DOS: Directional Object Separation in Text Embeddings for Multi-Object Image Generation

Dongnam Byun[1,3], **Jungwon Park**[1], **Jungmin Ko**[2], **Changin Choi**[2,4], **Wonjong Rhee**[1,2]
[1]Department of Intelligence and Information, Seoul National University
[2]Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3]Samsung Research, Samsung Electronics
[4]SAIT, Samsung Electronics
{east928, quoded97, jungminko, ci2015.choi, wrhee}@snu.ac.kr

Abstract

Recent progress in text-to-image (T2I) generative models has led to significant improvements in generating high-quality images aligned with text prompts. However, these models still struggle with prompts involving multiple objects, often resulting in *object neglect* or *object mixing*. Through extensive studies, we identify four problematic scenarios, Similar Shapes, Similar Textures, Dissimilar Background Biases, and Many Objects, where inter-object relationships frequently lead to such failures. Motivated by two key observations about CLIP embeddings, we propose DOS (Directional Object Separation), a method that modifies three types of CLIP text embeddings before passing them into text-to-image models. Experimental results show that DOS consistently improves the success rate of multi-object image generation and reduces object mixing. In human evaluations, DOS significantly outperforms four competing methods, receiving 26.24%-43.04% more votes across four benchmarks. These results highlight DOS as a practical and effective solution for improving multi-object image generation.

Code — https://github.com/dongnami/DOS
Extended version — https://arxiv.org/abs/2510.14376

1 Introduction

Recent progress in text-to-image (T2I) generative models has led to substantial improvements in generating high-quality images that closely align with text prompts (Podell et al. 2023; Stability AI 2024; Black Forest Labs 2025). These models leverage text embeddings from multiple CLIP (Radford et al. 2021) encoders, and optionally the T5 (Raffel et al. 2020) encoder, to enhance image-text alignment. However, they still struggle to accurately generate images containing multiple objects. In this work, we identify four scenarios where this issue becomes particularly severe and demonstrate that modifying the CLIP text embeddings, without altering the image generation process within the text-to-image models, can substantially mitigate the problem.

Previous work on multi-object image generation has primarily focused on the relationships between nouns and their associated attributes (Rassin et al. 2023; Zhuang, Hu, and Gao 2024; Hu et al. 2024a), often overlooking challenges that arise from interactions between multiple noun entities. However, through extensive experiments with state-of-the-art T2I models, we find that failures, specifically *object neglect* or *object mixing*, occur frequently in scenarios involving inter-object relationships, even when no attributes are present. To investigate these issues more thoroughly, we focus on prompts that primarily contain objects and avoid descriptive attributes. Following extensive exploration and experimentation, we identify four types of problematic scenarios. As shown in Figure 1, the first three categories represent specific types of inter-object relationships where failures become particularly severe: *Similar Shapes*, *Similar Textures*, and *Dissimilar Background Biases*. *Background bias* refers to the natural background typically associated with each object (e.g., 'land' for 'deer' vs. 'sea' for 'sea turtle'). The fourth, *Many Objects*, captures the compounded effect of numerous inter-object relationships, which may include but are not limited to the first three types.

The motivation for our method stems from two previously reported observations regarding CLIP text embeddings used in T2I models: (1) CLIP employs a causal masking mechanism, where information from earlier tokens is mixed into later token embeddings, leading to *information mix-ups* in the embeddings of later tokens (Zhuang, Hu, and Gao 2024; Chen et al. 2024), (2) the difference between two CLIP text embeddings appears to encode directional information, which can be indirectly observed through the behavior of T2I models (Hu et al. 2024a). Our method aims to mitigate information mix-ups in CLIP text embeddings by constructing directional vectors, referred to as *separation vectors*, derived from the difference between two CLIP embeddings. These vectors are added to the original text embeddings to inject directional information that encourages the separation of object-specific information.

Our method, called DOS (Directional Object Separation), constructs a separation vector for each object pair and each type of CLIP text embedding, namely semantic token embeddings of all object nouns, EOT embeddings, and pooled embeddings, to encode directional information that promotes separation. For each embedding type, the separation vectors of all object pairs are combined through a weighted average, where the weights reflect the difficulty of separating each corresponding pair. This weighted combination forms

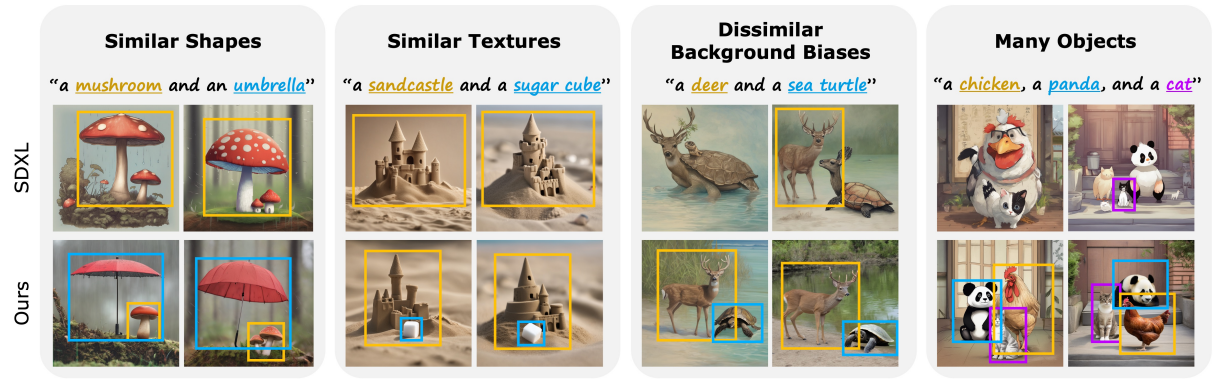

Figure 1: Text-to-image generative models often struggle to generate images with multiple objects, especially in four scenarios where relationships between objects, referred to as *inter-object relationships*, frequently lead to object neglect or object mixing: Similar Shapes, Similar Textures, Dissimilar Background Biases, and Many Objects. Our method mitigates these failures by modifying text embeddings before feeding them into the text-to-image model.

the DOS vector, which is then added to the original text embeddings.

