STEDiff: Strengthening Text Embedding for Text-to-Image Alignment in Diffusion Model

Hailan Zhang[1], Haipeng Liu[1,*], Bo Fu[2], Yang Wang[1]
[1]*School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China*
[2] *School of Computer and Artificial Intelligence, Liaoning Normal University, Dalian, China*
zhhl@mail.hfut.edu.cn, hpliu_hfut@hotmail.com, fubo@lnnu.edu.cn, yangwang@hfut.edu.cn

***Abstract*—Although pretrained text-to-image (T2I) generation models can produce high-quality images, they often fail to faithfully reflect the semantic intent of complex prompts due to stochastic noise and inherent model limitations. This issue frequently manifests as the model overlooking specific objects or failing to correctly bind attributes to their corresponding entities—a challenge referred to as *semantic alignment*. Unlike existing approaches that rely on computationally expensive fine-tuning or labor-intensive layout priors, we propose *STEDiff*, a training-free method designed to enhance semantic representations directly within the *text-embedding space*. Specifically, we introduce a method that primarily leverages [EOT] tokens to strengthen the relevant semantics of sub-sentences, and then replaces the corresponding tokens in the original prompt. Furthermore, a novel semantic enhancement loss is incorporated to enforce spatial constraints, ensuring that the semantics of each entity are precisely mapped to their respective image regions. Extensive quantitative and qualitative evaluations on the T2I-CompBench demonstrate that our method notably improves semantic consistency and generation integrity in complex scenarios.**

***Index Terms*—text embedding, semantic alignment, text-to-image, diffusion model**

I. Introduction

In recent years, T2I diffusion models [1]–[7] have led to remarkable improvements in generating high-quality, detail-rich images from text prompts. These models are usually trained on large text-image datasets using contrastive loss to establish alignment between different spaces. However, due to the limitations of the dataset and insufficient training, aligning the generated images with the prompts, which is referred to as *semantic alignment* [8], remains a significant challenge. Typical issues include: 1) binding objects with their attributes(*attribute binding*); 2) binding objects with their sub-objects(*sub-objects binding*); and 3) the inability to generate noun phrases from the prompts(*noun missing*). For example, as shown in Fig. 1, even the state-of-the-art model such as stable-diffusion-xl-base-1.0 [9], Stable Diffusion 3.5 [10] struggles to properly handle the relationships between different terms in the text prompt, leading to issues of text-image inconsistency.

To improve the alignment between image and text during the generation process, several methods have been proposed. Among these approaches, some focus on optimizing latent representations [11]–[15] to enhance consistency or fine-tuning T2I models [16]–[20]. In parallel, others introduce additional constraints, such as layout guidance [21], [22], or incorporate Large Language Models (LLMs) [23], [24] to improve the model's prompt comprehension. Furthermore, several methods [25]–[28] prioritize the optimization of text embeddings to strengthen semantic binding. However, these methods still encounter limitations. For example, [16] requires substantial computational resources, [21] sacrifices the diversity of generated images, and [26] performs simple token summation, which leads to low coherence in the images.

Fig. 1. **Comparison of Different Methods.** Facing complex prompts, T2I models often suffer from semantic binding issues. For example, attributes associated with "woman" may be incorrectly bound to "bicycle", or the "man" entity may fail to be generated correctly. To address these challenges, we propose the *STEDiff* method.

In this paper, we rethink the role of the text embeddings generated by the CLIP text encoder. Specifically, we perform singular value decomposition on the text embeddings to obtain singular values that represent the primary semantic components of the prompt. Building on the premise that noun phrases constitute the core semantic backbone, we enhance these singular values to effectively mitigate the neglect of such entities during generation. Meanwhile, due to the causal nature of the text encoder, tokens with higher indices—culminating in the [EOT] symbol—accumulate the most comprehensive contextual information. By leveraging the [EOT] token as a global semantic anchor and enhancing the singular values, our method ensures strong semantic binding, as demonstrated by

*Haipeng Liu is the corresponding author

extensive empirical and theoretical evidence.

Based on the above observations, we propose a text embedding-based semantic activation to assist the model in better understanding text prompt. Although many current methods aim to address the issue of attribute binding, there has been relatively little exploration into resolving the sub-objects binding [26]. As an illustration, consider the prompt "a man riding a bicycle and a woman walking a dog", the man is expected to ride a bicycle rather than walk a dog. In order to address the above challenges, we propose *STEDiff*, a novel train-free image generation alignment method, it first split the complex prompts by different entity-nouns, such as *man* and *woman*, thereby obtaining the sub-sentences "*a man riding a bicycle*" and "*a woman walking a dog*", represented as man* and woman*, then strength the semantics of the sub-sentence separately, finally perform the replacements from the initial complex prompts. Instead of operating on the complete prompt, we treat objects and their sub-objects as a whole, avoiding semantic leakage issues from other entities. Using woman* as an example, we build a text embedding matrix that includes both the intended content(woman walking a dog) and end-of-token [EOT] embeddings, applying *soft-weighted regularization* to this matrix can further activate the target semantics.

Meanwhile, since the overall layout is largely determined at the early stages of the diffusion process, we optimize the text embeddings at inference-time using a semantic enhancement loss and an entropy loss. The semantic enhancement loss strengthens the associations between objects and their sub-objects, preventing certain objects from being omitted during generation, thereby improving the alignment among textual concepts and consequently enhancing the structural integrity of the generated images. The entropy loss encourages each prompt token to attend to its corresponding region, thereby mitigating semantic leakage and preventing sub-objects from being incorrectly bound.

