Easing Concept Bleeding in Diffusion via Entity Localization and Anchoring

Jiewei Zhang [1 2] **Song Guo** [3] **Peiran Dong** [1] **Jie Zhang** [1] **Ziming Liu** [1] **Yue Yu** [2] **Xiao-Ming Wu** [1]

Abstract

Recent diffusion models have manifested extraordinary capabilities in generating high-quality, diverse, and innovative images guided by textual prompts. Nevertheless, these state-of-the-art models may encounter the challenge of concept bleeding when generating images with multiple entities or attributes in the prompt, leading to the unanticipated merging or overlapping of distinct objects in the synthesized result. The current work exploits auxiliary networks to produce mask-constrained regions for entities, necessitating the training of an object detection network. In this paper, we investigate the bleeding reason and find that the cross-attention map associated with a specific entity or attribute tends to extend beyond its intended focus, encompassing the background or other unrelated objects and thereby acting as the primary source of concept bleeding. Motivated by this, we propose Entity Localization and Anchoring (ELA) to drive the entity to concentrate on the expected region accurately during inference, eliminating the necessity for training. Specifically, we initially identify the region corresponding to each entity and subsequently employ a tailored loss function to anchor entities within their designated positioning areas. Extensive experiments demonstrate its superior capability in precisely generating multiple objects as specified in the textual prompts.

1. Introduction

In recent years, text-to-image diffusion models (Nichol et al., 2021; Saharia et al., 2022; Ramesh et al., 2022; Rombach et al., 2022; Podell et al., 2023) have experienced remarkable advancements, capturing widespread public attention for their ability to generate diverse and realistic images. Users can engage natural language to govern the content of the generated images. Nonetheless, achieving a perfect match between the synthesized images and the user-provided text prompt is challenging. Simultaneously, during the generation of complex scenes comprising multiple objects, two concepts may encounter mutual penetration (Rombach et al., 2022; Mou et al., 2023; Podell et al., 2023).

Figure 1. The visual features of cats and frogs, as well as elephants and pandas, appear fused in the left two columns of images generated by Stable Diffusion and Stable Diffusion XL, a phenomenon commonly known as *Concept Bleeding* or *Entity Leakage*. Notably, the bottom-left image lacks a panda, namely missing entities. Our method can alleviate these two issues.

Envision a scenario where a user seeks to produce multiple objects within a scene, like a cat and a frog, entities that would be challenging to coexist simultaneously in reality. Text-to-image synthesis diffusion models facilitate the creation of these imaginative compositions. Nevertheless, achieving meticulous object generation remains a formidable task. Even with a simple prompt like "an elephant and a panda", the consequence may omit certain objects, or the visual features of the two entities might overlap and intertwine (see Figure 1). In this paper, we denote these two occurrences as "*missing entities*" and "*entity leakage*" or "*concept bleeding*" respectively. Some inherent issues in stable diffusion might be the cause of these challenges. Specifically, P2P (Hertz et al., 2022) affirms that the cross-attention map is intricately connected to the structural layout of the synthesized image. The overlap or merging of these maps associated with distinct entities may lead to the fusion of entities in the generated images. Additionally, the presence of causal attention masks introduces a blending of the semantics between backward and forward tokens in a CLIP (Radford et al., 2021) text embedding sequence, ultimately causing concept bleeding for diffusion models utilizing CLIP text embedding for guidance.

Some attempts have been made to tackle these challenges.

[1]The Hong Kong Polytechnic University. [2]Peng Cheng Laboratory. [3]The Hong Kong University of Science and Technology.. Correspondence to: Song Guo <songguo@cse.ust.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

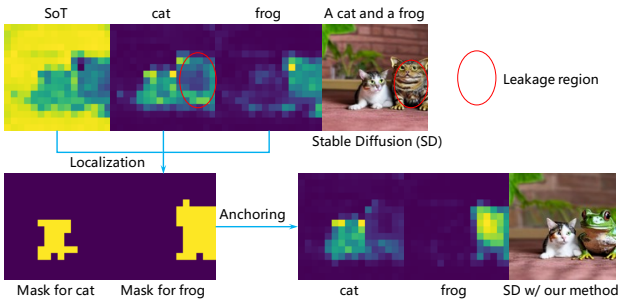

Figure 2. We display synthesized images alongside their corresponding cross-attention maps extracted from U-net. The first line reveals cross-attention map leakage, exemplified by the token "cat", leading to the unintended overlap of a cat and a frog. "SoT" designates the start of the prompt. The second line illustrates our technique, where we identify the entity's location and confine the respective cross-attention maps to localized regions.

DPL (Wang et al., 2023a) notes that the cross-attention map corresponding to an object often leaks into other regions, encompassing the background and adjacent objects (see Figure 2). DPL aims to ease entity leakage by learning embeddings related to entities in the prompt. However, it is primarily designed for image editing and may not be fully suitable for generation tasks. Additionally, the cross-attention maps related to the attributes may not entirely focus on the intended region, resulting in attribute binding errors. StructureDiffusion (Feng et al., 2023) neglected the influence of the cross-attention map and solely focused on addressing the binding errors through language structure guidance. Linguistic Binding (Rassin et al., 2023) emphasizes the alignment of cross-attention maps, yielding better outcomes on attribute binding. Attend-and-excite (Chefer et al., 2023) addresses missing entities by amplifying the activation value of the ignored object's cross-attention map. However, they disregard the mutual impact between entities.