Across four benchmarks corresponding to the problematic scenarios of *Similar Shapes*, *Similar Textures*, *Dissimilar Background Biases*, and *Many Objects*, DOS improves the success rate of multi-object image generation while reducing object mixing. In human evaluations, compared to four other models including the baseline, DOS received 26.24%-43.04% more votes than the second-best results across the four benchmarks. Since DOS does not alter the image generation process within the text-to-image model, its inference is approximately four times faster than well-known latent modification methods such as Attend-and-Excite (Chefer et al. 2023). These results demonstrate that DOS is both an effective and efficient algorithm for addressing object neglect and object mixing, which frequently arise from inter-object relationships in multi-object image generation.

2 Related Work

Multi-object image generation involves generating images containing multiple objects, optionally along with their associated attributes. However, during generation, interactions among multiple objects, and optionally their associated attributes, often lead to failures such as object neglect, object mixing, or inaccurate attribute binding. A substantial body of work has been proposed to address these issues (Chefer et al. 2023; Hu et al. 2024b; Jiang et al. 2024). Among them, Attend-and-Excite (Chefer et al. 2023) demonstrates that iteratively updating the intermediate image latent can reduce such failures. This inspired a series of follow-up methods based on latent modification, such as CONFORM (Meral et al. 2024). While effective to some extent, these methods typically require expensive iterative gradient updates during inference and often introduce visual artifacts due to deviations from the latent distribution the model was trained on. To overcome these limitations, more recent approaches have focused on modifying only the text embeddings–without altering the intermediate image latent (Chen et al. 2024; Hu et al. 2024a; Zhuang, Hu, and Gao 2024). Among these, TEBOpt (Chen et al. 2024) proposes modifying the semantic token embeddings corresponding to each object to better separate them. This approach significantly reduces information bias (i.e., cases where certain objects are more frequently generated than others), but shows only limited improvements in addressing object neglect and object mixing. In contrast, our method leverages *directional information* to modify all types of CLIP text embeddings–not only semantic token embeddings, but also EOT and pooled embeddings–in a way that promotes object separation, significantly reducing both object neglect and object mixing in multi-object image generation.

3 Preliminary Analysis

In Figure 1, we identify four types of problematic scenarios involving inter-object relationships: *Similar Shapes*, *Similar Textures*, *Dissimilar Background Biases*, and *Many Objects*. To examine how each of these scenarios intensifies object neglect or object mixing, we analyze four corresponding aspects in multi-object image generation. For each aspect, we construct two sets of multi-object prompts that contrast in condition. For example, under the shape aspect, the *dissimilar shapes* condition consists of 20 prompts featuring two objects with dissimilar shapes, following the template "a/an $\{object\ A\}$ and a/an $\{object\ B\}$." In contrast, the *similar shapes* condition includes 20 prompts with two similarly shaped objects. The conditions for the other three aspects are defined analogously. Full details on the construction of all prompt sets are provided in Table 6 of Appendix A.

We generate images for each prompt using 10 random seeds, resulting in 200 images per condition. We then compute the success rate, defined as the percentage of images that exhibit neither object neglect nor object mixing. Figure 2 presents both the quantitative results (top row) and a qualitative example (bottom row) for each condition across the four aspects. The observed differences in success rates between the two contrasting conditions range from 26.5% to 60.5%. These findings demonstrate that the four identified scenarios significantly intensify generation failures and motivate our focus on addressing them to reduce object neglect and mixing in multi-object image generation.

