Compositional Text-to-Image Synthesis with Attention Map Control of Diffusion Models

Ruichen Wang[1], Zekang Chen[2*], Chen Chen[1 #], Jian Ma[1], Haonan Lu[1 #], Xiaodong Lin[3]

[1]OPPO Research Institute
[2]South China University of Technology
[3]Rutgers University
wangruichen@oppo.com,chenzekang2018@163.com,{chenchen4,majian2,luhaonan}@oppo.com, lin@business.rutgers.edu

Abstract

Recent text-to-image (T2I) diffusion models show outstanding performance in generating high-quality images conditioned on textual prompts. However, they fail to semantically align the generated images with the prompts due to their limited compositional capabilities, leading to attribute leakage, entity leakage, and missing entities. In this paper, we propose a novel attention mask control strategy based on predicted object boxes to address these issues. In particular, we first train a BoxNet to predict a box for each entity that possesses the attribute specified in the prompt. Then, depending on the predicted boxes, a unique mask control is applied to the cross- and self-attention maps. Our approach produces a more semantically accurate synthesis by constraining the attention regions of each token in the prompt to the image. In addition, the proposed method is straightforward and effective and can be readily integrated into existing cross-attention-based T2I generators. We compare our approach to competing methods and demonstrate that it can faithfully convey the semantics of the original text to the generated content and achieve high availability as a ready-to-use plugin. Please refer to https://github.com/OPPO-Mente-Lab/attention-mask-control.

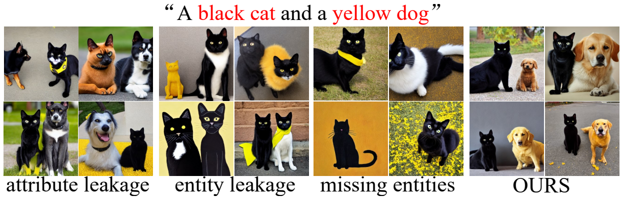

Figure 1: Example results from Stable Diffusion (first three sets of images) and Our method (last set). Our method aims to address three typical generation defects (*attribute leakage*, *entity leakage*, and *missing entities*) and generate images that are more semantically faithful to the image captions.

Introduction

Text-to-image (T2I) synthesis aims to generate realistic and diverse images conditioned on text prompts. Recently, diffusion models have achieved state-of-the-art results in this area (Rombach et al. 2022; Croitoru et al. 2023; Yang et al. 2023). Compared to previous generative models, such as generative adversarial networks (GANs) (Goodfellow et al. 2020) and variational autoencoder (VAE) (Doersch 2021), diffusion models exhibit superior performance with respect to image generation quality and diversity. They also enable better content control based on the input conditions such as grounding boxes, edge maps, or reference images, while avoiding the problems of training instability and mode collapse (Zhang, Rao, and Agrawala 2023; Li et al. 2023).

Despite their success, diffusion-model-based synthesis methods struggle to accurately interpret compositional text descriptions, especially those containing multiple objects or attributes (Feng et al. 2023; Han et al. 2023; Liu et al. 2023b; Chefer et al. 2023; Jiménez 2023). The generation defects of diffusion models such as Stable Diffusion (SD) (Rombach et al. 2022)fall into three categories: attribute leakage, entity leakage, and missing entities, as shown in Fig.1. Considering the prompt “a black cat and a yellow dog”, attribute leakage refers to the phenomenon where the attribute of one entity is observed in another (*e.g.,* a black dog). Entity leakage occurs when one entity overlays another (*e.g.,* two cats, one black and one yellow). Missing entities indicate that the model fails to generate one or more of the subjects mentioned in the input prompt (*e.g.,* only one black cat).

We attribute the infidelity issues in T2I synthesis to inaccurate attention regions, *i.e.,* the cross-attention regions between text tokens and image patches, as well as the self-attention regions within image patches themselves. Each entity and its attribute should, ideally, correspond to a coherent image region in order to generate multiple entities in a single image correctly. Existing T2I diffusion models, such as SD, lack explicit constraints on the attention regions and boundaries, which may lead to overlapping attention activations. To address these issues, we attempt to use parsed entities with attributes and their predicted object boxes to provide explicit attention boundary constraints for compositional generations. Specifically, predicted object boxes define the interest areas on images, while entities with attributes depict the interest text spans where each text token shares a common cross-attention region. By incorporating these boundary constraints, we achieve high-fidelity T2I synthesis while addressing the aforementioned problems.

[*]The author did his work during internship at OPPO Research Institute. # denotes corresponding authors.

