Be Yourself: Bounded Attention for Multi-Subject Text-to-Image Generation

Omer Dahary[1], Or Patashnik[1,2], Kfir Aberman[2], and Daniel Cohen-Or[1,2]

[1] Tel Aviv University
[2] Snap Research

Fig. 1: Our method bounds the attention to enable layout control over a pre-trained text-to-image diffusion model. *Bounded Attention* effectively reduces the impact of the innate semantic leakage during denoising, encouraging each subject to be itself. Our method can faithfully generate challenging layouts featuring multiple similar subjects with different modifiers (e.g., *ginger* and *gray* kittens).

Abstract. Text-to-image diffusion models have an unprecedented ability to generate diverse and high-quality images. However, they often struggle to faithfully capture the intended semantics of complex input prompts that include multiple subjects. Recently, numerous layout-to-image extensions have been introduced to improve user control, aiming to localize subjects represented by specific tokens. Yet, these methods often produce semantically inaccurate images, especially when dealing with multiple semantically or visually similar subjects. In this work, we study and analyze the causes of these limitations. Our exploration reveals that the primary issue stems from inadvertent semantic leakage between subjects in the denoising process. This leakage is attributed to the diffusion model's attention layers, which tend to blend the visual features of different subjects. To address these issues, we introduce Bounded Attention, a training-free method for bounding the information flow in the sampling process. Bounded Attention prevents detrimental leakage among subjects and enables guiding the generation to promote each subject's individuality, even with complex multi-subject conditioning. Through extensive experimentation, we demonstrate that our method empowers the generation of multiple subjects that better align with given prompts and layouts.

Keywords: Image Generation · Diffusion Models · Semantic Alignment

1 Introduction

In recent years, text-to-image generation has undergone a significant shift with the integration of conditional diffusion models [?, ?, ?, ?], allowing for the facile generation of high-quality and diverse images. The use of attention layers in architectures of such generative models has been identified as a major factor contributing to improved quality of generated images [?, ?]. However, these models struggle to accurately generate scenes containing multiple subjects, especially when they are semantically or visually similar.

In this work, we study the problem of multi-subject image generation in attention-based diffusion models. Our contribution is twofold. First, we recognize the underlying reasons for the difficulty in generating images containing multiple subjects, especially those sharing semantic similarities. Second, building on our insights, we present a method aimed at mitigating semantic leakage in the generated images, allowing control over the generation of multiple subjects (see Figure 1).

We demonstrate an innate bias within the common attention-based architectures utilized in diffusion models, which predisposes them to leak visual features between subjects. In particular, the functionality of attention layers is designed to blend features across the image. Therefore, they inherently lead to information leakage between subjects. This phenomenon is particularly noticeable when subjects are semantically similar and, therefore, attend to each other (Figure 2).

A plethora of works tries to mitigate the cross-subject leakage issue, either by modifying the sampling process to better follow different subjects in the prompt [?, ?], or by coupling the global prompt with layout information via segmentation maps or bounding boxes labeled with subject classes or local prompts [?, ?, ?, ?]. However, the majority of these methods still encounter difficulties in accurately aligning to input layouts, particularly in scenarios involving two or more semantically similar subjects.

In our approach, we guide the image generation with a spatial layout [?, ?]. To address cross-subject leakage, we introduce the *Bounded Attention* mechanism, utilized during the denoising process to generate an image. This method bounds the influence of irrelevant visual and textual tokens on each pixel, which otherwise promotes leakage. By applying this mechanism, we encourage each subject *to be itself*, in the sense that it hinders the borrowing of features from other subjects in the scene. We show that bounding the attention is needed both in the cross- and self-attention layers. Moreover, we find additional architectural components that amplify leakage, modify their operation, and present remedies to them.

We show that our method succeeds in facilitating control over multiple subjects within complex, coarse-grained layouts comprising numerous bounding boxes with similar semantics. Particularly challenging examples are demonstrated in Figure 1, where we successfully generate five kittens with a mix of adjectives. We conduct experiments on both Stable Diffusion [?] and SDXL [?] architectures and demonstrate the advantage of our method compared to previous ones, both supervised and unsupervised.