To evaluate the effectiveness of *STEDiff*, we perform a quantitative analysis on the widely adopted T2I-CompBench benchmark [29]. Through a comparative evaluation across various baselines, datasets, and metrics, it is clear that *STEDiff* outperforms the others, particularly in scenarios involving multi-object and multi-attribute generation. Notably, *STEDiff* is train-free and user-friendly, primarily modifying in the text embedding space without requiring additional input conditions, such as layout information, which further highlights the superiority of our approach. Overall, the main contributions of our work are as follows:

- We conduct a comprehensive analysis of the text embeddings, highlighting that the [EOT] token contains redundant semantic information(in Fig. 4). By enhancing the semantic information of [EOT], we can improve the semantic features in generated images(in Fig. 6).
- We propose *STEDiff*(Fig. 2), a train-free method that improves the semantic alignment capability of existing T2I models when dealing with complex prompts, through simple enhancement and replacement of text embeddings.
- We conducted extensive experiments(Fig. 7, Tab. I) to demonstrate the effectiveness of our method compared to current SOTA T2I models.

II. Related Work

A. Text-to-Image Diffusion Model

Diffusion models, such as Stable Diffusion [30], DALL-E 2 [4], and Imagen [3], have emerged as the leading approach [1], [2], [9], [10], [30]–[32] in T2I generation area. These models operate as stochastic diffusion processes [33], beginning with a simple distribution $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and gradually adding noise through the forward process, then removes noise over T steps using a denoising network $\boldsymbol{\epsilon_\theta}$ to generate samples in the reverse process. By injecting "external information" such as text embeddings $\boldsymbol{p}$ obtained by a text encoder such as CLIP [34] into the iterative denoising process, the model can be guided [35] to generate the desired content by $\boldsymbol{\epsilon_\theta}(\boldsymbol{x}_t, \boldsymbol{p}, t)$.

B. Text-to-Image Alignment

To improve T2I alignment in diffusion models, numerous studies [36]–[39] have been proposed. Layout-based methods leverage predefined layouts from users [21], [22], [40], [41] or LLMs [23], [24], [42], [43] to guide generation within specified regions. Other approaches [8], [11], [12], [44] employ test-time optimization to align cross-attention maps between attributes and subjects. Fine-tuning-based methods [16]–[18], [20] mitigate semantic incoherence by retraining models on large-scale paired datasets. However, these methods often suffer from limited diversity, difficulties with multi-object or multi-attribute prompts, or substantial computational overhead.

C. Text Embedding Approaches for Text-to-Image

Recent some research has focused on the text embedding space in T2I tasks [45]–[47]. [27] introduced that the causal processing of the CLIP text encoder leads to the accumulation of text information, resulting in biases and losses in the generated images. [48] suppresses the generation of certain text information by processing the [EOT] information within the text embeddings. Magnet addresses attribute bias in the text embedding space by introducing positive and negative binding vectors and highlights the contextual issues in padding embeddings that entangle different concepts. ToMe [26] leverages the additivity of text embeddings to incorporate all attributes into the objects token.

III. Method

We aim to enhance semantic coherence in T2I tasks within diffusion models. By analyzing the text embeddings, we found that in addition to the object token, [EOT] tokens accumulate a significant amount of semantic information that aids in image generation(in Fig. 4). By activating the target semantic information, we can facilitate the correct alignment of each target objects with their attributes. In this section, we first introduce some preliminary concepts in diffusion models(Sec.III-A). Subsequently, we analyze the importance