In this paper, we propose a novel compositional T2I approach based on SD (Rombach et al. 2022) with explicit control of cross- and self-attention maps to ensure that the attention interest areas are located within the predicted object boxes, as shown in Fig.2. Specifically, we first train a BoxNet applied to the forward process of SD on the COCO dataset (Lin et al. 2014) to predict object boxes for entities with attributes parsed by a constituency parser (Honnibal et al. 2020). We then enforce unique attention mask control over the cross- and self-attention maps based on the predicted boxes (image regions) and entities with attributes (text spans). The objective of BoxNet is to provide entity-bounding boxes for subsequent attention mask control. During the diffusion model's inference process, random sampling is performed at each step, so the box positions of each entity are constantly changing (as shown in Fig.3). Consequently, the BoxNet predicts the entity box positions at each step based on the diffusion model's intermediate features to avoid excessive conflict between the current sampling result and the mask control target. Our approach produces a more semantically accurate synthesis by constraining the attention region of each text token in the image. Furthermore, using the trained BoxNet, our method can guide the diffusion inference process on the fly without fine-tuning SD. We conduct comprehensive experiments on the publicly available COCO and open-domain datasets, and the results show that our method generates images that are more closely aligned with the given descriptions, thereby improving fidelity and faithfulness. The main contributions of our work can be concluded as follows:

- We propose BoxNet, an object box prediction module capable of estimating object locations at any timestep during the forward diffusion process. The predicted object boxes closely match the locations of the entities generated by the original SD.
- We develop an effective attention mask control strategy based on the proposed BoxNet, which constrains the attention areas to lie within the predicted boxes.
- The trained BoxNet and attention mask control of our method can be easily incorporated into existing diffusion-based generators as a ready-to-use plugin. We demonstrate our model's capability by integrating it into the original SD and two variants: Attend-and-Excite (Chefer et al. 2023) and GLIGEN (Li et al. 2023).

Related Work

Text-to-Image Diffusion Models. Diffusion models are becoming increasingly popular in the T2I synthesis area due to their exceptional performance in generating high-quality images (Ramesh et al. 2021; Esser, Rombach, and Ommer 2021; Ramesh et al. 2022; Balaji et al. 2022; Saharia et al. 2022). Generally, these models take a noisy image as input and iteratively denoise it back to a clean one while semantically aligning the generated content with a text prompt. SD (Rombach et al. 2022) uses an autoencoder to create a lower-dimensional space and trains a U-Net model (Ronneberger, Fischer, and Brox 2015) based on large-scale image-text datasets in this latent space, balancing algorithm efficiency and image quality. However, diffusion models have limited expressiveness, resulting in generated content that cannot fully convey the semantics of the original text. This issue is exacerbated when dealing with complex scene descriptions or multi-object generation (Chefer et al. 2023; Feng et al. 2023; Ma et al. 2023).

Compositional Generation. Recent studies have explored various approaches to enhance the compositional generation capacity of T2I diffusion models without relying on additional bounding box input. StructureDiffusion (Feng et al. 2023) uses linguistic structures to help guide image-text cross-attention. However, the results it produces frequently fall short of addressing semantic issues at the sample level. Composable Diffusion (Liu et al. 2022) breaks down complex text descriptions into multiple easily-generated snippets. A unified image is generated by composing the output of these snippets. Yet, this approach is limited to conjunction and negation operators. AAE (Chefer et al. 2023) guides a pre-trained diffusion model to generate all subjects mentioned in the text prompt by strengthening their activations on the fly. Although AAE can address the issue of missing entities, it still struggles with attribute leakage and may produce less realistic images when presented with an atypical scene description. Unlike previous methods, our work proposes a novel two-phase method of BoxNet and Attention Mask Control, gradually controlling the generation of multiple entities during the diffusion model sampling process.

Layout to Image Generation. Through the use of artificial input conditions such as bounding boxes, shape maps, or spatial layouts, some existing methods can generate controllable images. For instance, GLIGEN (Li et al. 2023) adds trainable gated self-attention layers to integrate additional inputs, such as bounding boxes, while freezing the original model weights. Chen *et al.* (Chen, Laina, and Vedaldi 2023) propose a training-free layout guidance technique for guiding the spatial layout of generated images based on bounding boxes. Shape-Guided Diffusion (Huk Park et al. 2022) leverages an inside-outside attention mechanism during the generation process to apply the shape constraint to the attention maps based on a shape map. However, these works require prior layout information to be provided as input, which is fixed during the generation process. In order to directly control the generation results of diffusion models, our work aims to provide a pure text-to-image generation method that does not require users to specify bounding boxes or layouts. Instead, BoxNet estimates such information at each sampling step.

Layout-based Generation. Some work can directly generate the layout based on user input text information and further generate images based on layouts. Wu *et al.* (Wu et al. 2023) address the infidelity issues by imposing spatial-temporal attention control based on the pixel regions of each object predicted by a LayoutTransformer (Yang et al. 2021). However, their algorithm is time-consuming, with each generation taking around 10 minutes. Also, Lian *et al.* (Lian et al. 2023) propose to equip diffusion models with off-the-shelf pretrained large language models (LLMs) to enhance their prompt reasoning capabilities. But this approach is highly dependent on LLMs, which are hard to control and prohibitively expensive to deploy.

