

INFERENCE-TIME CLIP EMBEDDING MANIPULATION FOR COMPOSITIONAL TEXT-TO-IMAGE ALIGNMENT

Anonymous authors

Paper under double-blind review

ABSTRACT

Text-to-image (T2I) diffusion models have advanced notably, but still fail to satisfy prompt conditions, resulting in attribute mismatches and missing objects. These errors normally fall into two categories: *concept loss*, when an object or attribute (i.e., concept) goes missing, and *concept confusion*, when an attribute is assigned to the wrong object or multiple objects blend together. Recent studies suggest that these challenges are caused by the limited capacity of CLIP text encoder to capture fine-grained semantic details. Although several methods have been proposed, concept loss and concept confusion persist in recent T2I models when handling multi-object, multi-attribute prompts. In this paper, we conduct an embedding-level analysis to develop an inference time, model independent solution, addressing concept loss and concept confusion. We find that (1) concepts mentioned later exhibit higher embedding entropy, indicating higher uncertainty and making them more vulnerable to concept loss, and (2) the first CLIP attention layer captures the strength of binding between each object and its attribute. Guided by our findings, we introduce TIE, a method that improves semantic alignment through a single text-embedding update. TIE addresses concept loss via entropy-aware singular value amplification and resolves concept confusion through interpolation–extrapolation binding based on CLIP attention scores, all in a training-free manner. Extensive experiments demonstrate that TIE enhances semantic fidelity in multi-concept scenarios with minimal sampling overhead.

1 INTRODUCTION

Text-to-image (T2I) generation has advanced substantially through the development of diffusion models Podell et al. (2023); Rombach et al. (2022); Esser et al. (2024); Saharia et al. (2022), achieving notable performance in generating images from general prompts Rombach et al. (2022); Saharia et al. (2022); Ramesh et al. (2022). However, they continue to struggle with prompts involving multiple objects and attributes (e.g., color, shape), often resulting in semantic inconsistencies Ramesh et al. (2022); Clark & Jaini (2023). For example, given a simple prompt "a brown boat and a blue cat", the generated image sometimes contains only a brown boat, or falsely depicts a brown cat and a blue boat. As prompt complexity increases, such failures occur more frequently and manifest in more varied forms.

In text-to-image (T2I) synthesis, the text prompt itself specifies the desired constraints on the output, in which objects should appear and which attributes should be bound to each object. Failures in controlled generation can therefore be viewed as violations of these prompt-induced constraints. For clear understanding, these failures can be classified into two categories: *concept loss* and *concept confusion*, as illustrated in Fig. 1. *Concept loss* occurs when some of the intended objects or attributes specified in the prompt are missing from the generated image. This concept loss includes cases where objects appear without their corresponding attributes (Fig. 1a) or where an object is entirely omitted (Fig. 1b). In contrast, *concept confusion* arises when the binding between attributes

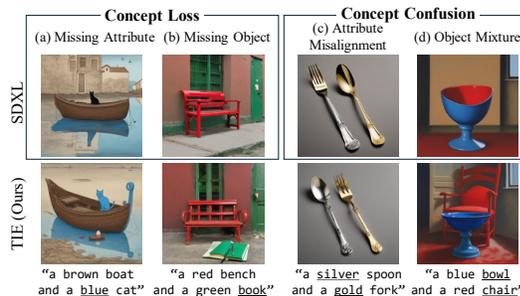


Figure 1: Failure cases in simple prompts.

054 and objects are incorrectly aligned. Such misalignment leads attributes to be assigned to unrelated
055 objects (Fig. 1c) or merges objects (Fig. 1d). These failures indicate that even large-scale T2I models
056 still struggle with multi-object and multi-attribute prompts.

057 Several studies Podell et al. (2023); Chefer et al. (2023); Feng et al. (2022); Liu et al. (2022) have
058 attributed the failures to the limitations of text embeddings used in T2I models. Although the widely-
059 used CLIP text encoder Radford et al. (2021) provides rich textual representations, it often struggles
060 to capture fine-grained details Yuksekogonul et al. (2022); Tang et al. (2023); Zarei et al. (2025); Jing
061 et al. (2024). To address this limitation, previous works have proposed methods to improve text em-
062 beddings suitable for T2I synthesis Radford et al. (2021). Some approaches merge tokens Hu et al.
063 (2024a) or adaptively bind embeddings Zhuang et al. (2024) to mitigate concept confusion. Others
064 optimize embeddings iteratively Chen et al. (2024) or utilize additional supervision with prompt
065 parsing Feng et al. (2022) to tackle concept loss. However, existing methods address only part of the
066 problem by focusing on *either* concept loss or concept confusion, often relying on precomputation
067 or additional supervision Feng et al. (2022); Hu et al. (2024a); Zhuang et al. (2024). In this work,
068 we perform a concept-level analysis of CLIP text encoder and its embeddings to devise a general
069 and efficient solution to concept loss and concept confusion.

070 To address these challenges, we identify two indicators of concept loss and concept confusion. First,
071 to address concept loss, we observe a positional bias in which concepts appearing later in the prompt
072 are more likely to be omitted in the generated image. Based on this finding, we look for a quantitative
073 measure which estimates the likelihood of concept loss. Through various observations, we find that
074 *token entropy*, the entropy of each token embeddings, reflects the degree of uncertainty and serves as
075 a quantitative indicator of concept loss. Second, to mitigate concept confusion, we look into the first
076 layer attention scores of the CLIP text encoder, referred to as *L0 attention scores*. We demonstrate
077 that L0 attention scores effectively show the strength of binding between concept tokens and how
078 much it should be strengthened to reduce concept confusion. These two indicators allow to predict
079 the chance of *both* concept loss and concept confusion using only the internal mechanisms of the
080 CLIP text encoder.

081 Based on our findings, we propose two methods to address the limitations of T2I models: Adaptive
082 Embedding Preservation (AEP) for mitigating concept loss, and Interpolation-Extrapolation Bind-
083 ing (IEB) for resolving concept confusion. AEP adaptively amplifies tokens according to token
084 entropy to encourage their presence in the generated image. IEB enhances the semantic binding
085 strength between objects and their corresponding attributes based on the L0 layer, while simultane-
086 ously suppressing the binding strength of irrelevant attribute-object pairs. We refer to the combined
087 framework of AEP and IEB as **TIE**. TIE improves text-image semantic alignment across multi-
088 attribute and multi-object scenarios. In addition, our model updates embeddings in a single step
089 and in a training-free manner, making it applicable to existing T2I models and enabling reliable
090 generation.

090 Our main contributions are as follows:

- 091 • We identify two indicators that jointly address concept loss and concept confusion. First,
092 the token entropy reflects uncertainty and serves as a cue for preventing concept loss. Sec-
093 ondly, the L0 attention score effectively captures attribute-object binding and serves as a
094 indicator for how much the binding should be adjusted to address concept confusion.
- 095 • We propose TIE, a post-training, inference-time framework that enhances text-image align-
096 ment via a single-step embedding update using AEP and IEB. AEP amplifies concept to-
097 kens with high embedding entropy to prevent concept loss, while IEB adjusts inter-concept
098 relations using CLIP attention to clarify associations between each object and its attribute,
099 thereby addressing concept confusion.
- 100 • Experimentally, TIE improves prompt-constraint satisfaction on multi-attribute and multi-
101 object benchmarks with minimal inference-time overhead compared to baselines.

103 2 RELATED WORKS

104 Text-to-image (T2I) diffusion models have advanced remarkably, enabling the generation of realistic
105 images from a given prompt Podell et al. (2023); Rombach et al. (2022); Esser et al. (2024); Zhang
106 et al. (2023). To generate text conditional images, T2I diffusion models integrate embeddings from
107

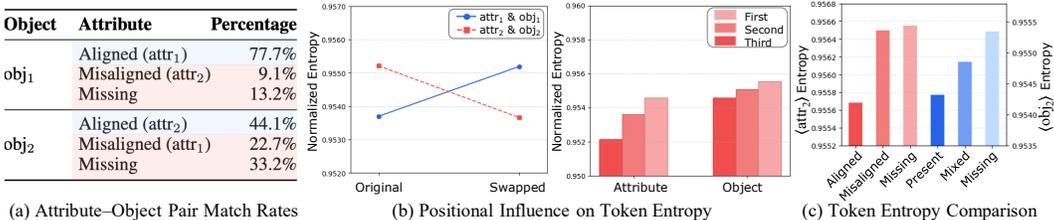


Figure 2: **Analysis of positional bias and token entropy on concept loss.** (a) When issuing the prompt "a <attr₁> <obj₁> and a <attr₂> <obj₂>", <attr₂> and <obj₂> is far more prone to being misaligned or missing. (b) Left: The attribute-object pair mentioned later in the prompts shows higher entropy than the pair mentioned earlier, regardless of the tokens themselves. Right: Same holds for 3-pair prompts. (c) Comparison of entropy for <attr₂> (red) and <obj₂> (blue) tokens each for when <attr₂> is Aligned, Misaligned or Missing, and when <obj₂> is Present, Mixed, or Missing, demonstrating that higher entropy predicts concept loss.

a text encoder into the denoising network through mechanisms such as cross-attention Saharia et al. (2022); Ramesh et al. (2022); Li et al. (2023); Nichol et al. (2022); Chen et al. (2023); Balaji et al. (2022). However, they often struggle to preserve the semantic alignment between prompts and generated images when multiple objects or attributes are involved Ramesh et al. (2022); Clark & Jaini (2023); Leivada et al. (2022); Rassin et al. (2022).

Latent-based Methods. To address this fundamental challenge, recent studies propose refining the cross-attention mechanism. A&E Chefer et al. (2023) is the first work to propose a technique for modifying cross attention during the sampling process. SynGen Rassin et al. (2023) and EBAMA Zhang et al. (2024) enhance attribute and object correspondence through attention alignment and energy-based conditioning. InitNO Guo et al. (2024) refines the initial noise distribution to improve semantic precision, while CONFORM Meral et al. (2024) employs contrastive learning techniques to achieve higher fidelity in the generated images. However, since these methods are applied during the iterative sampling process of the diffusion model, they come with a practical drawback of increased latency.