We have discussed three categories of generation defects: *missing entities*, the omission of one or more objects; *entity leakage*, the merging of visual elements from disparate entities; and *attribute binding errors*, the unintended association of specified properties with incorrect objects. Concept Bleeding (Podell et al., 2023) occurs when distinct visual elements unintentionally merge or overlap, summarizing both entity leakage and attribute binding errors. Cross-attention leakage potentially serves as the primary cause of concept bleeding, with missing entities primarily stemming from the overlooked objects having relatively low activation values in their cross-attention map (Chefer et al., 2023). To address the issue of concept bleeding, we introduce Entity Localization and Anchoring (ELA), a dedicated approach aimed at alleviating cross-attention leakage (see Figure 2). We initially determine the object's position by analyzing the cross-attention map difference between the two objects. Subsequently, we engage the cross-attention map of SoT with rich semantic information to eliminate the background. Then, a loss function is formulated to slightly adjust the latent during each denoising step, aiming to reinforce the cross-attention map within the localization region while attenuating it outside the designated area. Considering the fact that the cross-attention map undergoes minimal alteration in the advanced stages of denoising (Balaji et al., 2022; Mou et al., 2023), the aforementioned corrections are solely implemented at the early sampling stage. While our primary focus lies on addressing entity leakage, we aim to highlight activation within the localized area, which is a measure that can partially alleviate missing entities. The contributions of our paper are summarized as follows:

- We present a novel Entity Localization and Anchoring (ELA) approach to tackle cross-attention leakage. To the best of our knowledge, this marks the first training-free method designed to ease entity leakage.
- Our approach not only addresses concerns related to entity leakage and attribute binding errors but also instills confidence in the faithful representation of each object in the generated image.
- Experimental results illustrate that our approach excels in accurately generating multiple objects. Additionally, a series of experiments were conducted to analyze the factors contributing to entity leakage.

2. Related work

Text-to-Image Diffusion Model. Early unconditional Diffusion Models (Ho et al., 2020; Song et al., 2021) exhibit the capacity to produce high-quality images without resorting to adversarial training. (Dhariwal & Nichol, 2021) utilizes classifiers for conditional guidance in diffusion models, while (Ho & Salimans, 2022) showcases effective guidance for diffusion models via cross-attention mechanisms without classifiers. Other approaches, such as GLIDE (Nichol et al., 2021), DALL·E 2 (Ramesh et al., 2022), and Imagen (Saharia et al., 2022), achieve photorealistic image generation through text guidance. Stable Diffusion (SD) (Rombach et al., 2022) conducts diffusion and denoising in the latent space of a robust pre-trained autoencoder for efficient training. Stable Diffusion XL (SDXL) (Podell et al., 2023) integrates resolution and cropping as supplementary conditions. SDXL enhances the overall image quality through multi-aspect training and stacked diffusion model modules. Despite their impressive generation capabilities, accurately capturing the text prompt's semantics in the generated images remains a challenge. Additionally, SD and SDXL may also encounter issues with concept bleeding. Some personalized models strive for precise control over the generation to match prompts, providing a slight relief of these issues.

Personalized Diffusion Model. In pursuit of fine-grained control over the generated images, T2I-adapter and ControlNet (Mou et al., 2023; Zhang et al., 2023) introduce support for various conditions, such as segmentation, sketches, and keyposes. They leverage auxiliary networks to effectively align control signals with implicit information embedded in the diffusion model. Despite facilitating the accurate generation of each entity without overlap via segmentation or sketches guidance, they neglect the impact of cross-attention leakage. In addition, generating user-specific objects poses a significant challenge (Gal et al., 2023; Ruiz et al., 2023). According to (Kumari et al., 2023), this specific generation can be achieved by fine-tuning only a subset of parameters associated with the special identifier linked to the specific subject. (Ma et al., 2023) transforms referenced images into pseudo-words to facilitate subject-specific generation. While they also can combine specific objects, issues of concept bleeding inherited from Stable Diffusion persist.

Specialized Models for Easing Concept Bleeding. In recent years, numerous endeavors have sought to address the issue of concept bleeding. StructureDiffusion (Feng et al., 2023) utilizes linguistic structure guidance to ensure that objects align with specified attributes. StructureDiffusion breaks down the prompt into noun phrases, encoding each one individually to replace the corresponding original embeddings. However, the outcomes produced by StructureDiffusion often parallel those of Stable Diffusion, and the challenge of concept bleeding sustains, primarily attributed to overlooking the influence of an inaccurate cross-attention map. Our strategy updates the latent to rectify the cross-attention map through a specifically designed loss function. (Wang et al., 2023b) incorporates the auxiliary network BoxNet to acquire object masks, which are then employed to constrain the cross-attention map. However, training the auxiliary network is a time-consuming process. Attend-and-excite (Chefer et al., 2023) strengthen activation of the most neglected objects in the corresponding cross-attention map, overlooking the issue of cross-attention leakage. Entity leakage can also occur in some text-to-image editing works (Hertz et al., 2022; Tumanyan et al., 2023; Brooks et al., 2023; Parmar et al., 2023). DPL (Wang et al., 2023a) amends the cross-attention map by updating the corresponding embeddings of entities in the prompt. DPL requires the preservation of an embedding at each sampling step to prevent the merging of visual features from different objects, making it impractical for generation tasks. Our method aims to ease concept bleeding in the generation.

3. Preliminaries

Stable Diffusion. We implement our approach on Stable Diffusion (SD) (Rombach et al., 2022), a two-stage image synthesis model. Initially, the given image $x \in \mathbb{R}^{H\times W\times 3}$ is mapped into a latent code $z_0 = \mathcal{E}(x)$ through an encoder $\mathcal{E}$, where $z_0 \in \mathbb{R}^{\frac{H}{f}\times\frac{W}{f}\times c}$ and f is downsampling factor. A decoder $\mathcal{D}$ is employed to reconstruct the image from the latent code. The autoencoder, which comprises the encoder $\mathcal{E}$ and the decoder $\mathcal{D}$, is trained to ensure $\mathcal{D}(\mathcal{E}(x)) \approx x$. Diffusion and denoising operations are executed in the latent space learned by the autoencoder.

More precisely, during the diffusion process, Gaussian noise is progressively injected into the latent code until it approaches the standard normal distribution, denoted as $z_T \in \mathcal{N}(0, \mathbf{I})$. At timestep t, the perturbed latent code can be defined as $z_t = \sqrt{\overline{\alpha}_t} z_0 + \sqrt{1-\overline{\alpha}_t}\epsilon$, where $\overline{\alpha}_t$ is a scheduled hyperparameter and ϵ denotes the noise. Denoising is achieved by training neural networks to predict the noise added to the latent variable at every timestep and subsequently removing the noise gradually from z_T. This process is essentially the inverse of the diffusion process. Specifically, a U-Net (Ronneberger et al., 2015) architecture is employed to predict the noise, where U-Net can be conditioned on a text embedding. U-net is optimized by minimizing the loss function specified in Eq. (1):

$$\mathcal{L}_{SD} = \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\epsilon - \epsilon_\theta(z_t, c(y), t)\|_2^2\right], \quad (1)$$

where c denotes a fixed CLIP text encoder (Radford et al., 2021). This encoder translates the text description y into an embedding integrated into the cross-attention layers.