Current layout-based approaches typically split image generation into two completely disrelated stages: prompt-to-

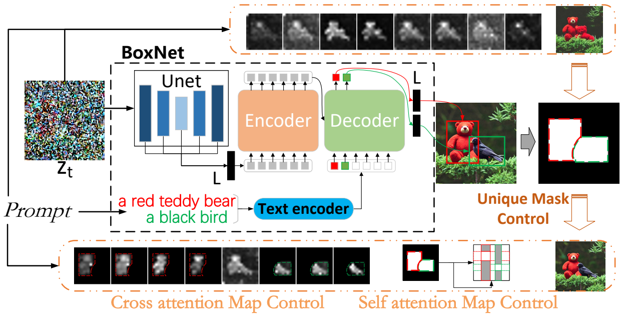

Figure 2: Overview of our BoxNet-based T2I generation pipeline. BoxNet consists of a text encoder and a U-Net followed by an encoder-decoder transformer, as shown in black dashed box. BoxNet takes as input a text prompt, a noisy image, and a timestep and outputs boxes that specify objects' locations. The orange dashed box shows the attention mask control strategy enforced over the cross-attention maps conditioned on the boxes (image regions) and phrases (text spans) as well as the self-attention maps.

layout and layout-to-image, while our method optimizes the diffusion model itself by performing step-wise box prediction and generation control at each sampling step to maintain the original capability of the model while improving the entity properties.

Method

Algorithm 1 shows the overall pipeline of our method, which contains two main parts: BoxNet that predicts a box for each entity with attributes, and attention mask control that ensures the generation of accurate entities and attributes. A single denoising step of our model is illustrated in Fig.2, in which we use BoxNet to predict the bounding box for each entity parsed from the input text and obtain unique masks. We then perform explicit unique mask control over cross- and self-attention maps on each attention layer of the SD (Rombach et al. 2022), which enables to generate entities with their attributes inside the unique mask areas.

The U-Net (Ronneberger, Fischer, and Brox 2015) denoiser contains both cross- and self-attention layers. Each cross-attention layer generates a spatial attention map that indicates the image region to which each textual token is paying attention. Similarly, each self-attention layer produces a spatial attention map that represents the interdependence of each patch and all patches. We assume the aforementioned infidelity defects are caused by the inaccurate cross- and self-attention regions in the U-Net. To alleviate the issues, we enforce an attention mask control strategy over attention maps based on the BoxNet during the diffusion backward process, as shown in Fig.2. In the original SD, attention regions for the entities "bear" and "bird" overlap, with the attention of "bird" being significantly weaker than that of "bear", leading to entity leakage (*i.e.,* generation of two bears). However, after using our method, the prompt "a red teddy bear is sitting next to a black bird" is generated correctly.

Algorithm 1: Denoising Process of Our Method

Input: A text prompt p, a trained BoxNet B, sets of each parsed entity's token indices $\{s_1, s_2, ..., s_N\}$, a trained diffusion model SD
Output: Denoised latent z_0.

```
for $t \leftarrow T, T-1, ..., 1$ do
    $boxes \leftarrow B(SD, z_t, p, t)$
    for $(cx, cy, h, w)$ in boxes do
        Convert box to zero-one masks $m_n$
        $G_n \leftarrow Gaussian\_distribution\_2D((cx, cy), h, w)$
    $M \leftarrow argmax(G_n)$
    $m'_n \leftarrow (M = n) \odot m_n, \quad n = 1, 2..., N$    ▷ unique masks
    $SD' \leftarrow SD$
    for each cross attention layer in $SD'$ do    ▷ cross attention mask control
        Obtain Cross Attention Map $C$
        $C_i \leftarrow C_i \odot m'_n \quad \forall\, i \in s_n, n = 1, 2..., N$
    for each self attention layer in $SD'$ do    ▷ self attention mask control
        Obtain Self Attention Map $S$
        $S_i \leftarrow S_i \odot flatten(m'_n) \quad \forall\, i \in \{i | flatten(m'_n)_i = 1\}, n = 1, 2..., N$
    $z_{t-1} \leftarrow SD'(z_t, p, t)$
```

BoxNet Architecture

Our BoxNet consists of a U-Net feature extractor, a text encoder, and an encoder-decoder transformer (Carion et al. 2020) as shown in Fig.2. When training the BoxNet, the U-Net and the text encoder are initialized and frozen from a pretrained SD checkpoint. At each timestep t of the SD denoising process, the U-Net takes as input a noisy image z_t, a text prompt p, and a timestep t, and then we extract the output feature maps from each down- and up-sampling layer of the U-Net. All the extracted feature maps are interpolated to the same size and concatenated together. A linear transformation is then applied to acquire a feature tensor f that represents the current denoised latent z_t.

After that, we use a standard encoder-decoder transformer to generate entity boxes. Note that the encoder expects a sequence as input; hence, we flatten f to fit the size, refer to (Carion et al. 2020). The decoder decodes boxes with input entity queries. To acquire entity queries, the text prompt input by a user is first parsed into N entities with attributes manually or by an existing text parser (Honnibal et al. 2020). Then, the entity phrases are encoded into embeddings by the text encoder. Entity embeddings are pad with a trainable placeholder tensor into a max length of M, and only the first N of the output sequences are used to calculate entity boxes by a weighted shared linear projection layer.