Text Embedding-based Methods. Another line of work aims to improve the CLIP text embeddings Radford et al. (2021), which is widely used but tends to capture broad semantic meanings while missing fine-grained details Yuksekogonul et al. (2022); Tang et al. (2023). In addition, it is necessary to address the problem that the CLIP text encoder is biased toward objects mentioned earlier in a sentence Abbasi et al. (2025b;a). Magnet Zhuang et al. (2024) explores modifying only the text embeddings to enhance semantic coherence. ToMe Hu et al. (2024a) and TeeMo Seo et al. (2025) further refine embedding representations by investigating token merging and geometric properties, respectively. However, such embedding-based approaches tend to show weaker text-image alignment compared to cross-attention-based ones. In this paper, we analyze the characteristics of the commonly used CLIP text encoder and its embeddings, and explore the potential of embedding-based methods.

3 ANALYSIS OF CONCEPT LOSS AND CONCEPT CONFUSION

CLIP text encoder Radford et al. (2021), which is commonly used in text-to-image (T2I) models, provides rich textual representations but often struggles to capture fine-grained details Yuksekogonul et al. (2022); Tang et al. (2023). This limitation often leads to concept loss and concept confusion in T2I tasks Zarei et al. (2025); Jing et al. (2024); Meral et al. (2024). To address this, we analyze the CLIP text encoder and identify two proxy signals of these issues: *token entropy* as a proxy for concept loss (Section 3.1) and *L0 attention score* as a proxy for concept confusion (Section 3.2). For clarity and controlled analysis, we first study these proxies using simple multi-attribute prompts. Importantly, since the proposed proxies are computed solely from the internal representations of the text encoder, they naturally extend to longer and syntactically more complex prompts as well.

Preliminary. T2I diffusion models Podell et al. (2023); Rombach et al. (2022) are typically composed of a text encoder and a denoising network. Specifically, we consider models that adopt the CLIP text encoder primarily used in the Stable Diffusion series Podell et al. (2023); Rombach et al.

(2022); Esser et al. (2024). The CLIP text encoder τ produces text embeddings from the input prompt \mathcal{P} through the causal mask and self-attention layers, where each token embedding c_i can only attend to the embeddings of previous or current tokens c_j for all $j \leq i$. The resulting text embedding is represented as $C = \tau(\mathcal{P}) \in \mathbb{R}^{N \times M}$, where N denotes the number of tokens and M is the dimension of each token embedding. To accommodate variable-length prompts, padding is applied to standardize the length N to a fixed value of tokens.

3.1 TOKEN ENTROPY REFLECTS CONCEPT LOSS

Understanding when concept loss occurs is a necessary first step toward addressing it. Previous work Chen et al. (2024) shows that objects mentioned later in a prompt, referred to as later objects, are more likely to disappear from the generated image. This object-level *positional bias* is driven by the causal mask of the CLIP text encoder, thereby making it difficult for later objects to appear entirely in the generated image. In this case, since the causal mask influences not only objects but also attributes, we first investigate whether positional bias also arises with respect to attributes. We then verify that *token entropy* exhibits positional bias and, furthermore, serves as an indicator of concept loss.

Positional Bias in Attribute Tokens. First, to test whether attributes suffer the same positional bias as objects, we sample captions of the form "a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ and a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ " from Concept Conjunction 500 (CC-500) dataset Feng et al. (2022) and retain 220 images in which both objects are clearly present. Then, for each image, we manually check whether the corresponding attribute of each object is (i) correct (Aligned), (ii) from another object (Misaligned), or (iii) Missing. As illustrated in Fig. 2a, we reveal that the attribute-object pair mentioned later in the prompt has a higher rate of missing or misaligned attributes compared to the pair mentioned earlier. Therefore, positional bias manifests in attributes as it does in objects, and concepts (i.e., objects and attributes) mentioned later in the prompt are more likely to undergo concept loss. However, the presence of such positional bias alone does not guarantee that concept loss will occur in a given sample, so we explore the entropy of each text embedding as a potential predictor of concept loss.

Higher Entropy Observed in Later Concepts. Next, to estimate the chances of concept loss, we analyze the token entropy \mathcal{H}_i of the token embedding $c_i \in \mathbb{R}^d$ derived from the CLIP text encoder:

$$p_{i,j} = \frac{|c_{i,j}|}{\sum_{k=1}^d |c_{i,k}|}, \quad \mathcal{H}_i = -\frac{1}{\log d} \sum_{j=1}^d p_{i,j} \log p_{i,j}. \quad (1)$$

Note that a low token entropy indicates that the token embedding is sparse, with information concentrated in a few dimensions and thus representing more distinctive features. In contrast, a high token entropy implies that the embedding is dense, with information distributed across many dimensions, resulting in broader and more general representations. We now evaluate whether token entropy can serve as an indicator of concept loss by first examining whether it exhibits positional bias. To this end, we compute token entropy on two semantically equivalent prompts with reversed orders: "a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ and a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ " and "a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ and a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ ". In addition, for prompts containing three $\langle \text{attr} \rangle$ - $\langle \text{obj} \rangle$ pairs, we synthetically generate 75 prompts and measure the token entropy for both attributes and objects according to their positions in the prompt. Fig. 2b shows that the average token entropy is consistently higher for later-mentioned $\langle \text{attr} \rangle$ and $\langle \text{obj} \rangle$, indicating the presence of positional bias in token entropy. Building on the observation that token entropy is consistently higher for later-mentioned concepts, we next examine how strongly it correlates with actual concept loss in generated images.

Token Entropy and Concept Loss. Finally, we investigate whether higher token entropy corresponds to concept loss in the generated images. Using the experimental setting from Fig. 2a, we first compare the average token entropy of the later attribute $\langle \text{attr}_2 \rangle$ across the three types—Aligned, Misaligned, and Missing (Fig. 2c). Using different samples from the same dataset, we also measure the token entropy of the later object $\langle \text{obj}_2 \rangle$, for when the object is Present, Mixed or Missing. We observe that for both attributes and objects, the token entropy is higher when concepts appear incorrectly (Misaligned, Missing or Mixed) compared to the when they are correctly Aligned or Present. This result suggests that higher token entropy is indicative of concept loss for both attributes and objects. Building on this finding, we propose a method to mitigate concept loss in Section 4.1. For more details, please refer to the appendix.

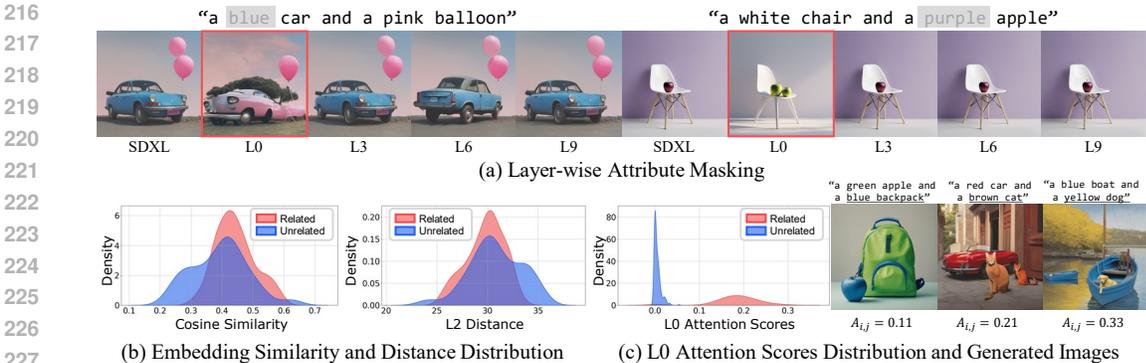


Figure 3: **Analysis of text embedding and self-attention in CLIP text encoder.** (a) Layer-wise attribute masking shows attribute binding primarily occurs at layer 0 (L0). (b) Cosine similarity and L2 distance between related and unrelated embeddings show no clear separation. (c) L0 self-attention clearly separates related from unrelated tokens. Underlined prompt words mark the corresponding concepts, showing that higher attention indicates stronger concept binding in the generated images.

3.2 TOKEN RELATIONSHIPS IN FIRST LAYER SELF-ATTENTION

To identify potential concept confusion, it is crucial to assess the attribute-object binding strength within the CLIP text encoder. Hereafter, a *related pair* is defined as a correct attribute-object association (e.g., $\langle attr_1 \rangle - \langle obj_1 \rangle$), while an *unrelated pair* refers to an incorrect association with an object from a different concept (e.g., $\langle attr_1 \rangle - \langle obj_2 \rangle$). We find that the *L0 attention scores*, the first self-attention score from the CLIP text encoder, reflects the relationship (related or unrelated) between $\langle attr \rangle$ and $\langle obj \rangle$.

Similarity Metrics between Embeddings. Initially, we analyze conventional similarity metrics between their embeddings to determine whether they can distinguish related attribute-object pairs. We compare the related and unrelated attribute-object pairs using the A&E dataset Chefer et al. (2023), where each prompt follows the form "a $\langle attr_1 \rangle$ $\langle obj_1 \rangle$ and a $\langle attr_2 \rangle$ $\langle obj_2 \rangle$ ". Fig. 3b shows the cosine similarity and L2 distance between the $\langle attr \rangle$ and $\langle obj \rangle$ embeddings. The distributions of these similarity metrics for related and unrelated pairs exhibit no significant difference, making it difficult to predict the relationship between $\langle attr \rangle$ and $\langle obj \rangle$ based on embedding similarity alone.

Layer-wise Self-Attention Analysis for Binding. Inspired by prior works that each layer in a Transformer-based model serves a different role Vig & Belinkov (2019); Vig (2019); Clark et al. (2019), we investigate the self-attention mechanism of the Transformer-based CLIP text encoder. Specifically, we conduct an analysis of layers by selectively masking an attribute token at each self-attention layer. Fig. 3a shows the generated images when the attribute key token is masked in the n -th-layer (L_n) self-attention. Interestingly, we observe that the attribute disappears from the generated image only when the self-attention in the first layer (L0) is masked. Additional qualitative examples of all layers are provided in the appendix.

L0 Attention Scores for Binding Strength. We now turn to analyzing the attention scores of this L0 layer to investigate the relevance of the L0 attention scores to the concept relationships between $\langle attr \rangle$ and $\langle obj \rangle$. We first compare the L0 attention scores between $\langle attr \rangle$ and $\langle obj \rangle$ tokens for related pairs and unrelated pairs. As shown in Fig. 3c (left), the distributions of related and unrelated attribute-object pairs are clearly separated according to their L0 attention scores. Furthermore, based on visual inspection of generated images across different attention score levels, as shown in Fig. 3c (right), higher L0 attention scores correspond to stronger semantic binding between $\langle attr_2 \rangle$ and $\langle obj_2 \rangle$. Thus, L0 attention scores serve as a reliable indicator of whether concept confusion has occurred. Please refer to the appendix for additional examples.