After training the autoencoder and U-Net, we can initiate image synthesis from Gaussian noise z_T. More precisely, we exploit either DDPM, DDIM, or PLMs (Ho et al., 2020; Song et al., 2021; Liu et al., 2022) to progressively denoise from z_T to acquire the latent representation z_0. Subsequently, this latent representation is decoded into a synthesized image using the decoder $\mathcal{D}$. The semantic information from text conditions is seamlessly incorporated into the predicted noise $\epsilon_\theta(z_t, c(y), t)$ through cross-attention.

Text Guidance via Cross-Attention. U-net comprises downsampling blocks, a middle block, and upsampling blocks, with semantic information from text embedding incorporated into the cross-attention layers across all blocks. Cross-attention is visually depicted in Figure 3.

We project the intermediate features $\varphi(z_t)$ from the U-net to the query $Q \in \mathbb{R}^{n\times hw\times d}$ using Eq. (2):

$$Q = f_Q(\varphi(z_t)) = W_Q^i \cdot \varphi(z_t). \quad (2)$$

The key $K \in \mathbb{R}^{n\times l\times d}$ and value $V \in \mathbb{R}^{n\times l\times d}$ are derived from the text embedding $c(y)$ through $f_K(c(y)) = W_K^{(i)} \cdot c(y)$ and $f_V(c(y)) = W_V^{(i)} \cdot c(y)$. Here, $W_Q^{(i)}$, $W_K^{(i)}$, and $W_V^{(i)}$ are learnable projection matrices from the cross-attention layer i. The parameters n, d, l, and t represent the number of attention heads, the feature dimension, the number of text tokens, and the timestep, respectively.

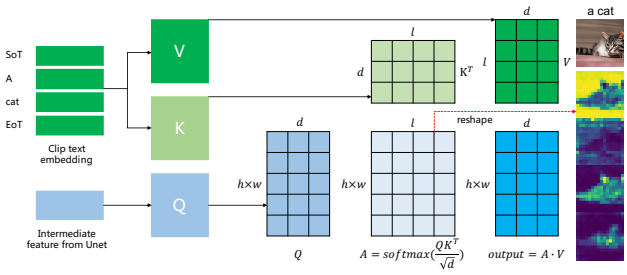

Figure 3. Overview of cross-attention. The query (Q) is projected from the intermediate features of the U-net, while the key (K) and value (V) are derived from the text embedding. Q is utilized to retrieve semantics that align with image features, accomplished through the dot product between Q and K. Following the application of the softmax function to the dot product results, a cross-attention map is generated, depicting the distribution of semantics across individual pixels in the feature map. The injection of textual semantics into image features is achieved by the dot product between the attention map (A) and the value (V). We employ grids to illustrate the attention mechanism. Each column in A corresponds to the cross-attention map of the respective token.

We compute the distribution of each token across pixels using Eq. (3):

$$A_t = softmax\left(\frac{QK^T}{\sqrt{d}}\right), \tag{3}$$

where the attention map A_t is multiplied with V to effectively incorporate semantic information from text prompts into image features, completing the process of text guidance.

4. Our Method

In this section, we introduce our propose Entity Localization and Anchoring (ELA) method to relieve cross-attention leakage for the accurate generation of all objects. Section 3 has provided a concise overview of Stable Diffusion and the text guidance mechanisms through cross-attention. In Section 4, we delve into a detailed presentation of ELA.

At the essence of our approach lies the anchoring of the entity's cross-attention map to a specific area without overlapping or merging. During the denoising step t with prompt y, as we predict noise, cross-attention maps are extracted to roughly determine the entity's position. By making subtle adjustments to the latent, we effectively confine the entity to diverse regions (see Figure 4). In the upcoming sections, we will delve into the extraction of the cross-attention map, entity localization, entity anchoring, and further elucidate additional details of the algorithm.

Extraction of Cross-Attention Maps. The spatial distribution of entities within generated images is intimately connected to cross-attention maps. Our attention is specifically on these maps, which delineate how each entity is distributed across the feature map. As depicted in Figure 3, the cross-attention maps denoted as A_t capture the distribution of all tokens in the given prompt. A_t results from the interplay between image features and text embeddings.

Applying a pre-trained Stable Diffusion model, we perform denoising on Gaussian noise z_T with the given prompt y. In the process of predicting noise at inference timestep t, we extract intermediate features $\varphi(z_t)$ from the upsampling blocks of the U-net. Initially chaotic, $\varphi(z_t)$ refines gradually as denoising progresses, leading to clearer features. Hence, we attempt to employ cross-attention maps following several rounds of sampling for localization and anchoring. The CLIP text encoder inserts special tokens, SoT and EoT, to mark the beginning and end of the text, respectively. We observe that the cross-attention maps for SoT and EoT are enriched with semantic information, especially SoT, which captures extensive background details. Our focus is on the entity, and it is essential to eliminate the influence of filling.

After a few denoising iterations, the spatial arrangement in the cross-attention map becomes marginally clearer. We derive A_t by averaging 16×16 cross-attention maps within the upsampling blocks. l_y and l_e signify the length of the prompt and the text embedding, separately. In the cross-attention map $A_t[0, 1, ..., l_y, ..., l_e - 1]$, serial numbers 0 and 1 to l_y corresponds to SoT and each word in the prompt, respectively, while the rest represent EoT. For simplicity, we use A_t^e and A_t^0 to denote the maps of entities and SoT, respectively, where $e \in E = \{e_1, ..., e_n\}$ and E is the entity set included in the prompt. Entities frequently extend into the background region, emphasizing the need to extract SoT corresponding maps to effectively eliminate this leakage.