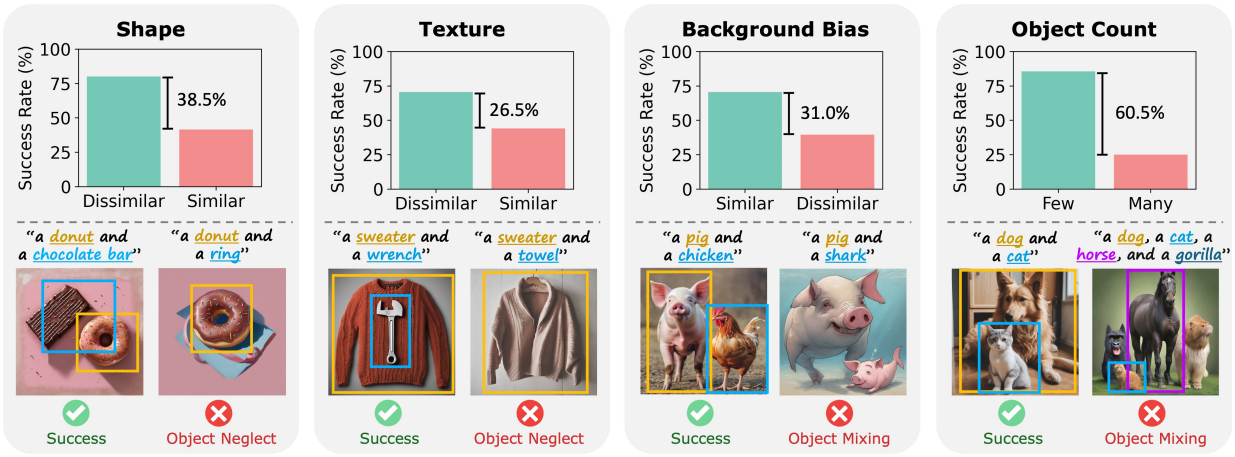

Figure 2: Analysis of four aspects of inter-object relationships in multi-object image generation. The top row shows quantitative comparisons between two contrasting conditions for each aspect, while the bottom row shows a qualitative example for each condition. A lower success rate indicates more frequent failures, such as object neglect or object mixing.

4 DOS: Directional Object Separation in Text Embeddings

Background on CLIP text embeddings in text-to-image generative models. Let P be the input prompt for image generation, which is passed through CLIP text encoders to produce the CLIP text embedding: $[\boldsymbol{c}_{\text{SOT}}, \boldsymbol{c}_{t_1}, \ldots, \boldsymbol{c}_{t_L}, \boldsymbol{c}_{\text{EOT}}, \boldsymbol{c}_{\text{PAD}}, \ldots, \boldsymbol{c}_{\text{PAD}}] \in \mathbb{R}^{77\times d}$, where SOT, EOT, and PAD denote special tokens indicating the start-of-text, end-of-text, and padding respectively; t_i denotes the i-th semantic token; and d is the embedding dimension. Recent T2I models such as SDXL also use a pooled embedding $\boldsymbol{c}_{\text{pool}}$, which is computed by applying a linear projection to the end-of-text (EOT) embedding $\boldsymbol{c}_{\text{EOT}}$. All these text embeddings are then used to condition the T2I model during image generation. Our method modifies these embeddings before passing them to the T2I model to mitigate object neglect and object mixing in multi-object image generation.

Overview of the proposed method. Suppose the input text prompt P contains N object nouns, where the n-th object noun is denoted by obj_n for all $n \in \{1, 2, \ldots N\}$. Let $\boldsymbol{c}^n_{\text{obj}}$ denote the semantic token embedding corresponding to obj_n. Additionally, let $\boldsymbol{c}_{\text{EOT}}$ and $\boldsymbol{c}_{\text{pool}}$ denote the EOT and pooled embeddings, respectively. Our method modifies all of these CLIP text embeddings by adding DOS vectors to each embedding. These DOS vectors are computed by aggregating separation vectors across all object pairs, where each separation vector defines a direction that distinguishes one object from the other in the pair. The following subsections describe: (1) how the separation vectors are computed for each object pair, (2) how the strength of each separation vector is adjusted prior to aggregation, and (3) how the adjusted vectors are combined to construct the final DOS vectors for each type of embedding. An overview of our method is illustrated in Figure 3.

Computing pairwise separation vectors. The separation vectors for each object pair (obj_n, obj_m) are designed to encode directional information that promotes separation from obj_m to obj_n. These vectors are computed differently depending on the type of text embedding. For the semantic token embedding $\boldsymbol{c}^n_{\text{obj}}$, the separation vector is defined as the difference between two object-specific embeddings, $\boldsymbol{c}^{\text{pure},n}_{\text{obj}}$ and $\boldsymbol{c}^{\text{pure},m}_{\text{obj}}$, each obtained from a pure prompt of the form $P_{\text{pure},i}$ = "a $\{obj_i\}$" for $i = n, m$, respectively:

$$\mathbf{s}^{(n,m)}_{\text{obj}} = \boldsymbol{c}^{\text{pure},n}_{\text{obj}} - \boldsymbol{c}^{\text{pure},m}_{\text{obj}} \tag{1}$$

Using pure prompts ensures that $\boldsymbol{c}^{\text{pure},n}_{\text{obj}}$ and $\boldsymbol{c}^{\text{pure},m}_{\text{obj}}$ capture clean, object-specific representations by avoiding undesirable information mix-ups across tokens. In contrast, for EOT and pooled embeddings, we intentionally exploit information mix-ups to define separation vectors. Specifically, we construct two contrastive prompts: $P_{\text{sep},(n,m)}$ = "a $\{obj_n\}$ separated from a $\{obj_m\}$" and $P_{\text{mix},(n,m)}$ = "a $\{obj_n\}$ mixed with a $\{obj_m\}$". From these, we extract the corresponding EOT and pooled embeddings, denoted as $\boldsymbol{c}^{\text{sep},(n,m)}_{\text{EOT/pool}}$ and $\boldsymbol{c}^{\text{mix},(n,m)}_{\text{EOT/pool}}$, and define the separation vectors as:

$$\mathbf{s}^{(n,m)}_{\text{EOT/pool}} = \boldsymbol{c}^{\text{sep},(n,m)}_{\text{EOT/pool}} - \boldsymbol{c}^{\text{mix},(n,m)}_{\text{EOT/pool}} \tag{2}$$

While we use the unified notation $\mathbf{s}^{(n,m)}_{\text{EOT/pool}}$ for brevity, these separation vectors are computed independently for the EOT and pooled embeddings. Overall, our separation vector design extracts directional signals that promote object separation: by avoiding information mix-ups for semantic token embeddings, and by deliberately exploiting contrastive prompts to leverage such mix-ups for EOT and pooled embeddings.

