

---

# Token Merging for Training-Free Semantic Binding in Text-to-Image Synthesis

---

Taihang Hu<sup>1</sup>, Linxuan Li<sup>1</sup>, Joost van de Weijer<sup>3</sup>, Hongcheng Gao<sup>4</sup>  
Fahad Shahbaz Khan<sup>5,6</sup>, Jian Yang<sup>1</sup>, Ming-Ming Cheng<sup>1,2</sup>, Kai Wang<sup>3\*</sup>, Yaxing Wang<sup>1,2\*</sup>

<sup>1</sup>VCIP, College of Computer Science, Nankai University, <sup>2</sup>NKIARI, Shenzhen Futian

<sup>3</sup>Computer Vision Center, Universitat Autònoma de Barcelona

<sup>4</sup>University of Chinese Academy of Sciences

<sup>5</sup>Mohamed bin Zayed University of AI, <sup>6</sup>Linköping University

{hutaihang00, linxuanli520, gaohongcheng2000}@gmail.com

{joost, kwang}@cvc.uab.es, fahad.khan@liu.se

{csjyang, cmm, yaxing}@nankai.edu.cn

## Abstract

Although text-to-image (T2I) models exhibit remarkable generation capabilities, they frequently fail to accurately bind semantically related objects or attributes in the input prompts; a challenge termed *semantic binding*. Previous approaches either involve intensive fine-tuning of the entire T2I model or require users or large language models to specify generation layouts, adding complexity. In this paper, we define semantic binding as the task of associating a given object with its attribute, termed *attribute binding*, or linking it to other related sub-objects, referred to as *object binding*. We introduce a novel method called *Token Merging (ToMe)*, which enhances semantic binding by aggregating relevant tokens into a single *composite token*. This ensures that the object, its attributes and sub-objects all share the same cross-attention map. Additionally, to address potential confusion among main objects with complex textual prompts, we propose *end token substitution* as a complementary strategy. To further refine our approach in the initial stages of T2I generation, where layouts are determined, we incorporate two auxiliary losses, an entropy loss and a semantic binding loss, to iteratively update the composite token to improve the generation integrity. We conducted extensive experiments to validate the effectiveness of *ToMe*, comparing it against various existing methods on the T2I-CompBench and our proposed GPT-4o object binding benchmark. Our method is particularly effective in complex scenarios that involve multiple objects and attributes, which previous methods often fail to address. The code will be publicly available at <https://github.com/hutaihang/ToMe>.

## 1 Introduction

Text-to-image generation has seen significant advancements with the recent introduction of diffusion models [57, 59, 62], with their capabilities of generating high-fidelity images from text prompts. Despite these achievements, aligning the generated images with the text prompts, which is referred to as *semantic alignment* [30, 43], remains a notable challenge. One of the most common issues observed in existing text-to-image (T2I) generation models is the lack of proper *semantic binding*, where a given object is not properly binding to its attributes or related objects. For example, as illustrated in Fig. 1, even a state-of-the-art T2I model such as SDXL [53] can struggle to generate content that accurately reflects the intended nuances of text prompts. To address the persistent challenges of aligning T2I diffusion models with the intricate semantics of text prompts, a variety of enhancement

---

\*: Co-corresponding authors



Figure 1: Current state-of-the-art T2I models often struggle with semantic binding in generated images according to textual prompts. For example, hats and sunglasses are placed on incorrect objects. We introduce a novel method *ToMe* to address these challenges.

strategies [35, 46, 87] are proposed, either by optimizing the latent representations [69, 82, 83], guiding the generation by layout priors [54, 71, 85] or fine-tuning the T2I models [21, 34]. Despite these advancements, these methods still encounter limitations, particularly in generating high-fidelity images involving complex scenarios where an object is binding with multiple objects or attributes.

In this paper, we categorize *semantic binding* into two categories. First, *attribute binding* involves correctly associating objects with their attributes, a topic that has been studied in prior work [58]. Second, *object binding*, which entails effectively linking objects to their related *sub-objects* (for example, a ‘hat’ and ‘glasses’), is less explored in the existing literature. Previous methods often struggled to address this aspect of semantic binding. One of the main problems is the misalignment of objects with their corresponding sub-objects. Existing solutions address this through an explicit alignment process of the attention maps [7, 43] or by factorizing the generation projects into layout phases and generation phase [55]. In this paper, we propose a simple solution to the attention alignment problem called *token merging (ToMe)*. Instead of multiple attention maps, which can be misaligned, we join these objects in a single *composite token* that represents the object and its attributes and sub-objects. This composite token has a single cross-attention map that ensures semantic alignment. The composite token is simply constructed by summing the CLIP text embeddings of the various tokens it represents. For example, the phrase “a dog with hat” is abbreviated as “a dog\*” by aggregating the text embeddings corresponding to the last three words, as shown in Fig. 4. To justify the applied embedding addition in *ToMe*, we experimented with the semantic additivity of the text embeddings (in Fig. 3). Furthermore, to mitigate potential semantic misalignment in the end tokens from the long sequences, we propose *end token substitution (ETS)* technique.

As the T2I generation predominantly determines the layout during earlier phases [27], we introduce an entropy loss and a semantic binding loss to update the token embeddings in early steps, integrating *ToMe* with an iterative update for the composite tokens. The entropy loss is defined as the entropy of the cross-attention map corresponding to the updated composite token. This loss aims to enhance generation integrity by ensuring diverse attention across relevant areas of the image, thereby preventing focusing on non-essential regions. The semantic binding loss encourages the new learned token to infer the same noise prediction as the original corresponding phrase. This alignment further reinforces the semantic coherence between the text and the generated image.

Our final method *ToMe* is quantitatively assessed using the widely adopted T2I-CompBench [31] and our proposed GPT-4o [1] *object binding* benchmark. Comparative evaluations against various types of approaches reveal that *ToMe* outperforms them by a significant margin. Remarkably, our approach is user-friendly, requiring no dependence on large language models or specific layout information. In qualitative evaluations, we notably achieve superior generation quality, particularly in scenarios involving multi-object multi-attribute generation. This further underscores the superiority of our method. In summary, the main contributions of this paper are as follows:

- We analyze the problem of semantic binding, and highlight the role of the [EOT] token (Fig. 2), and the problems with misaligned cross-attention maps (Fig. 7). In addition, we explore token additivity as a possible solution (Fig. 3).

- We introduce a *training-free* approach called *Token Merging* (Fig. 4), denoted as *ToMe*, as a more efficient and robust solution for semantic binding. It is further enhanced by our proposed *end token substitution* and iterative *composite token* updates techniques.
- In experiments conducted on the widely used T2I-CompBench benchmark and our GPT-4o object binding benchmark, we compared *ToMe* with various state-of-the-art approaches and consistently outperformed them by significant margins.

## 2 Related works

A critical drawback of current text-to-image models is related to their limited ability to faithfully represent the precise semantics of input prompts, commonly referred to as *semantic alignment*. Various studies have identified common semantic failures and proposed mitigation strategies. They can be roughly categorized into four main streams.

**Optimization-based methods** primarily adjust text embeddings [20, 65] or optimize noisy signals to strengthen attention maps [26, 48, 63, 69, 82, 83]. These methods are basically inspired by the observations from text-based image editing methods [27, 40, 64, 66], suggesting that the layouts of objects are determined by self-attention and cross-attention maps from the UNet of the T2I diffusion models. For example, Attend-and-Excite [7] improves object existence by exciting the attention score of each object. Divide-and-Bind [43] improves by maximizing the total variation of the attention map to prompt multiple spatially distinct attention excitations. SynGen [58] syntactically analyzes the prompt to identify entities and their modifiers, and then uses attention loss functions to encourage the cross-attention maps to agree with the linguistic binding reflected in the syntax. A-star [2] proposes to minimize concept overlap and change in attention maps through iterations. Composable Diffusion [45] decomposes complex texts into simpler segments and then composes the image from these segments. Structure Diffusion [20] attempts to address this by leveraging linguistic structures to guide the cross-attention maps. Rich-Text [24] enriches textual prompts by incorporating various formatting controls and decomposes the generation task into merging inferences from multiple region-based diffusions. However, these methods often fail in complex scenarios that generate multiple objects or multiple attributes.

**Layout-to-Image methods** [4, 9, 14, 17, 25, 32, 36, 47] are widely using layouts, particularly in the form of bounding boxes or segmentation maps, as a popular intermediary to bridge the gap between text input and the generated images. For example, BoxDiff [73] encourages the desired objects to appear in the specified region by calculating losses based on the maximum values in cross-attention maps. Similarly, Attention-Refocusing [52] modifies both cross-attention and self-attention maps to control object positions. BoxNet [67] first trains a network to predict the box for each entity that possesses the attribute specified in the prompt, and then force the generation to follow the attention mask control. Additionally, InstanceDiffusion [68] enhances text-to-image models by providing extra instance-level control. There are also finetuning methods [5, 42, 50, 79] allow for additional layout conditions after fine-tuning over pair images, which are not specifically designed to solve the *semantic alignment* problem. Despite their promise, these methods obviously prolong the training time. Furthermore, the application of layout priors is challenging when it comes to global background descriptions or abstract elements. This limitation constrains the versatility of these techniques, making it difficult to deploy them effectively across real scenarios where non-specific spatial arrangements are crucial.