As to the training phase, we train the BoxNet in the forward process of SD on the COCO dataset. It's worth noting that the primary goal of our BoxNet is to assign each entity a reasonable bounding box during generation steps, which can improve the attention map control to modify entity generation throughout the whole process. We don't concern much

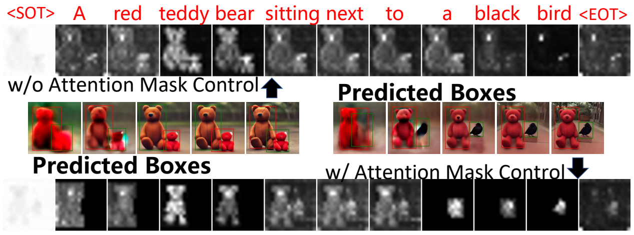
<SOT> A red teddy bear sitting next to a black bird <EOT>
w/o Attention Mask Control
Predicted Boxes
Predicted Boxes
w/ Attention Mask Control

Figure 3: BoxNet predicted box results and corresponding cross-attention maps are presented. We can generate better multi-entity images by controlling the attention map.

about achieving high object recognition accuracy, which differs from DETR. Since one input image may have multiple instance-level ground-truth boxes in the same category, it is necessary to define a proper loss function to constrain our predicted boxes with ground-truth. Inspired by (Carion et al. 2020), we first produce an optimal bipartite matching between predicted and ground-truth boxes, and then we optimize entity box losses. Let us denote by b the ground-truth set of N objects, and b' the set of top N predictions. To find a bipartite matching between these two sets, we search for a permutation of N elements $\sigma \in P_N$ with the lowest cost:

$$\hat{\sigma} = \underset{\sigma \in P_N}{\arg\min} \sum_{i}^{N} \mathcal{L}_{\text{match}}(b_i, b'_{\sigma(i)}), \tag{1}$$

where $\mathcal{L}_{\text{match}}(b_i, b'_{\sigma(i)})$ is a pair-wise matching cost. This optimal assignment is computed efficiently with the Hungarian algorithm, following prior works (Carion et al. 2020; Stewart, Andriluka, and Ng 2016). Different from (Carion et al. 2020), since our BoxNet aims to assign a reasonable bounding box to each object, a precise bounding box with a mismatched category is meaningless. Therefore, we prioritize classification accuracy over location accuracy by modifying the matching cost to include an extremely high penalty for bounding boxes with class mismatches:

$$\mathcal{L}_{\text{match}}(b_i, b'_{\sigma(i)}) = \lambda \cdot \mathbb{1}\{c_i \neq c_{\sigma(i)}\} + \mathcal{L}_{\text{box}}(b_i, b'_{\sigma(i)}) \tag{2}$$

where c_i is the target class label, $c_{\sigma(i)}$ the predicted class label, and $\mathcal{L}_{\text{box}}(\cdot,\cdot)$ the entity box loss described below. We assign λ an extremely high value to avoid class mismatches. The next step is to compute the loss function of BoxNet. We use a linear combination of the $L1$ loss and the generalized IoU loss $\mathcal{L}_{\text{box}}(\cdot,\cdot)$ from (Rezatofighi et al. 2019).

$$\mathcal{L}_{\text{box}}(b_i, b'_{\hat{\sigma}(i)}) = \lambda_{iou}\mathcal{L}_{\text{iou}}(b_i, b'_{\hat{\sigma}(i)}) + \lambda_{L1}\left|b_i - b'_{\hat{\sigma}(i)}\right| \tag{3}$$

where λ_{iou}, λ_{L1} are hyperparameters.

Though the BoxNet is trained on the COCO dataset with finite entity classification, we observe that it can also generalize well to unseen entities beyond the COCO dataset (NON-COCO dataset), which implies that the BoxNet, trained on the COCO dataset, establishes a mapping relationship between entity name embeddings and the SD model's intermediate generation results. In addition, as shown in Fig.3, the prediction results of the BoxNet match the location of entities with attributes generated by the original diffusion model even when infidelity problems occur. This provides us with the possibility to control the interest area of each entity on attention maps through predicted boxes.

Attention Mask Control

Before performing attention mask control, the predicted boxes need to be converted into zero-one masks. However, for those entity boxes with severe overlap, it is hard to limit each entity to its own area of interest, which may degrade the multi-entity controllability. So we introduce a unique mask algorithm that generates unique zero-one masks for attention map control. This ensures that each entity has its own area of interest and does not interfere with each other.

Unique Mask Algorithm. Assume we have predicted entity boxes, and they are converted to zero-one masks m_n, $n = 1, 2, ..., N$. For each entity box (c_x, c_y, w, h), we employ an independent 2-dimensional Gaussian distribution probability function G_n with two variances $\nu_1 = w/2$ and $\nu_2 = h/2$, where c_x, c_y means the center coordinate of the box and w, h means the width and height of the box.

$$G_n(x,y) = \frac{1}{\sqrt{2\pi\nu_1\nu_2}} \exp\left[-\frac{1}{2}\left(\frac{(x-c_x)^2}{\nu_1} + \frac{(y-c_y)^2}{\nu_2}\right)\right] \tag{4}$$

$x = 1, 2, ..., W; y = 1, 2, ..., H$ where W, H represent the spatial width and height of attention maps. Then we can get the max index map M by

$$M(x,y) = \underset{i=1,2,...,N}{\arg\max}\ (G_i(x,y)) \tag{5}$$

The unique attention masks can be further computed with:

$$m'_n(x,y) = \mathbb{1}(M(x,y) = n) \odot m_n(x,y), \quad n = 1, 2..., N \tag{6}$$

Assume we have unique attention masks m'_n with shape (H, W) from Eq. 6, where $n = 1, 2..., N$ indicates the unique mask of the n-th entity.