Computing pairwise adaptive strengths. In Figure 2, we observed that certain object pairs are more susceptible to object neglect or object mixing than others. Motivated by this observation, we adjust the strength of the separation vectors for each object pair based on their visual (shape and texture) similarities and background bias dissimilarities. To

this end, we first select 42 representative words covering a diverse range of object shapes and textures through interactions with GPT-o4-mini-high (OpenAI 2025) and define prompts of the form $P_{\text{attr},(k,n)}$ = "a $\{attribute_k\}$ $\{obj_n\}$" for all $k \in \{1, 2, \ldots, 42\}$. Similarly, we select 36 representative background phrases and define prompts of the form $P_{\text{bg},(l,n)}$ ="$\{background_l\}$, there is a $\{obj_n\}$" for all $l \in \{1, 2, \ldots, 36\}$. We then compare the text embeddings $\boldsymbol{c}^{\text{pure},n}_{\text{obj/EOT/pool}}$, obtained from the pure prompt $P_{\text{pure},n}$ = "a $\{obj_n\}$", with embeddings $\boldsymbol{c}^{\text{attr},(k,n)}_{\text{obj/EOT/pool}}$ and $\boldsymbol{c}^{\text{bg},(l,n)}_{\text{obj/EOT/pool}}$ obtained from the attribute and background prompts, respectively. Cosine similarity is used for this comparison:

$$\mathbf{a}^{\text{attr},n}_{\tau}[k] = \text{sim}(\boldsymbol{c}^{\text{pure},n}_{\tau}, \boldsymbol{c}^{\text{attr},(k,n)}_{\tau}), \tag{3}$$

$$\mathbf{a}^{\text{bg},n}_{\tau}[l] = \text{sim}(\boldsymbol{c}^{\text{pure},n}_{\tau}, \boldsymbol{c}^{\text{bg},(l,n)}_{\tau}), \tag{4}$$

for $k \in \{1, 2, \ldots, 42\}$, $l \in \{1, 2, \ldots, 36\}$, and $\tau \in \{\text{obj}, \text{EOT}, \text{pool}\}$, where τ denotes the type of text embedding. The resulting lists of cosine similarities $\mathbf{a}^{\text{attr},n}_{\tau}$ and $\mathbf{a}^{\text{bg},n}_{\tau}$ reflect how closely obj_n aligns with the predefined visual attributes and background contexts. We compute the same lists for obj_m, and define the adaptive strength for the object pair (obj_n, obj_m) as:

$$\alpha^{(n,m)}_{\tau} = \max\Big\{\sigma\big(\rho(\mathbf{a}^{\text{attr},n}_{\tau}, \mathbf{a}^{\text{attr},m}_{\tau}); x^{\text{attr}}_{\tau,0}\big), \sigma\big(1 - \rho(\mathbf{a}^{\text{bg},n}_{\tau}, \mathbf{a}^{\text{bg},m}_{\tau}); x^{\text{bg}}_{\tau,0}\big)\Big\}, \tag{5}$$

for $\tau \in \{\text{obj}, \text{EOT}, \text{pool}\}$, where $\rho(\cdot, \cdot)$ denotes Pearson correlation, which is used to avoid mean and scaling effects, and $\sigma(x; x_{\tau,0})$ is a shifted tempered sigmoid function defined as:

$$\rho(x, y) = \frac{\sum_t (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_t (x_t - \bar{x})^2}\ \sqrt{\sum_t (y_t - \bar{y})^2}}, \tag{6}$$

$$\sigma(x; x_{\tau,0}) = \big(1 + \exp\{-(x - x_{\tau,0})/T\}\big)^{-1} \tag{7}$$

Here, the offset $x^{\text{attr}}_{\tau,0}$ is the mean of $\rho(\mathbf{a}^{\text{attr},i}_{\tau}, \mathbf{a}^{\text{attr},j}_{\tau})$ over all object pairs (obj_i, obj_j) from the MS-COCO (Lin et al. 2014) dataset. Similarly, $x^{\text{bg}}_{\tau,0}$ is the mean of $\big(1 - \rho(\mathbf{a}^{\text{bg},i}_{\tau}, \mathbf{a}^{\text{bg},j}_{\tau})\big)$ over the same object pairs. These adaptive strengths promote stronger separation for object pairs that are more likely to be mixed or neglected, thereby helping to reduce object neglect and mixing in multi-object image generation. In Appendix B.1, we provide the full list of 42 representative words describing object shapes and textures and 36 representative background phrases, which are used to construct the prompts $P_{\text{attr},(k,n)}$ and $P_{\text{bg},(l,n)}$, respectively.