**LLM-augmented methods** are mainly following text-to-layout-to-image generation pipelines [15, 23, 33, 44, 55, 65, 80, 81, 86], first to generate layouts from large language models (LLMs) and force the T2I generations to follow this guidance as the previous layout-guided methods. Some methods, such as RPG [75] and MuLan [39], harness the powerful chain of thought reasoning ability of multimodal LLMs to enhance the compositionality of text-to-image diffusion models.

**Finetuning-based methods** [13, 76] update the model parameters over huge datasets to augment the semantic alignment. Among them, CoMat [34] proposes an end-to-end fine-tuning strategy for text-to-image diffusion models by incorporating image-to-text concept matching. ELLA [30] equips text-to-image diffusion models with powerful Large Language Models (LLM) to enhance text alignment by bridging these two pre-trained models with trainable semantic alignment connectors. More recently, Ranni [21] improves T2I generation by bridging the text and image with a semantic panel with LLMs and is fine-tuned over an automatically prepared semantic panel dataset. There

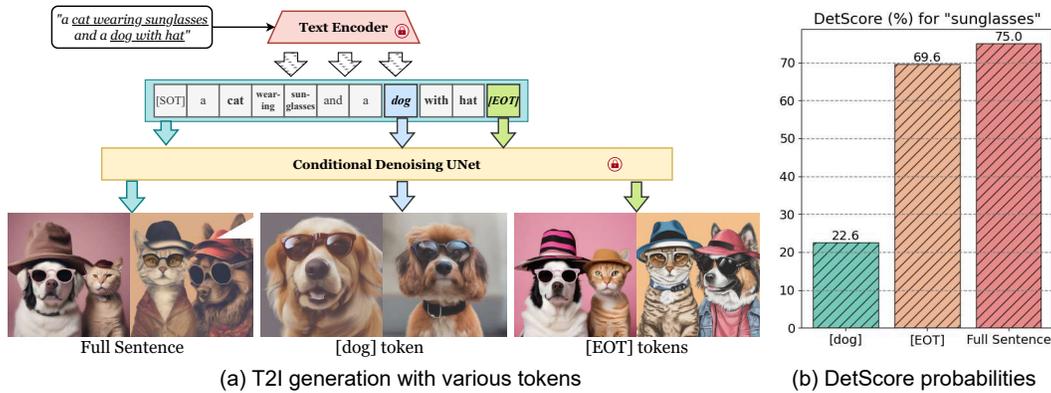


Figure 2: We generate images with various input prompts in (a): “a cat wearing sunglasses and a dog wearing a hat”; the single-token embedding [dog]; the end token [EOT] . (b) After that, we compute the probability of containing “sunglasses” in the generated images in subfigure .

are also improved T2I models [10, 11, 51] learning from scratch over huge datasets. These methods improve semantic alignment implicitly by better architecture design and larger amount of training data. They further demand marvelous computational resources to achieve the purpose.

In this paper, we tackle the *semantic binding* problem, which is a broad subcase of *semantic alignment*, in a training-free manner, neither needing the LLMs nor any training over additional datasets. Furthermore, we achieve better performance when facing complex T2I generation scenarios where users require multiple objects or multiple attributes related to a specific object.

### 3 Methods

Semantic binding in T2I generation refers to the crucial requirement of establishing accurate associations between objects and their relevant attributes or related sub-objects. This process avoids semantic misalignment in the generated images, ensuring that each visual element aligns correctly with its descriptive cues in the text. In this section, we begin by providing the preliminaries. Subsequently, we illustrate the motivation through a series of experimental analyses (Sec. 3.1). Finally, we elaborate on our methods in detail (Sec. 3.2). An illustration of our method *ToMe* is shown in Fig. 4.

**Latent Diffusion Models.** We build our novel approach for semantic alignment on the standard SDXL [53] model. The model is composed of two main parts: an autoencoder (i.e., an encoder  $\mathcal{E}$  and a decoder  $\mathcal{D}$ ) and a diffusion model (i.e.,  $\epsilon_\theta$  with parameter  $\theta$ ). The model  $\epsilon_\theta$  is updated by the loss:

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

where  $\epsilon_\theta$  is a UNet, conditioning a latent input  $z_t$ , a text embedding  $\tau_\xi(\mathcal{P})$  and a timestep  $t \sim \text{Uniform}(1, T)$ . More specifically, text-guided diffusion models aim to generate an image from random noise  $z_T$  and a conditional input prompt  $\mathcal{P}$ . To distinguish from the general conditions in LDMs, we itemize the textual condition as  $\mathcal{C} = \tau_\xi(\mathcal{P})$ , where  $\tau_\xi$  is the CLIP text encoder [56]<sup>†</sup>. The cross-attention map is obtained from  $\epsilon_\theta(z_t, t, \mathcal{C})$ . Let  $f_{z_t}$  be a feature map output of the network  $\epsilon_\theta$ . We get a query matrix  $Q_t = l_Q(f_{z_t})$  with projection network  $l_Q$ . Similarly, given a textual embedding  $\mathcal{C}$ , we compute a key matrix  $\mathcal{K} = l_K(\mathcal{C})$  with projection network  $l_K$ . Then the attention map is computed according to:  $\mathcal{A}_t = \text{softmax}(Q_t \cdot \mathcal{K}^T / \sqrt{d})$  where  $d$  is the latent dimension, and the cell  $[\mathcal{A}_t]_{ij}$  defines the weight of the  $j$ -th token on the  $i$ -th token.

#### 3.1 Text Embedding Analysis

To address the semantic binding problem, we concentrate on the text embeddings utilized during the diffusion model generation process, as they predominantly dictate the content of the generated images. For a given text prompt  $\mathcal{P}$ , it is tokenized by the CLIP text model by padding a start

<sup>†</sup>SDXL uses two CLIP text encoders and concatenate the two text embeddings as the final text embedding.

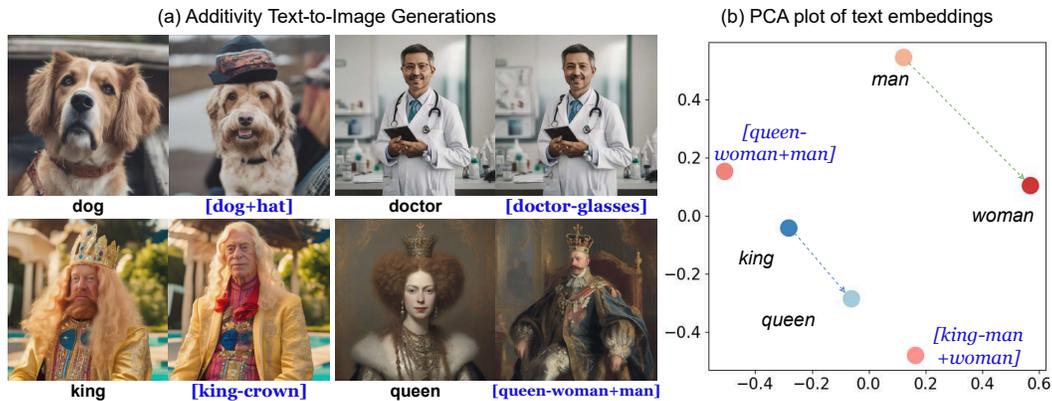


Figure 3: (a) Image generations with the property of token additivity. All images are generated by the prompt template “a photo of a {object}.” (b) PCA plot for additivity of text embeddings.

token [SOT] and several end tokens [EOT] to extend its length to  $M(=77$  by default). After the CLIP text encoder  $\tau_\xi$ , the condition is formulated as  $\mathcal{C} = \tau_\xi(\mathcal{P})$ . Each row in  $\mathcal{C}$  represents a corresponding token embedding after the CLIP text transformers. For example, the text embedding for the sentence  $\mathcal{P}$  = “a cat wearing sunglasses and a dog wearing a hat” is represented as:  $\mathcal{C} = [c_0^{SOT}, c_1^a, c_2^{cat}, \dots, c_7^{dog}, c_8^{wearing}, c_9^{hat}, c_{10}^{EOT}, \dots, c_{M-1}^{EOT}]$ . In the following analysis, we take this as a default example (except when defined differently).

**Information Coupling.** We begin by generating images conditioning on the textual embedding  $\mathcal{C}$ , as illustrated in the first two columns at the bottom of Fig. 2-(a). We observe that the attributes appear in a misalignment between the dog and the cat. Subsequently, we extract the token embedding  $c_7^{dog}$  from the textual embedding and input it to the UNet  $\epsilon_\theta$  (i.e.,  $\mathcal{C} = [c_7^{dog}]^\ddagger$ ). As depicted in the middle columns of Fig. 2-(a). The dog object is frequently wearing glasses, further highlighting the semantic leakage issue. Furthermore, when we take  $\mathcal{C}^{[EOT]} = [c_{10}^{EOT}, \dots, c_{M-1}^{EOT}]$  as input, the generated images closely resemble all information obtained using the entire textual embedding  $\mathcal{C}$ . As the [EOT] interacts with all tokens, it often encapsulates the entire semantic information [41, 72]. We further report the *DetScore* [12] to show the probability of containing the corresponding object (“sunglasses”) in the generated 100 images. As illustrated in Fig. 2-(b), for these three different cases, the DetScore is 22.6%, 69.6% and 75.0%, respectively. These findings also align with our observations above.