4 METHODS

Our method is motivated by the two key indicators derived from Section 3: *token entropy* for concept loss and *L0 attention score* for concept confusion. We improve T2I generation in multi-object

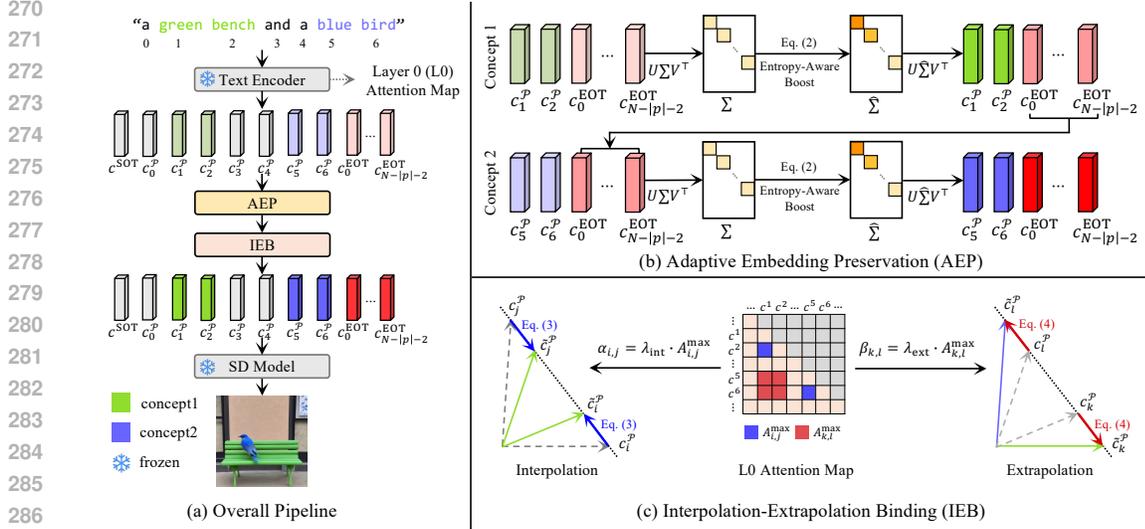


Figure 4: **Overview of the proposed TIE framework.** (a) Overall pipeline of TIE. (b) Adaptive Embedding Preservation (AEP), which boosts singular values of text embeddings based on entropy to address concept loss. (c) Interpolation-Extrapolation Binding (IEB), which leverages initial self-attention scores to mitigate concept confusion.

and multi-attribute scenarios by integrating two complementary components as shown in Fig. 4. To address concept loss, we first implement Adaptive Embedding Preservation (AEP), which exploits the tendency of tokens with higher token entropy. After recovering lost concepts, we mitigate concept confusion through Interpolation Extrapolation Binding (IEB), which estimates concept binding strength based on L0 attention scores.

4.1 ADAPTIVE EMBEDDING PRESERVATION

To preserve concepts that are more prone to concept loss using individual token embeddings, we introduce AEP. This method selectively amplifies concept token embeddings with end-of-text (EOT) tokens using entropy-aware singular value decomposition (SVD). Prior work has shown that EOT tokens also encode semantic information crucial for concept composition Zhuang et al. (2024); Toker et al. (2025). Thus, our AEP jointly strengthens explicit concept token embeddings and EOT tokens according to their entropy, successfully mitigating concept loss.

Singular Value Decomposition for Concepts. SVD decomposes a matrix into orthogonal directions ordered by singular values, concentrating semantic information into its leading components. Motivated by previous works Li et al. (2024); Liu et al. (2025), we leverage SVD to amplify singular values in concept token embeddings along with EOT tokens, thereby concentrating information for tokens with higher entropy to mitigate concept loss.

For simplicity, we explain with prompts in which only one attribute is paired with each object. For each attribute-object pair, we build the token-related matrix $X := [c_k^{\text{attribute}}, c_k^{\text{object}}, c_0^{\text{EOT}}, \dots, c_{N-|P|-2}^{\text{EOT}}]$, where c_k^{concept} is the embedding of the k -th pair (Fig. 4). We first gather the token embeddings $c = \{c^{\text{SOT}}, c_0^p, \dots, c_{|P|-1}^p, c_0^{\text{EOT}}, \dots, c_{N-|P|-2}^{\text{EOT}}\}$, where c^{SOT} is the start-of-text (SOT) token embeddings, c_i^p is the i -th prompt token, and c_i^{EOT} denotes the i -th EOT token. Decomposing X gives $X = U\Sigma V^T$, $\Sigma = \text{diag}(\sigma_0, \sigma_1, \dots, \sigma_{n_0-1})$, $n_0 = \min(M, N - |P| - 1)$ where U and V are orthogonal matrices. Each singular value is directly rescaled as $\hat{\sigma}_i = \delta e^{\gamma \sigma_i}$, with γ and δ as positive constants, concentrating the information in the concept embedding. This formulation updates the text embedding vectors to progressively amplify each concept based on its position. Specifically, rescaling embedding vectors allows the information to be concentrated along a small set of dominant directions. These singular value reduces token entropy by suppressing noisy or redundant information. Thus, the entropy of the concept tokens decreases, indicating reduced uncertainty and preservation of concept information. However, pre-

venting concept loss requires a more adaptive mechanism that can predict the risk of omission for individual concepts.

Entropy-aware Boost. As shown in Section 3.1, token entropy indicates the chances of concept loss. Accordingly, we apply an entropy-aware scaling that adaptively assigns larger weights to higher-entropy tokens rather than applying a direct scaling. Thus, later tokens with higher entropy receive proportionally stronger amplification while keeping earlier tokens modestly adjusted. We compute token entropy \mathcal{H}_i as in Eq. equation 1 and averaging over each embedding in matrix X with $N - |\mathcal{P}|$ tokens as $\tilde{\mathcal{H}} = \frac{1}{N-|\mathcal{P}|} \sum_{i=1}^{N-|\mathcal{P}|} \mathcal{H}_i$. This makes our entropy-aware scaling

$$\hat{\sigma}_i = \delta e^{\gamma \sigma_i \tilde{\mathcal{H}}}. \quad (2)$$

Thus, the entropy-aware boosting deliberately adjusts the embeddings to prevent potential loss for each concept.

4.2 INTERPOLATION-EXTRAPOLATION BINDING

After addressing concept loss through AEP, we now propose the Interpolation–Extrapolation Binding (IEB) mechanism to resolve concept confusion, ensuring that the relationships between concepts are accurately reflected in the generated image, as illustrated in Fig. 4c. Inspired by prior studies indicating that CLIP text embeddings exhibit compositional structure Trager et al. (2023), we incorporate concept tokens to both attribute and object tokens bidirectionally to bring their embeddings closer. In contrast, we extrapolate unrelated concepts by pushing them further apart while preserving their embedding norms. To reflect binding strength between tokens, IEB leverages the *L0 attention score* introduced in Section 3.2.

Interpolation for Binding. First, we enhance the semantic binding between related attribute-object pairs in the embedding space. We apply an interpolation mechanism guided by their mutual self-attention scores. Let $(c_i^{\mathcal{P}}, c_j^{\mathcal{P}})$ denotes a token embedding pair, where $c_i^{\mathcal{P}}$ corresponds to the attribute and $c_j^{\mathcal{P}}$ to the object. Their updated embeddings are computed as:

$$\tilde{c}_i^{\mathcal{P}} = (1 - \alpha_{i,j}) \cdot c_i^{\mathcal{P}} + \alpha_{i,j} \cdot c_j^{\mathcal{P}}, \quad \tilde{c}_j^{\mathcal{P}} = (1 - \alpha_{i,j}) \cdot c_j^{\mathcal{P}} + \alpha_{i,j} \cdot c_i^{\mathcal{P}}, \quad (3)$$

where $\alpha_{i,j} = \lambda_{\text{int}} \cdot A_{i,j}$. Here, $A_{i,j}$ denotes the attention score that exists under the constraints of the causal mask. The constant hyperparameter λ_{int} controls the overall strength of the binding effect.

Extrapolation for Repulsion. After binding the related pairs, we apply extrapolation to suppress semantic interference between unrelated pairs. Let $(c_k^{\mathcal{P}}, c_l^{\mathcal{P}})$ denote the token embeddings of an unrelated token pair, where $c_k^{\mathcal{P}}$ corresponds to the attribute and $c_l^{\mathcal{P}}$ to the object. The updated embeddings are computed as:

$$\tilde{c}_k^{\mathcal{P}} = (1 + \beta_{k,l}) \cdot c_k^{\mathcal{P}} - \beta_{k,l} \cdot c_l, \quad \tilde{c}_l^{\mathcal{P}} = (1 + \beta_{k,l}) \cdot c_l - \beta_{k,l} \cdot c_k^{\mathcal{P}}, \quad (4)$$

where $\beta_{k,l} = \lambda_{\text{ext}} \cdot A_{k,l}$. The constant hyperparameter λ_{ext} modulates the strength of this repulsive adjustment. By first binding attribute–object pairs through interpolation and then repelling unrelated concepts through extrapolation, IEB sequentially reduces concept confusion in the generated image.

5 EXPERIMENT

Baselines. We compare our model, TIE, with baselines including A&E Chefer et al. (2023), SynGen Rassin et al. (2023), Magnet Zhuang et al. (2024), and ToMe Hu et al. (2024a), all of which are built on the SDXL Podell et al. (2023) backbone. For fair comparison, we adapt A&E and SynGen to the SDXL base model, since their original implementations are not designed for it. When using SD3.5 as the backbone, architectural differences prevent direct adaptations of prior methods, and thus we compare only against the SD3.5 baseline. Further details on the adaptation are provided in the appendix.

Datasets & Metrics. We employ the BLIP-VQA in T2I-CompBench Huang et al. (2023) evaluation. This metric utilizes BLIP to perform visual question answering, enabling fine-grained assessment of various attributes in the generated images. We follow the evaluation protocol of prior works Hu et al. (2024a); Seo et al. (2025); Feng et al. (2024); Hu et al. (2024b); Jiang et al. (2024) based on

Table 1: **Quantitative results of concept binding performance.** All methods are evaluated with their corresponding backbones. BLIP-VQA Huang et al. (2023) scores are measured on five categories: Color, Shape, Texture, Spatial, and Complex prompts. Inference time is compared across methods to assess efficiency. Best results are shown in **bold** and second-best results are underlined. For the 3.5 backbone, only **bold** formatting is used due to the limited number of compared models.