Updating text embeddings. Finally, the DOS vectors for the semantic token, EOT, and pooled embeddings are computed as a weighted average of the separation vectors, where each weight corresponds to the adaptive strength of a given object pair:

$$\mathbf{v}^{\text{DOS},n}_{\tau} = \frac{1}{N-1} \sum_{m \neq n} \alpha^{(n,m)}_{\tau}\, \mathbf{s}^{(n,m)}_{\tau}, \tag{8}$$

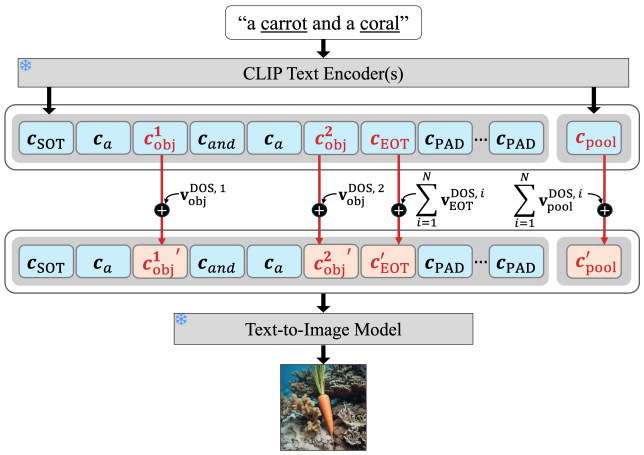

Figure 3: Method Overview. DOS vectors are computed for the semantic token embeddings corresponding to object nouns, the EOT embedding, and the pooled embedding. These vectors are added to the corresponding CLIP text embeddings, namely $\boldsymbol{c}^{n}_{\text{obj}}$ for $n = 1, \ldots, N$, $\boldsymbol{c}_{\text{EOT}}$, and $\boldsymbol{c}_{\text{pool}}$, before being passed into the text-to-image model. N denotes the number of object nouns in the prompt.

where $\tau \in \{\text{obj}, \text{EOT}, \text{pool}\}$, and $n \in \{1, 2, \ldots, N\}$. These DOS vectors are then added to the original text embeddings, which are subsequently passed into the T2I model. Specifically, semantic token embeddings corresponding to object nouns are updated as:

$$\boldsymbol{c}^{n}_{\text{obj}}{}' = \boldsymbol{c}^{n}_{\text{obj}} + \mathbf{v}^{\text{DOS},n}_{\text{obj}}, \tag{9}$$

for $n \in \{1, 2, \ldots, N\}$. If obj_n consists of multiple tokens, their semantic token embeddings are first averaged before computing the DOS vector, and the resulting DOS vector is then added to each of the semantic token embeddings that constitute obj_n. In addition, the EOT and pooled embeddings are updated by summing the DOS vectors across all objects:

$$\boldsymbol{c}'_{\text{EOT/pool}} = \boldsymbol{c}_{\text{EOT/pool}} + \sum_{i=1}^{N} \mathbf{v}^{\text{DOS},i}_{\text{EOT/pool}} \tag{10}$$

While we use the unified notation $\boldsymbol{c}_{\text{EOT/pool}}$ and $\boldsymbol{c}'_{\text{EOT/pool}}$ for brevity, the EOT and pooled embeddings are computed and updated independently. This aggregation across all objects is intended to mitigate the compounded effects of inter-object relationships in multi-object image generation.

5 Experiments

5.1 Experimental Setup

Implementation details. Our method is based on SDXL (Podell et al. 2023) and has also been implemented on the recently released Stable Diffusion 3.5 (SD3.5) (Stability AI 2024). We apply DOS updates to all CLIP text embeddings used by each model. For inference, we use a guidance scale of 5.0 with 50 denoising steps for SDXL, and a guidance scale of 7.0 with 28 steps for SD3.5, which are the official default configurations. The temperature T

Base Model	Method	Similar Shapes		Similar Textures		Dissimilar Background Biases		Many Objects	
		SR↑	MR↓	SR↑	MR↓	SR↑	MR↓	SR↑	MR↓
SDXL	Base	48.00%	6.50%	58.00%	7.50%	46.00%	22.50%	23.00%	27.50%
	TEBOpt	52.00%	5.00%	61.00%	4.00%	44.50%	23.00%	24.00%	27.00%
	A&E	60.50%	6.00%	67.50%	5.00%	53.50%	25.00%	28.50%	28.50%
	CONFORM	54.00%	6.50%	64.50%	9.00%	55.50%	22.00%	37.50%	26.50%
	Ours	**64.00%**	**3.50%**	**71.50%**	**3.50%**	**68.50%**	**17.00%**	**48.00%**	**15.50%**
SD3.5	Base	75.50%	4.00%	79.00%	6.00%	78.00%	17.50%	70.00%	16.50%
	Ours	**81.00%**	**3.00%**	**87.50%**	**3.00%**	**85.50%**	**13.50%**	**76.50%**	**10.50%**

Table 1: Quantitative comparison across four benchmarks: Similar Shapes, Similar Textures, Dissimilar Background Biases, and Many Objects. SR denotes Success Rate (↑ higher is better), and MR denotes Mixture Rate (↓ lower is better).

Method	Similar Shapes	Similar Textures	Dissimilar Background Biases	Many Objects
SDXL	7.59%	9.38%	3.59%	3.21%
TEBOpt	7.24%	7.03%	4.22%	4.82%
A&E	13.62%	14.84%	22.81%	11.43%
CONFORM	12.76%	16.88%	12.19%	12.14%
Ours	**50.52%**	**43.12%**	**53.44%**	**55.18%**
No winner	8.28%	8.75%	3.75%	13.21%

Table 2: Results of human preference studies across four benchmarks. All methods are based on SDXL.