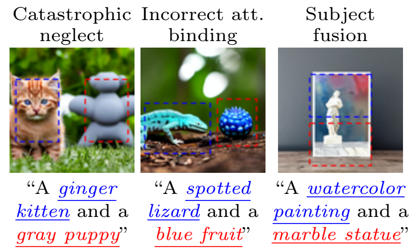

Fig. 2: Misalignment in layout-to-image generation include (i) *catastrophic neglect* [?] where the model fails to include one or more subjects mentioned in the prompt within the generated image, (ii) *incorrect attribute binding* [?, ?] where attributes are not correctly matched to their corresponding subjects, and (iii) *subject fusion* [?] where the model merges multiple subjects into a single, larger subject.

2 Related work

Text-to-image diffusion models. Diffusion models, trained on huge datasets [?], have demonstrated their power in learning the complex distributions of diverse natural images [?,?,?,?,?]. Augmenting attention layers into diffusion models and conditioning them on textual prompts by coupling them with visually-aligned text encoders [?] leads to powerful text-to-image models [?]. In this work, we specifically examine two such open-source text-to-image diffusion models: Stable Diffusion [?], and the more recent SDXL [?].

Semantic alignment in text-to-image synthesis. A critical drawback of current text-to-image models pertains to their limited ability to faithfully represent the precise semantics of input prompts. Various studies have identified common semantic failures and proposed mitigating strategies, such as adjusting text embeddings [?,?,?], or optimizing noisy signals to strengthen or align cross-attention maps [?, ?]. Nevertheless, these methods often fall short in generating multiple subjects, and do not adhere to positional semantics, such as subjects' number or location.

Layout-guided image synthesis. Addressing the semantic alignment concern, alternative approaches advocate for conditioning the diffusion process on layout information, either by training new models from scratch [?, ?] or fine-tuning an existing one [?, ?, ?, ?]. Despite their promise, these methods demand extensive computational resources and prolonged training times. Moreover, they are constrained by the layout distribution of the training data and the models' architectural bias to blend subject features, a limitation that our Bounded Attention aims to overcome.

To circumvent these challenges, numerous researchers explore training-free techniques, where the generation process itself is modified to enforce layout constraints. Several optimization-based works employ techniques similar to classifier-free guidance to localize the cross-attention [?, ?, ?, ?] and/or self-attention maps [?]. While effective in aligning random noise with the intended layout, guiding attention maps to distinct regions may inadvertently lead to undesired behavior, particularly when different subjects share similar semantics and visual features. Furthermore, these methods often exhibit a deteriorating effect on visual quality, thereby limiting their applicability to only the initial denoising steps and neglecting finer control over shape and visual details, which are determined only in later stages [?]. Bounded Attention addresses these shortcomings by regulating attention computation throughout the entire denoising process.

Another approach involves generating each subject separately in its own denoising process [?, ?]. While these methods inherently address catastrophic neglect, they tend to generate disharmonious images, and remain susceptible to leakage when merging subjects in subsequent stages. In contrast, masking attention maps to input bounding boxes [?] or attenuating attention in specific segments [?] represents a milder variant of this strategy, aimed at avoiding visible stitching. However, these methods often fall short of fully mitigating subject leakage, compromising semantic alignment. In comparison, Bounded Attention is able to carefully govern information propagation among subjects in a single denoising process.

While both the trained models and training-free techniques aim to generate numerous objects, they do not mitigate the inherent leakage caused by attention mechanisms. Unlike our Bounded Attention technique, these methods encounter challenges in effectively generating and controlling a multitude of subjects, especially when they share semantic similarity. Notably, existing techniques struggle to accurately generate even two semantically similar subjects, whereas our method, as demonstrated succeeds in generating five and even more subjects.