Backbone	Method	Opt-free	Inference Time (s) ↓	BLIP-VQA ↑				
				Color	Shape	Texture	Spatial	Complex
SDXL			9.82	0.5940	0.5068	0.5961	0.6249	0.4437
	A&E _{XL}	✗	62.72	0.5934	<u>0.5082</u>	<u>0.6040</u>	0.6230	0.4434
	SynGen _{XL}	✗	22.42	0.5948	0.5059	0.5955	0.6258	0.4426
	ToMe	✗	42.73	0.5161	0.3551	0.5621	–	0.2818
	Magnet	✓	10.66	0.6833	0.5076	0.5963	0.6239	0.4507
	+TIE (Ours)	✓	<u>11.86</u>	0.7004	0.5632	0.6703	0.6320	0.4577
SD3.5			15.62	0.7766	0.6117	0.7518	0.6837	0.5050
	+TIE (Ours)	✓	17.13	0.7796	0.6232	0.7551	0.6917	0.5168

Table 2: **Ablation study on T2I-CompBench.** The results demonstrate that AEP and IEB contribute significantly to concept binding, both individually and when combined for the SD3.5 backbone.

Exp. #	AEP	IEB	BLIP-VQA ↑				
			Color	Shape	Texture	Spatial	Complex
1	✗	✗	<u>0.7766</u>	0.6117	0.7518	0.6837	0.5050
2	✓	✗	0.7668	<u>0.6171</u>	0.7542	0.6910	0.5108
3	✗	✓	0.7756	0.6152	<u>0.7543</u>	0.6951	<u>0.5152</u>
4 (Ours)	✓	✓	0.7796	0.6232	0.7551	<u>0.6917</u>	0.5168

Color, Shape, and Texture categories, and extend it with Complex and Spatial categories for more comprehensive evaluation. Inference time is measured on an RTX A6000 GPU with a batch size of 1 and 50 timesteps. Details are provided in the appendix.

5.1 RESULTS

Quantitative Comparison. As shown in Tab. 1, we compare TIE with baselines built on the SDXL and SD3.5 backbones. For SDXL, our method achieves superior performance across all categories compared to other baseline models. We exclude ToMe Hu et al. (2024a) from the spatial category evaluation as it is specifically designed for attribute binding and does not incorporate spatial structure in its token merging process. For SD3.5, our method also demonstrates clear improvements over the baseline model. Since none of the comparison models are directly applicable to SD3.5, we restrict the comparison to the baseline. Notably, our method yields particularly large gains on challenging Complex prompts. In addition, we observe only marginal inference time overhead. Overall, these results demonstrate that our method effectively mitigates concept confusion and prevents concept loss with minimal overhead.

Qualitative Comparison. We qualitatively compare TIE with existing semantic binding methods, as shown in Fig. 5 and Fig. 6. The results demonstrate that TIE achieves improved semantic binding performance. In Fig. 5, applying our method to SDXL effectively alleviates concept loss while also improving spatial consistency compared to the baseline models. Fig. 6 further shows that TIE, which is built on SD3.5, consistently outperforms the baseline on Complex prompts involving multiple categories and concepts. These results highlight the enhanced robustness of TIE in preserving both attribute and object integrity during generation. Additional examples are provided in the appendix.

Ablation Study. We conduct an ablation study, as quantitatively shown in Tab. 2 for the SD3.5 backbone. By first addressing concept loss to preserve the attributes and objects in the input prompt and then establishing bindings among them, AEP and IEB serve complementary functions that are essential to the overall framework. Finally, our proposed method, TIE, demonstrates improved performance across the BLIP-VQA benchmarks. Additional ablation experiments are provided in the appendix.



446 **Figure 5: Qualitative comparison of TIE with SDXL-based models.**



466 **Figure 6: Qualitative comparison of TIE with SD3.5.**

467 6 CONCLUSION

468 Text-to-image (T2I) diffusion models struggle to generate images from prompts involving multiple
 469 objects and attributes, manifesting as concept loss and concept confusion. To address these chal-
 470 lenges, we conduct a detailed analysis of the inherent properties of CLIP text embeddings, providing
 471 a solution that applies across various model architectures through direct modification of the textual
 472 conditions. Our investigation reveals that tokens corresponding to later concepts exhibit higher en-
 473 tropy, increasing the risk of information loss. We further observe that early self-attention layers cap-
 474 ture clearer semantic bindings between concepts. Building on these insights, we propose Adaptive
 475 Embedding Preservation (AEP), which adjusts token embeddings based on their entropy by progres-
 476 sively amplifying their singular values. In addition, we introduce Interpolation-Extrapolation Bind-
 477 ing (IEB), which leverages early-layer attention to strengthen associations between related concepts
 478 while suppressing unrelated ones. Our approach mitigates both concept loss and concept confusion,
 479 enabling T2I diffusion models to generate high-quality images that fully reflect the input prompt

480 **Limitations.** Although our method mitigates concept confusion and concept loss by analyzing the
 481 intrinsic properties of the CLIP text encoder, several challenges still remain. In particular, our ap-
 482 proach does not fully overcome the inherent limitations of the text-to-image models. For instance,
 483 when the training data is biased, the generated images often reflect such bias to some extent. Ad-
 484 dressing these limitations through further analysis and extension of our method will lead to enhanced
 485 image generation.

REFERENCES

- 486
487
488 Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Ro-
489 hban, and Mahdieh Soleymani Baghshah. Analyzing clip’s performance limitations in multi-
490 object scenarios: A controlled high-resolution study, 2025a. URL [https://arxiv.org/](https://arxiv.org/abs/2502.19828)
491 [abs/2502.19828](https://arxiv.org/abs/2502.19828).
- 492 Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Ro-
493 hban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of
494 multi-object representation. *arXiv preprint arXiv:2502.19842*, 2025b.
- 495
496 Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika
497 Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-
498 to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*,
499 2022.
- 500 Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite:
501 Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on*
502 *Graphics (TOG)*, 42(4):1–10, 2023.
- 503
504 Chieh-Yun Chen, Chiang Tseng, Li-Wu Tsao, and Hong-Han Shuai. A cat is a cat (not a dog!): Un-
505 raveling information mix-ups in text-to-image encoders through causal analysis and embedding
506 optimization. *arXiv preprint arXiv:2410.00321*, 2024.
- 507 Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James
508 Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer
509 for photorealistic text-to-image synthesis, 2023.
- 510
511 Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero shot classifiers. *Advances*
512 *in Neural Information Processing Systems*, 36:58921–58937, 2023.
- 513
514 Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look
515 at? an analysis of bert’s attention. *arXiv preprint arXiv:1906.04341*, 2019.
- 516 Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam
517 Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers
518 for high-resolution image synthesis. In *Forty-first international conference on machine learning*,
519 2024.
- 520
521 Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato
522 Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for
523 compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022.
- 524
525 Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-
526 image diffusion for accurate instruction following. In *Proceedings of the IEEE/CVF Conference*
on Computer Vision and Pattern Recognition, pp. 4744–4753, 2024.
- 527
528 Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting
529 text-to-image diffusion models via initial noise optimization. In *Proceedings of the IEEE/CVF*
Conference on Computer Vision and Pattern Recognition, pp. 9380–9389, 2024.
- 530
531 Taihang Hu, Linxuan Li, Joost van de Weijer, Hongcheng Gao, Fahad Shahbaz Khan, Jian Yang,
532 Ming-Ming Cheng, Kai Wang, and Yaxing Wang. Token merging for training-free semantic
533 binding in text-to-image synthesis. *Advances in Neural Information Processing Systems*, 37:
534 137646–137672, 2024a.
- 535
536 Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models
537 with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024b.
- 538
539 Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A compre-
hensive benchmark for open-world compositional text-to-image generation. *Advances in Neural*
Information Processing Systems, 36:78723–78747, 2023.

- 540 Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu,
541 and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept
542 matching. *Advances in Neural Information Processing Systems*, 37:76177–76209, 2024.
- 543
544 Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Guoxing Yang, Wei Wei, Huiwen Zhao, and
545 Zhiwu Lu. FineCLIP: Self-distilled region-based CLIP for better fine-grained understanding.
546 In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL
547 <https://openreview.net/forum?id=nExI4FuKWD>.
- 548 Evelina Leivada, Elliot Murphy, and Gary Marcus. Dall-e 2 fails to reliably capture common syn-
549 tactic processes, 2022. URL <https://arxiv.org/abs/2210.12889>.
- 550 Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and
551 Jian Yang. Get what you want, not what you don’t: Image content suppression for text-to-image
552 diffusion models. *arXiv preprint arXiv:2402.05375*, 2024.
- 553
554 Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li,
555 and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023.
- 556 Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual
557 generation with composable diffusion models. In *European Conference on Computer Vision*, pp.
558 423–439. Springer, 2022.
- 559
560 Tao Liu, Kai Wang, Senmao Li, Joost van de Weijer, Fahad Shahbaz Khan, Shiqi Yang, Yaxing
561 Wang, Jian Yang, and Ming-Ming Cheng. One-prompt-one-story: Free-lunch consistent text-
562 to-image generation using a single prompt, 2025. URL [https://arxiv.org/abs/2501.](https://arxiv.org/abs/2501.13554)
563 13554.
- 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. In *Proceedings of the IEEE/CVF*
Conference on Computer Vision and Pattern Recognition, pp. 9005–9014, 2024.
- 567
568 Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob
569 Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and
570 editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on*
571 *Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804.
PMLR, 2022.
- 572
573 Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe
574 Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image
575 synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- 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.
8748–8763. PmLR, 2021.
- 580
581 Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-
582 conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- 583
584 Royi Rassin, Shauli Ravfogel, and Yoav Goldberg. DALLE-2 is seeing double: Flaws in word-to-
585 concept mapping in Text2Image models. In *Proceedings of the Fifth BlackboxNLP Workshop on*
586 *Analyzing and Interpreting Neural Networks for NLP*, pp. 335–345, Abu Dhabi, United Arab Emi-
587 rates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/
2022.blackboxnlp-1.28. URL [https://aclanthology.org/2022.blackboxnlp-1.](https://aclanthology.org/2022.blackboxnlp-1.28/)
28/.
- 588
589 Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik.
590 Linguistic binding in diffusion models: Enhancing attribute correspondence through attention
591 map alignment. *Advances in Neural Information Processing Systems*, 36:3536–3559, 2023.
- 592
593 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 confer-
ence on computer vision and pattern recognition*, pp. 10684–10695, 2022.