	SDXL	TEBOpt	A&E	CONFORM	Ours
Inference Time (s)	12.97	13.50	58.83	58.48	13.87

Table 3: Inference time comparison (in seconds). All values are averaged over 10 runs. All methods are based on SDXL.

of the shifted tempered sigmoid functions in Eq. 7 is set to $T = 0.6$ for all base models, determined through a light validation process. Additional implementation details are provided in Appendix B.1.

Dataset. To focus on the four problematic scenarios in multi-object image generation, we construct four corresponding benchmarks: Similar Shapes, Similar Textures, Dissimilar Background Biases, and Many Objects. Each of the first three benchmarks consists of 50 two-object prompts, following the template: "a/an $\{object\ A\}$ and a/an $\{object\ B\}$." The object pairs for these benchmarks are selected through interactions with GPT-o4-mini-high (OpenAI 2025). For the Many Objects benchmark, we extend the animal list used in Attend-and-Excite to include 17 animals, from which we randomly sample to create 25 three-object prompts and 25 four-object prompts. For each method, we generate 4 images per prompt using different random seeds, resulting in 200 generated images per benchmark. While the main experiment for Many Objects focuses on the animal list, we present results for an extended benchmark in Table 11 of Appendix C.3 that also includes electronics, vehicles, and fruits. Detailed information about the construction of each benchmark is provided in Appendix B.2.

Evaluation metrics. We adjust the VLM-based metric from EnMMDiT (Wei et al. 2024) to measure both the success rate (SR) and mixture rate (MR) for each generated image. SR evaluates whether all the objects specified in the prompt are clearly present in the image, while MR assesses whether any of the objects appear mixed. Specifically, we provided GPT-4o-mini with the generated image and the text prompt used for image generation, asking it to classify each object in the prompt as (1) fully intact, (2) mixed, or (3) absent. SR is calculated as the ratio of images where all objects are classified as fully intact, and MR is calculated as the ratio of images where one or more objects are classified as mixed. A high SR score and a low MR score indicate superior image-text alignment. We further validate the consistency of the results across two additional evaluators in Table 13 of Appendix C.6. Additional details about the VLM-based metrics are provided in Appendix B.3.

5.2 Experimental Results

We compare our method with TEBOpt (Chen et al. 2024), Attend-and-Excite (A&E) (Chefer et al. 2023), and CONFORM (Meral et al. 2024) using SDXL as the base model. For SD3.5, we only compare our method with the baseline model in our main experiment. Further comparisons on SD3.5 against recent methods, TEBOpt (Chen et al. 2024) and Self-Cross (Qiu, Wang, and Tang 2025), are provided in Table 14 of Appendix C.7.

Qualitative comparison. Qualitative comparisons for each problematic scenario are shown in Figure 4. Across all scenarios and base models, our method successfully generates all objects specified in the prompts, while other methods often exhibit object neglect or object mixing. For example, given the prompt "a carrot and an ice cream cone," our method successfully generates both objects without mixing, whereas the other methods fail to do so. In Figure 5, we also compare results on more complex prompts, where our method continues to show better generations. Additional qualitative results, including examples with complex prompts, are provided in Figures 12-17 of Appendix C.8.

Quantitative comparison. Table 1 presents quantitative comparisons across all four benchmarks in multi-object image generation. Across all benchmarks and base models, our method achieves the highest success rate (SR) (higher is better) and the lowest mixture rate (MR) (lower is better). Specifically, our method outperforms TEBOpt, which is the other text embedding modification method, by 10.50%-