3 Preliminaries

Latent diffusion models. In this work, we examine Stable Diffusion [?] and SDXL [?], which are both publicly available latent diffusion models. These models operate in the latent space of a pretrained image autoencoder, and are thus tasked with denoising a latent representation of the image, where each latent pixel corresponds to a patch in the generated image. Starting from pure random noise $\mathbf{z}_T$, at each timestep t, the current noisy latent $\mathbf{z}_t$ is passed to a denoising UNet ϵ_θ, trained to predict the current noise estimate $\epsilon_\theta\left(\mathbf{z}_t, y, t\right)$ using the guiding prompt encoding y.

Attention layers. At each block, the UNet utilizes residual convolution layers, producing intermediate features $\phi^{(l)}(\mathbf{z}_t)$, where l denotes the layer's index. These, in turn, are passed to attention layers, which essentially average different values $\mathbf{V_t^{(l)}}$ according to pixel-specific weights:

$$\phi^{(l+1)}(\mathbf{z}_t) = \mathbf{A}_t^{(l)}\mathbf{V}_t^{(l)},\ \text{where } \mathbf{A}_t^{(l)} = \text{softmax}\left(\mathbf{Q}_t^{(l)}\mathbf{K}_t^{(l)\intercal}\right). \tag{1}$$

Here, the keys $\mathbf{K}_t^{(l)} = f_K^{(l)}\left(\mathbf{C}_t^{(l)}\right)$ and values $\mathbf{V}_t^{(l)} = f_V^{(l)}\left(\mathbf{C_t^{(l)}}\right)$ are linear projections of context vectors $\mathbf{C}_t^{(l)}$. In the cross-attention layers we inject semantic context from the prompt encoding $\mathbf{C}_t^{(l)} \equiv y$, while self-attention layers utilize global information from the latent itself $\mathbf{C}_t^{(l)} \equiv \phi^{(l)}(\mathbf{z}_t)$.

The weighting scheme is determined by the attention map $\mathbf{A}_t^{(l)}$ which represents the probability of each pixel to semantically associate with a key-value pair. This is done by linearly projecting the latent noisy pixels to queries $\mathbf{Q}_t^{(l)} = f_Q^{(l)}\left(\phi^{(l)}(\mathbf{z}_t)\right)$ and computing their inner product with the keys. It has been widely demonstrated that the cross-attention maps are highly indicative of the semantic association between the image layout and the prompt tokens [?]. Meanwhile, the self-attention maps govern the correspondence between pixels, and thus form the image's structure [?].

4 Semantic Leakage

We begin by studying the causes of semantic leakage in Stable Diffusion [?], and examine the limitations of existing layout-to-image approaches.

4.1 On Subject Similarity

Figure 2 illustrates various misalignment failures observed in state-of-the-art layout-to-image training-free methods. As we shall show, these failures are more prevalent for subjects that share semantic or visual similarity.

Let $x_{s_1}, x_{s_2} \in \mathbb{R}^2$ be 2D latent coordinates corresponding to two semantically similar subjects s_1, s_2 in the generated image. Intuitively, we expect that along the denoising process, the queries corresponding to these pixels, $\mathbf{Q}_t^{(l)}\left[x_{s_1}\right], \mathbf{Q}_t^{(l)}\left[x_{s_2}\right]$, will be similar and hence also their attention responses. This, in turn, also implies that they will share semantic information from the token embeddings through the cross-attention layers or visual information via the self-attention layers.

To explore this hypothesis, we investigate the model's behavior when tasked with generating two subjects and analyze their attention features in both cross- and self-attention layers. Subsequently, we meticulously examine these features and demonstrate how their behavior sheds light on the leakage observed in generated images.

4.2 Cross-Attention Leakage

To analyze the leakage caused by cross-attention layers, we examine the cross-attention queries. We depict these queries in the plots in Figure 3, where each point corresponds to a single query projected to 2D with PCA. To label each point with its corresponding subject, we compute the subject masks by averaging cross-attention maps [?] and color each projected query according to the subject's text color. The leftmost plot, in which the two subjects were generated separately, serves as a reference point to the relation between the queries when there is no leakage. For comparative analysis, we also present results for Layout-guidance [?], as a simple representative of current training-free layout-guided approaches, and Bounded Attention, which we shall cover in the sequel.