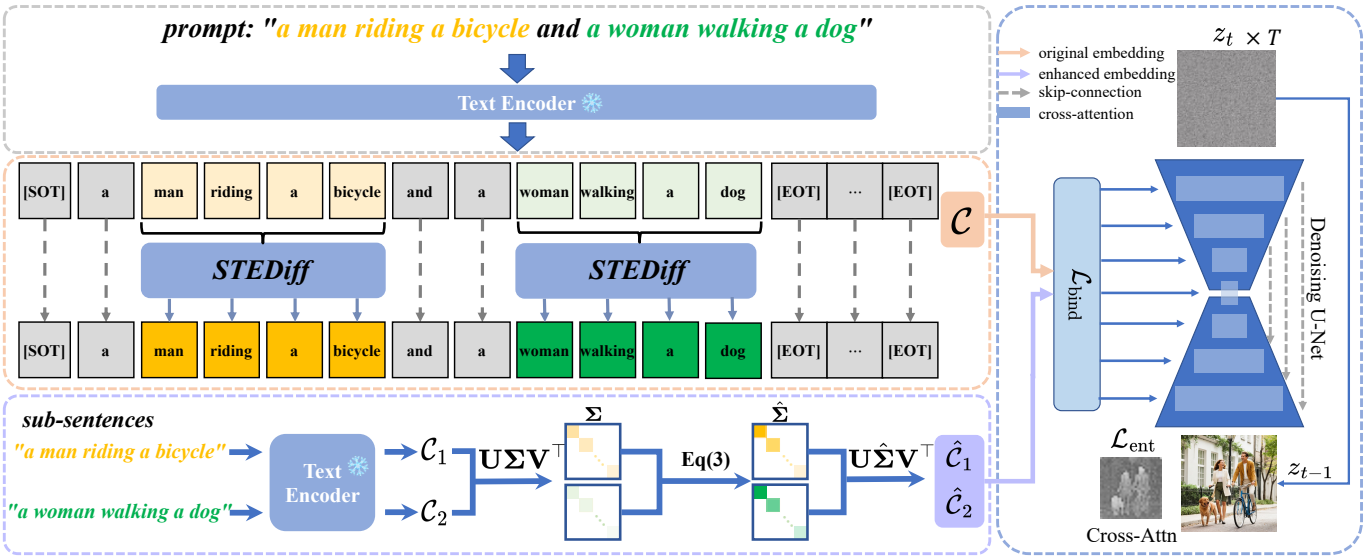

Fig. 2. **Overview of *STEDiff***. Our method starts by splitting complex prompts and treating each sub-sentence as a clean, prompt-level supervision signal. During the denoising process, we apply ***STEDiff*** to the resulting sub-sentences to obtain enhanced embeddings for improved image feature representation. In the early stages of denoising, the $\mathcal{L}_{\text{bind}}$ and $\mathcal{L}_{\text{ent}}$ are used in tandem to update and replace tokens between the original and enhanced embeddings.

of text embeddings in the generation process through a series of experiments and introduce our motivation(Sec.III-B). Finally, we provide a detailed explanation of the specifics of the semantic enhancement techniques and semantic enhancement loss of our method(Sec.III-C). Fig. 2 presents an illustration of our *STEDiff* method.

A. Preliminary

Our method is implemented using stable diffusion(SD) model, which belongs to the family of latent diffusion models, it usually includes a text encoder, a variational autoencoder(i.e.,a encoder $\mathcal{E}$ and a decoder $\mathcal{D}$), and a denoising UNet(i.e.,ϵ_θ with parameter θ). In the forward process, noise ϵ is added to the original latent representation z_0. Then a denoising function ϵ_θ is trained to predict the noise ϵ added to z_0, following the objective:

$$L_{LDM} := \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t\sim(1,T)}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\xi(\mathcal{P}))\|_2^2\right] \quad (1)$$

where τ_ξ is a pre-trained CLIP text encoder, $\mathcal{P}$ is the text prompt, and z_t is a noisy latent sample at timestep $t \sim [1, T]$. The final decoder reverses the latent representation into an image $\hat{x} = \mathcal{D}(z_0)$.

By introducing the text embeddings obtained from the CLIP text encoder, the SD model incorporates cross-attention layers. Cross-attention maps can be extracted internally from $\epsilon_\theta(z_t, t, \tau_\xi(\mathcal{P})$ as high-dimensional tensors, ensuring that the generated images are consistent with the text prompts. The i-th cross-attention map $A^{(i)} \in \mathbb{R}^{h\times w}$ is computed as

$$A^{(i)} = \text{Softmax}(\frac{Q^{(i)}(K^{(i)})^T}{\sqrt{d}}) \quad (2)$$

where $\sqrt{d}$ is a scaling factor, the features map of size is $h \times w$. By applying a linear transformation, the model converts the latent and text embeddings into queries Q and keys K for computing attention weights and identifying the most relevant information.

After visualizing the extracted cross-attention maps, Fig. 3 observes that each token within the prompt corresponds to its own attention map. However, when the prompt contains multiple objects or attributes, the attention distributions of different subjects become intertwined, showing obvious entanglement, which is the main cause of the *semantic alignment* issue.

B. Analysis of Text Embeddings

As one of the most critical components in the SD, the CLIP text encoder τ_ξ transforms the input prompt into a fixed-dimensional text embedding. For a given text prompt $\mathcal{P}$, the τ_ξ tokenizes it by adding a start token [SOT] at the beginning and [EOT] at the end, while padding with additional [EOT] tokens as necessary to extend it to a fixed length of N, where $N = 77$ in our experiments. For example, when the prompt is "a man riding a bicycle", it is first tokenized and then processed by the τ_ξ into a sequence such as $\boldsymbol{c} = \{\boldsymbol{c}^{SOT}, \boldsymbol{c}_0^a, \boldsymbol{c}_1^{man}, \boldsymbol{c}_2^{riding}, \boldsymbol{c}_3^a, \boldsymbol{c}_4^{bicycle}, \boldsymbol{c}_{|\boldsymbol{p}|}^{EOT}, \cdots, \boldsymbol{c}_{N-|\boldsymbol{p}|-2}^{EOT}\}$. Due to the causal manner of the CLIP text encoder, each token can only "see" the preceding tokens. Consequently, the $\boldsymbol{n}^{th}$ token becomes entangled with the previous $\boldsymbol{n}-1$ tokens, causing the later tokens to carry a substantial amount of accumulated semantic information. Through observation, we find that similar images can be generated not only when all [EOT] tokens($\{\boldsymbol{c}_{|\boldsymbol{p}|}^{EOT}, \cdots, \boldsymbol{c}_{N-|\boldsymbol{p}|-2}^{EOT}\}$) are used, but even when only a single [EOT] token is provided. Fig. 4 indicates that a single [EOT] token has already accumulated a substantial amount of semantic information.