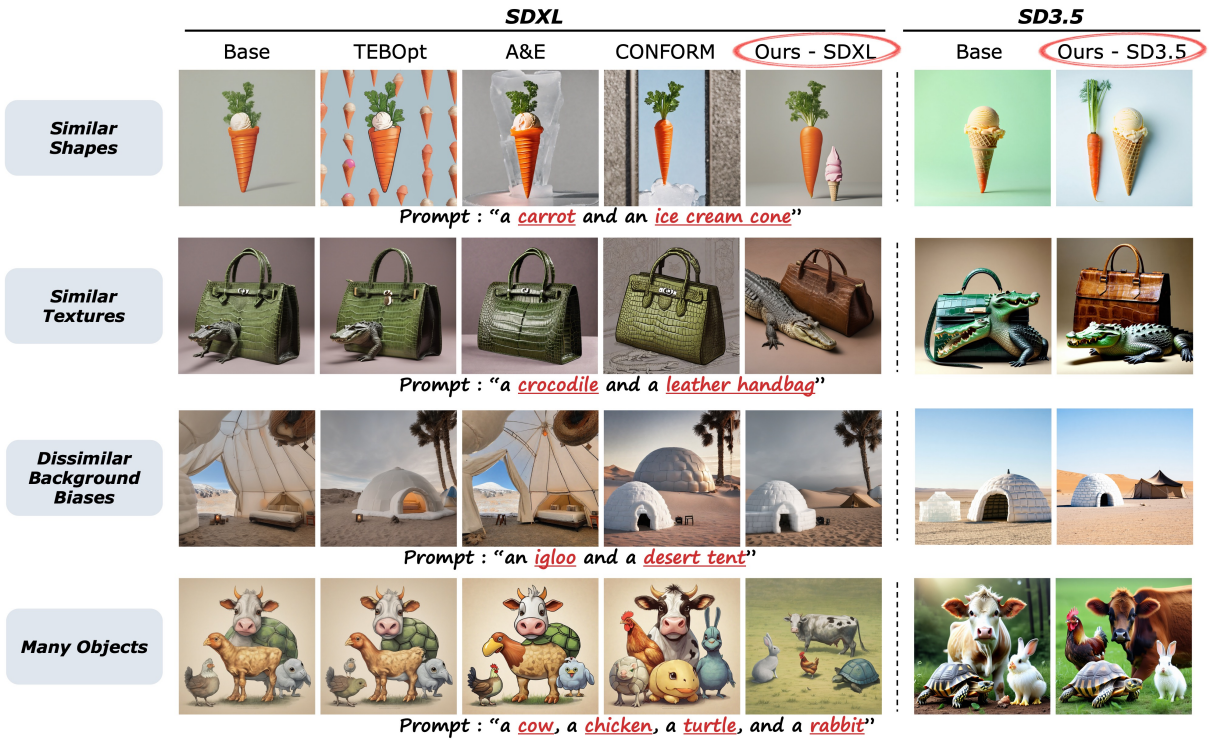

Figure 4: Qualitative comparison across four benchmarks: *Similar Shapes*, *Similar Textures*, *Dissimilar Background Biases*, and *Many Objects*. More results are shown in Figures 14-16 of Appendix C.8.

24.00% in SR. It also surpasses the latent modification methods A&E and CONFORM, which generally require much higher inference time compared to ours. On the recently released SD3.5 model, our method further improves over the baseline, demonstrating that its effectiveness extends to newer architectures.

Human preference study. We conducted human preference studies across all four benchmarks in multi-object image generation. In each survey, participants were presented with five images generated from the same text prompt but using different SDXL-based methods, and asked to select the one that best represents all the objects specified in the prompt. Participants were instructed to avoid images with object neglect or object mixing. Using Amazon Mechanical Turk (MTurk), we collected 580, 640, 640, 560 responses from 29, 32, 32, 28 valid participants for each survey, respectively. As shown in Table 2, our method significantly outperforms the others, receiving 26.24%-43.04% more votes than the second-best results across the four benchmarks. Additional details on the human preference studies, including survey questions and participant filtering criteria, are provided in Appendix B.4.

5.3 Analysis

Inference time comparison. To demonstrate the time efficiency of our method, we compare the inference times of various SDXL-based methods. As shown in Table 3, text embedding modification methods, including TEBOpt and our method, exhibit inference times similar to the baseline model, indicating minimal time overhead for modifying text embeddings. In contrast, latent modification methods A&E and CONFORM require approximately 4-5 times more computational time than the baseline, due to the iterative gradient updates during image generation. This highlights that our method offers both an efficient and effective solution to the challenges of multi-object generation.

Ablation on text embedding types. Our method applies DOS updates to three types of CLIP text embeddings: semantic token embeddings corresponding to object nouns, EOT embeddings, and pooled embeddings. In Table 4, we ablate the type of embeddings to which DOS is applied. The results show that applying DOS to either the semantic token embeddings or the EOT/pooled embeddings leads to notable improvements in SR, while MR is more significantly reduced when DOS is applied to EOT/pooled embeddings. This is likely because EOT and pooled embeddings tend to capture the overall semantics of the prompt (Wu, Yang, and Wang 2024), influencing the overall spatial structure of the generated images when modified. Finally, applying DOS to all three types of embeddings results in the best performance on both SR and MR. Qualitative examples for this ablation study are shown in Figure 10 of Appendix C.2.

Ablation on adaptive strengths. To examine the effect of the adaptive strengths defined in Eq. 5, we perform an ablation by setting all adaptive strengths to a fixed value of 0.5. The results, shown in Table 5, indicate that using the fixed strength for all separation vectors already leads to substan-

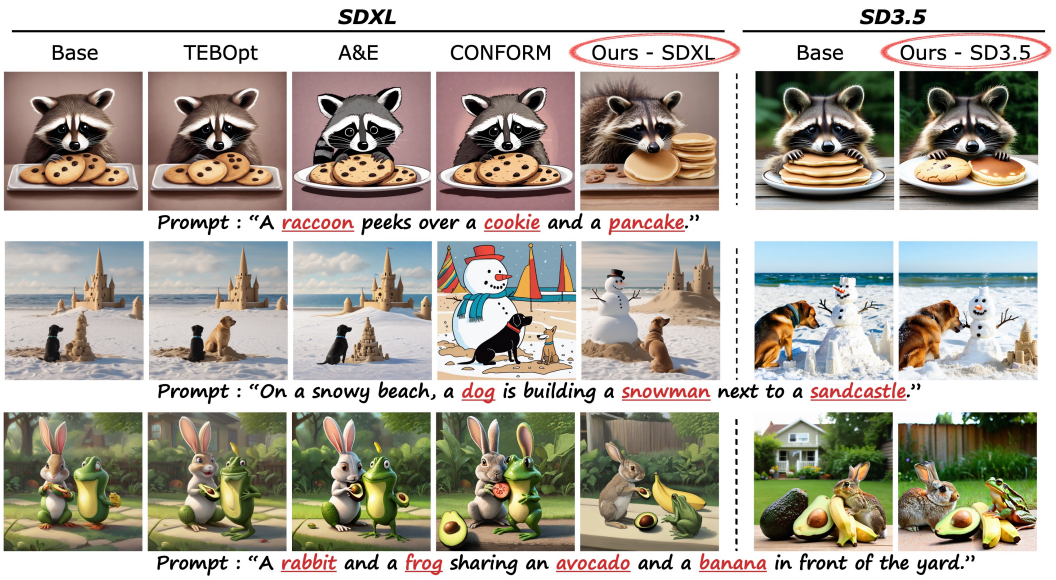

Figure 5: Qualitative comparison on complex prompts. More results are shown in Figure 17 of Appendix C.8.

