Text Embedding is Not All You Need: Attention Control for Text-to-Image Semantic Alignment with Text Self-Attention Maps

Jeeyung Kim*, Erfan Esmaeili*, and Qiang Qiu
Purdue University
{jkim17, efakhabi, qqiu}@purdue.edu

Abstract

In text-to-image diffusion models, the cross-attention map of each text token indicates the specific image regions attended. Comparing these maps of syntactically related tokens provides insights into how well the generated image reflects the text prompt. For example, in the prompt, "a black car and a white clock", the cross-attention maps for "black" and "car" should focus on overlapping regions to depict a black car, while "car" and "clock" should not. Incorrect overlapping in the maps generally produces generation flaws such as missing objects and incorrect attribute binding. Our study makes the key observations investigating this issue in the existing text-to-image models: (1) the similarity in text embeddings between different tokens—used as conditioning inputs—can cause their cross-attention maps to focus on the same image regions; and (2) text embeddings often fail to faithfully capture syntactic relations already within text attention maps. As a result, such syntactic relationships can be overlooked in cross-attention module, leading to inaccurate image generation. To address this, we propose a method that directly transfers syntactic relations from the text attention maps to the cross-attention module via a test-time optimization. Our approach leverages this inherent yet unexploited information within text attention maps to enhance image-text semantic alignment across diverse prompts, without relying on external guidance. Our project page and code are available at: https://t-sam-diffusion.github.io/

1. Introduction

Recent diffusion models can generate images from diverse text prompts [4, 7, 29]. However, semantic discrepancies often arise between the text and generated images, raising problems such as missing objects—where certain elements are overlooked—and attribute mis-binding—where attributes are incorrectly assigned to subjects.

*Equal contribution.

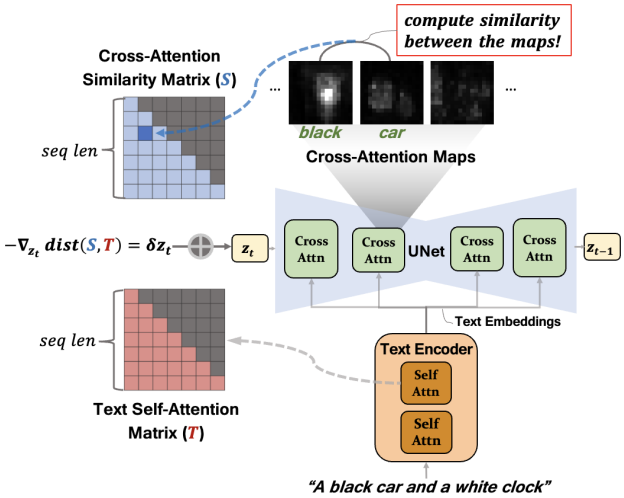
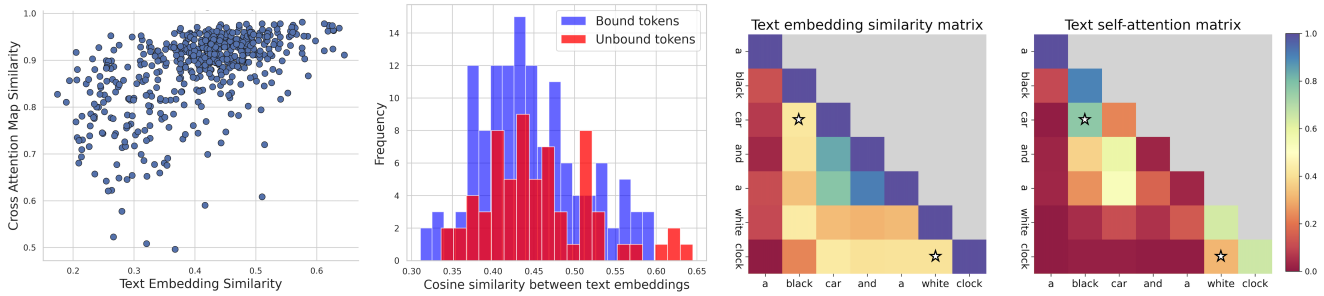


Figure 1. The overview of our method. We leverage text self-attention matrix and optimize the latent noise (z_t) by minimizing the distance between the cross-attention similarity matrix (S) and the text self-attention matrix (T). This encourages integrating syntactic relationships into text-to-image diffusion models.

Prior studies [10] demonstrated that the cross-attention map of each token in text-to-image (T2I) diffusion models highlights the attended regions in the image and provides clues about the spatial placement of elements corresponding to the tokens. In particular, [1, 21, 28] suggest that spatial alignment in cross-attention maps among related words influences the fidelity of images to text prompts in Stable Diffusion (SD) [29], as we also demonstrate in Section 4.1. For instance, in the prompt *a black car and a white clock*, if the cross-attention maps for *car* and *clock* overlap, unique token contributions can dilute, potentially omitting one object. Conversely, if the maps for *black* (or *white*) and *car* (or *clock*) diverge too much, attribute mis-binding can occur. This implies syntactically related words should ideally have spatially aligned cross-attention maps, as discussed in [8, 28]. However, the factors determining this spatial alignment between the maps remain poorly understood.

In our study, we first investigate the factors that con-



(a) (b) (c)

Figure 2. For the analysis, we use SD v1.5 as the base model and adopt the prompt sets from [3], structured as "[attribute$_1$] [object$_1$] and [attribute$_2$] [object$_2$]", "[object$_1$] and/with [object$_2$]" or "[object$_1$] and [attribute$_2$] [object$_2$]". (a) Comparison of the cosine similarity of text embeddings with that of the corresponding cross-attention maps at denoising step 1, with pairs of tokens (object$_i$, object$_j$), where $i \neq j$, and pairs of tokens (attribute$_m$, object$_n$) for both $m = n$ and $m \neq n$. As text embeddings become more similar, their cross-attention maps get similar. (b) The distributions of text embedding similarity between i) *Bound tokens*— (attribute$_i$, object$_i$) for $i = 1, 2$, and ii) *Unbound tokens*—(object$_1$, object$_2$). The distributions show no discernible difference, indicating text embeddings do not effectively represent the syntactic relationships. (c) Comparison of text embedding similarity (left) and the text self-attention map power by 3 (right) for the prompt *a black car and a white clock*. In the self-attention maps ($\top$), *clock* attends more to *white*, unlike the text embeddings.

tribute to the spatial alignment of cross-attention maps across different tokens, which can ultimately affect the accuracy of T2I generation. We reveal that text embeddings, which function as keys in the cross-attention modules, play a pivotal role in determining the similarity of these maps, as illustrated in Figure 2a. Specifically, as the text embeddings for different words become more similar, the cosine similarity of their corresponding cross-attention maps increases. This effect is noticeable from the initial denoising step.

Considering that (i) cross-attention maps should capture syntactic relationships between words and that (ii) the spatial alignment of cross-attention maps is influenced by the text embeddings, we question whether text embeddings accurately capture the linguistic structure in text prompts. Our findings suggest they do not, as illustrated in Figure 2b and the left of Figure 2c. Syntactically related words in prompts (*e.g.*, *black-car* and *white-clock*) do not necessarily yield similar text embeddings, as detailed in Section 4.1. This finding is further supported by [40], which argues that CLIP [25] often behaves like a bag-of-words model. As a result, the cross-attention maps, derived from text embeddings, can likely fail to faithfully reflect the syntactic relationships. In the end, the T2I diffusion models can often generate images that inaccurately represent the prompts.

To address the above issue, prior studies [1, 8, 21, 28] resort to external sources to obtain syntactic information within text prompts and regularize cross-attention maps to incorporate these relationships. Specifically, [8, 28] employ external text parsers to obtain linguistic structure, while [1, 21] rely on human intervention to manually group tokens based on syntactic relationships. However, these methods are limited by their dependence on external inputs.

Do we really need to rely on external sources to obtain syntactic relations within a sentence? Remarkably, T2I diffusion models inherently capture syntactic relationships within prompts through the self-attention maps in their text encoder. In the encoder's self-attention module, each token pays more attention to related words, leading to encoding the entire sentence effectively. As shown in the right of Figure 2c, the text self-attention maps exhibit higher attention scores between the syntactically related words. However, this information is weakly encoded in the text encoder outputs (*i.e.*, text embeddings). We conjecture this is due to the text attention module's strong focus on the <bos> token, known as *attention sink* [32, 35], which minimizes the influence of other tokens, as further discussed in Section 4.1.

As illustrated in Figure 1, our approach reuses the self-attention maps from the text encoder and directly transfers the syntax information to the diffusion models. We first define a similarity matrix whose values indicate the similarity between pairs of cross-attention maps. Then, we update the latent noise in diffusion models to minimize the distance between the cross-attention similarity matrix and the text self-attention matrix during inference. This allows for the seamless integration of syntactic relationships into the diffusion process. By guiding cross-attention modules to better capture contextual relationships, our method ultimately produces images accurately reflecting the intended meaning of prompts. We leverage overlooked information already embedded in T2I models, offering two key advantages: **(i) Self-contained**: no need for external inputs like text parsers or manual token indices; and **(ii) Generalizable**: effective across diverse sentence structures.

2. Related Work

Text-to-image generation. Text-to-image (T2I) generation [4, 5, 7, 23, 27, 30] is commonly based on latent diffusion models [29] where text data is processed via a text encoder

[25, 26]. An important problem in T2I models is the lack of full correspondence between the input text and the image. Although diverse methods [2, 8, 12, 19, 24, 31, 38] are proposed to diagnose and fix this problem, a common theme in prior works is adjusting the cross-attention maps in inference time, following the seminal work [10].

Cross-attention control for improved sematic alignment. Several defects in cross-attention maps hinder sematic alignment, including attention dominance [41]—one token getting huge attention weight, and attention leakage/overlap [20, 36]—attention weights not respecting the spatial boundaries of the intended objects. In particular, one line of work employs contrastive objectives to guide attention optimization [1, 14, 21, 28]. For example, *CONFORM* [21] uses prompts with distinct groups, where tokens in the same (opposite) group are treated as positive (negative) pairs. Similarly, [28] uses text parser models to identify such pairs. Other approaches, like those in [15, 34], optimize attention maps with spatial guidance (e.g., segmentation maps, masks, or layouts). However, these methods rely on external resources for text-to-image generation.

Attention sinks. The attention weights in pretrained transformers tend to be heavily biased towards special tokens such as <bos>, <eos>, and punctuation tokens, as studied in (vision-)language models [9, 32, 35]. The T2I diffusion model literature [3, 37] also notes the focus of attention scores on the <bos> token in the CLIP text encoder.