**Additivity Property.** Inspired by the semantic additivity of the text embeddings in previous research[6, 49], we experiment the additive property of the CLIP textual embedding. We represent the textual embedding corresponding to the prompt “a photo of a dog” as  $\mathcal{C}_1 = [c_0^{SOT}, c_1^a, \dots, c_5^{dog}, c_6^{EOT}, \dots, c_{M-1}^{SOT}]$ . The textual embedding for the prompt “a photo of a hat” is represented as  $\mathcal{C}_2 = [c_0^{SOT}, c_1^a, \dots, c_5^{hat}, c_6^{EOT}, \dots, c_{M-1}^{EOT}]$ . Next, we perform element-wise addition between the object tokens (i.e.,  $c_5^{dog}$  and  $c_5^{hat}$ ) and the corresponding [EOT] tokens. Specifically, the resulting new embedding is  $\mathcal{C}' = \text{Concat}(\mathcal{C}_1[0 : 4], \mathcal{C}_1[5 : M - 1] + \mathcal{C}_2[5 : M - 1])$ . Afterward, the textual embeddings  $\mathcal{C}'$  are input into the diffusion UNet to generate the images shown in Fig. 3-(a). We can observe that this additivity property allows adding objects (up-left), removing objects (up-right, down-left) and even complex semantic computations (down-right). To explore the mechanism behind this phenomenon, we conducted PCA dimensionality reduction visualization on the token representations of each prompt, as illustrated in Fig. 3-(b). The directional vector obtained from “queen-king” is approximately identical to that of “woman-man” with the cosine similarity of 0.998.

**In conclusion,** our analysis shows that the semantic content of text tokens is coupled and entangled, resulting in attribute confusion across different subjects. Moreover, we found that in diffusion models, text embeddings exhibit semantically additive properties. This implies that the diffusion model is capable of interpreting a composite token, derived from the summation of multiple individual tokens, integrating the semantic attributes of the combined tokens.

<sup>‡</sup>Note in this case, the size of the input textual embedding is  $1 \times 2048$  instead  $77 \times 2048$ .

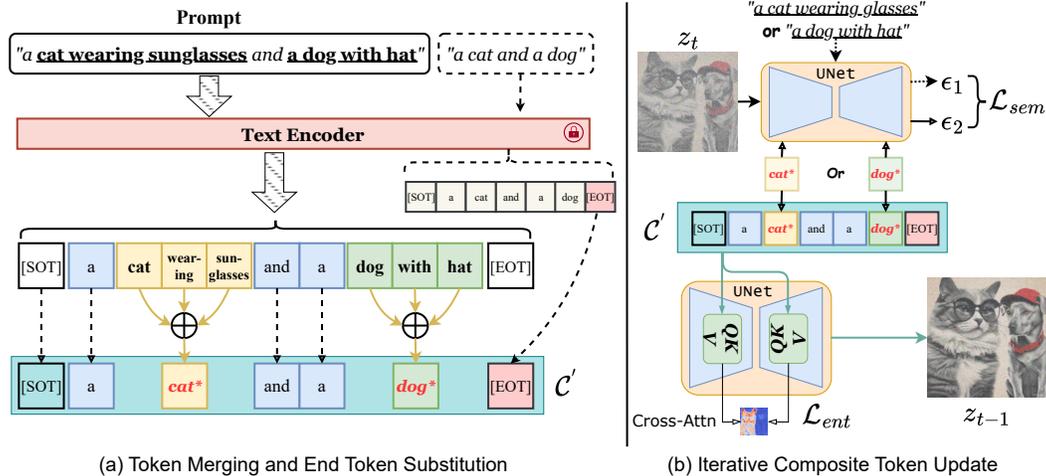


Figure 4: *ToMe* is composed of two parts: one with Token Merging and end token substitution, and the other token updating part with two auxiliary losses for iterative *composite token* update.

### 3.2 *ToMe*: Token Merging

Suppose the initial prompt  $\mathcal{P}$  contains  $K$  entities indicated by noun words and their corresponding tokens as  $\{n^1, \dots, n^k, \dots, n^K\}$ . Each entity is often related to a token with relevant objects or attributes set as  $(n^k, a^k)$ . For example, in the sentence “a cat wearing glasses and a dog with a hat”,  $n^1 = \langle \text{cat} \rangle$ ,  $a^1 = \{\langle \text{wearing} \rangle, \langle \text{glasses} \rangle\}$ ,  $n^2 = \langle \text{dog} \rangle$ ,  $a^2 = \{\langle \text{with} \rangle, \langle \text{a} \rangle, \langle \text{hat} \rangle\}$ .

#### 3.2.1 Token Merging techniques

The semantic additivity of token embeddings inspires us to achieve co-expression of entities and attributes by explicitly binding tokens together. We employ element-wise addition to accomplish semantic merging of tokens. For a prompt  $\mathcal{P}$  containing  $K$  entities, we fuse each subject-attribute pair  $(n^k, a^k)$  into  $\hat{c}_k = n^k + \sum a^k$ , referred to as a *composite token*. This innovative approach introduces an additional benefit by utilizing a single composite token to condense a lengthy prompt sequence, resulting in a unified cross-attention map, thus avoid semantic misalignment. Such observations are further shown in the ablation study and appendix.

**End Token Substitution (ETS).** Meanwhile, as the semantic information contained in [EOT] can interfere with attribute expression, we mitigate this interference by replacing [EOT] to eliminate attribute information contained within them, retaining only the semantic information of each subject. For instance, when the prompt is “a cat wearing hat and a dog wearing sunglasses,” we use the [EOT] obtained from the prompt “a cat and a dog” to replace the original [EOT]. As illustrated in Fig. 4-a, the final text embedding after subject-attribute enhancement and EOT replacement is  $\mathcal{C} = [c_0^{SOT}, c_1^a, c_2^{dog*}, \dots, c_5^{cat*}, c_6^{EOT*}, \dots, c_{76}^{EOT*}]$ . Here,  $dog^*$  and  $EOT^*$  respectively denote tokens after token merging and end token substitution.

#### 3.2.2 Iterative composite Token Update

**Semantic binding loss.** As stated in section 3.1, the semantic information of each token embedding is inherently linked. After strengthening the relationship between subjects and their attributes, it becomes crucial to eliminate any irrelevant semantic information within the composite tokens to prevent misrepresentation of attributes. As illustrated in Fig. 4-(b), to ensure that the semantics of the composite tokens correspond accurately to the noun phrases they are meant to represent, we employ a clean prompt as a supervisory signal. Specifically, for a composite token embedding  $\hat{c}^{dog}$ , which corresponds to the noun phrase “a dog wearing hat”, we aim for the diffusion model to exhibit consistent noise prediction for this composite token and the full phrase. In mathematical terms, this objective can be expressed as ensuring that  $\epsilon_\theta(z_t, \hat{c}^{dog}, t) \approx \epsilon_\theta(z_t, \mathcal{C}, t)$ . This effectively aligns  $\nabla_{z_t} \log P_\theta(z_t | \hat{c}^{dog}) \approx \nabla_{z_t} \log P_\theta(z_t | \mathcal{C})$  [18, 28]. At time step  $t$ , we use the semantic binding loss to align token semantics  $\mathcal{L}_{sem} = \sum_{k \in [1, K]} \|\epsilon_\theta(z_t, \hat{c}_k, t) - \epsilon_\theta(z_t, \mathcal{C}, t)\|_2^2$ .

**Entropy loss.** Following that, we calculate the information carried by each token embedding through entropy statistics. As shown in Fig. 7, we extract the cross-attention map  $\mathcal{A}_k$  corresponding to the  $k$ -th token [27]. After normalizing the cross-attention map as  $\sum_{p_i \in \mathcal{A}_k} p_i = 1$ , we compute the entropy of each token as  $entropy(\mathbf{token}_k) = \sum_{p_i \in \mathcal{A}_k} -p_i \log(p_i)$ . Decreasing the entropy of the cross-attention maps can help ensure that tokens focus exclusively on their designated regions, thereby preventing the cross-attention map from becoming overly divergent. This is further depicted in Fig. 7, where we observe instances of attribute confusion, characterized by different tokens inappropriately influencing the same image region. The entropy regularization loss is defined as  $\mathcal{L}_{ent} = \sum_{k \in [1, K]} \sum_{p_i \in \mathcal{A}_k} -p_i \log(p_i)$  during time step  $t$ .

Finally, the overall  $\mathcal{L} = \mathcal{L}_{ent} + \lambda \cdot \mathcal{L}_{sem}$  is computed by these two novel losses to update the *composite token* during each time  $t < T_{opt}$  and  $\lambda$  is the trade-off hyperparameter.

## 4 Experiments

### 4.1 Experimental Setups

**Evaluation Benchmarks and Metrics.** We evaluate the effectiveness of *ToMe* over T2I-CompBench [31], a comprehensive benchmark for open-world compositional T2I generations, encompassing attribute binding and object relationships. We focus on the semantic binding problem, where T2I-CompBench predominantly evaluates through three attribute subsets (i.e., color, shape, and texture). We follow the evaluation protocol [21, 30, 34] that using 300 validation prompts for evaluation under each subset and the BLIP-VQA score [31] as the evaluation metrics. Following that, we adopt the ImageReward [74] model to evaluate human preference scores, which comprehensively measure image quality and prompt alignment. To comprehensively evaluate *object binding* performance, we introduce a new **GPT-4o Benchmark** of 50 prompts using the template “a [objectA] with a [itemA] and a [objectB] with a [itemB]”. For example, objectA and objectB are objects like “cat” and “dog” while itemA and itemB are associated items “hat” and “glasses”. Afterward, we used the multimodal model GPT-4o [1] to compute the consistency score between the generated images and the prompts for objective assessment. More details are available in the Appendix C.5.