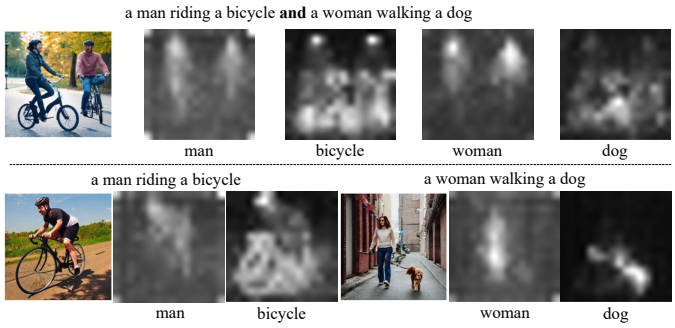

Fig. 3. **Attention visualization.** The cross-attention map for each token of the prompt "a man riding a bicycle and a woman walking a dog" and its sub-sentences is visualized. When faced with a complex prompt, there is a noticeable entanglement between different subjects, whereas simple sub-sentences do not exhibit this behavior.

C. Analysis of [EOT] tokens

To further reveal the issues that arise when the prompt contains multiple subject attributes, we calculated the information entropy of each token in the prompt and found that tokens appearing earlier in the sequence tend to exhibit higher entropy, as illustrated in Fig. 5(a). As the token index increases, the entropy sharply decreases at the [EOT] position of each sentence, after which the information entropy of subsequent tokens gradually diminishes. Early tokens dominate the overall semantic representation, whereas the [EOT] token as a semantic boundary, thereby revealing an inherent limitation of the text encoder in modeling prompts.

Due to the padding embedding mechanism, embeddings at later indices gradually deviate from the true semantic space, thereby leading to semantic degradation. Furthermore, As shown in Fig. 5(b), we compute the cosine similarity between the first [EOT] token and the subsequent padded [EOT] tokens under both the original prompts and their corresponding sub-sentences. The curve drops much more sharply for complex concepts than for simple ones, suggesting that, in complex clauses, the subsequent padded embeddings are more prone to forgetting part of the contextual information retained in the first [EOT].

However, prior studies [48], [49] have shown that these padded embeddings are crucial for the quality of image generation, and that the paddings are entangled with one another, making it impossible to manipulate any single concept in isolation.

D. STE: Strength Text Embedding

As observed in the preceding section, the SD model is more prone to forgetting when faced with complex prompts, and the [EOT] token contains significant semantic information. To mitigate this issue, we propose strengthening the text–image alignment by enhancing the embedding of the [EOT] tokens.

1) Text Embedding-Based Semantic Activation in diffusion model: Inspired by [48], [50], we analyze the text-embedding matrix of each sub-sentence, under the assumption that its dominant singular values represent the fundamental information of the prompt. Taking the sub-sentence "a man riding a bicycle" as an illustrative example, we apply SVD to its text-embedding matrix, i.e., $\boldsymbol{c} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{\Sigma} = diag(\sigma_0, \sigma_1, \cdots, \sigma_n)$, and $\sigma_0 \geq \cdots \geq \sigma_n$.

Fig. 4. **Analysis of Text Embeddings.** [EOT] tokens contain a significant amount of information, and even a single [EOT] token can generate images with semantically similar content to those produced by the full prompt.

To enhance the representation, we introduce a set of amplification coefficients for the singular values to strengthen its expressiveness. The formulation is given by:

$$\hat{\sigma} = \beta e^{\alpha\sigma} * \sigma \tag{3}$$

where e denotes the base of the natural exponential function, and α and β are positive scalar parameters. After obtaining the updated $\hat{\mathbf{\Sigma}} = \text{diag}(\hat{\sigma}_0, \hat{\sigma}_1, \cdots, \hat{\sigma}_n)$, the enhanced text-embedding matrix $\hat{\boldsymbol{c}} = \mathbf{U}\hat{\mathbf{\Sigma}}\mathbf{V}^\top$ can be reconstructed accordingly.

After obtaining the enhanced text embeddings, we generate images conditioned on these embeddings and then extract visual features using the DINOv3 model. The extracted features are subsequently projected into a 2D space via PCA for dimensionality reduction and visualization, as illustrated in Fig. 6(a). This is further depicted in Fig. 6(b), this pipeline yields richer and more informative visual representations, making it easier for the model to retain and reflect the information encoded in the text embeddings.

2) Semantic Bind Loss: As stated in Section III-D1, enhancing semantic information enables the generation of images with richer and more discriminative features. When dealing with complex prompts $\mathcal{P}$ of "a man riding a bicycle and a woman walking a dog". Let $\mathcal{C} = \{\boldsymbol{c}^{\text{SOT}}, \boldsymbol{c}_0^a, \boldsymbol{c}_1^{\text{man}}, \ldots, \boldsymbol{c}_7^{\text{woman}}, \boldsymbol{c}_{|\boldsymbol{p}|}^{\text{EOT}}, \ldots, \boldsymbol{c}_{N-|\boldsymbol{p}|-2}^{\text{EOT}}\} = \{\boldsymbol{C}^{\boldsymbol{man}}, \boldsymbol{C}^{\boldsymbol{woman}}\}$. It is necessary to strengthen the association between each sub-sentence's object and its corresponding sub-object to prevent incorrect bindings.