3. Preliminaries

In this section, we briefly review the structure of text-to-image diffusion models including the text encoder and the cross-attention module [4, 29].

Text encoder. To condition the text data for the diffusion process, a text encoder is needed. The input text is first tokenized and then converted into dense vectors. The dense vectors are then processed via a series of multi-head self-attention layers [33].

At a given layer ℓ in the text encoder, let us denote the *key* by $\mathbf{e}_i^{(\ell)} \in \mathbb{R}^{H_\mathrm{e} D_\mathrm{e}}, i = 1, \cdots, s$, where s is the key sequence length, D_e is the embedding dimension per head and H_e is the number of heads in the text encoder. The corresponding self-attention matrix $T^{(\ell,h)} \in \mathbb{R}^{s \times s}$ is given by

$$T_{ij}^{(\ell,h)} = \frac{\exp(\omega_{ij})}{\sum_k \exp(\omega_{ik})}, \ \omega_{ij} := \mathbf{e}_i^{(\ell)\top} W_\mathrm{en}^{(\ell,h)} \mathbf{e}_j^{(\ell)}, \quad (1)$$

where $W_\mathrm{en}^{(\ell,h)} \in \mathbb{R}^{H_\mathrm{e} D_\mathrm{e} \times H_\mathrm{e} D_\mathrm{e}}$ is a pretrained matrix at head h in layer ℓ of the text encoder.

In our proposed method, we use the self-attention matrix averaged over layers and heads:

$$\mathsf{T}' = \frac{1}{L_\mathrm{e} H_\mathrm{e}} \sum_{\ell=1}^{L_\mathrm{e}} \sum_{h=1}^{H_\mathrm{e}} T^{(\ell,h)}. \quad (2)$$

Due to the high attention probabilities assigned to <bos> and <eos> tokens, we remove their corresponding values and re-normalize each row as follows:

$$\mathsf{T}_{ij} = \frac{\mathsf{T}'_{ij}}{\sum_{m=2}^{i} \mathsf{T}'_{im}}. \quad (3)$$

We denote the final output of the text encoder by $\mathbf{k}_i \in \mathbb{R}^{H_\mathrm{e} D_\mathrm{e}}$ where $i = 1, \cdots, s$, which used as conditioning inputs in T2I diffusion models. We call $\mathbf{k}_i$ as *text embeddings*.

Denoising latent variables. *Latent diffusion models* generates latent tensors (z_0) of an image. A latent diffusion model D_θ learns to simulate the *denoising* process: starting from Gaussian noise z_τ, the model iteratively reconstruct z_0 by predicting the noise at step t:

$$z_{t-1} = z_t - D_\theta(z_t; \{\mathbf{k}_i\}), \quad 0 < t \leq \tau. \quad (4)$$

Cross-attention module. In T2I latent diffusion models, the interaction between text and image data is performed by multi-head cross-attention. In the cross-attention layer ℓ, we define the *query*, $\mathbf{q}_a^{(\ell)} \in \mathbb{R}^{H_\mathrm{c} D_\mathrm{c}}$, where $a = 1, \cdots, N_\mathrm{c}$, and N_c is the query sequence length at the given cross-attention layer. D_c is the hidden dimension per head, and H_c is the number of heads of the cross-attention layer.

The cross-attention map $A^{(\ell,h)} \in \mathbb{R}^{N_\mathrm{c} \times s}$ at head h with elements $A_{ai}^{(\ell,h)}$ is defined as

$$A_{ai}^{(\ell,h)} := \frac{\exp(\Omega_{ai})}{\sum_{j=1}^{s} \exp(\Omega_{aj})}, \ \Omega_{ai} := \mathbf{q}_a^{(\ell)\top} W_\mathrm{c}^{(\ell,h)} \mathbf{k}_i^{(\ell)}. \quad (5)$$

Here, $W_\mathrm{c}^{(\ell,h)} \in \mathbb{R}^{H_\mathrm{c} D_\mathrm{c} \times H_\mathrm{c} D_\mathrm{c}}$ is a pretrained matrix at head h in layer ℓ.

Cross-attention similarity matrix. In our study, we use the cross-attention maps averaged over heads and over all layers for which $N_\mathrm{c} = M$:

$$\mathsf{A} = \frac{1}{L_M H_\mathrm{c}} \sum_{\ell=1}^{L_M} \sum_{h=1}^{H_\mathrm{c}} A^{(\ell,h)}. \quad (6)$$

Here, L_M is the number of cross-attention layers with $N_\mathrm{c} = M$. Based on the average cross-attention map above, we define the similarity matrix $\mathsf{S} \in \mathbb{R}^{s \times s}$ as follows:

$$\mathsf{S}_{ij} := \frac{\mathsf{C}_{ij}}{\sum_{k=1}^{s} \mathsf{C}_{ik}}, \mathsf{C}_{ij} := \frac{\sum_{a=1}^{N_\mathrm{c}} \mathsf{A}_{ai} \mathsf{A}_{aj}}{\left(\sum_{a=1}^{N_\mathrm{c}} \mathsf{A}_{ai}^2\right)^{\frac{1}{2}} \left(\sum_{a=1}^{N_\mathrm{c}} \mathsf{A}_{aj}^2\right)^{\frac{1}{2}}}. \quad (7)$$

C_{ij} denotes cosine similarity between maps corresponding to i and j.

4. Understanding and Resolving Text-to-Image Semantic Discrepancy

In this section, we examine the cross-attention maps in Stable Diffusion (SD) [29] to identify factors that lead to semantic discrepancies between generated images and text

prompts. We provide additional results with PixArt-α [4] in Appendix B.1. Based on these findings, we introduce an approach to enhance the fidelity of T2I diffusion models by leveraging text self-attention maps to regularize cross-attention maps.

4.1. Why Do Generated Images Misrepresent Text?

Our findings are based on the premise that cross-attention maps should capture syntactic relationships between words for accurate T2I generation, as supported by prior works [1, 8, 21, 28]. We identify a key factor contributing to incorrect relations in cross-attention maps, resulting in less faithful image generation: While text embedding ($\mathbf{k}_i$) similarity strongly correlates with the cross-attention similarity matrix C, it insufficiently reflects syntactic bindings in the text prompt. This highlights a fundamental issue in text-image semantic discrepancy: *the similarity in text embeddings lack the syntax information necessary for accurate image generation*. We reveal that text self-attention maps (T) effectively capture syntactic information, but this is not sufficiently encoded in text embeddings. The absence of the information in text embeddings can be due to an artifact in the attention module, known as *attention sink* [35], where attention weights are biased toward the <bos> token as detailed later. These findings motivate the solution presented in the following section.

For the following empirical analysis, we use the prompt sets introduced in [3], containing prompts structured as the following three categories:
(i) $[\texttt{attribute}_1][\texttt{object}_1]$ *and* $[\texttt{attribute}_2][\texttt{object}_2]$,
(ii) $[\texttt{object}_1(\texttt{animal})]$ *with* $[\texttt{object}_2]$,
(iii) $[\texttt{object}_1(\texttt{animal})]$ *and* $[\texttt{attribute}_2][\texttt{object}_2]$.
Prompt sets in the above formats enable us to focus on two key cases of text-image semantic discrepancies: missing objects and attribute mis-binding. In the following, we refer to the group of *syntactically bound words (tokens)* as pairs of $([\texttt{attribute}_i], [\texttt{object}_i])$, $i = 1, 2$, and the group of *syntactically unbound words (tokens)* as $([\texttt{attribute}_i], [\texttt{object}_j])$ and $([\texttt{object}_1], [\texttt{object}_2])$, where $i \neq j$.

We revisit the role of cross-attention maps in text-image semantic coherence, where spatial overlap or separation is crucial, as discussed in [1, 8, 21, 28]. Figure 3 illustrates this effect: On the left, overlapping attention maps for syntactically unbound words lead to object missing, while greater overlap for syntactically bound words enhances attribute binding. This suggests that syntactic associations should be reflected in cross-attention maps.

To statistically assess cross-attention map's impact, we analyze categories (ii) and (iii) of the aforementioned prompt sets. Assuming spatial overlap is measurable by the cosine similarity matrix (C) in eq.(7), we compute cosine similarities for the attention maps of $object_1$ and $object_2$ in missing objects, and for $attribute_2$ and $object_2$ in attribute

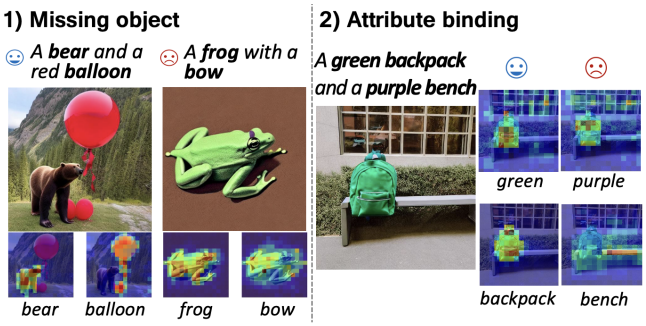
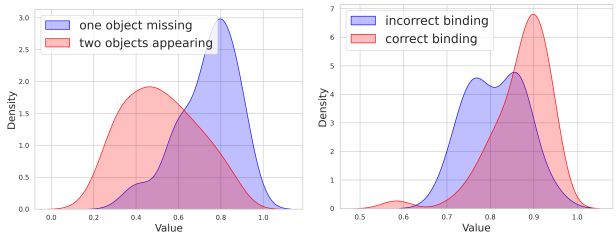


Figure 3. The generated images and cross-attention maps (A) for the specific tokens from SD v1.5. This illustrate the importance of spatial alignment in cross-attention maps for accurate image generation. Divergent (overlapping) cross-attention maps for syntactically unbound (bound) words enhances text-to-image fidelity.



(a) (b)

Figure 4. Comparison for the distributions of cosine similarity between cross-attention maps (at denoising step 10). (a) The cases with one missing object–incorrect and two objects present–correct. (b) The cases with incorrect and correct attribute binding. Correct instances are more frequent when the cosine similarity is low for *objects presence* and high for *attribute binding*.

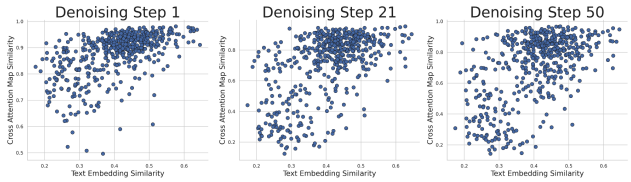


Figure 5. Correlation between the cosine similarity of text embeddings and that of cross-attention maps across denoising steps ($t = 1, 21, 50$). Similar text embeddings generally lead to similar cross-attention maps, with the correlation weakening over time.

binding cases. Then, we compare cosine similarity distributions for correct vs. incorrect images on both cases. See Appendix B.1 for setup details. Figure 4 shows that lower cosine similarity tends to correlate with object presence, while higher similarity supports more accurate bindings.