We consider the cross-attention queries in two examples: "a kitten" and "a puppy", and "a hamster" and a "squirrel". As can be seen in the reference plots, the kitten and puppy queries share some small overlap, and the hamster and squirrel queries are mixed together. The level of separation between the red and blue dots in the plots reveals the semantic similarity of the two forming subjects.

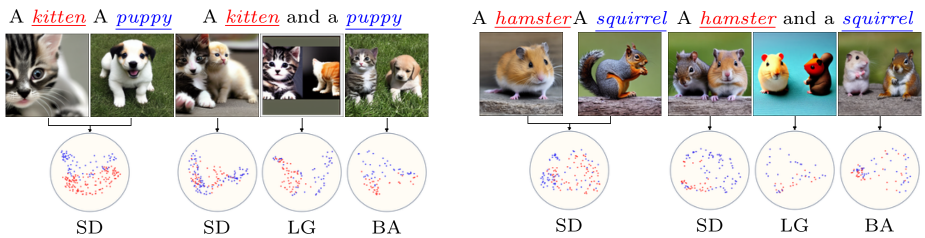

Fig. 3: Cross-Attention Leakage. We demonstrate the emergence of semantic leakage at the cross-attention layers. We show two examples: a puppy a kitten, and a hamster and a squirrel. In the two leftmost columns, the subjects were generated separately using Stable Diffusion (SD). In the right three columns, we generate a single image with the two subjects using three different methods: Stable Diffusion (SD), Layout Guidance (LG), and Bounded Attention (BA, ours). Under each row, we plot the two first principal components of the cross-attention queries. As can be seen, the separation of the queries (blue and red) reflects the leakage between the subjects in the generated images.

Clearly, vanilla Stable Diffusion struggles to adequately generate the two subjects within the same image. This is apparent by the visual leakage between the two subjects, as the model cannot avoid averaging their distinct visual properties. For example, the puppy has the visual features of a kitten, like raised ears and a triangular nose, while the squirrel loses its distinct ear shape and its feet are now pink like the hamster. Respectively, it can be seen that the queries of the two subjects are mixed, even for the kitten and the puppy which are more separated in the reference plot.

Meanwhile, Layout Guidance (LG), which optimizes $\mathbf{z}_t$ to have attention peaks for each noun token at their corresponding region, exhibits interesting results. Its optimization objective implicitly encourages the separation between subjects' cross-attention queries. This can have positive effects, like the hamster and squirrel having unique colors, but at the same time yields unwanted artifacts, like the puppy losing its face altogether. Moreover, it can inadvertently push the latent signal out of distribution, causing quality degradation, as evident by the hamster's and squirrel's cartoonish texture. When it overly pushes the latent out of distribution, it leads to the catastrophic neglect phenomenon (Figure 2).

In comparison, when examining the plots of our method alongside the reference plots, our approach preserves the feature distribution of the subjects' queries, and successfully generates the two subjects, even when the queries are as mixed as in the hamster and the squirrel.

The above analysis yields two immediate conclusions: (i) Semantic similarity between subjects is reflected by their queries proximity, and leads to mixed queries when the subjects are generated together. This in turn leads to leakage between the subjects in the generated images, and (ii) enforcing semantic separation by modifying the semantic meaning of the cross-attention queries is harmful. The former observation represents a crucial architectural limitation in

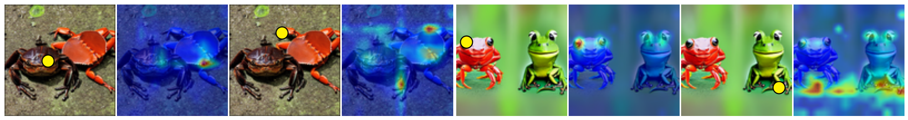

Fig. 4: Self-Attention Leakage. We demonstrate the emergence of semantic leakage at the self-attention maps of two generally semantically dissimilar subjects: a crab and a frog. The images are generated by Stable Diffusion (SD) and Layout-guidance (LG). We highlight specific pixels, such as those of a subject's eye or leg, and present their respective self-attention maps.

current diffusion models, and the latter pinpoints to a previously unexplored weakness in the widely used latent optimization methodology. Bounded Attention is designed to overcome these limitations.