Entity Localization. For simplicity, we let $E = \{e1, e2\}$. Commencing denoising from Gaussian noise, the initial prediction noise generates a cross-attention map that poses challenges in extracting meaningful spatial information. As denoising advances, the implicit spatial information becomes increasingly distinct. Nevertheless, the spatial layout of entities often extends beyond the expected area, reaching into the background or overlapping with other entities.

To position entities and avoid overlap, minimizing the impact of leakage is crucial. Successfully generated entities will manifest as prominently highlighted regions in the cross-attention map. For $A_t^{e_1}$, it is imperative to eliminate potentially overlapping components, including the background and entity e_2. The background mask can be obtained by binarizing the attention map of SoT using the equation $M_{bk} = B(A_t^0)$. Subsequently, we mitigate overlap between entities by calculating the difference between $A_t^{e_1}$ and $A_t^{e_2}$, with the binarized difference accentuating entity e_1. Following that, removing the background to identify the mask of

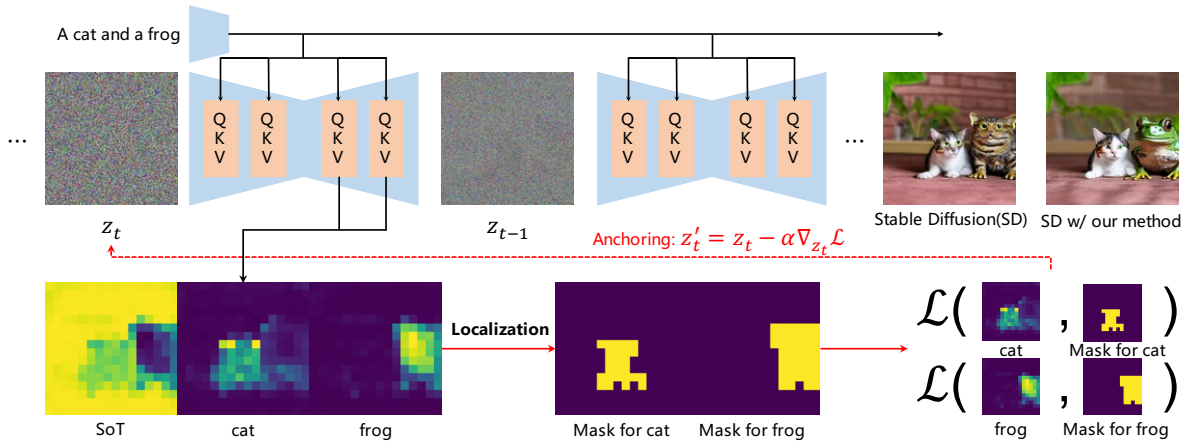

Figure 4. Overview of our proposed Entity Localization and Anchoring method. Utilizing the prompt "a cat and a frog", we extract cross-attention maps (A_t^0, A_t^2, A_t^5) for the Start of Text (SoT) and entities (cat, frog). Entity positions are estimated by analyzing the disparity between A_t^2 and A_t^5, followed by background removal to obtain entity masks. We minimize a tailored loss by slightly adjusting the latent z_t to constrain A_t^2 and A_t^5 within the masks. The cross-attention mapping of the End of Text (EoT) corresponds to the foreground in the generated image, while the cross-attention map of SoT A_t^0 contains significant semantic information about the background.

entity e_1 becomes straightforward through Eq. (4):

$$M_t^{e_1} = B(D_{12} - M_{bk}), D_{12} = B(A_t^{e_1} - A_t^{e_2}), \quad (4)$$

where B represents the binary function.

Entity Anchoring. Once the entities have been allocated to various regions, it is essential to firmly associate them with their respective designated positioning regions. We will bifurcate the anchoring process into two distinct parts: concentration and attenuation. Concentration involves directing the distribution of entities onto the designated mask region within the cross-attention map, while attenuation aims to diminish the distribution beyond the mask.

Instinctively, the maximization of $\frac{\sum M_t^e A_t^e}{\sum A_t^e}$ seeks to concentrate the distribution predominantly within the localized area, where $\sum A_t^e$ represents the sum of pixel activations in matrix A_t^e. Our optimization objective is expressed by Eq. (5):

$$\mathcal{L}_c = \frac{1}{n} \sum_{e \in E} (1 - \frac{\sum M_t^e A_t^e}{\sum A_t^e})^2. \quad (5)$$

By combining Eq. (2), (3), and (4), we can formulate this objective as a function of the latent z_t, denoted as $\mathcal{L}_c = f_c(z_t)$. However, this concentration might result in greater activation values outside the designated mask area.

To further mitigate overlap, the activation values beyond the mask region can be decreased by minimizing Eq. (6):

$$\mathcal{L}_a = \frac{1}{n} \sum_{e \in E} \|(1 - M_t^e) A_t^e\|_2^2. \quad (6)$$

However, exclusively minimizing $\mathcal{L}_a$ might cause a decrease in the activation value across the entire attention map, potentially resulting in a failure to highlight the entities within the mask region. The optimization of $\mathcal{L}_c$ is instrumental in emphasizing the presence of entities within the mask region. Combining these two optimization objectives is achieved through Eq. (7):

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_c + \lambda \mathcal{L}_a, \quad (7)$$

where λ serves as the weighting factor.

Integrating Eq. (2), (3), (4), (5), and (6), the optimization objective is expressed as $\mathcal{L} = f_L(z_t, y)$. Minimizing $\mathcal{L}$ involves making subtle adjustments to the latent variable z_t, directing the cross-attention map of the entity to align with the expected area. Additionally, optimizing the objective by making amendments to the text embedding of prompt y can alter the semantic guidance, potentially causing the generated result to deviate from the user-specified prompt. Furthermore excessive modifications to z_t may result in unconventional generation outcomes. Hence, we implement updates gradually over a few steps using Eq. (8):

$$z'_t = z_t - \nabla_{z_t} \mathcal{L}. \quad (8)$$

We outline our method in Algorithm 1. During the later stage of sampling, the cross-attention map undergoes marginal changes, whereas, in the early stages, meaningful layout information is not significantly present (see Figure 6). Hence, we configure the start and end timesteps (T_{start}, T_{end}) to establish meaningful constraints on entities. We will choose specific timesteps $T_s = \{T_1, ..., T_k\}$. Specifically, when $t \in T_s$, we will iteratively execute Steps 4 to 11 in algorithm 1 until the loss reaches the threshold. However, at timestep t, we exclusively employ the initially calculated mask to ensure anchoring stability.