Cross attention mask control. For cross attention, we get the attention map C by:

$$C = softmax\left(\frac{QK^T}{\sqrt{d}}\right) \tag{7}$$

In cross-attention, Q comes from the flattened feature map of the model (the Unet of Stable Diffusion); assume the shape is (L, C), where $L = H * W$. While K has the shape (K, C), which represents the embedding of the input prompt p, K here is the token number. $\frac{1}{\sqrt{d}}$ is the scaling factor. The self-attention map C has shape (L, K).

For each n-th entity, its token indices in the tokenization of p is s_n. We can aplly cross attention map control on the cross attention map C by:

$$C[:, i] = C[:, i] \odot flatten(m'_n) \quad \forall\, i \in s_n \tag{8}$$

Self attention mask control. As the Self Attention Mask Control method. First of all, we get the self attention map S as the same equation in eq. 7. Differently, in self-attention, both Q and K are from the flattened feature map with shape

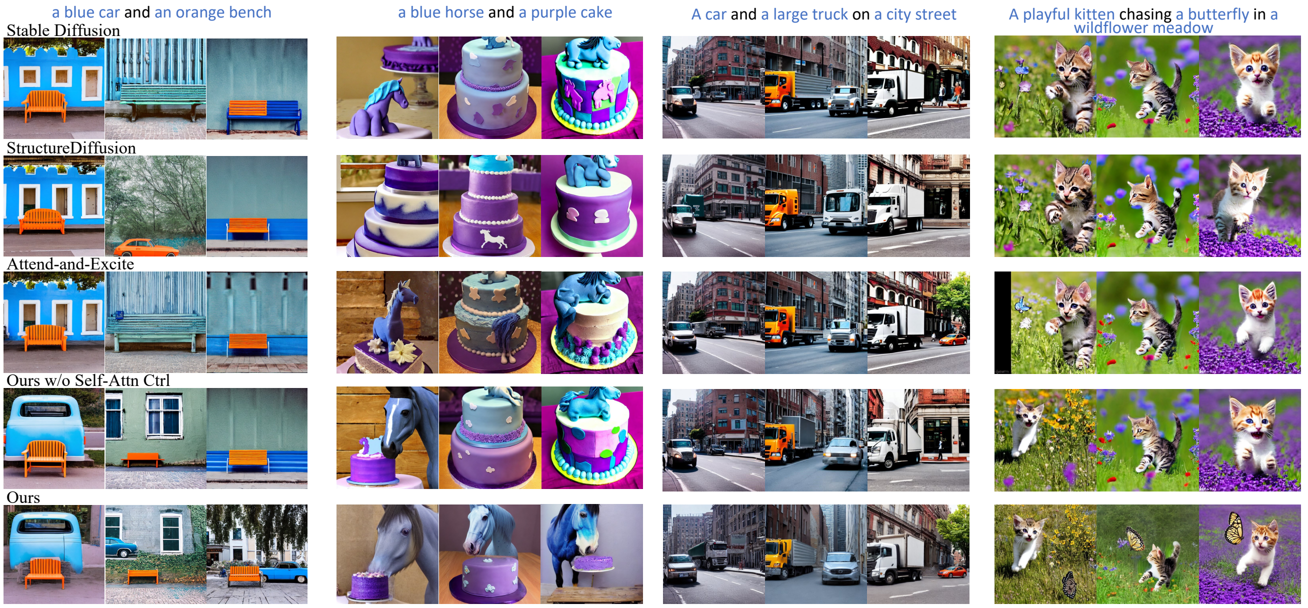
a blue car and an orange bench
a blue horse and a purple cake
A car and a large truck on a city street
A playful kitten chasing a butterfly in a wildflower meadow
Stable Diffusion
StructureDiffusion
Attend-and-Excite
Ours w/o Self-Attn Ctrl
Ours

Figure 4: Qualitative comparison of self-built prompts in fixed format (first three columns) and complex prompts in COCO-style (last two columns) with more than two entities and complex attributes. We display four images generated by each of the five competing methods for each prompt, with fixed random seeds used across all approaches. The entities with attributes are highlighted in blue.

(L, C). And the attention map S has shape $(L*L)$. For each n-th entity, similar to cross-attention control, we need to choose indices of attention map to be masked. Since K represents the feature map itself, we use the unique mask m'_n to seek indices to be masked instead of s_n in cross-attention mask control. As Fig.6 in Appendix A.1 shows, the self-attention map can be controlled by:

$$S[:,i] = S[:,i] \odot flatten(m'_n) \ \ \forall\, i \in \{i | flatten(m'_n)_i = 1\} \tag{9}$$

Plugin Method

Once the BoxNet is trained, our method can act as a plugin to guide the inference process of diffusion-based models on the fly, improving the quality of multi-entity generation with attributes. Our BoxNet can provide input conditions for some layout-based generation models, reducing user input and optimizing the efficiency of large-scale data generation. Furthermore, the attention mask control based on predicted boxes can also be directly applied to other T2I generators to address the three infidelity issues. We introduce two plugin solutions using existing models as examples and compare their results. For more details, refer to Table 2.