This observation motivates us to examine the factor contributing to similarity in cross-attention maps across tokens.
Finding 1: The cosine similarity of text embeddings has a large correlation with cross-attention similarity matrix C. Figure 5 shows there is a correlation between the cosine similarity of text embedding and C, which persists throughout the final denoising steps. This indicates that similar text

embeddings can result in overlapping cross-attention maps.

Next, we justify this finding mathematically.

Proposition 1. *If $A^{(\ell,h)} \in \mathbb{R}^{N_c \times s}$ is a cross-attention map defined in eq. (5), then under the assumptions i, ii, and iii described in Appendix A, the cosine similarity matrix can be written in terms of key vectors $\mathbf{k}_i^{(\ell,h)} \in \mathbb{R}^{H_c D_c}$ as*

$$\cos(A_i^{(\ell,h)}, A_j^{(\ell,h)}) = \exp\Big(-\frac{1}{2}(\mathbf{k}_i - \mathbf{k}_j)^\top W^2 (\mathbf{k}_i - \mathbf{k}_j)\Big), \tag{8}$$

up to terms of at least $\mathcal{O}(1/\sqrt{N_c})$ and $\mathcal{O}(\epsilon)$, where $W^2 := W_c^{(\ell,h)\top} \Sigma^{(\ell)} W_c^{(\ell,h)}$ and $\Sigma^{(\ell)} \in \mathbb{R}^{H_c D_c \times H_c D_c}$ is the covariance matrix of query vectors. □

Refer to Appendix A for the proof.

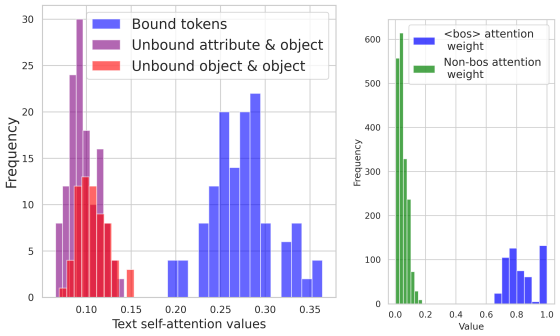
We empirically and mathematically showed that the similarity in text embeddings $\mathbf{k}_i$ influences cross-attention maps. Next, we evaluate whether the similarity of these embeddings reflects syntax information in text inputs.

Finding 2-1: There is no significant correlation between word syntactic bindings and text embedding similarity. Prior study [40] suggests CLIP embeddings [25], used in SD, behave like a bag-of-words model, ignoring word relationships. We also show in Figure 2b the similarity in CLIP text embeddings does not correlate with syntactic bindings. Specifically, we expect close embeddings for *syntactically bound tokens* and distant embeddings for *unbound tokens*, yet the distributions lack separation.

Finding 2-2: Text self-attention maps do have syntax information. We examine the text encoder that produces text embeddings. Interestingly, the self-attention maps in the encoder capture syntactic relationships, as shown in Figure 6a. These maps reveal higher similarity between *syntactically bound tokens* and lower similarity between *unbound tokens*. For more complex prompt structures, see the Appendix B.1.

Finding 2-3: The *attention sink* can contribute to why text embeddings lack the syntax information. *Why do text embeddings lack relational information despite being derived from multiple self-attention modules?* We attribute this gap to *attention sink* [32, 35], where attention scores are concentrated on a few tokens. In CLIP's text encoder, attention is mainly focused on the <bos> token, as discussed in [3, 37] and shown in Figure 6b. We hypothesize the focus on the <bos> token can limit the transfer of relational information from self-attention maps (T) to text embeddings, as attention scores for other tokens remain much smaller, minimizing their influence in each self-attention layer.

To provide a mathematical justification, we express the difference between the output and input of one block of the



(a) (b)

Figure 6. (a) The distributions of text self-attention value (T'_{ij} in eq.(2)) for bound tokens (attribute$_m$, object$_m$) and unbound tokens (attribute$_m$, object$_n$) / (object$_1$, object$_2$), where $m, n \in \{1, 2\}$. The separate distributions indicate text self-attention maps can indeed represent the syntactic relationships. (b) Comparison of text self-attention probability histograms between <bos> token and non-<bos> tokens on 100 prompts: The probability allocated to <bos> is on average 20 times larger than that of other tokens.

self-attention module for a token e_i as:

$$\mathbf{o}_i^{(\ell,h)} = \sum_{j=1}^{i} T_{ij}^{(\ell,h)} W_{\mathrm{v}}^{(\ell,h)} \mathbf{e}_j^{(\ell)}, \tag{9}$$

where $W_{\mathrm{v}}^{(\ell,h)} \in \mathbb{R}^{D_{\mathrm{e}} \times H_{\mathrm{e}} D_{\mathrm{e}}}$ is a parameter, $\mathbf{e}_i \in \mathbb{R}^{H_{\mathrm{e}} D_{\mathrm{e}}}$, and $T_{ij}^{(\ell,h)}$ is the self-attention matrix in eq. (1). We consider the situation where *attention sink* occurs, that is the attention weights for the <bos> token is much higher than the rest of the sequence:

$$\epsilon = \frac{\sum_{j \neq 1}^{i} T_{ij}}{T_{i1}} \ll 1, \qquad i = 2, \cdots, s. \tag{10}$$

In Appendix A, we prove the following statement:

Proposition 2. *Define matrix $R \in \mathbb{R}^{s \times s}$ as*

$$R_{ij} = \mathbf{e}_i^{(\ell)\top} W_v^{(\ell,h)\top} W_v^{(\ell,h)} \mathbf{e}_j^{(\ell)}, \tag{11}$$

and suppose it has the property

$$\frac{|R_{mn}|}{R_{11}} \sim \mathcal{O}(1/\epsilon), \qquad \frac{|R_{1m}|}{R_{11}} \sim \mathcal{O}(1), \quad 1 < m, n \leq s, \tag{12}$$

where $\mathbf{e}_1^{(\ell)}$ is the bos embedding. Then the following holds:

$$\cos(\mathbf{o}_i^{(\ell,h)}, \mathbf{o}_j^{(\ell,h)}) = 1 - \mathcal{O}(\epsilon). \tag{13}$$

□

This suggests that cosine similarity between token vectors ($\mathbf{o}_i^{(\ell,h)}$)—potentially influencing cross-attention maps—remains barely changed across text self-attention layers due to attention sink on the <bos> token. In other words, the attention sink can hinder the accurate encoding of self-attention maps into embeddings.

Our findings reveal text embedding alone are insufficient

for generating semantically aligned images. *On the other hand, we notably show the potential of transferring neglected syntactic information from text self-attention maps to the cross-attention to enhance T2I semantic alignment.*

4.2. Text Self-Attention Maps (T-SAM) Guidance

In the previous section, we show text self-attention maps capture syntax information within a sentence. Building on this insight, we propose leveraging the self-attention maps within the text encoder—a component of diffusion models—to enhance cross-attention maps. By minimizing the distance between the similarity matrix of the cross-attention maps and the text self-attention matrix, our approach ensures that embedded syntactic relationships are effectively transferred to cross-attention.

Our method optimizes cross-attention maps during inference, adjusting their similarity matrix to align with the text self-attention matrix T. The normalized cosine similarity matrix, S (defined in eq.(7)), is used as the cross-attention similarity matrix. This is achieved by simply minimizing the loss function:

$$\mathcal{L}(z_t) = \sum_{i=1, j \leq i}^{s} \rho_i |\mathsf{T}_{ij}^{\gamma} - \mathsf{S}_{ij}(z_t)|, \tag{14}$$

where the exponent γ acts to amplify larger values and compress smaller ones so the effect of controlling temperature and $\rho_i = i/s$. For example, if two words in the prompt have negligible syntactic relation according to the text self-attention matrix, i.e. $\mathsf{T}_{ij} \approx 0$, we demand that their similarity of cross-attention maps must *not* be similar: $\mathsf{S}_{ij} \approx 0$.

In practice, this optimization will be applied only to z_t at a few denoising steps during inference as followed:

$$z_t' = z_t - \alpha \cdot \nabla_{z_t} \mathcal{L}(z_t). \tag{15}$$

5. Experiments

This section presents our empirical study on the effectiveness of T-SAM in enhancing semantic alignment across diverse T2I diffusion models, including SD v1.5 [29] and PixArt-α [4], evaluated both quantitatively and qualitatively. To the best of our knowledge, this is the first work to introduce attention control in PixArt-α, underscoring the novelty of our approach.

Prompt datasets. We evaluate our approach on diverse text prompts using two existing benchmarks. First, the *TIFA v1.0 benchmark* [13] is a large-scale text-to-image generation dataset featuring a wide range of sentence structures. This benchmark comprises 4,000 prompts, including 2,000 image captions from the COCO validation set [18], 161 prompts from DrawBench [30], 1,420 prompts from PartiPrompt used in Parti [39], and 500 texts from PaintSkill [6]. Second, we use structured prompt sets containing multiple objects and their corresponding attributes from *Attend-n-Excite* [3]. Prompts in this dataset are grouped into three categories: *Objects* (e.g., "[attribute$_1$] [object$_1$] and [attribute$_2$] [object$_2$]"), *Animals-Objects* (e.g., "[animal] with [object]" or "[animal] and/with [attribute] [object]"), and *Animals* (e.g., "[animal$_1$] and [animal$_2$]"). We exclude the *Animals* category, as it lacks the complex syntactic relations as discussed in [28].

Table 1. Evaluation results for the TIFA benchmark, including TIFA scores and CLIP similarity scores. *External Info.* indicates whether external information is used. CLIP-I (CLIP-T) refers to image-text (text-text) CLIP similarity. The base model is SD.

	External Info.	TIFA	CLIP-I	CLIP-T
SD	✗	0.79	0.33	**0.77**
LB	✓	0.80	0.33	0.76
T-SAM	✗	**0.83**	**0.34**	**0.77**

Implementation details. For SD, we use 50 sampling iterations, updating z_t at each denoising step from 1 to 25. We set $M = 256$ in eq.(6) and $\gamma = 2$ in eq.(14). For the TIFA benchmark, we generate one image per prompt with a shared seed across methods, while using a unique seed for each prompt. We set $\alpha = 40$ in eq.(15). For *Attend-n-Excite* prompts, we generate results using 64 seeds, in line with standard practice, setting $\alpha = 10$ and applying 20 iterative updates at $i \in \{0, 10, 20\}$. See Appendix B.2 for details on PixArt-α as well as SD.