**Implementation Details.** We used SDXL [53] as our base model. To automate image generation for evaluation, we employed SpaCy [29] for syntactic parsing of prompts to identify each object and its corresponding attributes for token merging. The iterative composite token update is performed during the first 20% of the denoising steps  $T_{opt} = 0.2T$ .

**Comparison Methods.** To evaluate our method’s effectiveness, we compared the current state-of-the-art methods. These primarily encompass: (1) state-of-the-art T2I diffusion models, including SDXL [53], Playground-v2 [37] (2) Finetuning-based methods, including CoMat [34], ELLA [30] (3) Optimization-based method SynGen [58] (4) LLM-augmented finetuning-based method Ranni [21]. More comparison results are shown in the Appendix E.

### 4.2 Experimental Results

**Quantitative Comparison.** As shown in Table 1, *ToMe* consistently outperforms or performs comparably to existing methods in BLIP-VQA scores across the color, texture, and shape attribute binding subsets, indicating its effectiveness in avoiding attribute confusion. Human-preference scores evaluated through the ImageReward [74] model (note that the model scores are logits and can be negative) suggest that images generated by *ToMe* can better align with prompts. Specifically, despite ELLA’s [30] use of LLama or T5-XL to replace the CLIP Text Encoder for stronger text embeddings, our method still achieves higher BLIP-VQA scores compared to ELLA. The significant improvement in GPT-4o scores also demonstrates the effectiveness of *ToMe* in *object binding*.

**Qualitative Comparison.** Following SynGen [58], we classify the failure cases of *attribute binding* into three main categories. (i) Semantic leak in prompt, where the attribute  $a^k$  is not corresponding to its entity  $n^k$ ; (ii) Semantic leak out of prompt, where the attribute  $a^k$  is describing the background or some entity not referred to in the prompt  $\mathcal{P}$ ; (iii) Attribute neglect, where the attribute  $a^k$  is totally ignored in the image generation. Fig. 5 presents our qualitative comparison results with other methods. The first three rows show more complex *object binding* results, while the last two rows demonstrate attribute binding results. The semantic binding errors in images generated by SDXL [53]

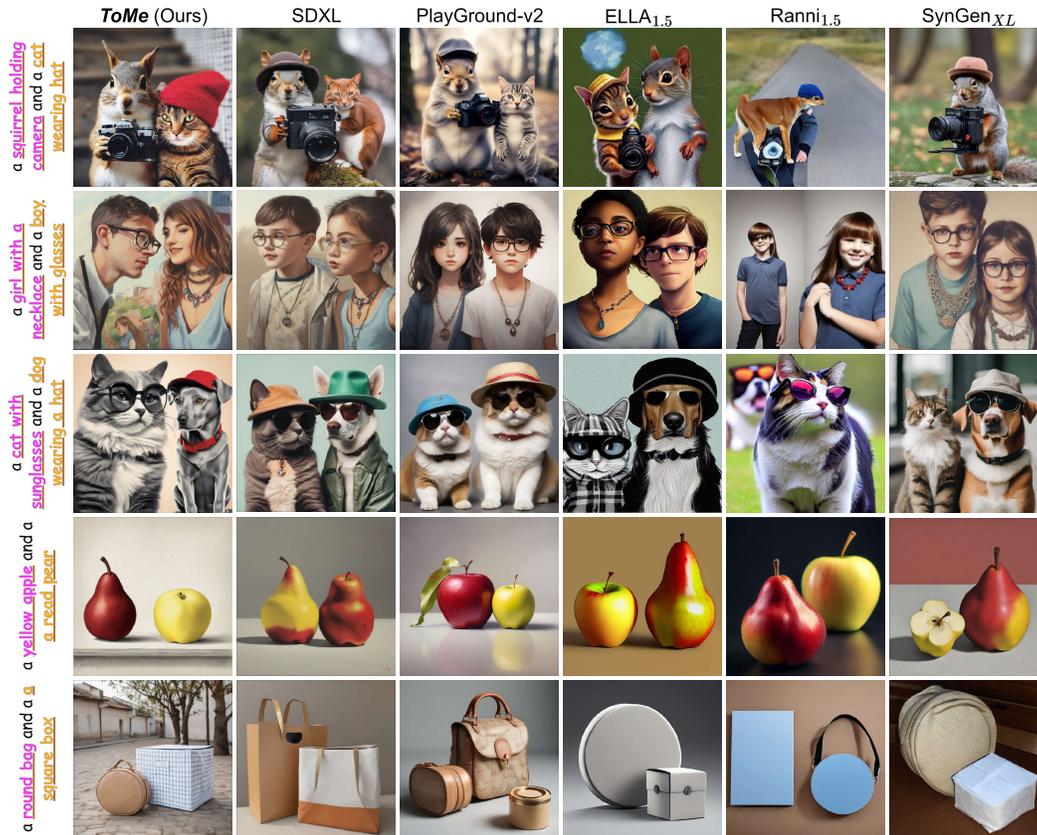


Figure 5: Qualitative comparison among various T2I generation methods with complex prompts.

can largely be attributed to (i) semantic leak in the prompt, as evidenced in the first and second row. Playground-v2[37] confronts similar semantic binding issue as SDXL. ELLA[30] can occasionally succeed in simple attribute binding as in the fifth row, but it frequently encounters (i) semantic leak in the prompt and (iii) attribute neglect errors as shown in the first three prompts. Ranni [21] generates images based on layouts created by a large language model, which can partially address more complex object binding (second row). However, layout-based methods may encounter constraints in achieving proper image layouts, such as shown in the first row with complex descriptions. SynGen [58], which focus on attribute binding problems, achieves good results in color and shape binding but fails in object binding, exhibiting varying degrees of (i) and (iii) failures. Compared to these methods, our

Table 1: 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	Base Model	Train	BLIP-VQA $\uparrow$			Human-preference $\uparrow$			GPT-4o $\uparrow$
			Color	Texture	Shape	Color	Texture	Shape	
SDXL[53]	-	✓	0.6369	0.5637	0.5408	0.7798	0.5140	0.4029	0.4907
PlayG-v2[37]	-	✓	0.6208	0.6125	0.5087	-	-	-	0.5417
Ranni[21]	SD1.5	✓	0.2414	0.3029	0.2857	-0.8554	-0.6853	-0.8051	0.4166
ELLA[30]		✓	0.6911	0.6308	0.4938	0.6586	0.2963	0.0565	0.6481
SynGen[58]		✗	0.6619	0.6451	0.4661	0.4326	0.5072	0.0426	0.5545
CoMat[34]		✓	0.6561	0.6190	0.4975	-	-	-	-
Ranni[21]	SDXL	✓	0.6893	0.6325	0.4934	-	-	-	-
ELLA[30]		✓	0.7260	0.6686	0.5634	-	-	-	-
SynGen[58]		✗	0.7010	0.6044	0.5069	1.016	0.7867	0.4016	0.6458
CoMat[34]		✓	0.7774	0.6591	0.5262	-	-	-	-
ToMe (Ours)	SDXL	✗	0.7656	0.6894	0.6051	1.074	0.9281	0.5916	0.9549

Table 2: Ablation Study conducted on the T2I-CompBench benchmark.

Conf.	<i>ToMe</i>	$\mathcal{L}_{ent}$	$\mathcal{L}_{sem}$	BLIP-VQA		
				Color	Texture	Shape
A	×	×	×	0.6369	0.5637	0.5408
B	✓	×	×	0.6577	0.5828	0.5437
C	✓	✓	×	0.7525	0.6775	0.5797
D	×	✓	✓	0.5881	0.6194	0.5386
E	×	✓	×	0.5983	0.5798	0.5125
F	✓	×	✓	0.6804	0.6263	0.5645
<b>Ours</b>	✓	✓	✓	<b>0.7656</b>	<b>0.6894</b>	<b>0.6051</b>



Figure 6: Text-to-Image generation with various configurations.

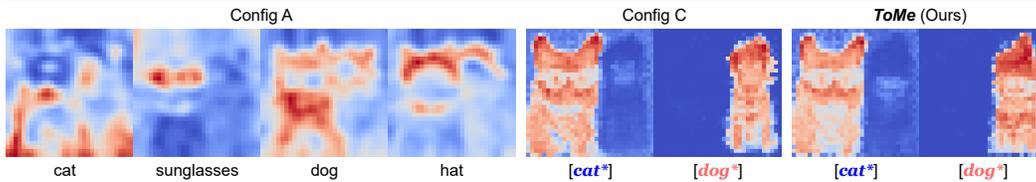


Figure 7: Cross-Attention maps visualization with various configurations.