Method	Similar Shapes		Similar Textures		Dissimilar Background Biases		Many Objects	
	SR↑	MR↓	SR↑	MR↓	SR↑	MR↓	SR↑	MR↓
SDXL	48.0%	6.5%	58.0%	7.5%	46.0%	22.5%	23.0%	27.5%
+obj	58.0%	5.0%	64.0%	6.5%	51.0%	23.0%	30.0%	29.5%
+EOT/pool	57.0%	4.5%	68.0%	5.0%	59.5%	17.5%	36.5%	16.0%
+obj+EOT/pool	**64.0%**	**3.5%**	**71.5%**	**3.5%**	**68.5%**	**17.0%**	**48.0%**	**15.5%**

Table 4: Ablation study on the types of text embeddings modified by DOS. We compare applying DOS to only the semantic token embeddings (obj) or only the EOT/pooled embeddings, against our default approach that modifies all.

Method	Similar Shapes		Similar Textures		Dissimilar Background Biases		Many Objects	
	SR↑	MR↓	SR↑	MR↓	SR↑	MR↓	SR↑	MR↓
SDXL	48.0%	6.5%	58.0%	7.5%	46.0%	22.5%	23.0%	27.5%
Ours ($\alpha = 0.5$)	60.5%	4.5%	67.5%	6.5%	67.0%	19.5%	41.0%	25.5%
Ours (adaptive α's)	**64.0%**	**3.5%**	**71.5%**	**3.5%**	**68.5%**	**17.0%**	**48.0%**	**15.5%**

Table 5: Ablation study comparing adaptive strengths against a fixed strength of $\alpha = 0.5$.

tial performance improvements compared to the baseline. This suggests that the modification directions specified by the separation vectors are effective on its own, even without adaptive strength modulation. However, applying adaptive strength results in even better performance, indicating that modulating the strength of separation differently for each object pair further enhances image generation with multiple objects. An additional ablation study is provided in Table 12 of Appendix C.4, examining sensitivity to predefined words and phrases used to compute adaptive strengths.

6 Limitation

We demonstrated that the adaptive strengths (defined in Eq. 5) in DOS effectively adjust the magnitude of separation vectors, leading to improved multi-object image generation (see Table 5). However, as shown in Figure 6, it does not always yield the optimal strengths. In this figure, we compare the default adaptive strength with the optimal strength obtained via grid search over 121 hyperparameter settings (an 11×11 grid with a 0.1 interval). Although such cases are rare, they suggest that there is still room to improve how the strength of each separation vector is determined, potentially by incorporating additional criteria beyond shape, texture, and background bias.

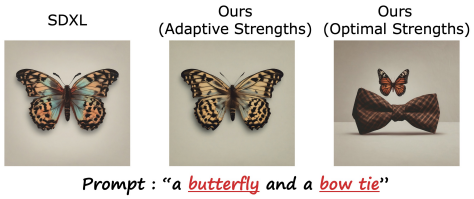

Figure 6: Example where adaptive strengths fail to properly scale separation vectors, leading to object neglect.

7 Conclusion

In this work, we propose DOS (Directional Object Separation), an efficient and effective method for modifying text embeddings to improve multi-object image generation. Through extensive studies, we identify four problematic scenarios where inter-object relationships frequently cause failures such as object neglect or object mixing. Motivated by two key observations about CLIP embeddings, we construct DOS vectors and add them to three types of CLIP text embeddings before passing them into T2I models. This approach not only offers a practical solution to challenges in multi-object image generation but also demonstrates how the properties of CLIP embeddings can be leveraged to address limitations in downstream models that rely on them.

Acknowledgments

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. RS-2022-NR070876) and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) ([NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], [No.RS-2023-00235293, Development of autonomous driving big data processing, management, search, and sharing interface technology to provide autonomous driving data according to the purpose of usage]).

References

Black Forest Labs. 2025. Announcing Black Forest Labs.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4): 1–10.

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

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

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

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Meral, T. H. S.; Simsar, E.; Tombari, F.; and Yanardag, P. 2024. 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*, 9005–9014.

OpenAI. 2025. GPT-o4-mini-high: An OpenAI Reasoning Model. Version gpt-o4-mini-high; accessed 07/16/2025.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv preprint arXiv:2307.01952*.

Qiu, W.; Wang, J.; and Tang, M. 2025. Self-Cross Diffusion Guidance for Text-to-Image Synthesis of Similar Subjects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23528–23538.

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

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

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

Stability AI. 2024. Introducing Stable Diffusion 3.5. https://stability.ai/news/introducing-stable-diffusion-3-5. Accessed: 07/16/2025.

Wei, T.; Chen, D.; Zhou, Y.; and Pan, X. 2024. Enhancing mmdit-based text-to-image models for similar subject generation. *arXiv preprint arXiv:2411.18301*.

Wu, Y.; Yang, X.; and Wang, X. 2024. Relation rectification in diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7685–7694.

Zhuang, C.; Hu, Y.; and Gao, P. 2024. Magnet: We never know how text-to-image diffusion models work, until we learn how vision-language models function. *Advances in Neural Information Processing Systems*, 37: 57115–57149.