Evaluation metrics. To quantitatively evaluate the accuracy of generated images, we use two metrics: TIFA scores [13] (for TIFA benchmark) and CLIP similarity scores. TIFA scores measure how well generated images reflect the text prompts. In TIFA, questions for each prompt are generated by GPT-3.5 [22], and a vision-language model [16] provides answers. For CLIP [25] scores, we follow the protocol in [3] to evaluate both image-text and text-text similarity. Additional details on CLIP scores are provided in Appendix B.2. Moreover, we conducted a user study with 10 random prompts in *Animals-Objects*, generating four images per prompt using different seeds. A total of 25 participants select the image that best matches the text.

Baselines. When evaluating our method on SD, we compare our method with SD, Linguistic-Binding (LB) [28], Attend-n-Excite (A&E) [3] and *CONFORM* [21]. LB depending on external parsers (SpaCy [11]) is limited to attribute-binding tasks. This method can be applied to both the TIFA benchmark and the Attend-n-Excite (*Objects* and *Animals-Objects*) prompt sets, though they are used only when the text parser identifies (modifier, entity-noun) pairs within the given prompts. A&E and *CONFORM* require manual token selection per prompt, restricting their use to the fixed-template prompts (*Objects* and *Animals-Objects*) due to the high cost of selecting token indices for diverse prompts. We reproduce the images with SD, *CONFORM*, and LB based on SD v1.5 and using the same seeds. For PixArt-α, we com-

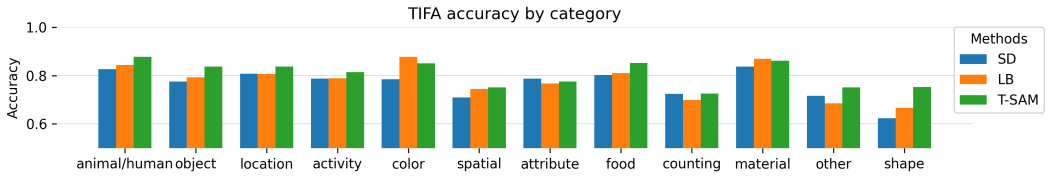


Figure 7. Accuracy for each question type in the TIFA benchmark. Our method achieves the best performance in most categories. The *attribute* includes properties such as large, small, young, etc. The base model is SD.

Figure 8. Comparison of our method (T-SAM) on SD with recent state-of-the-art methods on prompts from *Objects* and *Animals-Objects*. The images corresponding to the same position across different methods are generated using the same seed. Best in zoom.

pare only with the base model, as no prior attention-control methods have been applied to it.

5.1. Results

Quantitative results. Table 1 shows the evaluation results in TIFA benchmark and Figure 7 illustrates the breakdown of TIFA scores across question types, using SD as the base model. We highlight that our approach (T-SAM) outperforms SD and LB on complex syntactic prompts, where other baselines are inapplicable. While LB demonstrates improvements over SD in the *color*, *shape*, and *material* categories, which are closely related to attribute-binding tasks, it fails to enhance accuracy in the *activity* and *counting* categories. This underscores LB's inability to capture diverse word relationships, even when using external text parsers. In contrast, our method shows improvements over SD in nearly all categories including *color*, *shape*, and *activity*, except for *attribute* (e.g., properties such as large, small, young), which differ from the term "attribute" used in this study, demonstrating its versatility across diverse word relationships. Additionally, it achieves higher CLIP scores than the baselines, proving its superior semantic alignment.

Table 2. Average CLIP similarity scores between text prompts and images generated with 64 different seeds. The base model is SD.

		Image-Text Prompt				Prompt-Caption	
		Full Prompt		Minimum Object			
	External Info.	Objects	Animals-Objects	Objects	Animals-Objects	Objects	Animals-Objects
SD	✗	0.34	0.34	0.25	0.26	0.76	0.80
LB	✓	**0.36**	0.35	0.27	0.27	0.80	0.83
CONFORM	✓	**0.36**	0.36	**0.28**	**0.28**	**0.81**	**0.85**
A&E	✓	**0.36**	0.35	0.27	0.26	0.81	0.83
T-SAM	✗	**0.36**	**0.37**	**0.28**	**0.28**	**0.81**	**0.85**

Table 3. CLIP similarity scores on prompts from *Objects* and *Animals-Objects* using PixArt-α as the base model. CLIP-I (CLIP-T) refers to image-text (text-text) CLIP similarity.

Prompt Set	Method	CLIP-T	CLIP-I (Full, Min)
Objects	PixArt-α	0.81	0.36, 0.27
	T-SAM	**0.82**	**0.37, 0.28**
Animals-Objects	PixArt-α	0.85	0.36, 0.27
	T-SAM	**0.86**	**0.37, 0.28**

Figure 9. Qualitative comparison using prompts from MSCOCO contained in TIFA benchmark. Our approach successfully generates multiple elements, including *a sewing machine* in the first prompt, *people* in the second, *popsicle* in the third, and *cow* in the fourth. SD misses some elements and LB generates identical images to SD when no relations are extracted from the text parser.

In *Objects* and *Animals-Objects*, T-SAM on SD performs comparably to the state-of-the-art *CONFORM*, which, unlike our approach, requires manually defined token indices for positive and negative groups, as shown in Table 2. Our method outperforms LB and A&E, which also rely on external inputs. This demonstrates extracting syntactic information from text self-attention maps can be more effective than relying on text parsers or manually selecting tokens for optimization. An additional ablation study is provided in Appendix B.2. Moreover, Table 3 shows that T-SAM is also effective on PixArt-α with *Attend-n-Excite* prompts, demonstrating the generalizability of our method to a different text encoder (T5-XXL) and model architecture (DiT).

Qualitative results. In Figure 8, T-SAM either outperforms or performs comparably to the baselines in *Objects* and *Animals-Objects*. SD often omits objects (*e.g.*, *clock*, *apple*, or *bow*) or misbinds attributes (e.g., *blue clock*). *CONFORM* and LB improve over SD but have limitations. LB is restricted to attribute-binding and fails on prompts without modifiers (e.g., *An elephant with a bow*), producing results identical to SD *CONFORM* can impose overly rigid separations, as seen in the first example. T-SAM, by leveraging smoother linguistic cues from text attention maps—rather than binary positive/negative pairs—achieves broader applicability and more consistent improvements. See Figure 9 for more comparison of images generated by T-SAM and the baselines, showcasing effectiveness of T-SAM with complex prompts, such as MSCOCO captions from the TIFA benchmark.

For PixArt-α, Figure D illustrates the effectiveness of our method. Although the base model is already strong compared to SD, T-SAM further improves its performance.

User study. The user study result in Table D show that T-SAM on SD outperforms LB and achieves performance comparable to *CONFORM*. Additionally, Table E shows T-SAM on PixArt-α surpasses the base PixArt-α model.

6. Conclusion

To enhance fidelity in T2I diffusion models, we improved cross-attention maps by aligning their similarity matrix with text self-attention maps. Our approach is based on two insights: (1) similar text embeddings produce similar cross-attention maps, but (2) syntactic relations are missed in embeddings but captured by text self-attention maps. T-SAM enabled cross-attention to better capture syntactic structure, significantly improving text-to-image fidelity across diverse sentence structures, without requiring external tools. Nonetheless, our method has limitations such as inaccuracies in object counting and potential noise introduced by text attention maps, leaving room for future improvement.

Acknowledgement

This work was partially supported by NIH (1U01CA269192).

References

[1] Aishwarya Agarwal, Srikrishna Karanam, KJ Joseph, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. A-star: Test-time attention segregation and retention for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2283–2293, 2023. 1, 2, 3, 4

[2] Aishwarya Agarwal, Srikrishna Karanam, and Balaji Vasan Srinivasan. Alignit: Enhancing prompt alignment in customization of text-to-image models. *arXiv preprint arXiv:2406.18893*, 2024. 3

[3] 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)*, 42(4):1–10, 2023. 2, 3, 4, 5, 6

[4] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023. 1, 2, 3, 4, 6

[5] 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. *arXiv preprint arXiv:2403.04692*, 2024. 2

[6] Jaemin Cho, Abhay Zala, and Mohit Bansal. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3043–3054, 2023. 6

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2

[8] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1, 2, 3, 4

[9] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. *arXiv preprint arXiv:2410.10781*, 2024. 3

[10] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 1, 3

[11] 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. 6

[12] 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. 3

[13] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 6, 5

[14] Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, and Wangmeng Zuo. Mc2: Multi-concept guidance for customized multi-concept generation. *arXiv preprint arXiv:2404.05268*, 2024. 3

[15] 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. 3

[16] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 6, 5

[17] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 5

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[19] Oscar Mañas, Pietro Astolfi, Melissa Hall, Candace Ross, Jack Urbanek, Adina Williams, Aishwarya Agrawal, Adriana Romero-Soriano, and Michal Drozdzal. Improving text-to-image consistency via automatic prompt optimization. *arXiv preprint arXiv:2403.17804*, 2024. 3

[20] Arash Marioriyad, Mohammadali Banayeeanzade, Reza Abbasi, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Attention overlap is responsible for the entity missing problem in text-to-image diffusion models! *arXiv preprint arXiv:2410.20972*, 2024. 3

[21] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9005–9014, 2024. 1, 2, 3, 4, 6

[22] OpenAI. Openai gpt-3 api [gpt-3.5-turbo], 2024. Available at: https://platform.openai.com. 6, 5

[23] 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. 2

[24] Zipeng Qi, Lichen Bai, Haoyi Xiong, et al. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024. 3

[25] 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. 2, 3, 5, 6

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 3

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[28] 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, 2024. 1, 2, 3, 4, 6

[29] 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. 1, 2, 3, 6, 5

[30] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2, 6

[31] Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9370–9379, 2024. 3

[32] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024. 2, 3, 5

[33] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[34] Zirui Wang, Zhizhou Sha, Zheng Ding, Yilin Wang, and Zhuowen Tu. Tokencompose: Grounding diffusion with token-level supervision. *arXiv preprint arXiv:2312.03626*, 2023. 3

[35] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023. 2, 3, 4, 5

[36] Fei Yang, Shiqi Yang, Muhammad Atif Butt, Joost van de Weijer, et al. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *Advances in Neural Information Processing Systems*, 36:26291–26303, 2023. 3

[37] Mingyang Yi, Aoxue Li, Yi Xin, and Zhenguo Li. Towards understanding the working mechanism of text-to-image diffusion model. *arXiv preprint arXiv:2405.15330*, 2024. 3, 5