AAE (Chefer et al. 2023) guides the latent at each denoising timestep and encourages the model to attend to all subject tokens and strengthen their activations. As a denoising step-level control method, our method can be combined with AAE directly by adding AAE gradient control in our generation algorithm process (both cross- and seld-attention control based on BoxNet in Algorithm 1).

GLIGEN (Li et al. 2023) achieves T2I generation with caption and bounding box condition inputs. Based on GLIGEN, we apply two-stage generation. In the first stage, given the prompt input, we use BoxNet to predict the box for each entity mentioned in the prompt. In the second stage, the predicted entity boxes and captions are fed into the GLIGEN model, and then attention mask control is adopted during generation to obtain layout-based images.

Experiments

Training and Evaluation Setup

All the training details and hyper-parameter determination are presented in Appendix A.2. For evaluation, we construct a new benchmark dataset to evaluate all methods with respect to semantic infidelity issues in T2I synthesis. To test the multi-object attribute binding capability of the T2I model, the input prompts should preferably consist of two or more objects with corresponding attributes (*e.g.,* color). We come up with one unified template for text prompts: “a [colorA][entityA] and a [colorB][entityB]”, where the words in square brackets will be replaced to construct the actual prompts. Note that [entity#] can be replaced by an animal or an object word. We design two sets of optional vocabulary: the COCO category and the NON-COCO category (open domain). Every vocabulary contains 8 animals, 8 object items, and 11 colors, detailed in Appendix A.3. For color-entity pairs in one prompt, we select colors randomly without repetition. For each prompt, we generate 60 images using the same 60 random seeds applied to all methods. For ease of evaluation, our prompts are constructed of color-entity pairs and the conjunction “and”. Yet, our method is not limited to such patterns and can be applied to a variety of prompts with any type of subject, attribute, or conjunction.

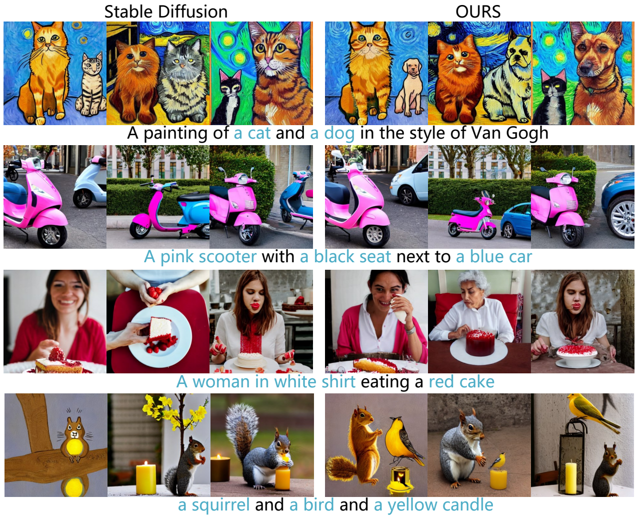
Stable Diffusion
OURS
A painting of a cat and a dog in the style of Van Gogh
A pink scooter with a black seat next to a blue car
A woman in white shirt eating a red cake
a squirrel and a bird and a yellow candle

Figure 5: Comparison with complex prompts of more than two entities or multiple attributes. For each prompt, we apply the same set of random seeds on all methods. The entity-attribute pairs are highlighted in blue.

Qualitative Comparisons

In Fig.4, we present the generated results using fixed format self-built prompts as well as complex ones with more than two entities or intricate attributes (*e.g.,* object actions, spatial relationships), which are taken from the AAE paper (Chefer et al. 2023) and the test split of COCO datset (Lin et al. 2014). For each prompt, we show three images generated by the SD, StructureDiffusion, AAE, *Ours* and *Ours w/o Self-Attn Ctrl*, respectively. *Ours* denotes the method with both cross- and self-attention mask control. As we can see, StructureDiffusion tends to generate images with missing entities and attribute leakage. For example, given "a blue car and an orange bench", its generated images may only contain an orange bench or an orange car that mixes the bench's color with the car's entity. As to AAE, its generated images still suffer from infidelity problems. Given "a blue horse and a purple cake", the AAE correctly generates the two mentioned entities in some cases but fails to bind each entity's color correctly (*e.g.*, generating a purple horse or a white cake). In contrast, our method generates images that faithfully convey the semantics of the original prompt, showing robust attribute binding capability. This is because we explicitly enforce cross- and self-attention mask control over the attention areas to effectively alleviate attribute and entity leakage. For instance, the generated images of *Ours* correctly correspond with the prompt "a blue car and an orange bench", where the colors of the car and bench do not leak or mix. Further more, Fig.5 and the last two columns of Fig.4 show more comparisons with more than two entities and complex backgrounds, demonstrating its effectiveness when dealing with complicated prompts. Additionally, we provide more generation results based on simple or complex prompt descriptions in Appendix C.