4.3 Self-Attention Leakage

We now turn to analyze the leakage caused by self-attention layers. It has been shown that self-attention features exhibit dense correspondences within the same subject [?] and across semantic similar ones [?]. Hence, they are suspected as another source of leakage, that we shall study next.

Here, we choose to examine the self-attention maps as a means to understand the leakage. In Figure 4 we focus on representative pixels (marked in yellow) associated with the subjects' eyes and legs, where visual leakage is most pronounced. As expected, features from one subject's eye or leg attend to the semantic similar body parts of the other. As a result, the features of each of the yellow points are directly affected by the features of the counterpart subject, causing leakage. In both images, the crab and the frog have similar appearances. In vanilla SD, both have a crab-like color and limbs with frog's toe pads. In LG, both have frog-like eyes and crab legs.

Notably, this tendency to rely on similar patches aids the model in denoising the latent signal and is essential for achieving coherent images with properly blended subjects and backgrounds. Nevertheless, it has the drawback of leaking features between disjointed subjects. Therefore, completely disjointing the visual features during denoising [?], or naively pushing self-attention queries apart through optimization [?], can lead to subpar results. Consequently, we introduce the Bounded Attention mechanism to mitigate leakage and guide the latent signal towards subject separability, while avoiding detrimental artifacts.

It's important to highlight that the leakage resulting from both cross- and self-attention layers is intertwined and mutually reinforced. Therefore, addressing the leakage caused by only one of these layers is insufficient to prevent the leakage in the generated images.

4.4 Levels of Similarity

In the two previous sections, our focus was primarily on subjects that share semantic similarity. Building upon the observation that the UNet's inner layers

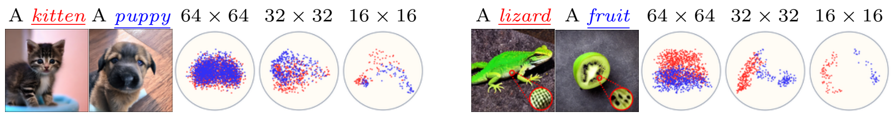

Fig. 5: We generate different subjects, and plot the first two principal components of the cross-attention queries at different layers of the UNet, where each layer is of different resolution. The high semantic similarity between the kitten and the puppy is expressed by the proximity of their queries through all layers. Meanwhile, the lizard and fruit share similar texture, and hence only their high-resolution queries are entangled.

dictate the subject's semantics and shape, while the outer layers control its style and appearance [**?**], we analyzed the inner UNet layers. However, leakage can also occur prominently when generating subjects that share visual similarity rather than semantic. We now turn to explore this scenario and demonstrate that, in such cases, the leakage originates from the UNet's outer layers.

In Figure 5, we visualize cross-attention queries at different decoder's layers, when generating the kitten, puppy, lizard, and fruit in isolation. As can be seen, the queries of the kitten and the puppy are mixed across all UNet's layers, aligning with the visual and semantic similarity between these animals. On the other hand, the queries of the lizard and the fruit are overlapped only in the highest-resolution layer, aligning with the lack of semantic similarity between them. A closer look reveals that the lizard and the fruit share surprisingly similar textures, which explains the overlap of the queries in the highest-resolution layer. As explained in the previous sections, this overlap causes leakage between the two subjects, in this case, a visual rather than a semantic leakage (see Figure 2).

5 Bounded Attention

Our method takes as input n distinct textual subjects $S = \{s_i\}_{i=1}^n$ contained within a global prompt y, along with their corresponding bounding boxes $B = \{b_i\}_{i=1}^n$. Our objective is to condition the generation process on y, S, and B, while preserving the intended semantics of each subject, all without requiring any training or fine-tuning.