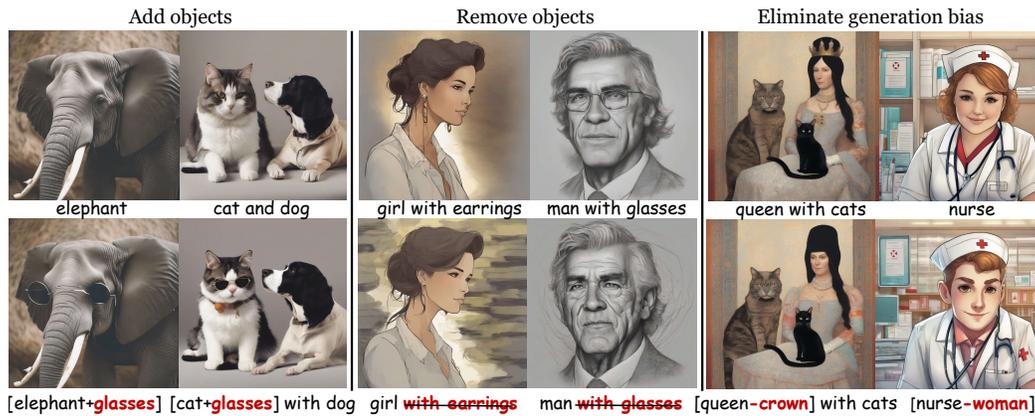


Figure 8: Additional applications of semantic additivity in text embedding.

approach *ToMe* shows improved performance in both object and attribute binding scenarios, which is consistent with the quantitative metrics reflected in Table 1.

**Ablation Study** over each component is quantitatively shown in Table 2. We can observe that using only token merging techniques (with *ToMe* and ETS as config.B) results in a slight performance improvement, which is consistent with the qualitative results in Fig. 6. However, token merging serve as the foundation for subsequent optimizations. When they are combined with the entropy loss  $\mathcal{L}_{ent}$  as config.C, the performance improves significantly. We hypothesize that is partly due to the more regularized cross-attention maps as shown in Fig. 7. Nevertheless, config.C without the semantic binding loss still leads to worse generation performance in Fig. 6, as the dog on the right side still exhibits cat-like features. Incorporating the semantic alignment loss  $\mathcal{L}_{sem}$  (as our default configuration) ensures that the two subjects correctly bind to their respective attributes without appearance confusion, achieving the best results quantitatively and qualitatively. Suppose token merging is ignored, and we only apply the optimization (Config D and Config E), the performances are only comparable to the baseline. Removing  $\mathcal{L}_{ent}$  from *ToMe* (Config F) can also improve over the baseline, but the generation is with noticeable artifacts, which is mainly due to the less regularized cross-attention map. In conclusion, each element of these three novel techniques in *ToMe* contributes to achieving state-of-the-art performance. See Appendix D for more detailed ablation experiments.

**Additional Applications** of *ToMe* are shown in Fig. 8. *ToMe* can not only successfully address the semantic binding problem, it can also be applied to other problems widely exist in T2I generations, including adding objects [84, 70], removing objects [3, 22] and even bias mitigation [16, 61, 77, 78].

## 5 Conclusion

In this paper, we investigate a critical issue in text-to-image (T2I) generation models known as *semantic binding*. This phenomenon refers to instances where T2I models struggle to accurately interpret and visually bind the related semantics. Recognizing that previous methods often entail extensive fine-tuning of the entire T2I model or necessitate explicit specification of generation layouts by large language models, we introduce a novel training-free approach called Token Merging, denoted as *ToMe*, to tackle semantic binding issues in T2I generation. *ToMe* incorporates innovative techniques by stacking up the object token with its relevant tokens into a single *composite token*. This mechanism eliminates the semantic misalignment by unifying the cross-attention maps. Furthermore, we assist the *ToMe* with end token substitution, and iterative composite token updates technique to strengthen the semantic binding. In extensive experiments, we quantitatively compare it against various existing methods using the T2I-Compbench and our proposed GPT-4o benchmarks. The results demonstrate its ability to handle intricate and demanding generation tasks more effectively than current methods, especially for *object binding* cases that are ignored in previous research.

## Acknowledgements

We acknowledge project PID2022-143257NB-I00, financed by the Spanish Government MCIN/AEI/10.13039/501100011033 and FEDER. We acknowledge project "Science and Technology Yongjiang 2035" key technology breakthrough plan project (2024Z120). The Supercomputing Center of Nankai University supports computation.

## References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [2] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasani Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the International Conference on Computer Vision*, pages 2283–2293, 2023.
- [3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.
- [4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qinsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [5] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 843–852, June 2023.
- [6] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. Sega: Instructing text-to-image models using semantic guidance. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 25365–25389. Curran Associates, Inc., 2023.
- [7] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 2023.
- [8] Chieh-Yun Chen, Li-Wu Tsao, Chiang Tseng, and Hong-Han Shuai. A cat is a cat (not a dog!): Unraveling information mix-ups in text-to-image encoders through causal analysis and embedding optimization. *arXiv preprint arXiv:2410.00321*, 2024.

- [9] Hongyu Chen, Yiqi Gao, Min Zhou, Peng Wang, Xubin Li, Tiezheng Ge, and Bo Zheng. Enhancing prompt following with visual control through training-free mask-guided diffusion. *arXiv preprint arXiv:2404.14768*, 2024.
- [10] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\sigma$ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024.
- [11] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- $\alpha$ : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023.
- [12] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.
- [13] Kai Chen, Enze Xie, Zhe Chen, Yibo Wang, Lanqing HONG, Zhenguo Li, and Dit-Yan Yeung. Geodiffusion: Text-prompted geometric control for object detection data generation. In *The Twelfth International Conference on Learning Representations*, 2024.
- [14] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, January 2024.
- [15] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for text-to-image generation and evaluation. In *Advances in Neural Information Processing Systems*, 2023.
- [16] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.
- [17] Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2174–2183, October 2023.
- [18] Bradley Efron. Tweedie’s formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.
- [19] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.
- [20] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations*, 2023.
- [21] Yutong Feng, Biao Gong, Di Chen, Yujun Shen, Yu Liu, and Jingren Zhou. Ranni: Taming text-to-image diffusion for accurate instruction following. *arXiv preprint arXiv:2311.17002*, 2023.
- [22] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436, 2023.
- [23] Hanan Gani, Shariq Farooq Bhat, Muzammal Naseer, Salman Khan, and Peter Wonka. Llm blueprint: Enabling text-to-image generation with complex and detailed prompts. *International Conference on Learning Representations*, 2024.
- [24] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023.

- [25] Biao Gong, Siteng Huang, Yutong Feng, Shiwei Zhang, Yuyuan Li, and Yu Liu. Check, locate, rectify: A training-free layout calibration system for text-to-image generation. *arXiv preprint arXiv:2311.15773*, 2023.
- [26] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- [27] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *International Conference on Learning Representations*, 2023.
- [28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.
- [29] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017.
- [30] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.
- [31] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 78723–78747. Curran Associates, Inc., 2023.
- [32] Vikram Jamwal and S Ramaneswaran. Composite diffusion: whole  $\geq$   $\sigma$  parts. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7206–7215. IEEE, 2024.
- [33] Yuhao Jia and Wenhan Tan. Divcon: Divide and conquer for progressive text-to-image generation. *arXiv preprint arXiv:2403.06400*, 2024.
- [34] Dongzhi Jiang, Guanglu Song, Xiaoshi Wu, Renrui Zhang, Dazhong Shen, Zhuofan Zong, Yu Liu, and Hongsheng Li. Comat: Aligning text-to-image diffusion model with image-to-text concept matching. *arXiv preprint arXiv:2404.03653*, 2024.
- [35] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. If at first you don't succeed, try, try again: Faithful diffusion-based text-to-image generation by selection. *arXiv preprint arXiv:2305.13308*, 2023.
- [36] Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, and Jun-Yan Zhu. Dense text-to-image generation with attention modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7701–7711, 2023.
- [37] Daiqing Li, Aleks Kamko, Ali Sabet, Ehsan Akhgari, Linmiao Xu, and Suhail Doshi. Playground v2.
- [38] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.
- [39] Sen Li, Ruochen Wang, Cho-Jui Hsieh, Minhao Cheng, and Tianyi Zhou. Mulan: Multimodal-llm agent for progressive multi-object diffusion. *arXiv preprint arXiv:2402.12741*, 2024.
- [40] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Stylediffusion: Prompt-embedding inversion for text-based editing, 2023.
- [41] Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *International Conference on Learning Representations*, 2024.

- [42] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, June 2023.
- [43] Yumeng Li, Margret Keuper, Dan Zhang, and Anna Khoreva. Divide & bind your attention for improved generative semantic nursing. *Proceedings of the British Machine Vision Conference*, 2023.
- [44] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. Llm-grounded diffusion: Enhancing prompt understanding of text-to-image diffusion models with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024.
- [45] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.
- [46] Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Correcting diffusion generation through resampling. *arXiv preprint arXiv:2312.06038*, 2023.
- [47] Wan-Duo Kurt Ma, Avisek Lahiri, JP Lewis, Thomas Leung, and W Bastiaan Kleijn. Directed diffusion: Direct control of object placement through attention guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4098–4106, 2024.
- [48] Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. *arXiv preprint arXiv:2312.06059*, 2023.
- [49] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.
- [50] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024.
- [51] Pablo Pernias, Dominic Rampas, Mats Leon Richter, Christopher Pal, and Marc Aubreville. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In *International Conference on Learning Representations*, 2024.
- [52] Quynh Phung, Songwei Ge, and Jia-Bin Huang. Grounded text-to-image synthesis with attention refocusing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- [53] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [54] Zipeng Qi, Guoxi Huang, Zebin Huang, Qin Guo, Jinwen Chen, Junyu Han, Jian Wang, Gang Zhang, Lufei Liu, Errui Ding, et al. Layered rendering diffusion model for zero-shot guided image synthesis. *arXiv preprint arXiv:2311.18435*, 2023.
- [55] Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the ACM International Conference on Multimedia*, pages 643–654, 2023.
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