To this end, we treat each sub-sentence separately, i.e., $\mathcal{C}_1 = \{\boldsymbol{c}^{\text{SOT}}, \boldsymbol{c}_0^a, \boldsymbol{c}_1^{\text{man}}, \ldots, \boldsymbol{c}_4^{\text{bicycle}}, \boldsymbol{c}_{|\boldsymbol{p}_1|}^{\text{EOT}}, \ldots, \boldsymbol{c}_{N-|\boldsymbol{p}_1|-2}^{\text{EOT}}\}$ and $\mathcal{C}_2 = \{\boldsymbol{c}^{\text{SOT}}, \boldsymbol{c}_0^a, \boldsymbol{c}_1^{\text{woman}}, \ldots, \boldsymbol{c}_4^{\text{dog}}, \boldsymbol{c}_{|\boldsymbol{p}_2|}^{\text{EOT}}, \ldots, \boldsymbol{c}_{N-|\boldsymbol{p}_2|-2}^{\text{EOT}}\}$, we applied the *STEDiff* to enhance the text embeddings of $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively, thereby obtaining the enhanced representations of $\hat{\mathcal{C}}_1 = \{\hat{\boldsymbol{C}}^{man}\}$ and $\hat{\mathcal{C}}_2 = \{\hat{\boldsymbol{C}}^{woman}\}$.

We treat each sub-sentence as a clean prompt-level supervision signal, with the aim of making each component of $\mathcal{P}$ match its corresponding sub-sentence as closely as possible. In other words, the objective is to make $\hat{\boldsymbol{C}}^{\boldsymbol{man}}$ as close as possible to $\boldsymbol{C}^{\boldsymbol{man}}$ in the text embedding space, and the same

applies to woman. Here, we aim to encourage the binding between each subject and its corresponding attributes. To this end, we introduce a positive binding loss:

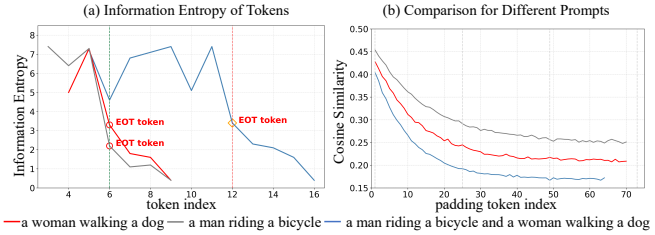

Fig. 5. **Analysis of [EOT] tokens.** (a)Earlier tokens exhibit higher information entropy, which sharply decreases upon reaching the first [EOT] token as the token index increases. (b)When faced with complex prompts, the text embeddings are more likely to deviate from the true semantic space.

$$\mathcal{L}_{\text{bind}} = \sum_{k=1}^{K} \left\| \epsilon_\theta(z_t, \hat{C}_k, t) - \epsilon_\theta(z_t, C_k, t) \right\|_2^2. \tag{4}$$

As illustrated in Fig. 3, each token in the textual prompt contains a substantial amount of information, which may easily lead to semantic leakage. To enhance the concentration of attention, inspired by [26], [51], we combine the previously defined loss term with an entropy-based loss to construct the overall loss function $\mathcal{L}_{\text{sel}}$. The entropy-based loss $\mathcal{L}_{\text{ent}}$ is defined as:

$$\mathcal{L}_{\text{ent}} = -\sum_{k \in K} \sum_{p_i \in \mathcal{A}_k} p_i \log p_i, \tag{5}$$

where $\mathcal{A}_k$ represents the cross-attention map associated with the k-th token, and p_i denotes the attention probability at the i-th spatial position in $\mathcal{A}_k$.

By combining the loss functions $\mathcal{L}_{\text{bind}}$ and $\mathcal{L}_{\text{ent}}$, the final loss function $\mathcal{L}_{\text{sel}}$ is defined as:

$$\mathcal{L}_{\text{sel}} = \mathcal{L}_{\text{bind}} + \lambda \mathcal{L}_{\text{ent}} \tag{6}$$

where λ is a trade-off hyperparameter. During the early diffusion steps t, we update the latent vectors via gradient descent:

$$\mathbf{z}_t' \leftarrow \mathbf{z}_t - \eta \cdot \nabla_{\mathbf{z}_t} \mathcal{L}_{\text{sel}} \tag{7}$$

where η is the step size.

It is worth noting that existing studies have not explicitly demonstrated that activating semantic information can enhance the binding between subjects and attributes. During optimization, for the text embedding of each subject, we apply *STEDiff* to emphasize its corresponding semantics.

IV. Experiments

A. Evaluation Benchmarks and Baseline Methods

We carry out extensive experiments on T2I-CompBench [29], a widely used benchmark designed to evaluate text-to-image generation under complex compositional prompts. The benchmark offers a broad and diverse collection of prompts for comprehensive performance assessment. Specifically, we focus on the color, shape, and texture categories to measure the model's attribute-binding capability. In addition, we employ ImageReward [52] to further examine the perceptual quality of the generated images and to estimate human preference scores. As the first general-purpose human-preference reward model for text-to-image generation, ImageReward supports reliable scoring and ranking of generated outputs. We follow the evaluation protocol [53] using the BLIP-VQA score and human preference as evaluation metrics.

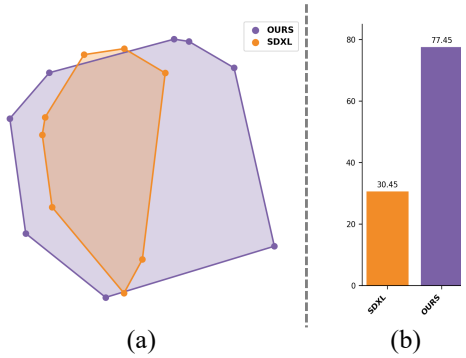

Fig. 6. **Comparison of image features.** (a)Image features are extracted using DINOv3 and visualized through PCA reduction. Compared to the SDXL, our method demonstrates more diverse and rich image features, which contribute to enhanced semantic binding. (b)Statistics of the average feature area calculated from the dimensionality-reduced areas demonstrate that our method reveals more image features.