- 594 Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kam-
595 yar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Sal-
596 imans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image dif-
597 fusion models with deep language understanding, 2022. URL [https://arxiv.org/abs/](https://arxiv.org/abs/2205.11487)
598 [2205.11487](https://arxiv.org/abs/2205.11487).
- 599 Hoigi Seo, Junseo Bang, Haechang Lee, Joohoon Lee, Byung Hyun Lee, and Se Young Chun. On
600 geometrical properties of text token embeddings for strong semantic binding in text-to-image
601 generation. *arXiv preprint arXiv:2503.23011*, 2025.
- 602 Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the
603 concept association bias of vision-language models. In *Proceedings of the 2023 Conference on*
604 *Empirical Methods in Natural Language Processing*, pp. 14333–14348, 2023.
- 605 Michael Toker, Ido Galil, Hadas Orgad, Rinon Gal, Yoad Tewel, Gal Chechik, and Yonatan Be-
606 linkov. Padding tone: A mechanistic analysis of padding tokens in t2i models, 2025. URL
607 <https://arxiv.org/abs/2501.06751>.
- 608 Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Ste-
609 fano Soatto. Linear spaces of meanings: compositional structures in vision-language models. In
610 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15395–15404,
611 2023.
- 612 Jesse Vig. A multiscale visualization of attention in the transformer model. *arXiv preprint*
613 *arXiv:1906.05714*, 2019.
- 614 Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language
615 model. *arXiv preprint arXiv:1906.04284*, 2019.
- 616 Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and
617 why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint*
618 *arXiv:2210.01936*, 2022.
- 619 Arman Zarei, Keivan Rezaei, Samyadeep Basu, Mehrdad Saberi, Mazda Moayeri, Priyatham Kat-
620 takinda, and Soheil Feizi. Improving compositional attribute binding in text-to-image genera-
621 tive models via enhanced text embeddings, 2025. URL [https://arxiv.org/abs/2406.](https://arxiv.org/abs/2406.07844)
622 [07844](https://arxiv.org/abs/2406.07844).
- 623 Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image
624 diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*
625 *(ICCV)*, pp. 3836–3847, October 2023.
- 626 Yasi Zhang, Peiyu Yu, and Ying Nian Wu. Object-conditioned energy-based attention map align-
627 ment in text-to-image diffusion models. In *European Conference on Computer Vision*, pp. 55–71.
628 Springer, 2024.
- 629 Chenyi Zhuang, Ying Hu, and Pan Gao. Magnet: We never know how text-to-image diffusion mod-
630 els work, until we learn how vision-language models function. *arXiv preprint arXiv:2409.19967*,
631 2024.
- 632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

Appendix

A BROADER IMPACTS

Our research on TIE aims to enhance semantic fidelity in complex, multi-concept prompts, enabling more accurate alignment between text and image. This improvement supports a wide range of applications, including educational visualizations, assistive design tools, and creative authoring for artists and storytellers. At the same time, by making generative models more reliable, we also lower the barrier to malicious uses such as highly realistic deepfakes for disinformation, fraud, and privacy or copyright infringement. Even when functioning as intended, subtle misbindings (e.g., swapping or omitting attributes) in critical contexts such as medical illustrations or legal diagrams could mislead users, and erroneous outputs may go unnoticed in black-box deployments.

B IMPLEMENTATION DETAILS

Analysis and Method Details. We conduct all experiments on a single NVIDIA RTX A6000 GPU. Our TIE builds upon SDXL Podell et al. (2023) and SD3.5 Esser et al. (2024). SDXL relies on CLIP text embeddings, which enables clear analysis of the effects of CLIP text embedding manipulation. Therefore, we use SDXL as the primary backbone in our analysis. SD3.5 employs multiple text encoders, including OpenCLIP ViT-G, CLIP ViT-L, and T5-XXL. In this setting, we apply our method only to the CLIP text embeddings. We also employ the Stanza package for automatic parsing.

Hyperparameters. We set the hyperparameters as follows: $\delta = 1.2$, $\gamma = 0.001$, $\lambda_{\text{int}} = 0.1$, and $\lambda_{\text{ext}} = 0.2$. Each hyperparameter is validated through Section 5 and in the Section E. In addition, we set the number of inference steps to 50 and the guidance scale to 7.5. We also evaluated our method on other diffusion models, using different hyperparameters to align with each text embedding extraction method. For SD3.5 Esser et al. (2024), which utilizes a T5 encoder with two CLIP encoders, we adjust only the CLIP text embedding and set $\gamma = 0.0002$, $\lambda_{\text{int}} = 0.3$, and $\lambda_{\text{ext}} = 0.1$ with 28 inference steps and 3.5 guidance scale. To mitigate the influence of previously processed concepts, we subtract 0.5 and 0.8 times the embeddings of mean prior concepts from subsequent concept embeddings in TIE based on SDXL Podell et al. (2023) and SD3.5 Esser et al. (2024), respectively.

C DATASETS

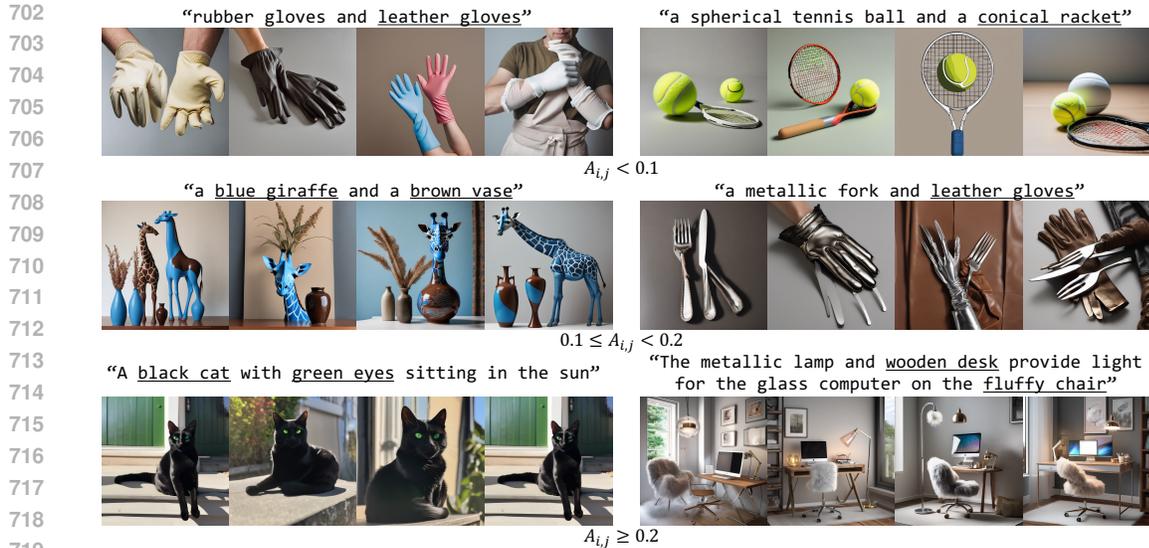
We use four types of datasets for our analysis and experiments.

Synthetic Dataset. We conduct our token-entropy analysis (Section 3.1b) on synthetic prompts sampled from the A&E Chefer et al. (2023) dataset. The left plot used 64 prompts of the object category from the A&E dataset and switched position of the attribute and objects, to compare the entropy of the tokens when the prompt takes the form of "a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ and a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ " with when it takes the form of "a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ and a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ ". For the right plot, we sampled attributes and objects from the same 64 prompts and made prompts in the form of "a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ and a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ and a $\langle \text{attr}_3 \rangle \langle \text{obj}_3 \rangle$ ".

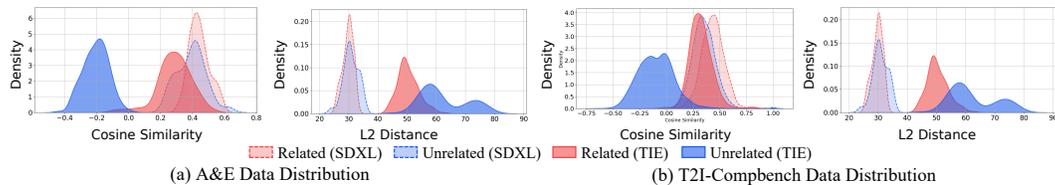
A&E. For Section 3.2, we use the A&E Chefer et al. (2023) dataset consisting of 64 prompts across various object categories. Each prompt follows the format: "a $\langle \text{attr}_1 \rangle \langle \text{obj}_1 \rangle$ and a $\langle \text{attr}_2 \rangle \langle \text{obj}_2 \rangle$ ". To further investigate with more diverse sentence structures and datasets, we additionally conduct experiments as shown in Section D.1.

T2I-Compbench. We evaluate the attribute types of Color, Shape, and Texture, Spatial and Complex using the T2I CompBench Huang et al. (2023). Specifically, Complex category consists of prompts containing more than two attributes and objects and involving at least two of the following subcategories: Color, Shape, Texture and object relationships. Following prior works Hu et al. (2024a); Seo et al. (2025); Feng et al. (2024); Hu et al. (2024b); Jiang et al. (2024), we randomly select 60 long prompts and 240 short prompts for each category, resulting in 300 prompts per attribute type. For each prompt, we generate images using five different random seeds for evaluation.

Concept Conjunction 500 (CC-500). Each prompt consists of two subjects, with one attribute assigned to each. Following the setup in Zhuang et al. (2024), we categorize the prompts into three



720 Figure 8: **L0 attention scores and generated images.** The underline indicates the concept corresponding to each L0 attention score range.



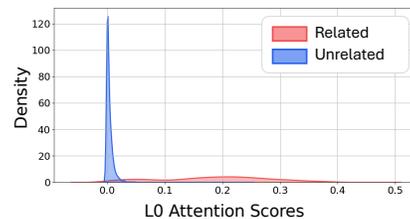
730 Figure 9: **Comparison of traditional similarity metrics between SDXL Podell et al. (2023) and TIE**

733 types: two living subjects, one living and one non-living subject, and two non-living subjects. We randomly select 80 prompts per category, resulting in 240 prompts, and generate 10 images for each.