[38] Hu Yu, Hao Luo, Fan Wang, and Feng Zhao. Uncovering the text embedding in text-to-image diffusion models. *arXiv preprint arXiv:2404.01154*, 2024. 3

[39] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 6

[40] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*, 2022. 2, 5

[41] 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. 3

Text Embedding is Not All You Need: Attention Control for Text-to-Image Semantic Alignment with Text Self-Attention Maps

Supplementary Material

A. Proofs

A.1. Notation

Symbol	Definition and Properties
D_{c}	embedding dimension per head in cross-attention layers
H_{c}	number of heads in cross-attention layers
N_{c}	query sequence length in cross-attention layers
D_{e}	embedding dimension per head in text encoder
H_{e}	number of heads in text encoder
s	text sequence length
$\mathbf{q}_a^{(\ell)}$	$\in \mathbb{R}^{H_{\mathrm{c}}D_{\mathrm{c}}}$, $a = 1, \cdots, N_{\mathrm{c}}$, query vectors at layer ℓ cross-attention layer
$\mathbf{k}_i^{(\ell)}$	$\in \mathbb{R}^{H_{\mathrm{e}}D_{\mathrm{e}}}$, $i = 1, \cdots, s$, text embeddings
$W_{\mathrm{c}}^{(\ell,h)}$	$\in \mathbb{R}^{H_{\mathrm{c}}D_{\mathrm{c}} \times H_{\mathrm{c}}D_{\mathrm{c}}}$, projection parameter matrix in cross-attention layer ℓ and head h. It is related to the product key and query projection parameters ($\in \mathbb{R}^{H_{\mathrm{c}} \times D_{\mathrm{c}} \times H_{\mathrm{c}}D_{\mathrm{c}}}$) via $W_{\mathrm{c}}^{(\ell,h)} = W_{\mathrm{q}}^{(\ell,h)\top} W_{\mathrm{k}}^{(\ell,h)}$
$W_{\mathrm{v}}^{(\ell,h)}$	$\in \mathbb{R}^{D_{\mathrm{e}} \times H_{\mathrm{e}}D_{\mathrm{e}}}$, value projection matrix in text encoder self-attention layer ℓ and head h
$W_{\mathrm{out}}^{(\ell)}$	$\in \mathbb{R}^{H_{\mathrm{e}}D_{\mathrm{e}} \times H_{\mathrm{e}}D_{\mathrm{e}}}$, out projection matrix in text encoder self-attention layer ℓ
$A^{(\ell,h)}$	$\in \mathbb{R}^{N_{\mathrm{c}} \times s}$, cross-attention maps at layer ℓ and head h. The elements are denoted by $A_{ai}^{(\ell,h)}$
$\mathbf{e}_i^{(\ell)}$	$\in \mathbb{R}^{H_{\mathrm{e}}D_{\mathrm{e}}}$, $i = 1, \cdots, s$, text dense vectors in the text-encoder layer ℓ
$T^{(\ell,h)}$	$\in \mathbb{R}^{s \times s}$, text self-attention matrix at layer ℓ and head h of the text encoder
ϵ	$\ll 1$, the inverse ratio of <bos> attention weight to the sum of attention weights of the rest of the sequence

Table A. Table of Notations

A.2. The Big O Notation

Based on the empirical observations, we consider the situation where *attention sink* occurs both in the text encoder and in the cross-attention layers of the diffusion model: the attention weights for the <bos> token are much higher than the rest of the sequence:

$$\text{self-attention in the text encoder}: \quad \frac{\sum_{j \neq 1}^{i} T_{ij}^{(\ell,h)}}{T_{i1}^{(\ell,h)}} < \epsilon \ll 1, \qquad i = 2, \cdots, s, \tag{1}$$

$$\text{cross-attention in the diffusion model}: \quad \frac{\sum_{j \neq 1}^{s} A_{aj}^{(\ell,h)}}{A_{a1}^{(\ell,h)}} < \epsilon \ll 1, \qquad a = 1, \cdots, N_{\mathrm{c}}. \tag{2}$$

In practice $\epsilon \sim 0.1$ or smaller in the middle layers of U-Net and the later layers of CLIP text encoder. Our approach for calculating the approximate quantities in the limit of small ϵ is *perturbation theory*: we assume that the variables of the problem, such as $T^{(\ell,h)}$ can be written as a power series in a small parameter ϵ:

$$T^{(\ell,h)} = T^{(\ell,h)(0)} + \epsilon T^{(\ell,h)(1)} + \epsilon^2 T^{(\ell,h)(2)} + \cdots . \tag{3}$$

In this context, $\mathcal{O}(\epsilon)$ mean terms that are linear or higher order in ϵ. If ϵ is sufficiently small, the first few term give a good approximation to the true variable.

A.3. Proof of Proposition 1

In the following, we suppress the (ℓ, h) superscript in quantities $A^{(\ell,h)}, \Sigma^{(\ell,h)}, \mu^{(\ell,h)}, W_{\mathrm{c}}^{(\ell,h)}, \mathbf{k}_i^{(\ell)}, \mathbf{q}_i^{(\ell)}$ defined below to reduce clutter. Consider the similarity matrix of the form:

$$\cos(A_{ai}, A_{aj}) := \frac{\sum_{a=1}^{N_{\mathrm{c}}} A_{ai} A_{aj}}{\big(\sum_{a=1}^{N_{\mathrm{c}}} A_{ai}^{(\ell,h)2}\big)^{\frac{1}{2}} \big(\sum_{a=1}^{N_{\mathrm{c}}} A_{aj}^{(\ell,h)2}\big)^{\frac{1}{2}}}, \tag{4}$$

where

$$A_{ai} := \frac{\exp(\Omega_{ai})}{\sum_{j=1}^{s} \exp(\Omega_{aj})}, \ \Omega_{ai} := \mathbf{q}_a^\top W_{\mathrm{c}} \mathbf{k}_i. \tag{5}$$

Proposition 1. *If $A \in \mathbb{R}^{N_c \times s}$ is a cross-attention map defined in eq. (5), then under the assumptions i, ii, and iii described below, the similarity matrix can be written in terms of key vectors $\mathbf{k}_i \in \mathbb{R}^{H_c D_c}$ as*

$$\cos(A_i, A_j) = \ \exp\Big(-\frac{1}{2}(\mathbf{k}_i - \mathbf{k}_j)^\top W^2 (\mathbf{k}_i - \mathbf{k}_j)\Big), \tag{6}$$

up to terms of at least $\mathcal{O}(1/\sqrt{N_c})$ and $\mathcal{O}(\epsilon)$, where $W^2 := W_c^\top \Sigma W_c$ and $\Sigma^{(\ell)} \in \mathbb{R}^{H_c D_c \times H_c D_c}$ is the covariance matrix of query vectors and $W_c \in \mathbb{R}^{H_c D_c \times H_c D_c}$ is a parameter. □

Proof. If queries are iid samples of some distribution $\mathbf{q}_a^{(\ell)} \sim p_{\mathbf{q}}$ with finite mean and variance, we can use the Central Limit Theorem to write the cosine similarity as

$$\cos(A_i, A_j) := \frac{\mathbb{E}[A_{ai} A_{aj}] + \mathcal{O}(\frac{1}{\sqrt{N_{\mathrm{c}}}})}{\big(\mathbb{E}[A_{ai}^{(\ell,h)2}] + \mathcal{O}(\frac{1}{\sqrt{N_{\mathrm{c}}}})\big)^{\frac{1}{2}} \big(\mathbb{E}[A_{aj}^{(\ell,h)2}] + \mathcal{O}(\frac{1}{\sqrt{N_{\mathrm{c}}}})\big)^{\frac{1}{2}}}, \tag{7}$$

$$= \frac{\mathbb{E}[A_{ai} A_{aj}]}{\big(\mathbb{E}[A_{ai}^2]\big)^{\frac{1}{2}} \big(\mathbb{E}[A_{aj}^2]\big)^{\frac{1}{2}}} + \mathcal{O}(\frac{1}{\sqrt{N_{\mathrm{c}}}}). \tag{8}$$

Assumption i. *The query sequence length N_c is large enough so that when summing over queries, deviations from true mean can be approximated by the first term in $1/N_c$ expansion, and the corrections from the dependence between samples appear at higher orders in the expansion.* □

As a first approximation to the distribution of queries, consider a statistical model where the query vectors are jointly normally distributed:

$$\mathbf{q}_a \sim \mathcal{N}(\mu, \Sigma), \qquad \mu \in \mathbb{R}^{H_{\mathrm{c}} D_{\mathrm{c}}}, \ \Sigma \in \mathbb{R}^{H_{\mathrm{c}} D_{\mathrm{c}} \times H_{\mathrm{c}} D_{\mathrm{c}}}. \tag{9}$$

This is strictly true at the first denoising step. Moreover, if the true distribution remains close to Gaussian, the corrections from the distribution can in principle be perturbatively calculated and added accordingly. Therefore, the assumption above may not be interpreted as a restriction, but as a first (and good) approximation to the true distribution.

Assumption ii. *Query vectors $\mathbf{q}_a \in \mathbb{R}^{H_c D_c}$ are jointly Gaussian as in (9).* □

Note that the *attention scores* $\Omega_{ai} = \mathbf{q}_a^\top W_{\mathrm{c}} \mathbf{k}_i$ are now gaussian variables with

$$\mathbb{E}[\Omega_{ai}] = \mu^\top W_{\mathrm{c}} \mathbf{k}_i := \mu_i, \qquad \mathrm{Var}[\Omega_{ai}] = \mathbf{k}_i^\top W_{\mathrm{c}}^\top W_{\mathrm{c}} \mathbf{k}_i := \sigma_i^2. \tag{10}$$

Assumption iii. *We empirically observe that i) $\mu_1 \gg \mu_i$, ii)$\mu_1 \gg \sigma_1$, iii)$\sigma_1 \approx \sigma_i$ for , $i = 2, \cdots, s$, such that*

$$e^{\mu_i - \mu_1} \sim \mathcal{O}(\epsilon). \tag{11}$$

□

Writing cross-attention probabilities in terms of attention scores, we have

$$A_{ai} = \frac{e^{\Omega_{ai}}}{e^{\Omega_{a1}} + \sum_{m=2}^{s} e^{\Omega_{am}}} = \frac{e^{\Omega_{ai} - \Omega_{a1}}}{1 + \sum_{m=2}^{s} e^{\Omega_{am} - \Omega_{a1}}}. \tag{12}$$