- [57] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [58] Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *Advances in Neural Information Processing Systems*, 36, 2023.
- [59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 06 2022.
- [60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [61] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. *International Conference on Learning Representations*, 2024.
- [62] Alex Shonenkov, Misha Konstantinov, Daria Bakshandaeva, Christoph Schuhmann, Ksenia Ivanova, and Nadiia Klokova. Deepfloyd-iff. <https://github.com/deep-floyd/IF>, 2023.
- [63] Maria Mihaela Trusca, Wolf Nuyts, Jonathan Thomm, Robert Honig, Thomas Hofmann, Tinne Tuytelaars, and Marie-Francine Moens. Object-attribute binding in text-to-image generation: Evaluation and control. *arXiv preprint arXiv:2404.13766*, 2024.
- [64] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [65] Hazarapet Tunanyan, Dejia Xu, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Multi-concept t2i-zero: Tweaking only the text embeddings and nothing else. *arXiv preprint arXiv:2310.07419*, 2023.
- [66] Kai Wang, Fei Yang, Shiqi Yang, Muhammad Atif Butt, and Joost van de Weijer. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 2023.
- [67] Ruichen Wang, Zekang Chen, Chen Chen, Jian Ma, Haonan Lu, and Xiaodong Lin. Compositional text-to-image synthesis with attention map control of diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5544–5552, 2024.
- [68] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. *arXiv preprint arXiv:2402.03290*, 2024.
- [69] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Grounding diffusion with token-level supervision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- [70] Navve Wasserman, Noam Rotstein, Roy Ganz, and Ron Kimmel. Paint by inpaint: Learning to add image objects by removing them first. *arXiv preprint arXiv:2404.18212*, 2024.
- [71] Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7766–7776, 2023.
- [72] Yinwei Wu, Xingyi Yang, and Xinchao Wang. Relation rectification in diffusion model, 2024.
- [73] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the International Conference on Computer Vision*, pages 7452–7461, 2023.

- [74] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.
- [75] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *International Conference on Machine Learning*, 2024.
- [76] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, June 2023.
- [77] Hidir Yesiltepe, Kiymet Akdemir, and Pinar Yanardag. Mist: Mitigating intersectional bias with disentangled cross-attention editing in text-to-image diffusion models. *arXiv preprint arXiv:2403.19738*, 2024.
- [78] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3969–3980, 2023.
- [79] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [80] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023.
- [81] Xincheng Zhang, Ling Yang, Yaqi Cai, Zhaochen Yu, Jiake Xie, Ye Tian, Minkai Xu, Yong Tang, Yujiu Yang, and Bin Cui. Realcompo: Dynamic equilibrium between realism and compositionality improves text-to-image diffusion models. *arXiv preprint arXiv:2402.12908*, 2024.
- [82] Yang Zhang, Teoh Tze Tzun, Lim Wei Hern, Tiviatis Sim, and Kenji Kawaguchi. Enhancing semantic fidelity in text-to-image synthesis: Attention regulation in diffusion models. *arXiv preprint arXiv:2403.06381*, 2024.
- [83] Yasi Zhang, Peiyu Yu, and Ying Nian Wu. Object-conditioned energy-based attention map alignment in text-to-image diffusion models. *arXiv preprint arXiv:2404.07389*, 2024.
- [84] Ziyue Zhang, Mingbao Lin, and Rongrong Ji. Objectadd: Adding objects into image via a training-free diffusion modification fashion. *arXiv preprint arXiv:2404.17230*, 2024.
- [85] Peiang Zhao, Han Li, Ruiyang Jin, and S Kevin Zhou. Loco: Locally constrained training-free layout-to-image synthesis. *arXiv preprint arXiv:2311.12342*, 2023.
- [86] Dewei Zhou, You Li, Fan Ma, Zongxin Yang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024.
- [87] Yupeng Zhou, Daquan Zhou, Zuo-Liang Zhu, Yaxing Wang, Qibin Hou, and Jiashi Feng. Maskdiffusion: Boosting text-to-image consistency with conditional mask. *arXiv preprint arXiv:2309.04399*, 2023.

## Appendix

### A Limitations

Since our method is optimized for inference based on SDXL, it inherits some inherent limitations of SDXL. For example, it may produce artifacts in generated images and is unable to create images with complex layouts. Additionally, the *ToMe* technique relies on the CLIP text encoder to generate text embeddings, which may be subject to the limitations of the encoder itself. For instance, the CLIP encoder might not fully capture all the subtle semantic nuances in the text, which could restrict the performance of *ToMe* when processing certain types of text prompts. Addressing these limitations and advancing our understanding in these areas will help improve image generation technology.

### B Broader Impacts

*ToMe* enhances the semantic binding capability in text-to-image synthesis by enhancing text embeddings. However, it also carries potential negative implications. It could be used to generate false or misleading images, thereby spreading misinformation. If *ToMe* is applied to generate images of public figures, it poses a risk of infringing on personal privacy. Additionally, the automatically generated images may also touch upon copyright and intellectual property issues.

### C Implementation Details

#### C.1 Method details

We extract the cross attention maps from the first three layers of the decoder in the UNet backbone, which contain rich semantic information, with a resolution of  $32 \times 32$ . For *Iterative composite Token Update*, since the early timesteps of the denoising process determine the layout of the image[27], we execute it only during the first 20% of the denoising process. All experiments were conducted on an NVIDIA-A40 GPU.

#### C.2 Baseline methods implementation

For the quantitative comparison in Tab. 1, we used the official implementations of Ranni[21], ELLA[30], SyGen[58], and CoMat[34]. Since the SDXL versions of the Ranni[21], ELLA[30], and CoMat[34] methods have not been open-sourced, we refer to the BLIP-VQA scores reported in their respective papers. SynGen[58], like our method, performs optimization during inference. To ensure a fairer comparison, we adapted SynGen to SDXL.

#### C.3 Text embedding analysis

Fig. 9's statistical analysis further demonstrates the information coupling property and semantic additivity of text embeddings. We employed MMDetection[12] and GLIP[38] to detect the probability of specified objects in images, referred to as *DetScore*, as shown in Fig. 9-(a). Fig. 9-(b) presents statistical results on 100 generated images, showing that the probability of detecting a hat in images generated from the text embedding corresponding to "a dog" is 0%. However, in images generated from the element-wise "[dog+hat]" additive embedding, the probability of detecting a hat is 68.61%, which is close to the probability of 73.12% for images generated using the prompt 'a dog wearing a hat'.

The information coupling of token embeddings is also reflected in the entropy of cross-attention for each token. Taking the prompt "a cat wearing sunglasses and a dog wearing a hat" as an example, we can extract the cross-attn map  $\mathcal{A}_k \in \mathbb{R}^{1024}$  for each token, averaged over 50 time steps and multiple heads. After normalizing each map to 1.0 (i.e.,  $\mathcal{A}_k[i] := \frac{\mathcal{A}_k[i]}{\sum_{i \in [1, 32]} \mathcal{A}_k[i]}$ ), we calculate the token's information entropy as  $\sum_{p_i \in \mathcal{A}_k} -p_i \log(p_i)$ . As shown in Fig. 9-(c), we conducted statistics on 100 generated images and found that tokens positioned later in the prompt tend to have higher entropy, indicating more dispersed cross-attn maps. This phenomenon might be attributed to CLIP's[56] masked attention mechanism, where each token can interact with all preceding tokens, and tokens

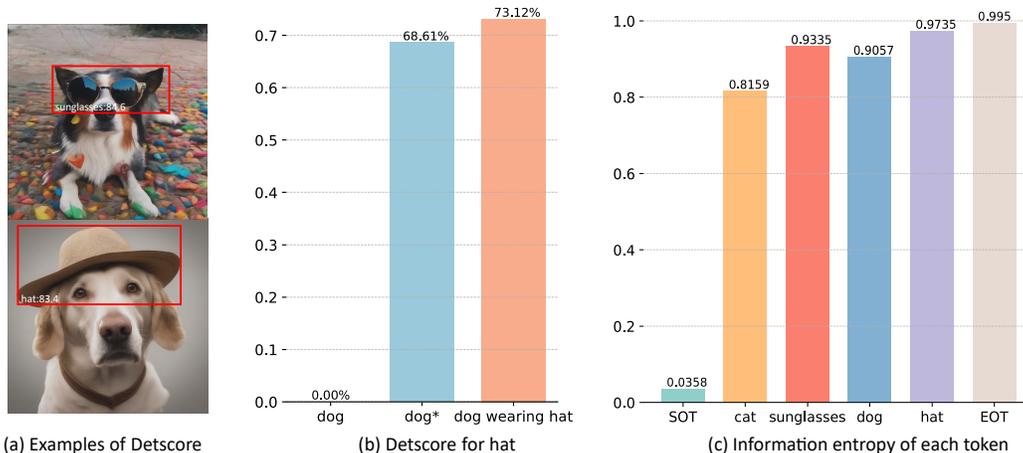


Figure 9: Additional statistical analyses, all statistical values are averaged results from 100 images. (a) An example of DetScore visualization. (b) By fusing the dog and hat token, we obtain dog\*, and the generated images often include a hat. The DetScore value for dog\* is close to the DetScore value obtained using the complete prompt “a dog wearing a hat”. (c) We calculated the entropy of the cross-attention maps for each token and found that tokens positioned later in the sequence generally have higher entropy, indicating that their cross-attention maps are more dispersed.

positioned later can interact with more tokens, thus containing more information. Consequently, we employ an entropy regularization loss to constrain each attention map to be as concentrated as possible, thereby reducing the amount of irrelevant information contained in each token embedding.