Quantitative Analysis

Firstly, we quantify the performance of every competing approach through Average CLIP image-text similarities and Average CLIP text-text similarities from AAE (Chefer et al. 2023). But since global CLIP image-text scores are insensitive to entity missing and attribute leaking issues, we further propose to use Grounding DINO score (Liu et al. 2023a) as a more fine-grained evaluation metric which focus on local object level. However, even Grounding DINO score takes into account issues of entity missing and entity leakage, it's still insensitive to entity attributes so that it does not reflect whether attributes such as color are generated correctly or not. To measure the overall generation performance of both entities and attributes, taking full account of the three infidelity issues, we futher conduct a user study. Additionally, we use FID (Heusel et al. 2017) to assess the overall quality of generated images on 10k samples of the COCO dataset by calculating the distance between feature vectors of generated and real images. All detailed analysis and descriptions of the evaluation metrics (both objective and subjective) are presented and discussed in Appendix B.

DINO Similarity Scores. Grounding DINO is an open-set object detection model, which accepts an image-text pair and predicts object boxes. Each predicted object box has similarity score ranging from 0 to 1 across all input words. We use the DINO score for the most neglected entity as the quantitative measure of multi-entity generation performance. To this end, we compute the DINO score between every entities exist in the original prompt of each generated image. Specifically, given the prompt "a [colorA] [EntityA] and a [colorB] [EntityB]", we extract the names of the entities (*e.g.,* "a [EntityA]" and "a [EntityB]"), and feed them with the generated image into the DINO model to obtain boxes and corresponding similarity scores. If one entity has multiple detected boxes, we adopt the highest similarity score across all boxes as its score. Conversely, if one entity has no detected boxes, we assign a score of 0 to it. Given all the entity scores (two in our case) for each image, we are more concerned with the smallest one as this would correspond to the issues of entity missing and entity leakage. The average of the smallest DINO scores across all seeds and prompts is taken as the final metric of each method, called *Minimum Object Score*.

User Study. We also perform a user study to analyze the fidelity of the generated images. 25 prompts on COCO or NON-COCO datasets are randomly sampled to generate 10 images, while each method shares the same set of random seeds. For the results of each prompt "a [colorA] [EntityA] and a [colorB] [EntityB]", we ask the respondents to answer two questions: (1) "is there a [colorA] [EntityA] in this picture?" and (2) "is there a [colorB] [EntityB] in this picture?". An answer of "YES" indicates both the color and entity can match the given text prompt. Only if the answer to both questions is yes, can this generated image be considered correct. We obtain *Subjective Fidelity Score* by counting the correct proportion of all 25×10 images on COCO or NON-COCO datasets.

Comparison to Prior Work. The quantitative results on the COCO and NON-COCO datasets are summarized in Table 1. We compare our method with three baselines (STA-

No.	Model	Avg CLIP image-text		Avg CLIP	Min. Object Score		Subj. Fidelity Score		FID
		Full Prompt	Min. Object	text-text	COCO	NON-COCO	COCO	NON-COCO	COCO
[1]	STABLE	0.337	0.241	0.778±0.092	0.397±0.002	0.400±0.005	0.302±0.076	0.370±0.100	17.79
	StructureDiffusion				0.373±0.004	0.372±0.004	0.277±0.057	0.302±0.082	-
	AAE	0.338	0.244	0.770±0.097	0.444±0.003	0.434±0.002	0.355±0.104	0.350±0.097	-
[2]	[1]+BoxNet &Cross Attn Mask Ctrl	-	-	-	0.401±0.003	0.431±0.004	-	-	-
[3]	[2]+Uniq Mask	0.334	0.241	0.767±0.099	0.446±0.004	0.478±0.006	0.414±0.109	0.398±0.100	18.11
[4]	[3]**+Self Attn Ctrl(OURS)**	**0.343**	**0.252**	**0.786±0.101**	**0.603±0.005**	**0.599±0.004**	**0.433±0.140**	**0.431±0.121**	**17.47**

Table 1: The quantitative evaluation results of five metrics for the six methods, including three baselines and three ablated variants of our method. Avg CLIP image-text/text-text and Min. Object Score measure multi-entity generation quality based on CLIP and the DINO score, respectively. Subj. Fidelity Score evaluates the correctness of entity and attribute generation through a user study. FID assesses the quality of generated images by measuring the feature distance between generated and real images.

State	COCO		
	STABLE	AAE	GLIGEN
BASE	0.397±0.002	0.444±0.003	0.505±0.002
BoxNet	-	-	0.579±0.001
w/ Cross-Attn Ctrl	0.446±0.004	0.483±0.003	0.620±0.001
w/ Cross&Self-Attn Ctrl	**0.603±0.005**	**0.626±0.006**	**0.672±0.005**
	NON-COCO		
BASE	0.400±0.005	0.434±0.002	0.458±0.006
BoxNet	-	-	0.559±0.002
w/ Cross-Attn Ctrl	0.478±0.006	0.496±0.002	0.633±0.001
w/ Cross&Self-Attn Ctrl	**0.599±0.004**	**0.592±0.003**	**0.684±0.002**

Table 2: Comparison of the Min. Object Scores for the proposed plugin solutions, split by evaluation datasets. The first column indicates different states of methods. We show the performance of the three methods after being plugged with our proposed techniques, respectively.