Note that since attention scores are Gaussian,

$$\mathbb{P}[e^{\Omega_{ai} - \Omega_{a1}} < \epsilon/s] = \Phi\big(\frac{\log(\epsilon/s) - \mu_i + \mu_1}{\sqrt{\sigma_1^2 + \sigma_i^2}}\big). \tag{13}$$

Therefore, if assumption iii holds, for some large enough μ_1, we can have $\mathbb{P}[e^{\Omega_{ai}-\Omega_{ai}} < \epsilon/s] > 1-\epsilon^3$. This means that the sum of attention probabilities of all non-<bos> tokens does not exceed ϵ with the probability of at least $1-\epsilon^3$. As a result, we have

$$A_{ai} = e^{\Omega_{ai}-\Omega_{a1}} + \mathcal{O}(\epsilon^2), \tag{14}$$

$$A_{ai}A_{aj} = e^{\Omega_{ai}+\Omega_{aj}-2\Omega_{a1}} + \mathcal{O}(\epsilon^3), \tag{15}$$

with high probability. To evaluate the cosine similarity, we need to compute expectations:

$$\frac{\mathbb{E}[A_{ai}A_{aj}]}{\left(\mathbb{E}[A_{ai}^2]\right)^{\frac{1}{2}}\left(\mathbb{E}[A_{aj}^2]\right)^{\frac{1}{2}}} = \frac{\mathbb{E}[e^{\Omega_{ai}+\Omega_{aj}}] + \mathcal{O}(\epsilon^3)}{\left(\mathbb{E}[e^{2\Omega_{ai}}] + \mathcal{O}(\epsilon^3)\right)^{\frac{1}{2}}\left(\mathbb{E}[e^{2\Omega_{aj}}] + \mathcal{O}(\epsilon^3)\right)^{\frac{1}{2}}}. \tag{16}$$

We can evaluate this expression using the well-known formula of the moment-generating function of Gaussian distribution:

Lemma 1. *If $\mathbf{q}_a \sim \mathcal{N}(\mu, \Sigma)$ and $\mathbf{r} \in \mathbb{R}^{H_c D_c}$, then*

$$\mathbb{E}[e^{\mathbf{q}\cdot\mathbf{r}}] = \exp(\mathbf{r}\cdot\mu + \frac{1}{2}\mathbf{r}\cdot\Sigma\cdot\mathbf{r}). \tag{17}$$

□

Define $\mathbf{r}_{ij} = W_c(\mathbf{k}_i + \mathbf{k}_j - 2\mathbf{k}_1)$:

$$\frac{\mathbb{E}[A_{ai}A_{aj}]}{\left(\mathbb{E}[A_{ai}^2]\right)^{\frac{1}{2}}\left(\mathbb{E}[A_{aj}^2]\right)^{\frac{1}{2}}} = \frac{\exp(\mu^\top\mathbf{r}_{ij} + \frac{1}{2}\mathbf{r}_{ij}^\top\Sigma\mathbf{r}_{ij})}{\exp(\frac{1}{2}\mu^\top\mathbf{r}_{ii} + \frac{1}{4}\mathbf{r}_{ii}\Sigma\mathbf{r}_{ii})\exp(\frac{1}{2}\mu^\top\mathbf{r}_{jj} + \frac{1}{4}\mathbf{r}_{jj}\Sigma\mathbf{r}_{jj})} + \mathcal{O}(\epsilon). \tag{18}$$

Here, we used the fact that each of exponentials are $\sim \mathcal{O}(\epsilon^2)$ to simplify the correction terms to $\mathcal{O}(\epsilon)$. When applying Lemma 1, one might worry that the integration includes regions of $\mathbb{R}^{H_c D_c}$ that the approximation (15) fails. Although this is a valid point, the total probability of such regions is ϵ^3 by assumption, which is at the order of correction terms.

Simplifying (18) gives

$$\frac{\mathbb{E}[A_{ai}A_{aj}]}{\left(\mathbb{E}[A_{ai}^2]\right)^{\frac{1}{2}}\left(\mathbb{E}[A_{aj}^2]\right)^{\frac{1}{2}}} = \exp(\frac{1}{2}\mathbf{r}_{ij}^\top\Sigma\mathbf{r}_{ij} - \frac{1}{4}\mathbf{r}_{ii}^\top\Sigma\mathbf{r}_{ii} - \frac{1}{4}\mathbf{r}_{jj}^\top\Sigma\mathbf{r}_{jj}) + \mathcal{O}(\epsilon). \tag{19}$$

Substituting the definition of $\mathbf{r}_{ij}$, we get

$$\frac{\mathbb{E}[A_{ai}A_{aj}]}{\left(\mathbb{E}[A_{ai}^2]\right)^{\frac{1}{2}}\left(\mathbb{E}[A_{aj}^2]\right)^{\frac{1}{2}}} = \exp(-\frac{1}{2}(\mathbf{k}_i - \mathbf{k}_j)^\top W_c^\top\Sigma W_c(\mathbf{k}_i - \mathbf{k}_j) + \mathcal{O}(\epsilon). \tag{20}$$

□

A.4. Proof of Proposition 2

Proposition 2. *Consider a self-attention layer with output $\mathbf{o}_i^{(\ell,h)} \in \mathbb{R}^{D_e}$ defined as*

$$\mathbf{o}_i^{(\ell,h)} = \sum_{j=1}^{i} T_{ij}^{(\ell,h)} W_v^{(\ell,h)} \mathbf{e}_j^{(\ell)}, \tag{21}$$

where $W_v^{(\ell,h)} \in \mathbb{R}^{D_e \times H_e D_e}$ is a parameter, $\mathbf{e}_i \in \mathbb{R}^{H_e D_e}$, and $T_{ij}^{(\ell,h)}$ is the self-attention matrix. Define $R \in \mathbb{R}^{s\times s}$ as

$$R_{ij} = \mathbf{e}_i^{(\ell)\top} W_v^{(\ell,h)\top} W_v^{(\ell,h)} \mathbf{e}_j^{(\ell)}, \tag{22}$$

and suppose it has the property

$$\frac{|R_{mn}|}{R_{11}} \sim \mathcal{O}(1/\epsilon), \qquad \frac{|R_{1m}|}{R_{11}} \sim \mathcal{O}(1), \quad 1 < m, n \le s, \tag{23}$$

where $\mathbf{e}_1^{(\ell)}$ is the bos embedding. Then the following holds:

$$\cos(\mathbf{o}_i^{(\ell,h)}, \mathbf{o}_j^{(\ell,h)}) = 1 - \mathcal{O}(\epsilon). \tag{24}$$

□

Proof. In terms of the matrix R, we have

$$\cos(\mathbf{o}_i^{(\ell,h)}, \mathbf{o}_j^{(\ell,h)}) = \frac{\mathbf{o}_i^{(\ell,h)} \cdot \mathbf{o}_j^{(\ell,h)})}{\|\mathbf{o}_i^{(\ell,h)}\|\|\mathbf{o}_j^{(\ell,h)}\|} = \frac{\sum_{m=1}^{i}\sum_{n=1}^{j} T_{im}^{(\ell,h)} T_{jn}^{(\ell,h)} R_{mn}}{\big(\sum_{m=1}^{i}\sum_{n=1}^{j} T_{im}^{(\ell,h)} T_{jn}^{(\ell,h)} R_{mn}\big)^{\frac{1}{2}} \big(\sum_{m=1}^{i}\sum_{n=1}^{j} T_{im}^{(\ell,h)} T_{jn}^{(\ell,h)} R_{mn}\big)^{\frac{1}{2}}}. \quad (25)$$

By separating the contributions from the <bos> token, we can write the numerator as

$$\sum_{m=1}^{i}\sum_{n=1}^{j} T_{im}^{(\ell,h)} T_{jn}^{(\ell,h)} R_{mn} = T_{i1}^{(\ell,h)} T_{j1}^{(\ell,h)} R_{11} + \sum_{m'=2}^{i} T_{im'}^{(\ell,h)} T_{j1}^{(\ell,h)} R_{m'1} + \sum_{n'=2}^{j} T_{i1}^{(\ell,h)} T_{jn'}^{(\ell,h)} R_{1n'} + \sum_{m'=2}^{i}\sum_{n'=2}^{j} T_{im'}^{(\ell,h)} T_{jn}^{(\ell,h)} R_{m'n'}, \quad (26)$$

where $m', n' = 2, \cdots, s$. Using $T_{im'}^{(\ell,h)} \sim \mathcal{O}(\epsilon)$, $T_{i1}^{(\ell,h)} = 1 - \mathcal{O}(\epsilon)$and the conditions on the R matrix stated in Proposition 2, we can show that the first term (which is necessarily positive) has a larger norm than the rest:

$$\sum_{m=1}^{i}\sum_{n=1}^{j} T_{im}^{(\ell,h)} T_{jn}^{(\ell,h)} R_{mn} = R_{11} + \mathcal{O}(\epsilon). \quad (27)$$

We can use this property to perform a Taylor expansion in the denominator and keep up to the linear term in ϵ. At this order of approximation, we can use $\frac{1}{1+x} \approx 1 - x$ since we are keeping only the first terms in Taylor expansion. After some algebra, the expression drastically simplifies to

$$\cos(\mathbf{o}_i^{(\ell,h)}, \mathbf{o}_j^{(\ell,h)}) \approx 1 + \sum_{m'=2}^{i}\sum_{n'=2}^{j} \Big(\frac{T_{im'}^{(\ell,h)} T_{jn'}^{(\ell,h)}}{T_{i1}^{(\ell,h)} T_{j1}^{(\ell,h)}} - \frac{1}{2}\frac{T_{im'}^{(\ell,h)} T_{in'}^{(\ell,h)}}{T_{i1}^{(\ell,h)2}} - \frac{1}{2}\frac{T_{jm'}^{(\ell,h)} T_{jn'}^{(\ell,h)}}{T_{j1}^{(\ell,h)2}} \Big) R_{m'n'}. \quad (28)$$

First, note that the expression in brackets is $\sim \mathcal{O}(\epsilon^2)$ and since $R_{m'n'} \sim \mathcal{O}(1/\epsilon)$ (both claims from empirical evidence), the total expression will be $1 + \mathcal{O}(\epsilon)$. Furthermore, matrix R is positive semi-definite by definition. Performing an SVD decomposition $R = U\kappa U^\top$ and absorbing the terms in brackets in U matrices, one can use the Cauchy-Schwarz inequality to show that the correction term above is negative. In conclusion, we have

$$\cos(\mathbf{o}_i^{(\ell,h)}, \mathbf{o}_j^{(\ell,h)}) = 1 - c\epsilon + \mathcal{O}(\epsilon^2), \quad (29)$$

for some positive $c \sim \mathcal{O}(1)$. □

A.5. An extension to Proposition 2

In this section, we extend the result of Proposition 2 by considering the full self-attention layer of the text encoder. Proposition 2 was only concerned with products of keys, queries, and values. However, a self-attention layer typically includes an output-projection linear layer and a skip connection. Here, we explore the effect of these two components and compare the cosine similarities of input text embeddings versus those of outputs. We show that based on practical assumptions that are valid in the later layers of CLIP text encoder, the output cosine similarities are close to input cosine similarities. Consider the output of an attention head at layer ℓ:

$$\mathbf{e}_i^{(\ell)\text{out}} = \mathbf{e}_i^{(\ell)} + W_{\text{out}}^{(\ell)} \text{concat}\big[\sum_{m=1}^{i} T_{im}^{(\ell,1)} W_{\text{v}}^{(\ell,1)} \mathbf{e}_m^{(\ell)}, \cdots, \sum_{m=1}^{i} T_{im}^{(\ell,H_{\text{e}})} W_{\text{v}}^{(\ell,H_{\text{e}})} \mathbf{e}_m^{(\ell)} \big] \quad (30)$$

in which $\mathbf{e}_i^{(\ell)\text{out}}, \mathbf{e}_i^{(\ell)} \in \mathbb{R}^{H_{\text{e}} D_{\text{e}}}$ for all $i = 1, \cdots, s$. Here, $W_{\text{out}}^{(\ell)} \in \mathbb{R}^{H_{\text{e}} D_{\text{e}} \times H_{\text{e}} D_{\text{e}}}$ is the out-projection layer.