Algorithm 1 Entity Localization and Anchoring

Input: A text prompt y, a trained stable diffusion model SD, a set of entities indices $E = \{e_1, ..., e_n\}$, a trained decoder $\mathcal{D}$, the start timestep T_{start}, the end timestep T_{end}.
Output: synthesized image x

Let $z_T \sim \mathcal{N}(0, \mathbf{I})$.
for $t = T$ **to** 1 **do**
  if $T_{start} \geq t \geq T_{end}$ **then**
    $A_t \leftarrow SD(z_t, y, t)$
    $A_t \leftarrow average(A_t)$
    Background mask $M_{bk} \leftarrow B(A_t^0)$
    while $e \in E$ **do**
      Get a mask M_t^e for entity e using Eq. (4)
    end while
    Compute the loss $\mathcal{L}$ using Eq. (5), (6), (7)
    $z_t \leftarrow z_t - \nabla_{z_t}\mathcal{L}$
  end if
  $z_{t-1} \leftarrow SD(z_t, y, t)$
end for
return $\mathcal{D}(z_0)$

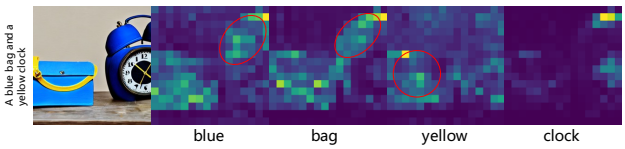

Figure 5. Not only will the cross-attention map corresponding to the entity leak into unexpected areas, but the cross-attention map corresponding to the attribute will also leak, leading to attribute binding errors.

Attribute binding. Figure 5 illustrates the problem of attribute binding errors in stable diffusion, caused by the leakage of the cross-attention map corresponding to the attribute, similar to the issue observed with entities. Our method not only alleviates entity leakage but also addresses attribute binding errors to some extent. We can constrain attributes to specific localization areas to ensure their correct pairing with entities. Considering attribute binding, in Equation (5) and Equation (6), the attribute's corresponding part needs to be added to set E, and the mask should match the entity corresponding to the attribute. In the experimental section, we provide more discussions on attribute binding.

5. Experiments

In this section, we will conduct a thorough qualitative and quantitative comparison of our method with existing approaches. We will also delve into the specific roles of each component of our algorithm in Section 4. Additionally, we will present an in-depth introduction to the implementation, evaluation metrics, and datasets.

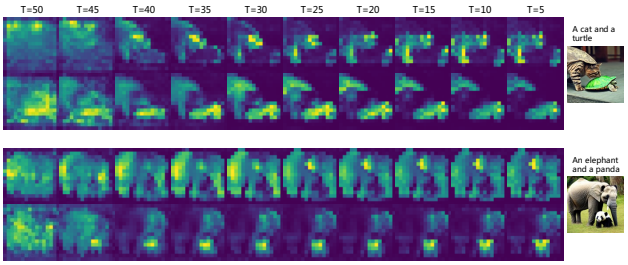

Figure 6. We present the generation results of Stable Diffusion V-1.4 and the cross-attention maps of entities at different denoising stages. These maps, extracted from the upsampling block in U-net with a 16×16 resolution, contain abundant spatial layout information. In the early denoising stages, these maps lack meaningful spatial details, while as denoising progresses, they show minimal alterations. Cross-attention leakage can be observed.

5.1. Experimental Setup

Implementation. Our algorithm is employed within the pre-trained stable diffusion V-1.4. Specifically, we concentrate on cross-attention maps associated with entities mentioned in the prompt. These maps are primarily extracted in the upsampling block with a resolution of 16×16. Additionally, we have the flexibility to upsample or downsample maps of various resolutions to the specified resolution. We implement Algorithm 1 during inference.

Evaluation metrics and datasets. We primarily employ text-image similarity and text-text similarity as evaluation metrics. The CLIP cosine similarity assesses the faithfulness between generated images and prompts. Furthermore, we utilize a pre-trained BLIP (Li et al., 2022) to generate captions for synthesized results, allowing us to measure text-text similarity. We predominantly consider two types of text prompts: 1. "a [entity A] and a [entity B]", 2. "a [attribute A] [entity A] and a [attribute B] [entity B]". For the first type, our attention is drawn to the problem of entity leakage, and we examine three pairs: animal-animal, animal-object, and object-object. In the second type, our concern revolves around the challenge of attribute binding. Additionally, in both cases, there is the potential issue of missing entities. We generate 64 images for each prompt.

5.2. Qualitative Comparisons

We present the qualitative performance of our method on entity leakage in Figure 7. Given prompts structured as "a [entity A] and a [entity B]", Stable Diffusion V-1.4 (Rombach et al., 2022) typically produces images where entities overlap, merge, or one of the entities is ignored. In the case of "a turtle and a clock", the generated image either combines the turtle and the clock or overlooks the clock, akin to the outcome of "a cat and a frog", where the visual characteristics of frogs and cats blend. StructureDiffusion