736 D ADDITIONAL ANALYSIS OF THE CLIP TEXT ENCODER

738 D.1 L0 ATTENTION SCORES ANALYSIS

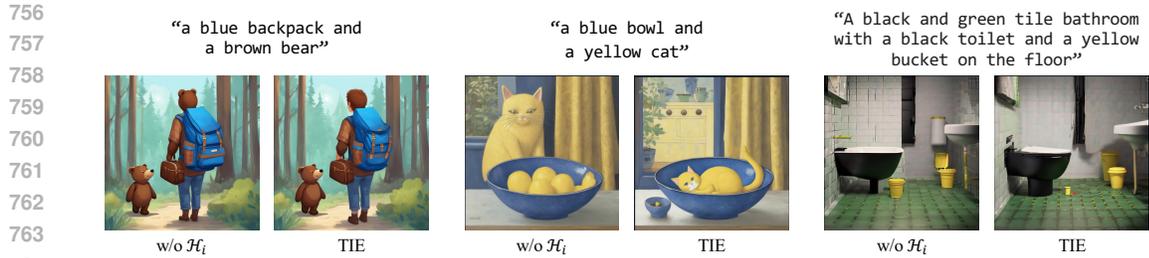
739 **Masking L0 Attention Scores.** We find that L0 attention scores play a critical role in attribute binding across various T2I diffusion models in Section 3.2. To investigate the consistency of this behavior across various architectures, we extend the L0 attention score masking to an additional model. As shown in Fig. 12, we mask each key token corresponding to attributes in the first layer self-attention map of the SDXL model. This leads to the disappearance of the corresponding attributes in the generated images. Interestingly, the model frequently shows the removal of the attribute together with the associated object. This indicates that L0 attention scores are essential for establishing attribute binding as well as for preserving the presence of the related object, even in complex sentence structures. Furthermore, we find that these scores play a consistent role across various types of attributes, including color, shape, and texture.



750 Figure 7: **L0 attention scores in T2I-Compbench Huang et al. (2023).**

751
 752
 753
 754
 755

L0 Attention Scores and Generated Images. We observe a correlation between L0 attention scores and the degree of semantic binding in the generated images, as shown in Fig. 3c. The L0 attention scores of related pairs are primarily distributed between 0 and 0.4, as illustrated in Fig. 3c and



765 Figure 10: Comparison between our method without entropy and TIE.



780 Figure 11: Comparison between our method without L0 attention scores and TIE.

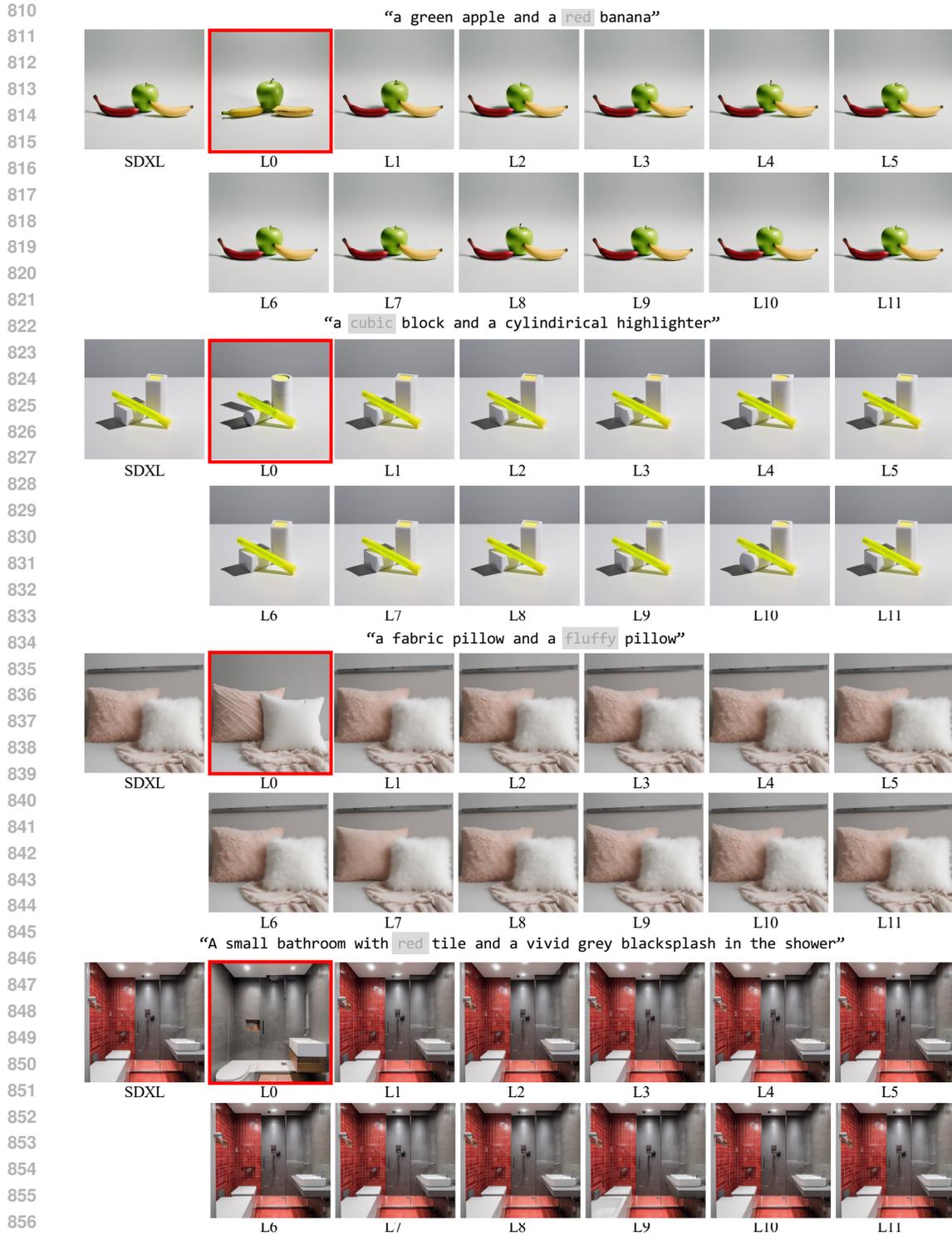
783 Fig. 9. Based on this observation, we analyze the relationship between L0 attention scores and
784 the generated images across different score intervals in Fig. 9. When the attention score $A_{i,j} <$
785 0.1 , the corresponding concept often fails to match semantically, or the object does not appear.
786 In the range $0.1 \leq A_{i,j} < 0.2$, the attribute may appear, but it is often entangled or ambiguously
787 associated with the object. For higher scores, semantic binding between the attribute and object
788 tends to improve substantially. These findings suggest that higher L0 attention scores generally
789 indicate stronger concept binding, and that the degree of binding should be adaptively adjusted
790 based on each concept’s attention strength.

791 **Comparison of SDXL and TIE Using Traditional Similarity Metrics.** In Section 3.2, we reveal
792 that CLIP text embeddings indicate only a subtle distinction between related and unrelated concepts
793 using traditional similarity metrics. To further examine this limitation, we analyze the distributions
794 of cosine similarity and L2 distance across additional datasets and assess how they change when
795 applying our TIE. As shown in Fig. 9, TIE text embeddings more clearly differentiate between
796 related and unrelated concepts compared to the original SDXL embeddings across both datasets.

797 E ADDITIONAL ABLATION STUDIES

798
799 **Adaptive Embedding Preservation (AEP) without Entropy-aware Boost.** In Fig. 10, some qual-
800 itative examples show the difference between our method without taking the token entropy into
801 account. This often causes an over-manipulation of the tokens as the model is not aware of whether
802 the concepts are preserved or not. Thus, it is important for the model to be entropy-aware, so that it
803 preserves concepts effectively.

804 **Interpolaton-Extrapolation Binding (IEB) without L0 Attention Scores.** As shown in Fig. 11,
805 applying IEB without incorporating L0 attention scores leads to two major issues. First, it can
806 result in overmanipulation. When constant scaling is applied, the generated image may deviate from
807 the intended semantics, causing artifacts such as additional objects or an unnatural painting-like
808 appearance. Next, concept confusion persists where objects are merged or attribute bindings remain
809 incorrect. These observations underscore the importance of incorporating L0 attention scores into
IEB to enable more precise manipulation.



858 **Figure 12: Results of masking L0 attention scores in SDXL.** The gray box in the prompt indicates
 859 the masked attribute, and the red outline shows the result after applying masking at Layer 0.

860
861
862
863

Table 3: **The effect of operation sequence of TIE.** Config.1 refers to Sequential IEB \rightarrow AEP and Config.2 is Sequential Extrapolation \rightarrow Interpolation. We measure BLIP-VQA Huang et al. (2023) for when the order of AEP and IEB are switched and when the order of interpolation and extrapolation is switched. Best results are shown in **bold** and second-best results are underlined.

Method	BLIP-VQA Huang et al. (2023) \uparrow			
	Color	Shape	Texture	Avg.
Config.1	<u>0.6995</u>	0.5636	<u>0.6700</u>	<u>0.6444</u>
Config.2	0.6253	0.5329	0.6232	0.5938
TIE (Ours)	0.7004	<u>0.5632</u>	0.6703	0.6447

F ADDITIONAL EXPERIMENTS

We look into the order of the methods in TIE and present our results in Table 3. In TIE, AEP is applied before IEB and interpolation is always applied ahead of extrapolation. We switch the order of these methods to show the effectiveness of our method TIE. This experiment was conducted with SDXL Rombach et al. (2022) backbone.

Reverse the Order of AEP and IEB, Swapping AEP and IEB (Config.1) yields a minor decrease in Color and Texture scores and a slight uptick in Shape. This demonstrates that beginning with AEP to strengthen each token to prevent concept loss *before* binding the related tokens to alleviate concept confusion is effective.

Reverse the Order of Interpolation and Extrapolation When extrapolation is applied before interpolation (Config.2), all three metrics drop sharply. This shows that applying interpolation first is essential to establish a stable attribute-specific embedding distribution before repelling the tokens that are unrelated to that specific token.

G ADDITIONAL QUALITATIVE RESULTS

We provide additional results on the CC-500 dataset in Fig. 13 and Fig. 14. We also provide additional results on the T2I-Compbench dataset in Figs. 15–17 for SDXL and Figs. 18–22 for SD3.5.