Define the average attention probabilities:

$$\tau_i^{(\ell,h)} = \frac{1}{i} \sum_{m=2}^{i} T_{im}^{(\ell,h)}, \qquad i = 2, \cdots, s. \quad (31)$$

By adding and subtracting terms proportional to $\tau_i^{(\ell,h)}$ we have

$$\mathbf{e}_i^{(\ell)\text{out}} = \mathbf{e}_i^{(\ell)} + \underbrace{W_{\text{out}}^{(\ell)} \text{concat}\Big[T_{i1}^{(\ell,h)} W_{\text{v}}^{(\ell,1)} \mathbf{e}_1^{(\ell)} + \tau_i^{(\ell,1)} \sum_{m=2}^{i} W_{\text{v}}^{(\ell,1)} \mathbf{e}_m^{(\ell)}, \cdots, T_{i1}^{(\ell,H_{\text{e}})} W_{\text{v}}^{(\ell,H_{\text{e}})} \mathbf{e}_1^{\ell} + \tau_i^{(\ell,H_{\text{e}})} \sum_{m=2}^{i} W_{\text{v}}^{(\ell,H_{\text{e}})} \mathbf{e}_m^{(\ell)} \Big]}_{:= \mathbf{e}_i'^{(\ell)}}$$

$$+ \underbrace{W_{\text{out}}^{(\ell)} \text{concat}\Big[\sum_{m=2}^{i} (T_{im}^{(\ell,h)} - \tau_i^{(\ell,1)}) W_{\text{v}}^{(\ell,1)} \mathbf{e}_m^{(\ell)}, \cdots, \sum_{m=2}^{i} (T_{im}^{(\ell,H_{\text{e}})} - \tau_i^{(\ell,H_{\text{e}})}) W_{\text{v}}^{(\ell,H_{\text{e}})} \mathbf{e}_m^{(\ell)} \Big]}_{:= \delta \mathbf{e}_i^{(\ell)}}. \quad (32)$$

We empirically observe that

$$\|\mathbf{e}_i^{(\ell)}\| \sim \mathcal{O}(1/\epsilon), \qquad \|\mathbf{e}_i'^{(\ell)}\| \sim \mathcal{O}(1), \qquad \|\delta\mathbf{e}_i^{(\ell)}\| \sim \mathcal{O}(\epsilon). \tag{33}$$

The result of the above conditions is

$$\begin{aligned}\left|\mathbf{e}_i^{(\ell)\text{out}} \cdot \mathbf{e}_j^{(\ell)\text{out}} - (\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}) \cdot (\mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)})\right| &= (\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}) \cdot \delta\mathbf{e}_j^{(\ell)} + (\mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)}) \cdot \delta\mathbf{e}_i^{(\ell)} + \delta\mathbf{e}_i^{(\ell)} \cdot \delta\mathbf{e}_j^{(\ell)} \\ &\sim \mathcal{O}(1). \end{aligned} \tag{34}$$

Thus, the cosine similarity will be

$$\cos(\mathbf{e}_i^{(\ell)\text{out}}, \mathbf{e}_j^{(\ell)\text{out}}) = \frac{(\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}) \cdot (\mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)}) + \mathcal{O}(1)}{\left(\|\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}\|^2 + \mathcal{O}(1)\right)^{\frac{1}{2}} \left(\|\mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)}\|^2 + \mathcal{O}(1)\right)^{\frac{1}{2}}} \tag{35}$$

$$= \frac{(\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}) \cdot (\mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)})}{\left(\|\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}\|^2\right)^{\frac{1}{2}} \left(\|\mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)}\|^2\right)^{\frac{1}{2}}} + \mathcal{O}(\epsilon^2) \tag{36}$$

$$= \cos(\mathbf{e}_i^{(\ell)} + \mathbf{e}_i'^{(\ell)}, \mathbf{e}_j^{(\ell)} + \mathbf{e}_j'^{(\ell)}) + \mathcal{O}(\epsilon^2). \tag{37}$$

This result shows that the similarity between output text embeddings corrected by terms that depend only on the averages of attention probabilities (τ_i).

B. Details of the Empirical Study

B.1. Additional details on Section 4

The experiment setup for Figure 4 in Section 4. We use 144 prompts from the *Animals-Objects* in [3], featuring two distinct objects, to address the missing object scenario illustrated in Figure 4a. In addition, we use 107 prompts from the same dataset, incorporating both attributes and objects, to explore the attribute binding case shown in Figure 4b. To assess image correctness, we use TIFA [13], which evaluates how well the generated images reflect the text prompts. In TIFA, questions (*e.g.*, "*is there a green backpack?*") for each text prompt are generated by GPT-3.5 [22] and a vision-language model [16] is used to provide binary or multiple-choice answers. A case is considered incorrect if the TIFA score for any question regarding object presence or attribute binding is incorrect.

Text embeddings vs. text self-attention maps on prompts with complex sentence structure. To further highlight the generalizability of using the text self-attention matrix, we extend our analysis to more complex prompts, including those with *relative pronouns* (*e.g.*, who, which etc.). Interestingly, Figure A shows that the text self-attention matrix effectively captures the syntactic role of words like *whose*, emphasizing the preceding vocabulary (e.g., *valley*). In contrast to the text self-attention matrix, the text embedding similarity rarely exhibits this pattern. In addition, Figure B demonstrates the text self-attention maps for the MSCOCO captions included in TIFA benchmark.

Pixart-α text embedding and text self-attention maps. We apply our method to PixArt-α, which uses T5-XXL (T5) as its text encoder and DiT as its backbone. Figures C (a) and (b) show that the T5 text encoder exhibits patterns similar to those of the CLIP text encoder: the text embeddings do not capture word relationships, whereas the text self-attention maps do—bound tokens tend to have higher attention values compared to unbound ones.

B.2. Experiments details on Section 5

Evaluation metric-CLIP similarity score. Image-text similarity includes two measures: *full-prompt similarity*, which evaluates overall alignment with the prompt, and *minimum object similarity*, defined as the lowest similarity score between the generated image and the two main subjects in the prompt. Additionally, we use BLIP [17] to generate captions for the images, comparing the input prompt with these captions using CLIP to assess text-text similarity.

Experiment details in T-SAM on SD. Our method builds upon Stable Diffusion (SD) v1.5 [29]. To enhance processing, we apply Gaussian smoothing to the cross-attention maps before computing cosine similarity, as discussed in [3]. Additionally,

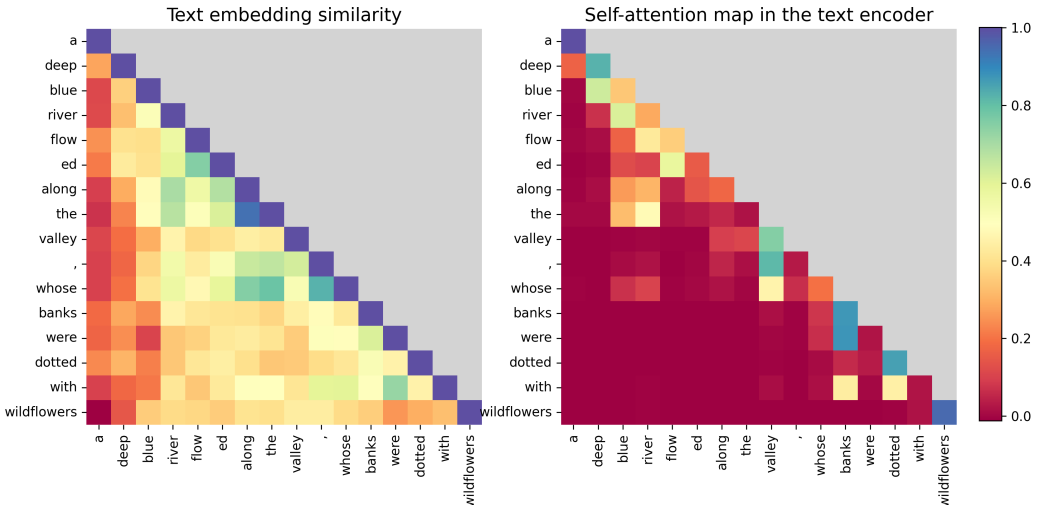


Figure A. Comparison of text embedding cosine similarity (left) and text self-attention maps (right) on the complex prompt, *"A deep blue river flowed along the valley, whose banks were dotted with wildflowers"*. The prompt includes relative pronouns *whose*.

Table B. CLIP scores over varying α with $\gamma = 4$.