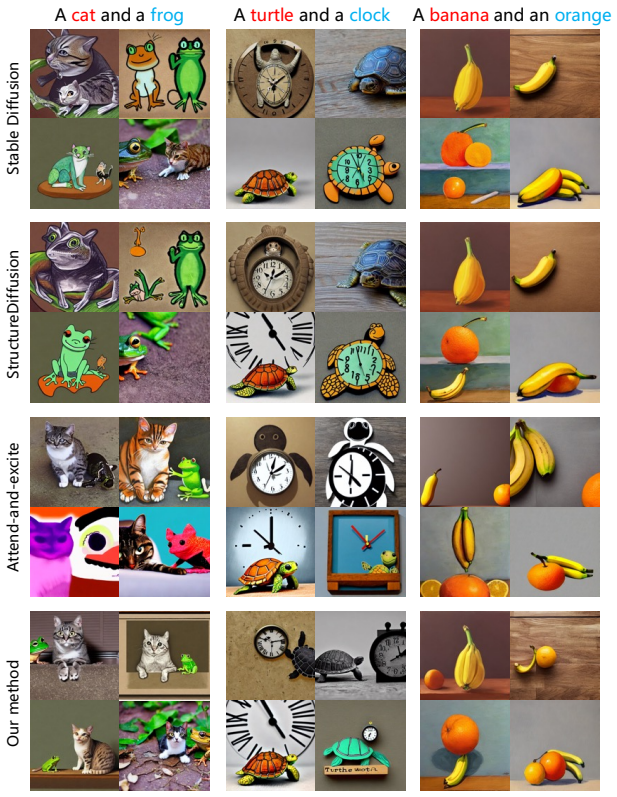

Figure 7. We provide a qualitative comparison of our approach with StructureDiffusion (Feng et al., 2023) and Attend-and-Excite (Chefer et al., 2023). We present generated images separately for animal-animal, animal-object, and object-object pairs. We employ a consistent seed across all approaches and showcase four images for each prompt. The first two columns showcase our approach's capability to relieve entity leakage, while the third column verifies its effectiveness in mitigating missing entities.

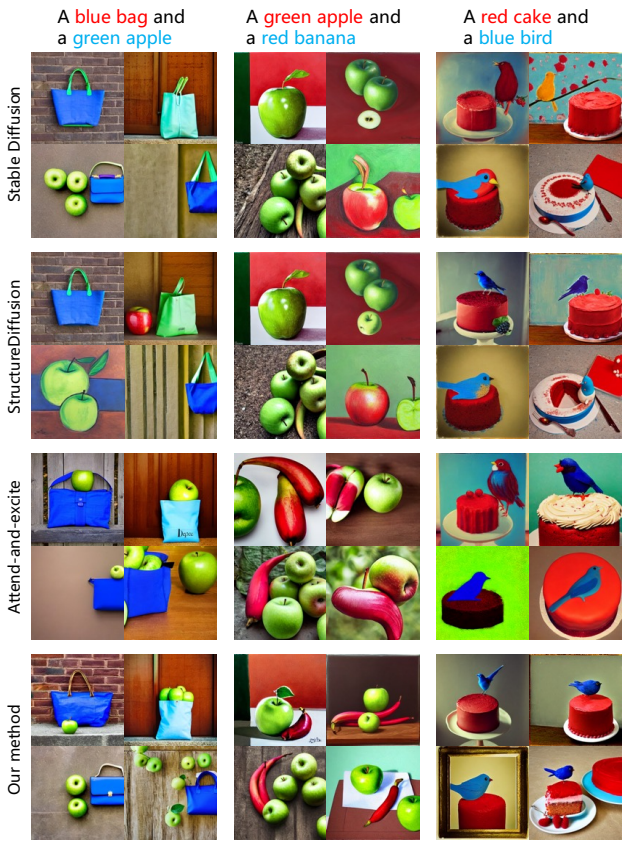

Figure 8. Here, we conduct an additional qualitative comparison, focusing on prompts of the form "a [color A] [entity A] and a [color B] [entity B]". Utilizing a consistent seed, we generate four images for each prompt. Our method effectively eases both missing entities and attribute binding errors.

(Feng et al., 2023) yields comparable results to the former, lacking significant corrections to the cross-attention maps and consequently failing to rectify the semantic errors. Attend-and-excite (Chefer et al., 2023) excels in generating all entities (See the third column in Figure 7); however, it struggles with entity leakage, as illustrated by instances such as the clock being embedded in the turtle shell. Our method successfully disentangles interwoven entities, as seen in the case where a mixture of turtles and clocks is effectively separated into independent entities. As well this approach promotes the generation of all objects.

In Figure 8, we focus on attribute binding. Stable Diffusion encounters difficulties in accurately connecting attributes to specific objects. In the prompt "a blue bag and a green apple", the color "green" is mistakenly linked to the bag's strap, and the apples are mostly disregarded. In the other two prompts, colors are incorrectly associated with unintended regions. StructureDiffusion achieved accurate color binding only for the cake and bird; however, in the other two prompts, incorrect semantics inherited from Stable Diffusion persist. Attend-and-Excite, while able to associate the correct attributes with entities (refer to the first column in Figure 8), faces challenges in handling interactions between entities (see the second column of Figure 8). Our method, in contrast, does not encounter this issue.

Overall, our method successfully tackles the generation defects mentioned in Section 1. However, StructureDiffusion neglects cross-attention leakage and struggles to effectively rectify both semantic errors, providing only limited relief for attribute binding errors. Attend-and-excite, on the other hand, encounters difficulties in addressing entity leakage. In the appendix, we present additional qualitative results, where the prompt is no longer limited to the single forms of "a [entity A] and a [entity B]".

5.3. Quantitative Comparisons

The CLIP text-image similarity can partially indicate the faithfulness between the generated results and the prompts. We compute the average this similarity across all prompts and random seeds, but this metric only captures the simi-

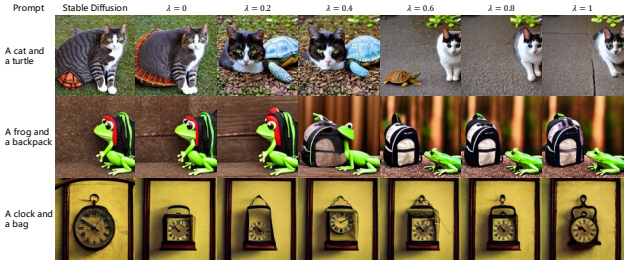

Figure 9. In the qualitative comparison of different λ values in our anchoring loss function, we increment λ from left to right while maintaining the same prompt and random seed. As λ increases, we observe a gradual improvement in accurately generating each entity. However, excessive λ values can result in missing entities.

larity between images and texts in their joint feature space. Consequently, there may be a high similarity score for images with entity leakage, as they contain the semantics of each entity in the feature space. Similar to Attend-and-excite (Chefer et al., 2023), we decompose the full prompt into sub-prompts, each containing a single entity. Subsequently, we assess the similarity between the clauses and the image to verify the accurate generation of each entity. The smallest one, denoted as "Min. Entity", reflects the performance. Our method demonstrated superior performance in "Full Prompt" evaluations (refer to Table 1). Parallel to the qualitative analysis, StructureDiffusion exhibits performance akin to Stable Diffusion, while Attend-and-excite demonstrates slightly better results. This is attributed to the fact that the former is similar to Stable Diffusion regarding cross-attention maps, while the latter focuses exclusively on amplifying neglected objects.