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

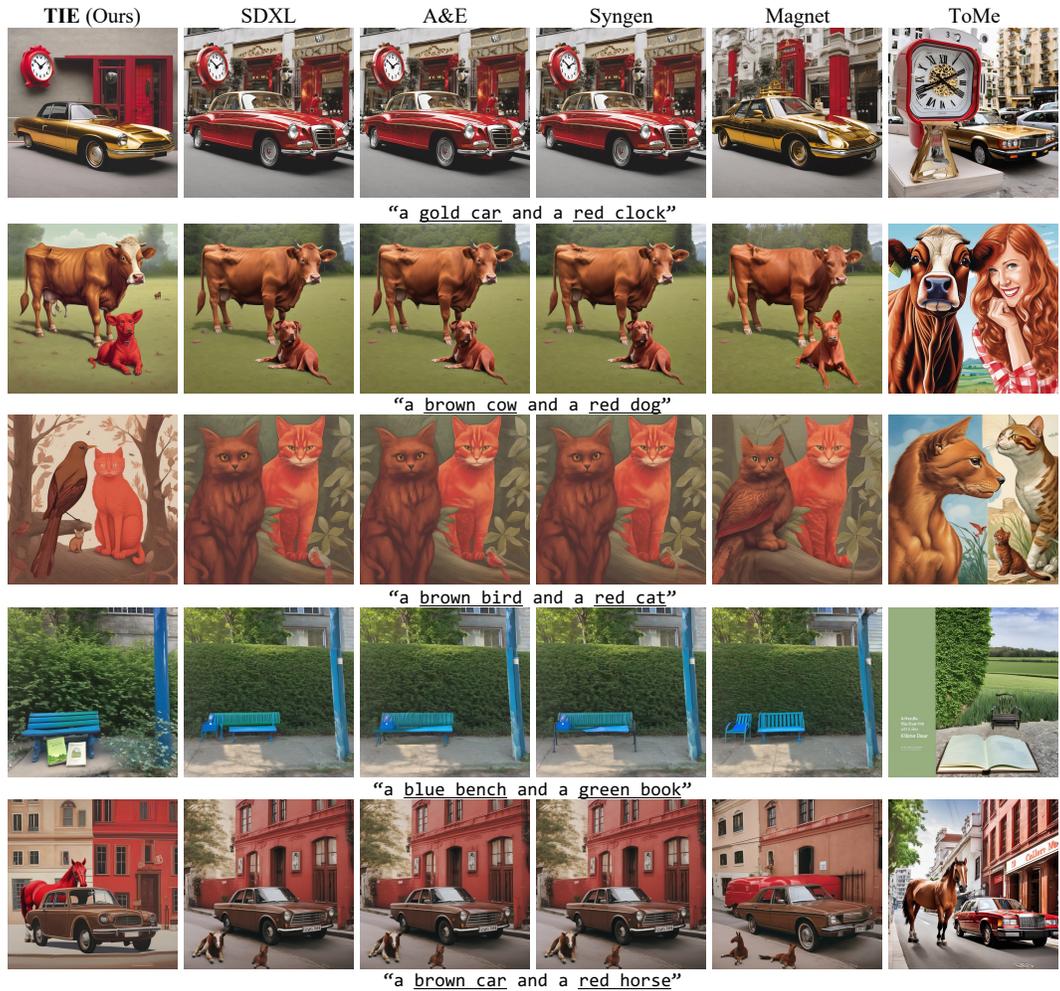


Figure 13: Qualitative results based on SDXL Podell et al. (2023) from the CC-500 dataset.

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
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025

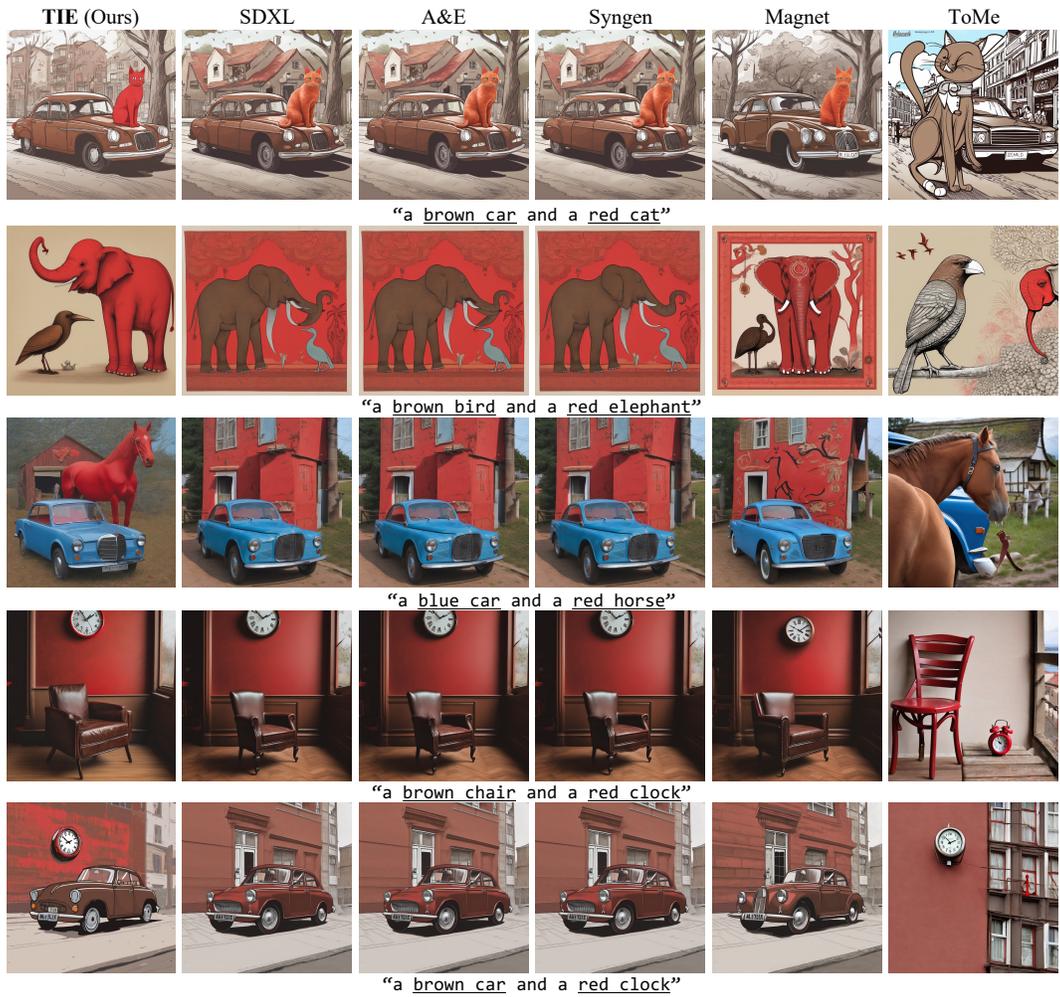


Figure 14: Qualitative results based on SDXL Podell et al. (2023) from the CC-500 dataset.

1026
 1027
 1028
 1029
 1030
 1031
 1032
 1033
 1034
 1035
 1036
 1037
 1038
 1039
 1040
 1041
 1042
 1043
 1044
 1045
 1046
 1047
 1048
 1049
 1050
 1051
 1052
 1053
 1054
 1055
 1056
 1057
 1058
 1059
 1060
 1061
 1062
 1063
 1064
 1065
 1066
 1067
 1068
 1069
 1070
 1071
 1072
 1073
 1074
 1075
 1076
 1077
 1078
 1079

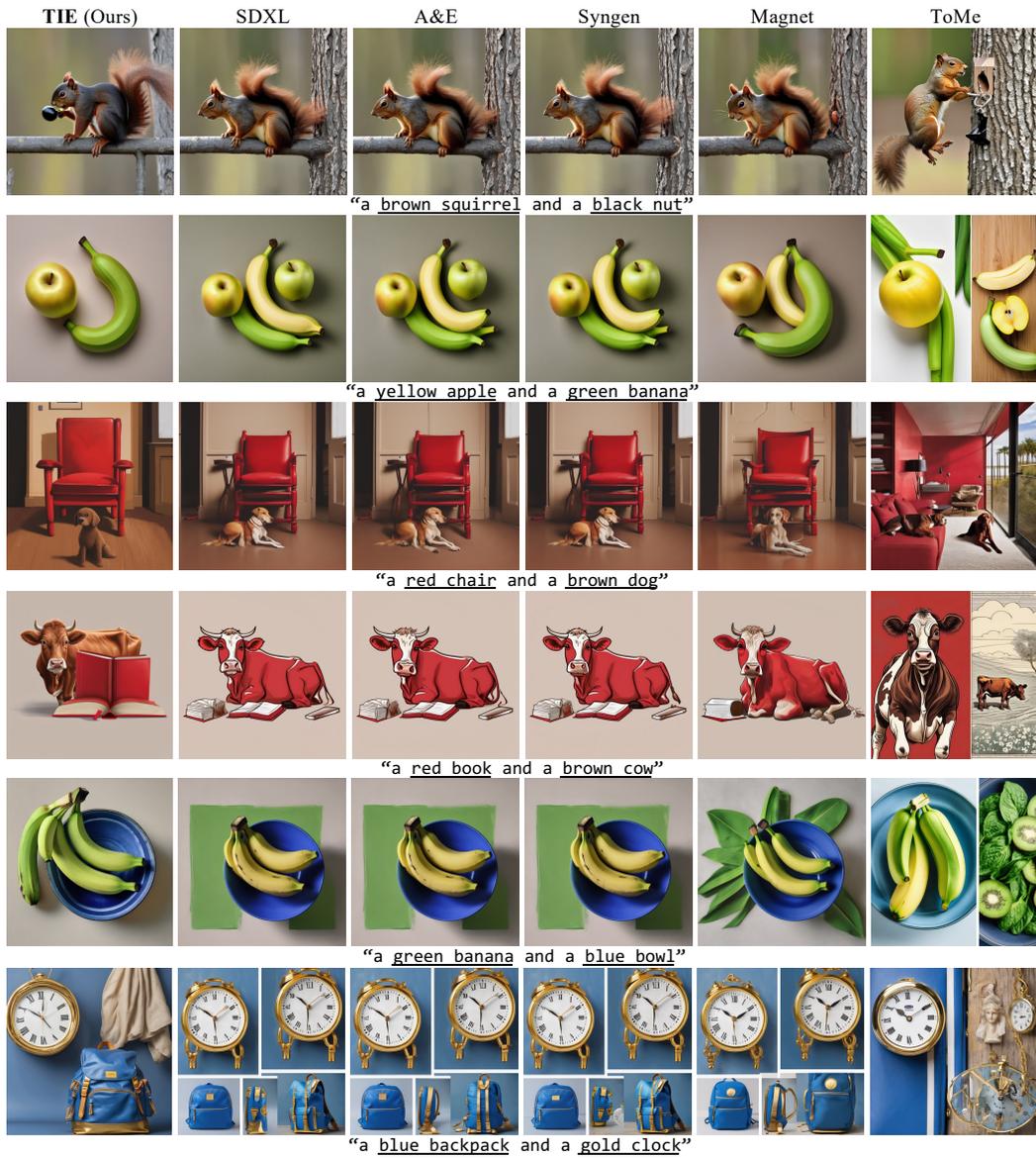


Figure 15: Qualitative results based on SDXL Podell et al. (2023) from the T2I-Compbench dataset Color prompts.

