEMENT**  
865

866 Following ICLR 2026 guidelines, we disclose that a Large Language Model (LLM) was utilized  
867 for assistance with grammar correction, text polishing, and the generation of prompts for additional  
868 experiments. All research contributions, experimental results, and scientific claims are entirely the  
869 work and responsibility of the authors.

870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917