To further evaluate the precision of the generated results, we utilize BLIP to generate a caption for the synthesized image. Subsequently, we use text-text similarity to gauge the faithfulness of the image to the user-specified prompt. If each entity in the image is accurately depicted, the caption generated by BLIP should encompass all entities. Images that encounter entity leakage or missing entities will have lower scores. Our examination specifically targets three types of entity pairs: animal-animal, animal-object, and object-object. Combinations that are challenging to observe in reality are particularly prone to entity leakage or missing entities. Our method achieved the best results in all pairs.

5.4. Ablation Study

We will explore the ablation study from two aspects. Firstly, we will analyze the roles of component concentration and attenuation in the loss $\mathcal{L}$ by adjusting the parameter λ. Secondly, we investigated the effects of different cross-attention layers on Entity Localization and Anchoring.

In Figure 9, we qualitatively analyze the influence of the weighting factor λ. When $\lambda = 0$, the anchoring loss $\mathcal{L}$ reduces to its concentration part $\mathcal{L}_c$, which primarily emphasizes each entity in the mask region. However, handling overlapping or merging between entities becomes challenging, as shown in Figure 9 ($\lambda = 0 \rightarrow 0.4$). As λ increases, the fusion of visual features from different entities gradually diminishes, with the attenuation part $\mathcal{L}_a$ of the loss playing an increasingly crucial role. However, exclusively minimizing $\mathcal{L}_a$ might lead to a decrease in the activation value across the entire attention map, resulting in the omission of the object, as depicted in Figure 9 ($\lambda = 0.8 \rightarrow 1$).

U-net consists of downsampling blocks, a middle block, and upsampling blocks. In this context, we focus on selected blocks with suitable resolutions for computational efficiency and meaningful spatial distribution. While combining multiple blocks can yield favorable outcomes, it introduces additional semantic information beyond spatial distribution. Among the upsampling blocks, the cross-attention maps from block 1 most accurately capture the spatial distribution of entities. As shown in Table 2, utilizing the cross-attention map of the first upsampling block for Entity Localization and Anchoring yielded the most favorable results.

5.5. Human evaluation

Linguistic binding(Rassin et al., 2023) effectively solves the problem of attribute binding but does not address entity leakage. Table 3 presents the results of our human assessment, showing that our method achieved the best results in addressing physical leaks.

Figure 10. It's a challenging scenario; both methods struggle to generate faithful images in complex scenes featuring multiple entities.

6. Limitation

In complex scenarios involving the generation of four or more entities, the mutual influence between entities becomes increasingly intricate. The introduction of attributes further exacerbates this complexity. As shown in Figure 10, neither our method nor others can accurately generate each entity under these conditions.

Table 1. The method we employ to combine entities in our prompt resembles Attend-and-excite, but we involve 20 animals and objects. In terms of text-text similarity, the prompts follow the format: "a [entity A] and a [entity B]". This format allows us to specifically address concerns related to entity leakage and missing entities.

Method	Avg CLIP text-image		Avg CLIP text-text		
	Full Prompt	Min. Entity	Animal-animal	Animal-object	Object-object
Stable Diffusion (Rombach et al., 2022)	0.3258	0.2362	0.7571 ± 0.0942	0.7837 ± 0.1030	0.7350 ± 0.1003
StructureDiffusion (Feng et al., 2023)	0.3255	0.2306	0.7549 ± 0.0893	0.7921 ± 0.0738	0.7318 ± 0.0877
Attend-and-excite (Chefer et al., 2023)	0.3268	**0.2509**	0.7638 ± 0.1021	0.8312 ± 0.0869	0.8000 ± 0.1044
Entity Localization and Anchoring (ours)	**0.3307**	0.2429	$\mathbf{0.7964 \pm 0.0822}$	$\mathbf{0.8559 \pm 0.0791}$	$\mathbf{0.8135 \pm 0.0919}$

Table 2. We conduct an ablation study of ELA on various attention layers, using the CLIP text-text score as the evaluation metric. Our attention is solely on prompts featuring animal-animal pairs.

DOWN-2	DOWN-3	UP-1	UP-2	SCORE
√				0.7651
	√			0.7724
	√	√		0.7930
		√		**0.7964**
		√	√	0.7901
			√	0.7681

Table 3. Human evaluation. Here, we have two types of datasets. The first type, ALL, comprises two kinds of prompts: one featuring [entity A] and entity B, and another featuring [attribute A] [entity A] and [attribute B] [entity B]. The prompts containing attributes constitute only one-third of the total. The second type, Binding, exclusively includes the latter kind of prompt. We both selected 15 prompts, each capable of generating 64 images.

MODEL	ALL	BINDING
STABLE DIFFUSION	1.80	2.11
LINGUISTIC BINDING	17.92	**31.76**
ATTEND AND EXCITE	18.05	12.08
OURS	**34.02**	24.20
NO MAJORITY WINNER	28.21	29.85

7. Conclusion and Future work

When provided with a prompt containing multiple objects, the diffusion model experiences challenges such as entity overlap or merging. In this paper, we introduce Entity Localization and Anchoring, a method aimed at constraining entities in distinct regions. The spatial layout of entities is closely related to the distribution of the corresponding token in the cross-attention maps. However, it frequently spills into unintended regions. Specifically, we prevent the overlap of these maps by manipulating the latent. Our findings confirm that the implementation of our approach effectively eases the previously mentioned issue of concept bleeding.