To further validate the effectiveness of the proposed *STEDiff* method, we compare it with a variety of baseline approaches: ToMe [26], ELLA [17], Attention Regulation [54], Syngen [55], Attend and Excite [44], PixArt [20] and SDXL [9]. For each prompt, we employ the NLP toolkit SpaCy [56] to identify noun-related subjects. Meanwhile, $\alpha = 0.004$ and $\beta = 1.1$ in our experiments. All experiments were run on an Nvidia A40 GPU.

B. Quantitative Comparison

As summarized in Tab. I, based on extensive quantitative evaluation against diverse semantic binding methods, *STEDiff* attains consistently superior or comparable BLIP-VQA scores on the color, shape, and texture subsets, highlighting its capability to alleviate attribute confusion. Meanwhile, *STEDiff* achieves favorable human-preference scores on ImageReward [52], which serve as a proxy for real human preferences, demonstrating stronger alignment with human judgments. Overall, our method demonstrates competitive performance and, on average, outperforms the existing baseline approaches.

C. Qualitative Comparison

Meanwhile, to more intuitively demonstrate the effectiveness of *STEDiff*, we qualitatively compare it with existing semantic binding methods across multiple text-to-image generation models. Fig. 7 presents qualitative comparison results between our method and other approaches. The first row shows the results on *object binding*, while the second row presents the results on *attribute binding*. For complex prompts (rows 3–5), we report the outputs produced by our method, demonstrating its effectiveness in handling challenging compositional instructions. The results clearly demonstrate that *STEDiff* achieves

Fig. 7. **Qualitative comparison of *STEDiff* against semantic binding baselines in different scenarios.** Whether handling objects and their attributes(second row) or sub-objects(first row), *STEDiff* consistently maintains high alignment. Moreover, when faced with complex prompts, it reliably generates images that align accurately with the text description.

superior semantic binding performance in both object binding and attribute binding scenarios involving two objects, as well as under complex prompting conditions.

Owing to its reliance on composite tokens and simple summation, ToMe [26] is prone to entity fragmentation. PixArt [20] achieves high image quality but fails to maintain semantic consistency in binding tasks. ELLA [17] improves text understanding with LLMs yet struggles with complex prompts, while Attend-and-Excite [44] and SDXL [9] are limited in handling attribute binding and often overlook prompt attributes. SynGen [55] and Attention Regulation [54] focus on attribute binding but fail to generalize to object-level or multi-subject scenarios. In contrast, *STEDiff* produces images with more coherent compositions and better alignment with the input prompts.

D. Ablation Study

We conduct ablation experiments to evaluate the effectiveness of the proposed components. The ablation study of each component is quantitatively reported in Tab. II. We observe that using either $\mathcal{L}_{\text{ent}}$ or $\mathcal{L}_{\text{bind}}$ alone leads to only a slight improvement in performance, whereas combining them yields a significant performance gain. Moreover, for the same

Fig. 8. **Ablation study of different optimization terms in the attention mechanism.** Cross-attention visualization of the first [EOT] token shows that using any single optimization term alone does not yield ideal results.

prompt, we generate one image under each configuration to enable a controlled comparison. Since our method performs activation and replacement by leveraging information from the [EOT] token, visualizing the cross-attention of the first few prompt tokens is not informative. The first [EOT] token often aggregates substantial semantic information accumulated from the preceding context; therefore, we visualize the cross-

TABLE I
QUANTITATIVE RESULTS FOR SEMANTIC BINDING ASSESSMENT ON VARIOUS BENCHMARKING SUBSETS. WE DENOTE THE BEST SCORE IN BLUE, AND THE SECOND-BEST SCORE IN GREEN.

Method	Train	BLIP-VQA ↑			Human-preference ↑		
		Color	Texture	Shape	Color	Texture	Shape
SD1.5 [30]	✓	0.4719	0.4334	0.3898	-0.3600	-0.689	-0.483
ELLA$_{1.5}$ [17]	✓	0.6911	0.6308	0.4938	-	-	-
CoMat$_{1.5}$ [16]	✓	0.6561	0.6190	0.4975	-	-	-
SynGen$_{1.5}$ [55]	✗	0.6619	0.6451	0.4661	0.4326	0.5072	0.0426
SDXL [9]	✓	0.6369	0.5637	0.5408	0.7330	0.1240	1.184
Attention Regulation [54]	✗	0.5860	0.5173	0.4672	0.2680	-0.603	1.118
PlayG-v2 [57]	✓	0.6208	0.6125	0.5087	-	-	-
Ranni$_{xl}$ [18]	✓	0.6893	0.6325	0.4934	-	-	-
PixArt [20]	✓	0.6886	0.7044	0.5582	0.2420	-0.788	0.1650
ToMe [26]	✗	0.6583	0.6371	0.5517	-0.0280	-0.851	0.4540
STEDiff (Ours)	✗	0.7010	0.6971	0.6096	0.6232	0.7215	1.1285

TABLE II
ABLATION STUDY CONDUCTED ON THE T2I-COMPBENCH BENCHMARK.