BLE, AAE, Structure) in terms of five metrics. We have replicated the test dataset and metrics used in (Chefer et al. 2023), which are recorded in Table 1 as the Average CLIP image-text similarities and Average CLIP text-text similarities. Further, the Min. Object Score, Subj. Fidelity Score, and FID distance are calculated for better comparison. As shown, our method consistently outperforms all competing methods, with significant improvements in the fidelity of multi-entity generation and the correctness of attribute bindings between colors and entities. StructureDiffusion obtains scores similar to those of SD (even slightly lower), which is consistent with (Chefer et al. 2023). And AAE gains scores slightly higher than SD. Although trained on the COCO dataset, our method still performs well in the NON-COCO (open-domain) dataset, exhibiting good generalization ability. Additionally, our method achieves a slightly better FID than SD, indicating that the generation quality does not decrease after applying our attention mask control strategy.

Ablation Study. For the ablation study, we start with the original SD model and gradually add constitutive elements until we reach the complete OURS method. Whereas [1] represents the SD model, [2] applies BoxNet and non-uniq cross-attention mask control and can obtain experimental results that are comparable to (slightly better than) those of [1]. [3] applies uniq mask control based on [2], and can achieve similar metric results to AAE. By finally adding the self-attention control, we have the OURS method, marked as [4].Table 1 shows the contribution of different components of our model to the compositional T2I synthesis.

Plugin Experiments

In this section, we verify the effectiveness of our proposed two plugin solutions by comparing the results of existing models (AAE and GLIGEN) with and without our method. The experiment results are shown in Table 2. The first column indicates different states of methods. The **BASE** indicates the original state of each method as described in their papers. Note that in this state, we randomly generate object boxes as additional input conditions for GLIGEN. In the **BoxNet** state, the predicted boxes of BoxNet are used to replace the input random boxes for GLIGEN, while the remaining two states represent the results after imposing our attention mask control strategy on the three methods. As we can see, the generation quality of AAE and GLIGEN is significantly improved after being plugged into our strategy. Both the cross- and self-attention control can alleviate the infidelity issues, while the self-attention control contributes more to the improvement of Min. Object Score. However, in the open-domain NON-COCO evaluation, *AAE w/ Cross- and Self-Attn Ctrl* unexpectedly perform worse than its counterpart in SD. We suspect that this is because the predicted boxes of the BoxNet on the NON-COCO dataset do not overlap with the region of interest in AAE, resulting in a conflict between these two methods. More qualitative results can be found in Appendix C.

Conclusion

In this paper, we present a novel attention mask control strategy based on the proposed BoxNet. We first train a BoxNet to predict object boxes when given the noisy image, timestep and text prompt as input. We then enforce unique mask control over the cross- and self-attention maps based on the predicted boxes, through which we alleviate three common issues in the current Stable Diffusion: attribute leakage, entity leakage, and missing entities. During the whole training process of BoxNet, the parameters of diffusion model are frozen. Our method guides the diffusion inference process on the fly, which means it can be easily incorporated into other existing diffusion-based generators when given a trained BoxNet.

References

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; Catanzaro, B.; et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 213–229. Springer.

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, M.; Laina, I.; and Vedaldi, A. 2023. Training-Free Layout Control with Cross-Attention Guidance. *arXiv preprint arXiv:2304.03373*.

Croitoru, F.-A.; Hondru, V.; Ionescu, R. T.; and Shah, M. 2023. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Doersch, C. 2021. Tutorial on Variational Autoencoders. *stat*, 1050: 3.

Esser, P.; Rombach, R.; and Ommer, B. 2021. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12873–12883.

Feng, W.; He, X.; Fu, T.-J.; Jampani, V.; Akula, A. R.; Narayana, P.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Training-Free Structured Diffusion Guidance for Compositional Text-to-Image Synthesis. In *The Eleventh International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. *arXiv preprint arXiv:2303.11305*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Honnibal, M.; Montani, I.; Van Landeghem, S.; and Boyd, A. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Huk Park, D.; Luo, G.; Toste, C.; Azadi, S.; Liu, X.; Karalashvili, M.; Rohrbach, A.; and Darrell, T. 2022. Shape-Guided Diffusion with Inside-Outside Attention. *arXiv e-prints*, arXiv–2212.

Jiménez, Á. B. 2023. Mixture of Diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22511–22521.

Lian, L.; Li, B.; Yala, A.; and Darrell, T. 2023. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655*.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, N.; Li, S.; Du, Y.; Torralba, A.; and Tenenbaum, J. B. 2022. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, 423–439. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023a. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*.

Liu, Z.; Feng, R.; Zhu, K.; Zhang, Y.; Zheng, K.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023b. Cones: Concept Neurons in Diffusion Models for Customized Generation. *arXiv preprint arXiv:2303.05125*.

Ma, W.-D. K.; Lewis, J.; Kleijn, W. B.; and Leung, T. 2023. Directed Diffusion: Direct Control of Object Placement through Attention Guidance. *arXiv preprint arXiv:2302.13153*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241. Springer.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Stewart, R.; Andriluka, M.; and Ng, A. Y. 2016. End-to-end people detection in crowded scenes. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, 2325–2333.

Wu, Q.; Liu, Y.; Zhao, H.; Bui, T.; Lin, Z.; Zhang, Y.; and Chang, S. 2023. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7766–7776.

Yang, C.-F.; Fan, W.-C.; Yang, F.-E.; and Wang, Y.-C. F. 2021. Layouttransformer: Scene layout generation with conceptual and spatial diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3732–3741.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.