Our method effectively mitigates concept blending without the need for training but introduces an additional load on the sampling process. Furthermore, in intricate scenarios with numerous user-specified entities, the interplay between entities becomes more complex. Moving forward, it is crucial to streamline computational costs and improve the handling of entity leakage in more intricate scenarios.

Acknowledgements

This research was supported by fundings from the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152228/23E).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

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

Brooks, T., Holynski, A., and Efros, A. A. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402, June 2023.

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

Dhariwal, P. and Nichol, A. Diffusion models beat gans

on image synthesis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

Feng, W., He, X., Fu, T.-J., Jampani, V., Akula, A. R., Narayana, P., Basu, S., Wang, X. E., and Wang, W. Y. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *The Eleventh International Conference on Learning Representations*, 2023.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., and Cohen-or, D. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2023.

Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., and Cohen-Or, D. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E., and Zhu, J.-Y. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1931–1941, June 2023.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Liu, L., Ren, Y., Lin, Z., and Zhao, Z. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

Ma, Y., Yang, H., Wang, W., Fu, J., and Liu, J. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint arXiv:2303.09319*, 2023.

Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., and Qie, X. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., and Zhu, J.-Y. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

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

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

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., and Chechik, G. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. *arXiv preprint arXiv:2306.08877*, 2023.

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

Ronneberger, O., Fischer, P., and Brox, T. 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*, pp. 234–241. Springer, 2015.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22500–22510, June 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic text-to-image diffusion models with deep language understanding. In Koyejo, S., Mohamed, S.,

Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 36479–36494. Curran Associates, Inc., 2022.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1921–1930, June 2023.

Wang, K., Yang, F., Yang, S., Butt, M. A., and van de Weijer, J. Dynamic prompt learning: Addressing cross-attention leakage for text-based image editing. *arXiv preprint arXiv:2309.15664*, 2023a.

Wang, R., Chen, Z., Chen, C., Ma, J., Lu, H., and Lin, X. Compositional text-to-image synthesis with attention map control of diffusion models. *arXiv preprint arXiv:2305.13921*, 2023b.

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

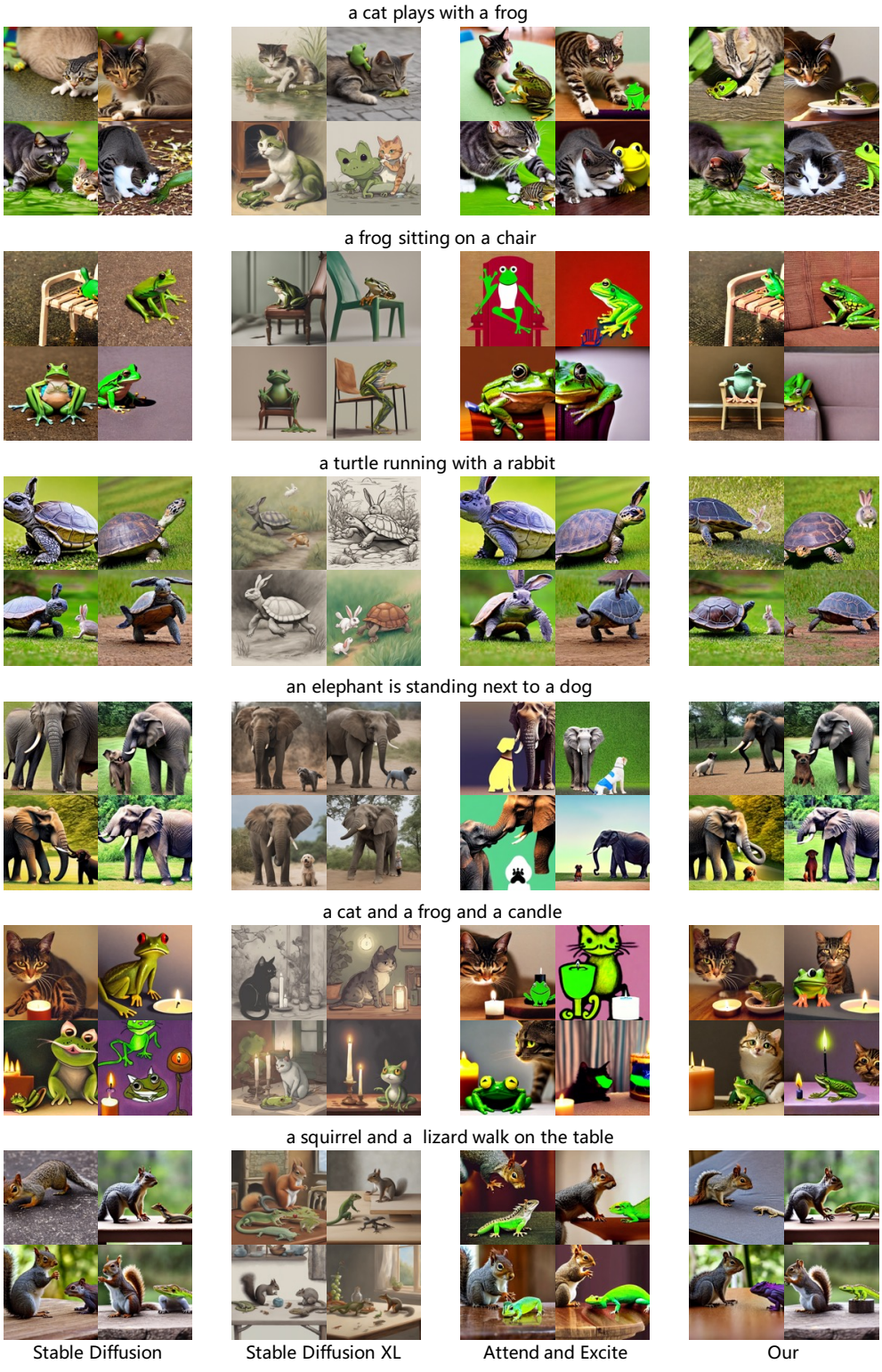

Figure 11. More cases.