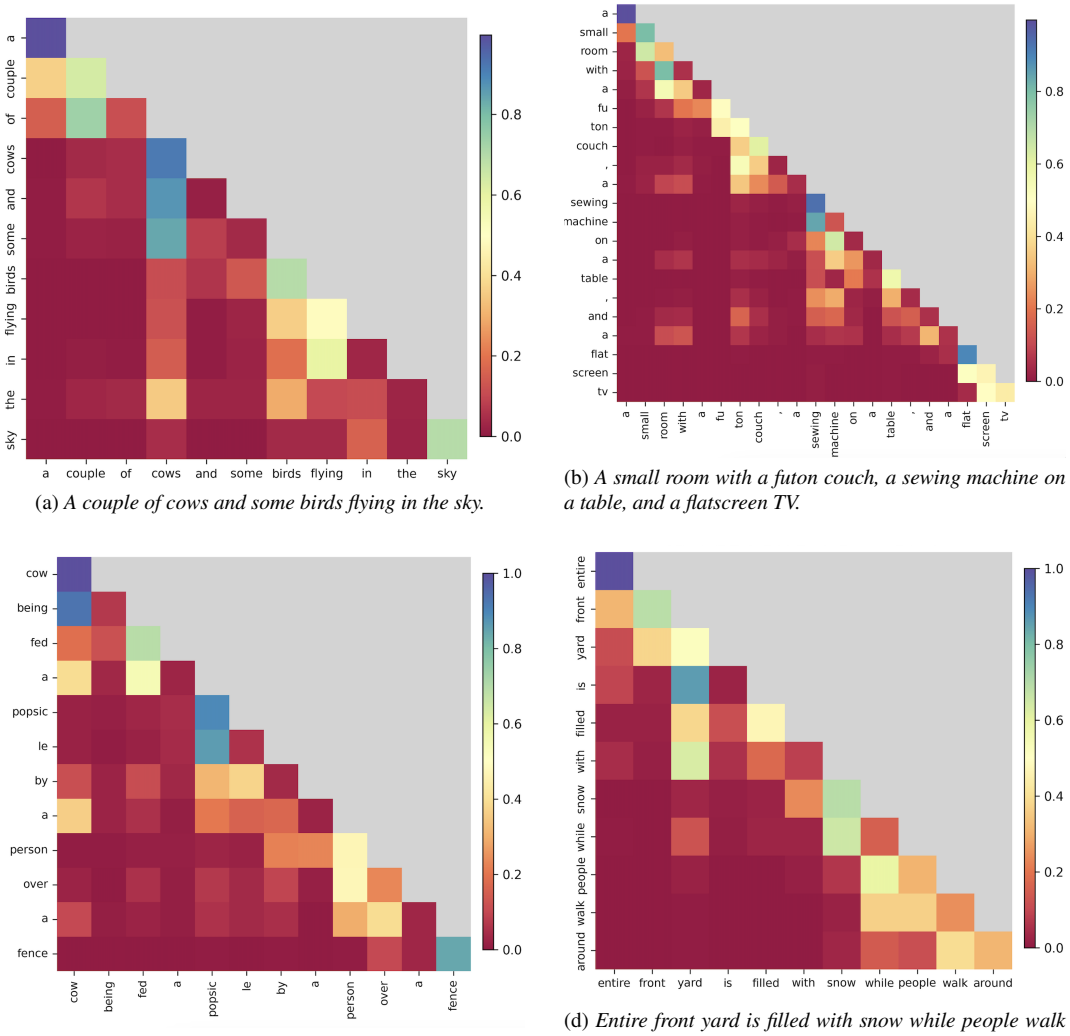
α	CLIP-T	CLIP-I (Full, Min)
5	0.80	0.36, 0.28
10	0.80	0.36, 0.28
15	0.79	0.36, 0.28
20	0.78	0.36, 0.28

Table C. CLIP scores over varying γ with $\alpha = 10$.

γ	CLIP-T	CLIP-I (Full, Min)
2	0.81	0.36, 0.28
3	0.80	0.36, 0.28
4	0.80	0.36, 0.28
5	0.80	0.36, 0.28

we renormalize the text self-attention maps, excluding the <BOS> and <EOS> tokens as clarified in Eq.(3). When computing loss function in Eq.(14), the first row, corresponding to the <BOS> token, is omitted from the computations. Regarding the prompt datasets, for *Objects*, we use 66 prompts structured as "[attribute$_1$][object$_1$] *and* [attribute$_2$][object$_2$]". For *Animals-Objects*, 144 prompts are employed. They are structured in two templates: "[animal] *with* [object]" and "[animal] *and* [attribute][object]".

Ablation study in SD. A larger α in our optimization process imposes stronger constraints on the latent variable (z_t), enhancing the regularization of cross-attention maps by aligning them more closely with the text self-attention maps. On the other hand, γ serves as an exponent, amplifying larger values and compressing smaller ones, thereby controlling the temperature. We conducted a grid search for α within the set $\{5, 10, 15, 25, 40\}$, and for γ, we explored values in $\{2, 3, 4\}$. The parameters that maximize the CLIP-full and CLIP-min similarity scores are chosen. Based on the grid search, we ultimately selected a scale factor of $\alpha = 10$ and $\gamma = 2$ for *Objects* and *Animals-Objects*, and $\alpha = 40$ and $\gamma = 4$ for the TIFA benchmark, achieving the optimal balance between performance and output quality. Table B and Table C show a sensitivity analysis on *Objects* prompts.



(a) *A couple of cows and some birds flying in the sky.*

(b) *A small room with a futon couch, a sewing machine on a table, and a flatscreen TV.*

(c) *Cow being fed a popsicle by a person over a fence.*

(d) *Entire front yard is filled with snow while people walk around.*

Figure B. Text self-attention maps power by 3 for the MSCOCO captions included in TIFA benchmark.

Experiment details in T-SAM on Pixart-α. We highlight key differences between the CLIP and T5-XXL text encoders. Unlike the CLIP encoder, which employs causal attention, T5-XXL uses bidirectional attention. Accordingly, when applying T-SAM to T5-XXL, we redefine T' in Equation 2 as $\mathsf{T}' = \text{tril}((\mathsf{T}' + \mathsf{T}'^{\dagger})/2)$ to ensure consistency with the CLIP-based formulation, where tril stands for "lower triangular part" of a matrix. Since PixArt-α is built on the DiT architecture, we compute the average cross-attention maps from layers 10–18. We set $\gamma = 4$ and $\alpha = 15$, and perform 50 inference steps. During denoising, we update z_t at each step from $t = 1$ to 25, with 10 iterative updates applied at steps 0, 10, and 20. For a fair comparison, we also use 50 inference steps for the original PixArt-α experiments.

Qualitative comparison between T-SAM and Pixart-α. Figure D demonstrates the effectiveness of T-SAM on PixArt-α using structured templates (*Objects* and *Animals-Objects*). These results show that extracting syntactic information from text self-attention maps offers an effective way to improve semantic alignment in PixArt-α.

User study. The user study result in Table D show that T-SAM on SD outperforms LB and achieves performance comparable to *CONFORM*. Note that LB and CONFORM rely on external information. Additionally, Table E shows T-SAM on PixArt-α surpasses the base PixArt-α model.

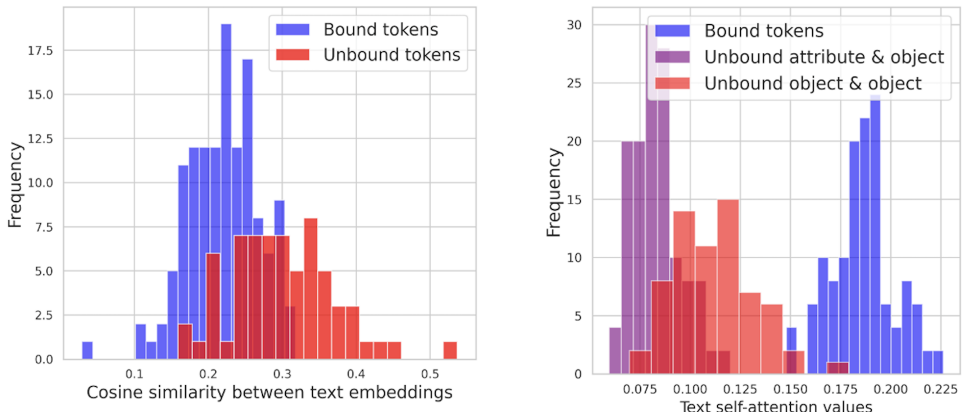


Figure C. Analysis of the PixArt-α text encoder. The text embeddings do not capture word relationships, whereas the text self-attention maps do. **Left:** The distributions of text embedding similarity between bound tokens and unbound tokens. **Right:** The distributions of text self-attention value for bound tokens and unbound tokens.

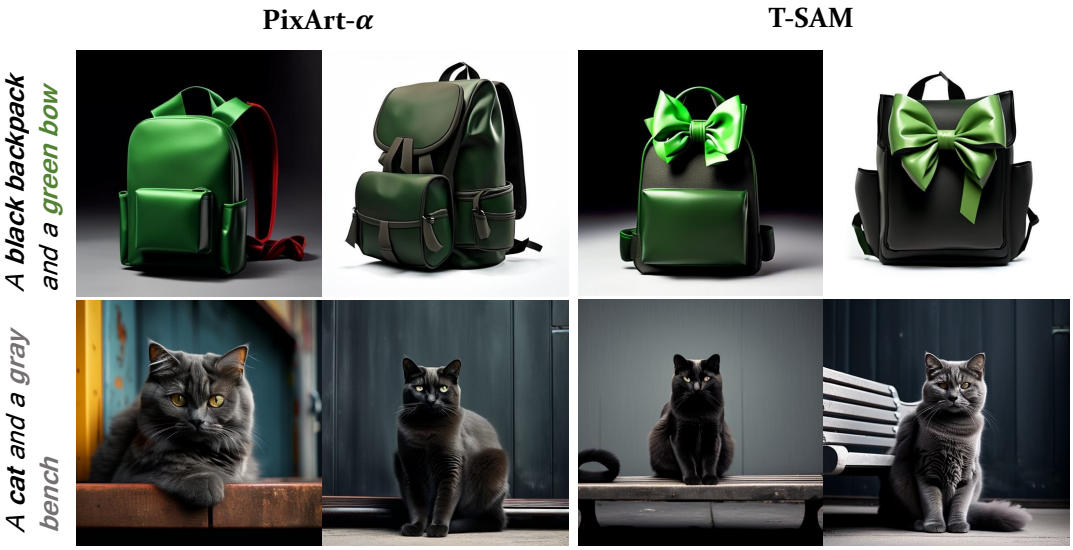


Figure D. Qualitative comparison between T-SAM (ours) and PixArt-α.

Table D. Results from a user survey with 22 respondents. SD is used as the base model.

Comparison	**Ours Wins (%)**	**Ours Loses (%)**	**Tie (%)**
Ours vs. LB	27.4	12.9	59.7
Ours vs. CONFORM	12.6	14.6	72.7

Table E. Results from a user survey with 25 respondents. Pixart-α is used as the base model.

Comparison	**Ours Wins (%)**	**Ours Loses (%)**	**Tie (%)**
Ours vs. PixArt-α	**25.0**	18.7	56.3