#### C.4 Time complexity

Table 3: Time Complexity of various methods. The results of our method are highlighted in bold.

Method	Inference Steps	Time Cost	Color	Texture	Shape
SDXL	20	18s	0.6136	0.5449	0.5260
ToMe (Config C)	20	23s	<b>0.7419</b>	<b>0.6581</b>	<b>0.5742</b>
ToMe (Ours)	20	45s	<b>0.7612</b>	<b>0.6653</b>	<b>0.5974</b>
Ranni (SDXL)	50	87s	0.6893	0.6325	0.4934
ELLA (SDXL)	50	51s	0.7260	0.6686	0.5634
SynGen (SDXL)	50	67s	0.7010	0.6044	0.5069
SDXL	50	42s	0.6369	0.5637	0.5408
ToMe (Config C)	50	56s	<b>0.7525</b>	<b>0.6775</b>	<b>0.5797</b>
ToMe (Ours)	50	83s	<b>0.7656</b>	<b>0.6894</b>	<b>0.6051</b>

Tab. 3 reports the inference time costs of various methods, all measured on a single NVIDIA-A40 GPU. We demonstrate that our method does not significantly increase inference time while improving semantic binding performance with 50 inference steps. We further extend this analysis by measuring the time cost with 20 inference steps and various ToMe configurations, as shown in the Tab. 3. We report the time cost (by seconds) along with BLIP-VQA scores across the color, texture, and shape attribute binding subsets. From this table, we can observe that using the token merging (ToMe) technique and entropy loss (Config.C), our method achieves excellent performance with minimal additional time cost. Additionally, even with only 20 inference steps, our method, ToMe, maintains high performance with very little degradation.

#### C.5 GPT-4o Score

In order to better demonstrate the binding ability of our model for complex prompts. We have constructed a set of high-difficulty prompts, where the content primarily uses nouns to describe the subject. We use OpenAI’s latest release, GPT-4o, to evaluate the quality of images generated by various models because GPT-4o excels in image discernment, allowing for precise evaluation of the generated outputs. As show in Fig. 10, We designed nine scoring levels, ranging from 0 to 100 points,

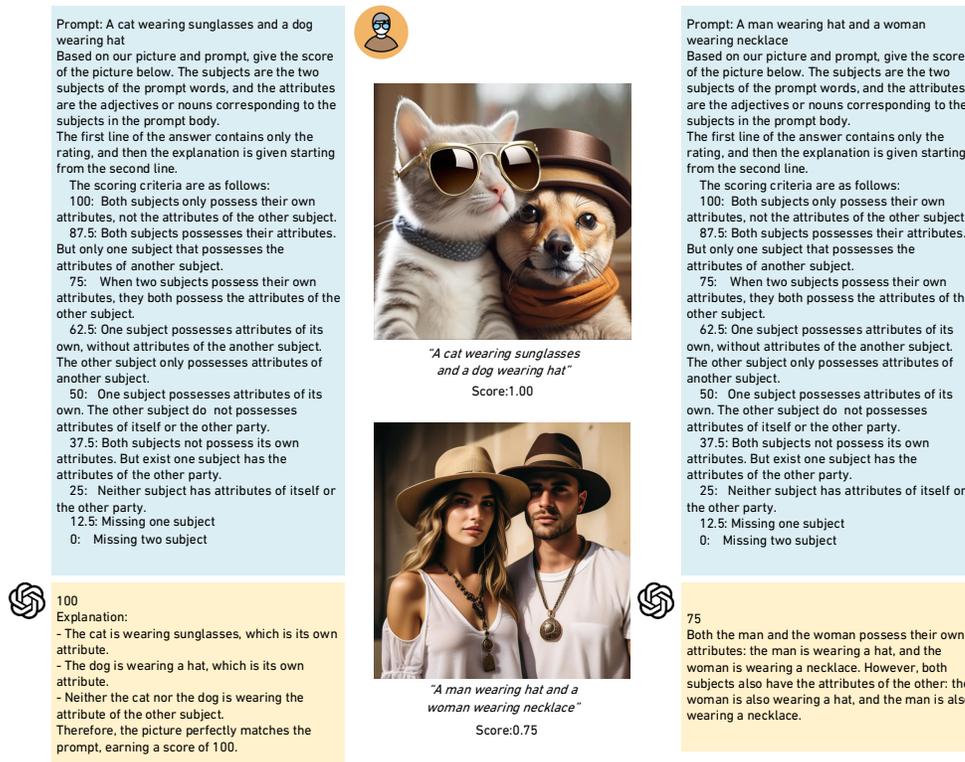


Figure 10: Evaluation Metric: GPT-4o

based on factors such as whether the objects correctly possess their attributes, the mixing of attributes between objects, and whether the objects are correctly generated, to distinguish different levels of generation quality.

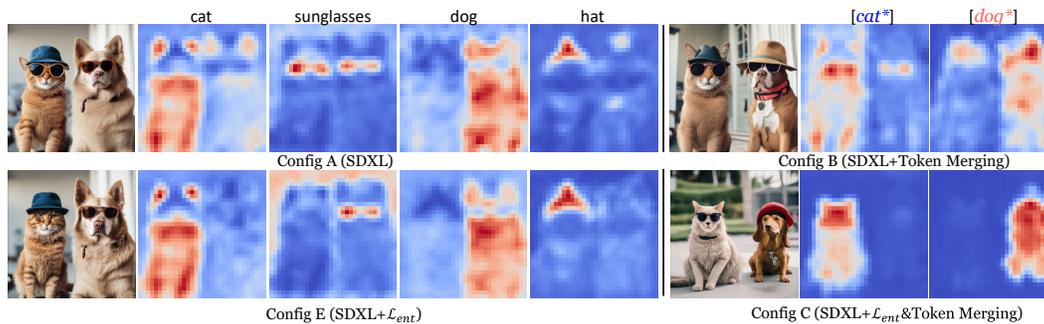


Figure 11: Cross-attention maps visualization with various configurations, with the input prompt “a cat wearing sunglasses and a dog wearing hat”

## D Additional Ablation Studies

### D.1 More Configures and ETS ablation

As an example in Fig. 11, the original SDXL (Config.A) suffered from attribute binding errors due to divergent cross-attention maps. When only applying token merging (Config B), the co-expression of entities and attributes resulted in a dog wearing a hat in the image, but the attribute leakage issue remained due to the divergent cross-attention maps. When only applying the entropy loss  $\mathcal{L}_{ent}$  (Config E), although the cross-attention maps corresponding to each token are more concentrated, they may focus on wrong regions. Only by applying both token merging and  $\mathcal{L}_{ent}$  techniques (Config

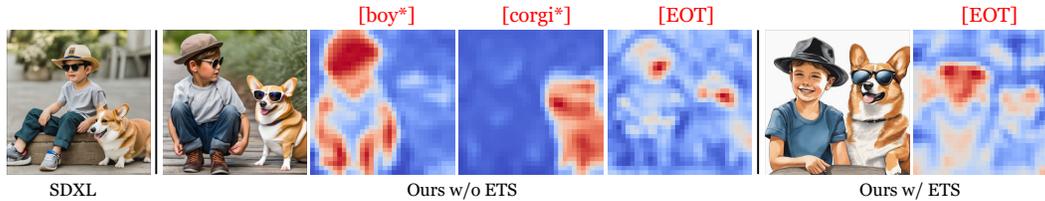


Figure 12: Ablation study of our proposed end token substitution (ETS) technique, with the input prompt “a boy wearing hat and a dog wearing sunglasses”

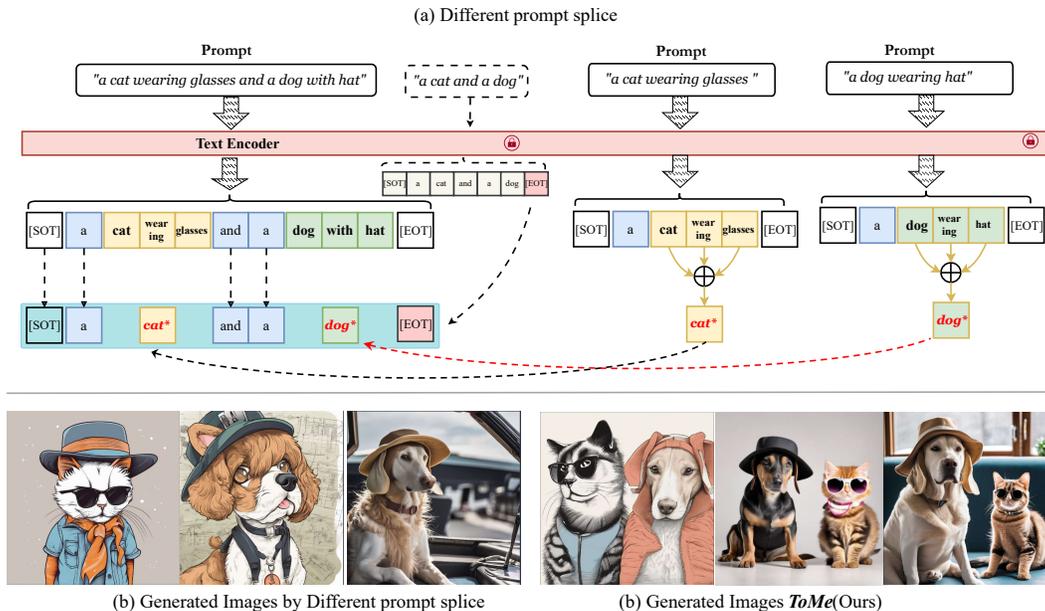


Figure 13: Comparison of images generated by different prompts splice

C), the cross-attention map of the composite token becomes better concentrated on the correct areas and thus leading to more satisfactory semantic binding of entities and attributes.