Conf.	$\mathcal{L}_{bind}$	$\mathcal{L}_{ent}$	BLIP-VQA		
			Color	Texture	Shape
A	×	×	0.5654	0.5815	0.5590
B	✓	×	0.6360	0.6139	0.5799
C	×	✓	0.6621	0.6274	0.5972
Ours	✓	✓	**0.7010**	**0.6971**	**0.6096**

attention maps associated with this [EOT] token. As shown in Fig. 8, incorporating our $\mathcal{L}_{\text{bind}}$ enables the model to generate images that better match the prompt, although fine-grained details remain imperfect. In contrast, using the $\mathcal{L}_{\text{ent}}$ alone is insufficient to accomplish the binding task.

V. CONCLUSION

In this paper, we conduct a comprehensive analysis of the semantic binding problem in T2I generation. We propose *STEDiff*, a novel training-free method designed to alleviate semantic inconsistency issues that arise when models are prompted with complex textual descriptions. *STEDiff* requires neither LLMs nor additional training. It activates text embeddings via [EOT] tokens by decomposing complex prompts into subsentences and replacing subject-related tokens, effectively reducing semantic inconsistency. Extensive experiments on T2I-CompBench validate its effectiveness over existing methods. The results demonstrate that, compared with prior approaches, *STEDiff* is more effective at alleviating incorrect semantic binding and object disappearance issues that commonly arise in T2I models when handling complex prompts.

Acknowledgments This research is supported by Institute of Advanced Medicine and Frontier Technology (2023IHM01080); The computation is completed on the HPC Platform of Hefei University of Technology.

REFERENCES

[1] Y. Balaji, S. Nah, X. Huang, A. Vahdat, J. Song, Q. Zhang, K. Kreis, M. Aittala, T. Aila, S. Laine, *et al.*, "ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers," *arXiv preprint arXiv:2211.01324*, 2022.

[2] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*, pp. 8821–8831, Pmlr, 2021.

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36479–36494, 2022.

[4] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[5] H. Liu, Y. Wang, and M. Wang, "One stone with two birds: A null-text-null frequency-aware diffusion models for text-guided image inpainting," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[6] H. Liu, Y. Wang, B. Qian, M. Wang, and Y. Rui, "Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8038–8047, 2024.

[7] Y. Wang, J. Peng, H. Wang, and M. Wang, "Progressive learning with multi-scale attention network for cross-domain vehicle re-identification," *Science China Information Sciences*, vol. 65, no. 6, p. 160103, 2022.

[8] Y. Li, M. Keuper, D. Zhang, and A. Khoreva, "Divide & bind your attention for improved generative semantic nursing," *arXiv preprint arXiv:2307.10864*, 2023.

[9] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[10] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *Forty-first international conference on machine learning*, 2024.

[11] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment," *Advances in Neural Information Processing Systems*, vol. 36, pp. 3536–3559, 2023.

[12] S. Ge, T. Park, J.-Y. Zhu, and J.-B. Huang, "Expressive text-to-image generation with rich text," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7545–7556, 2023.

[13] B. Qian, Y. Wang, H. Yin, R. Hong, and M. Wang, "Switchable online knowledge distillation," in *European Conference on Computer Vision*, pp. 449–466, Springer, 2022.

[14] B. Qian, Y. Wang, R. Hong, and M. Wang, "Adaptive data-free quantization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7960–7968, 2023.

[15] Y. Wang, B. Qian, H. Liu, Y. Rui, and M. Wang, "Unpacking the gap box against data-free knowledge distillation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 9, pp. 6280–6291, 2024.

[16] D. Jiang, G. Song, X. Wu, R. Zhang, D. Shen, Z. Zong, Y. Liu, and H. Li, "Comat: Aligning text-to-image diffusion model with image-to-text concept matching," *Advances in Neural Information Processing Systems*, vol. 37, pp. 76177–76209, 2024.

[17] X. Hu, R. Wang, Y. Fang, B. Fu, P. Cheng, and G. Yu, "Ella: Equip diffusion models with llm for enhanced semantic alignment," *arXiv preprint arXiv:2403.05135*, 2024.

[18] Y. Feng, B. Gong, D. Chen, Y. Shen, Y. Liu, and J. Zhou, "Ranni: Taming text-to-image diffusion for accurate instruction following," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4744–4753, 2024.

[19] H. Liu, G. Li, M. Gao, X. Zhen, F. Zheng, and Y. Wang, "Few-shot referring video single-and multi-object segmentation via cross-modal affinity with instance sequence matching," *International Journal of Computer Vision*, vol. 133, no. 8, pp. 5610–5628, 2025.

[20] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, and Z. Li, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," in *European Conference on Computer Vision*, pp. 74–91, Springer, 2024.

[21] J. Xie, Y. Li, Y. Huang, H. Liu, W. Zhang, Y. Zheng, and M. Z. Shou, "Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7452–7461, 2023.

[22] Q. Phung, S. Ge, and J.-B. Huang, "Grounded text-to-image synthesis with attention refocusing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7932–7942, 2024.

[23] L. Lian, B. Li, A. Yala, and T. Darrell, "Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models," *arXiv preprint arXiv:2305.13655*, 2023.

[24] H. Gani, S. F. Bhat, M. Naseer, S. Khan, and P. Wonka, "Llm blueprint: Enabling text-to-image generation with complex and detailed prompts," *arXiv preprint arXiv:2310.10640*, 2023.

[25] B. Qian, Y. Wang, R. Hong, and M. Wang, "Rethinking data-free quantization as a zero-sum game," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, pp. 9489–9497, 2023.