1080
 1081
 1082
 1083
 1084
 1085
 1086
 1087
 1088
 1089
 1090
 1091
 1092
 1093
 1094
 1095
 1096
 1097
 1098
 1099
 1100
 1101
 1102
 1103
 1104
 1105
 1106
 1107
 1108
 1109
 1110
 1111
 1112
 1113
 1114
 1115
 1116
 1117
 1118
 1119
 1120
 1121
 1122
 1123
 1124
 1125
 1126
 1127
 1128
 1129
 1130
 1131
 1132
 1133

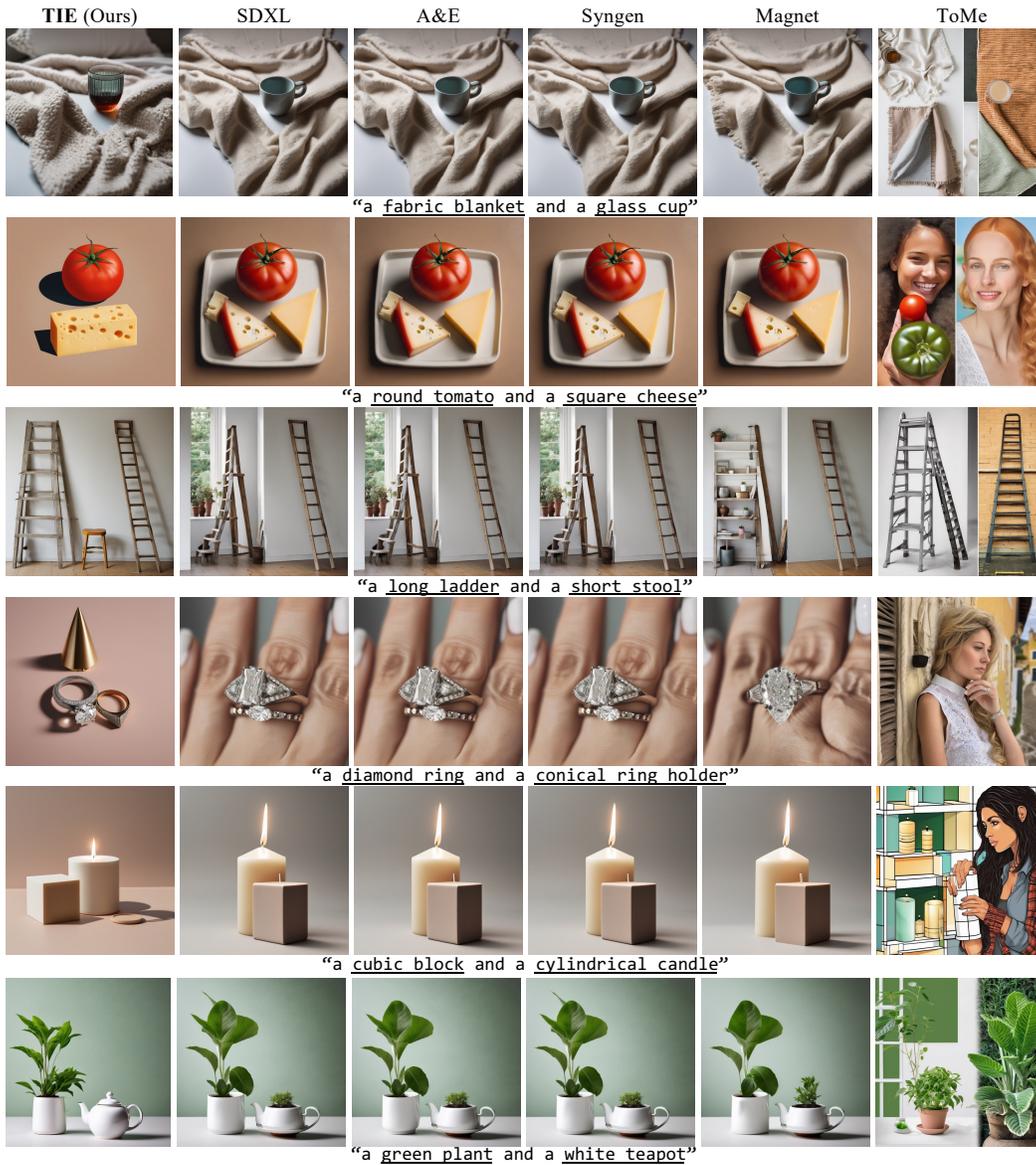


Figure 16: Qualitative results based on SDXL Podell et al. (2023) from the T2I Compench dataset Shape, Texture and Color prompts.

1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187



Figure 17: Qualitative results based on SDXL Podell et al. (2023) from the T2I Compench dataset Complex prompts.

1188
 1189
 1190
 1191
 1192
 1193
 1194
 1195
 1196
 1197
 1198
 1199
 1200
 1201
 1202
 1203
 1204
 1205
 1206
 1207
 1208
 1209
 1210
 1211
 1212
 1213
 1214
 1215
 1216
 1217
 1218
 1219
 1220
 1221
 1222
 1223
 1224
 1225
 1226
 1227
 1228
 1229
 1230
 1231
 1232
 1233
 1234
 1235
 1236
 1237
 1238
 1239
 1240
 1241



Figure 18: Qualitative results based on SD3.5 Esser et al. (2024) from the T2I Compench dataset Color prompts.

1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295



Figure 19: Qualitative results based on SD3.5 Esser et al. (2024) from the T2I Compench dataset Shape prompts.

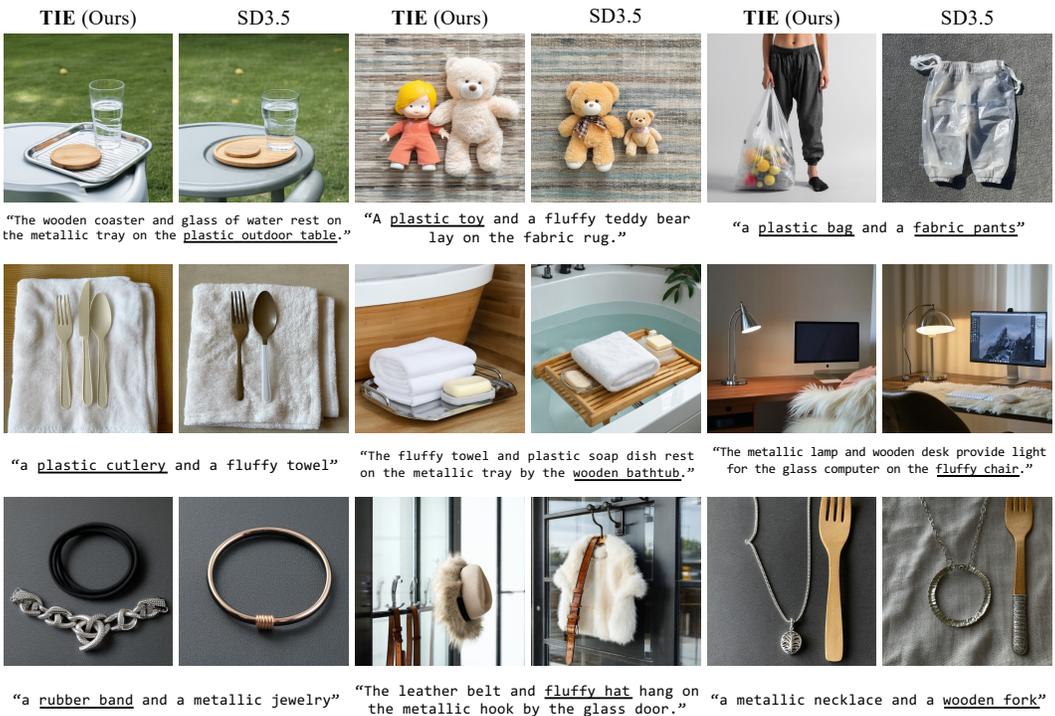


Figure 20: Qualitative results based on SD3.5 Esser et al. (2024) from the T2I Compench dataset Texture prompts.

1296

1297

1298

1299



1305

1306

1307

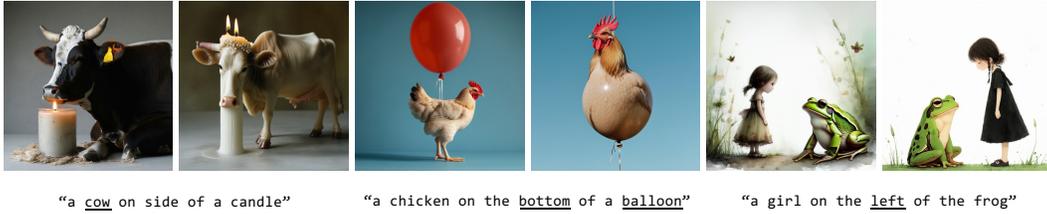
1308

1309

1310

1311

1312



1313

1314

1315

1316

1317

1318

1319



1320

1321

Figure 21: Qualitative results based on SD3.5 Esser et al. (2024) from the T2I Compench dataset Spatial prompts.

1322

1323

1324

1325

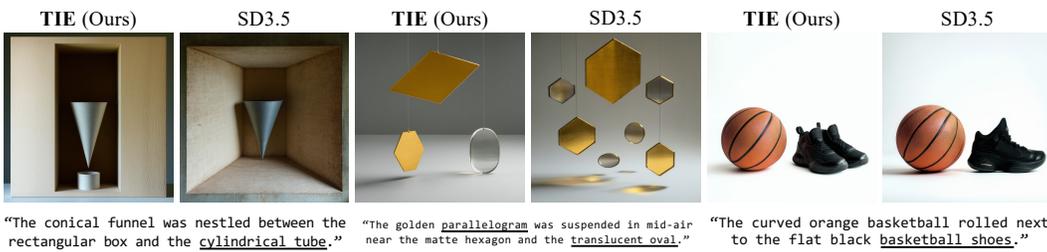
1326

1327

1328

1329

1330



1336

1337

1338

1339

1340

1341

1342

1343

1344

1345



1346

1347

Figure 22: Qualitative results based on SD3.5 Esser et al. (2024) from the T2I Compench dataset Complex prompts.

1348

1349