The end token substitution (ETS) technique is proposed to address potential semantic misalignment in the final tokens of long sequences. As the [EOT] token interacts with all tokens, it often encapsulates the entire semantic information, as shown in Fig. 2. Therefore, the semantic information in [EOT] can interfere with attribute expressions, we mitigate this by replacing [EOT] to remove the attribute information it contains from the original prompts, retaining only the semantic information for each subject.

For example, as the cross-attention maps and T2I generation performance shown in Fig. 12, when ToMe is not combined with the EST technique, the ‘sunglasses’ semantics contained in the EOT token cause the boy to incorrectly wear sunglasses. However, when combined with ETS, the unwanted semantic binding is relieved.

## D.2 Different prompts splice

In Sec. 3.2.1, we fuse each object and its corresponding attributes. At this stage, both the object token embedding and the attribute token embedding are derived from the text embedding obtained by processing the same prompt through the CLIP Text Encoder, potentially causing the information between them to be coupled. We also experimented with splicing token embeddings from different prompts, as illustrated in Fig. 13. While keeping other components of ToMe unchanged, the resulting images often exhibit a missing of the object. We hypothesize that this may be due to the lack of contextual semantics between token embeddings from different prompts[8].

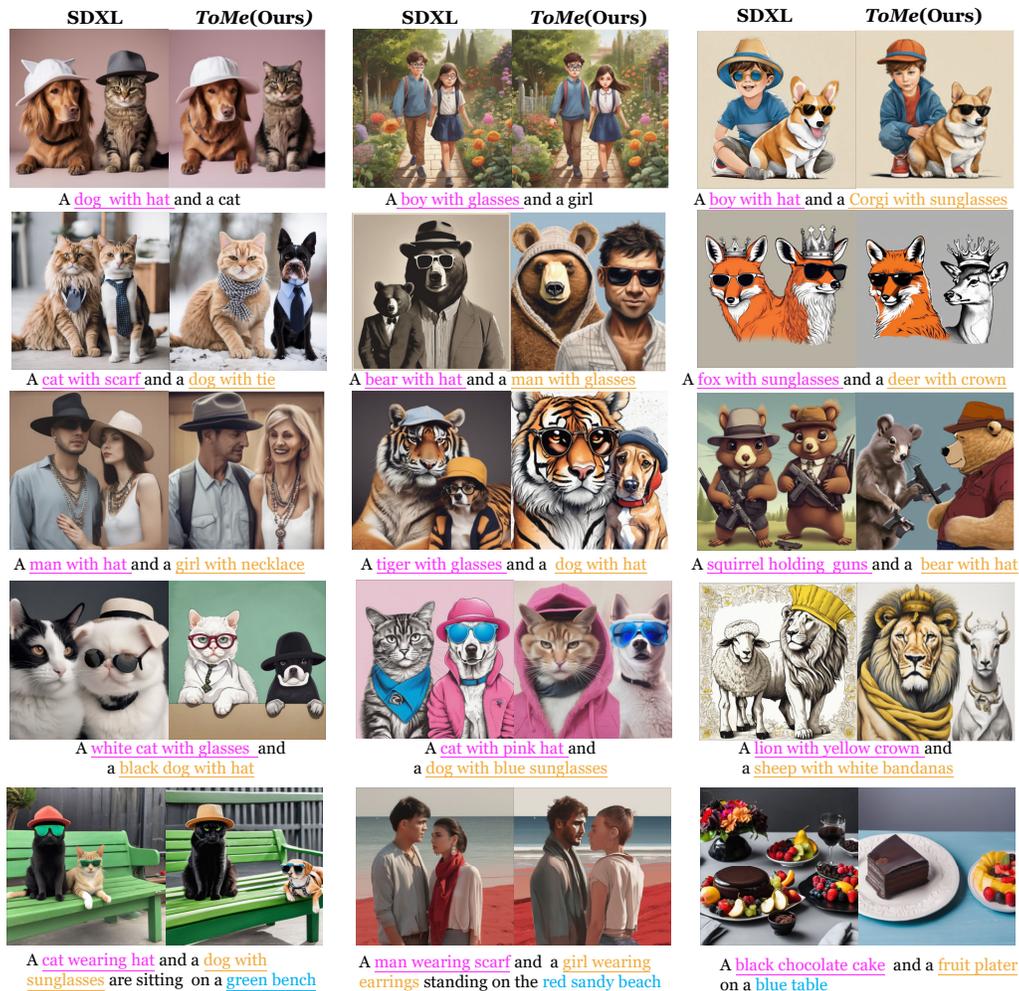


Figure 14: Additional semantic binding results. Our method not only achieves good results in object binding but is also effective for composite binding of objects and their adjective attributes.

## E Additional Results

As shown in Tab. 4, we have added quantitative comparison results with additional methods. Our method consistently outperforms or is on par with the existing methods. Fig. 14 presents more qualitative comparison results, demonstrating that our method achieves good performance in attribute binding, object binding, and the composite binding of attribute and objects. *ToMe* can also generate images with subjects or backgrounds featuring multiple attributes (Fig. 14, the last line), in this scenario, we find that using an additional positional loss [19] based on the attention map is effective.

We also conduct a user study with 20 participants to enrich the evaluation. Here we compare our method *ToMe* with SDXL [53], SynGen [58], Ranni [21] and ELLA [30]. As shown in Fig. 15, we ask the participants to rate the semantic binding into 4 levels and calculate the distribution of each comparison method over these four diverse levels. We can observe that our method better achieves semantic binding performance by mainly distributing in the highest level 1, while the other methods struggle to obtain user satisfactory results.

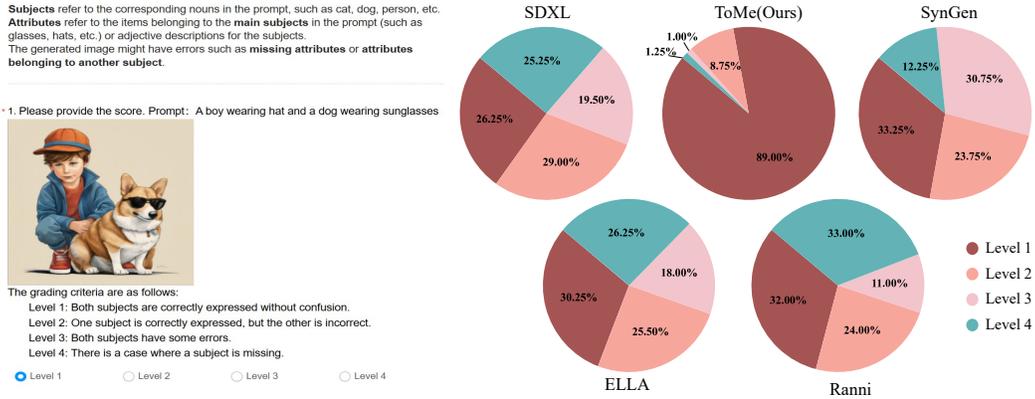


Figure 15: User study with 20 participants, we ask users to rate the semantic binding into four levels.

Table 4: Comparison of BLIP-VQA Scores

Method	Base Model	Train	BLIP-VQA ↑		
			Color	Texture	Shape
SD v1.5[60]	-	✓	0.3750	0.4159	0.3724
SD v2[60]	-	✓	0.5065	0.4922	0.4221
DALL-E2[57]	-	✓	0.5750	0.6374	0.5464
SDXL[53]	-	✓	0.6369	0.5637	0.5408
PlayG-v2[37]	-	✓	0.6208	0.6125	0.5087
Ranni[21]	SD1.5	✓	0.2414	0.3029	0.2857
ELLA[30]	SD1.5	✓	0.6911	0.6308	0.4938
SynGen[58]	SD1.5	×	0.6619	0.6451	0.4666
CoMat[34]	SD1.5	✓	0.6561	0.6190	0.4975
Composable v2[45]	SD2.0	×	0.4063	0.3645	0.3299
Structured v2[20]	SD2.0	×	0.4990	0.4900	0.4218
Attn-Exct v2[7]	SD2.0	×	0.6400	0.5963	0.4517
GORS[31]	SD2.0	×	0.6603	0.6287	0.4785
Ranni[21]	SDXL	✓	0.6893	0.6325	0.4934
ELLA[30]	SDXL	×	0.7260	0.6686	0.5634
SynGen[58]	SDXL	×	0.7010	0.6044	0.5069
CoMat[34]	SDXL	✓	0.7774	0.6591	0.5262
<i>ToMe</i> (Ours)	SDXL	×	0.7656	0.6894	0.6051

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Abstract and Sec. 1

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Appendix A

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Sec. 3

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

#### 4. Experimental Result Reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Sec. 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

#### 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Supplementary Material

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Sec. 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Sec. 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix C.1

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: We have carefully checked the NeurIPS code of ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Appendix B

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We politely cited the existing assets and read their usage license.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

#### 13. New Assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

#### 14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Appendix E

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

#### 15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.