[26] T. Hu, L. Li, J. van de Weijer, H. Gao, F. Shahbaz Khan, J. Yang, M.-M. Cheng, K. Wang, and Y. Wang, "Token merging for training-free semantic binding in text-to-image synthesis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 137646–137672, 2024.

[27] C.-Y. Chen, C. Tseng, L.-W. Tsao, and H.-H. Shuai, "A cat is a cat (not a dog!): Unraveling information mix-ups in text-to-image encoders through causal analysis and embedding optimization," *Advances in Neural Information Processing Systems*, vol. 37, pp. 57944–57969, 2024.

[28] H. Liu, Y. Wang, M. Wang, and Y. Rui, "Delving globally into texture and structure for image inpainting," in *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1270–1278, 2022.

[29] K. Huang, K. Sun, E. Xie, Z. Li, and X. Liu, "T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 78723–78747, 2023.

[30] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[31] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[32] B. F. Labs, S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, S. Kulal, K. Lacey, Y. Levi, C. Li, D. Lorenz, J. Müller, D. Podell, R. Rombach, H. Saini, A. Sauer, and L. Smith, "Flux.1 kontext: Flow matching for in-context image generation and editing in latent space," 2025.

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[35] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[36] O. Dahary, O. Patashnik, K. Aberman, and D. Cohen-Or, "Be yourself: Bounded attention for multi-subject text-to-image generation," in *European Conference on Computer Vision*, pp. 432–448, Springer, 2024.

[37] F.-Y. Wang, Z. Huang, A. Bergman, D. Shen, P. Gao, M. Lingelbach, K. Sun, W. Bian, G. Song, Y. Liu, *et al.*, "Phased consistency models," *Advances in neural information processing systems*, vol. 37, pp. 83951–84009, 2024.

[38] D. Shen, G. Song, Z. Xue, F.-Y. Wang, and Y. Liu, "Rethinking the spatial inconsistency in classifier-free diffusion guidance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9370–9379, 2024.

[39] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel, "Plug-and-play diffusion features for text-driven image-to-image translation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1921–1930, 2023.

[40] R. Wang, Z. Chen, C. Chen, J. Ma, H. Lu, and X. Lin, "Compositional text-to-image synthesis with attention map control of diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 5544–5552, 2024.

[41] X. Wang, T. Darrell, S. S. Rambhatla, R. Girdhar, and I. Misra, "Instancediffusion: Instance-level control for image generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6232–6242, 2024.

[42] L. Yang, Z. Yu, C. Meng, M. Xu, S. Ermon, and B. Cui, "Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms," in *Forty-first International Conference on Machine Learning*, 2024.

[43] S. Li, R. Wang, C.-J. Hsieh, M. Cheng, and T. Zhou, "Mulan: Multimodal-llm agent for progressive and interactive multi-object diffusion," *arXiv preprint arXiv:2402.12741*, 2024.

[44] H. Chefer, Y. Alaluf, Y. Vinker, L. Wolf, and D. Cohen-Or, "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models," *ACM transactions on Graphics (TOG)*, vol. 42, no. 4, pp. 1–10, 2023.

[45] H. Seo, J. Bang, H. Lee, J. Lee, B. H. Lee, and S. Y. Chun, "Geometrical properties of text token embeddings for strong semantic binding in text-to-image generation," *arXiv preprint arXiv:2503.23011*, 2025.

[46] H. Yu, H. Luo, F. Wang, and F. Zhao, "Uncovering the text embedding in text-to-image diffusion models," *arXiv preprint arXiv:2404.01154*, 2024.

[47] F. Wu, Y. Pang, J. Zhang, L. Pang, J. Yin, B. Zhao, Q. Li, and X. Mao, "Core: Context-regularized text embedding learning for text-to-image personalization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 8377–8385, 2025.

[48] S. Li, J. van de Weijer, T. Hu, F. S. Khan, Q. Hou, Y. Wang, and J. Yang, "Get what you want, not what you don't: Image content suppression for text-to-image diffusion models," *arXiv preprint arXiv:2402.05375*, 2024.

[49] W. Feng, X. He, T.-J. Fu, V. Jampani, A. Akula, P. Narayana, S. Basu, X. E. Wang, and W. Y. Wang, "Training-free structured diffusion guidance for compositional text-to-image synthesis," *arXiv preprint arXiv:2212.05032*, 2022.

[50] S. Gu, L. Zhang, W. Zuo, and X. Feng, "Weighted nuclear norm minimization with application to image denoising," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2862–2869, 2014.

[51] C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[52] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, "Imagereward: Learning and evaluating human preferences for text-to-image generation," 2023.

[53] L. Chen, J. Wang, Z. Pan, B. Zhu, X. Yang, and C. Zhang, "Detail++: Training-free detail enhancer for text-to-image diffusion models," *arXiv preprint arXiv:2507.17853*, 2025.

[54] Y. Zhang, T. T. Tzun, L. W. Hern, T. Sim, and K. Kawaguchi, "Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models," 2024.

[55] R. Rassin, E. Hirsch, D. Glickman, S. Ravfogel, Y. Goldberg, and G. Chechik, "Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment," 2024.

[56] M. Honnibal, "spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing," *(No Title)*, 2017.

[57] D. Li, A. Kamko, A. Sabet, E. Akhgari, L. Xu, and S. Doshi, "Playground v2."