Figure 6 illustrates an overview of our method. There, the input prompt y is "A kitten and a puppy", $S = \{\text{"kitten"}, \text{"puppy"}\}$, and the two corresponding bounding boxes $\{b_1, b_2\}$ are illustrated in the top left corner. Bounded Attention operates in two modes: Bounded Guidance and Bounded Denoising. Specifically, at the beginning of the denoising process, for $t \in [T, T_{guidance}]$, we perform a Bounded Guidance step followed by a Bounded Denoising step. In the guidance step, we utilize the Bounded Guidance loss. This interval of timesteps constitutes the optimization phase. Then, for $t \in [T_{guidance}, 0]$, we apply only Bounded Denoising steps. In both modes, we manipulate the model's forward pass by adopting an augmented weighting scheme in the attention layers, that safeguard the flow of information between the queries and the keys:

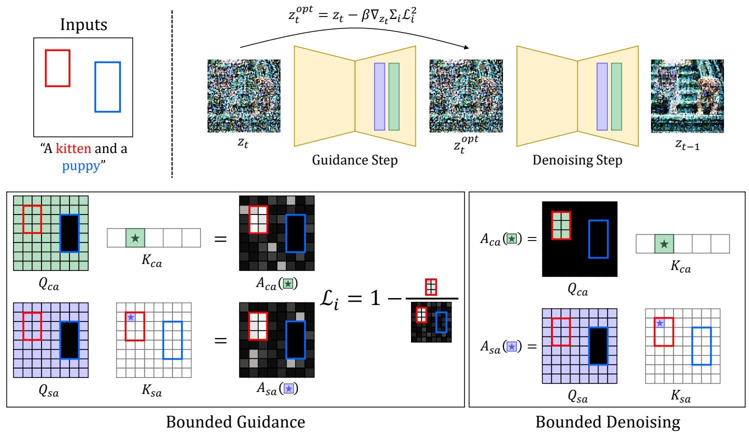

Fig. 6: Bounded Attention operates in two modes: guidance and denoising, each imposing strict constraints to bound the attention of each subject solely to itself and possibly to the background, preventing any influence from other subjects' features. In guidance mode, a loss is minimized to encourage each subject's attention to concentrate within its corresponding bounding box, with other bounding boxes masked to prevent artifacts. During the denoising step, we confine the attention of each subject solely to its bounding box, along with the background in the self-attention. This strategy effectively prevents feature leakage while maintaining the natural immersion of the subject within the image. Attention maps of a specific key, marked with $\star$, are shown for each mode: the cross-attention map displays the key corresponding to the "kitten" token, and the self-attention map shows a key within the kitten's target bounding box.

$$\mathbf{A}_t^{(l)} = \operatorname{softmax}\left(\mathbf{Q}_t^{(l)} \mathbf{K}_t^{(l)\intercal} + \mathbf{M}_t\right), \tag{2}$$

where l represents the layer index, t represents the diffusion timestep, and $\mathbf{Q}_t^{(l)}, \mathbf{K}_t^{(l)}$ are the queries and keys of the l-th attention layer. $\mathbf{M}_t$ represents time-specific masks composed of zeros and $-\infty$ elements. We refer to $\mathbf{A}_t^{(l)}$ as the *Bounded Attention map*. When indexing $\mathbf{A}_t^{(l)}$, we use pixel coordinates $\mathbf{x}$ for rows, and attention-type-specific context vectors $\mathbf{c}$ for columns. In locations $[\mathbf{x}, \mathbf{c}]$, where $\mathbf{M}_t[\mathbf{x}, \mathbf{c}] = -\infty$, it holds that $\mathbf{A}_t^{(l)}[\mathbf{x}, \mathbf{c}] = 0$. Therefore, these masks prevent harmful information flow between pixels in self-attention layers, and between pixels and token embeddings in cross-attention layers.

5.1 Bounded Guidance

In Bounded Guidance, we backpropagate through the diffusion model to steer the latent signal toward the desired layout, using Gradient Descent. Our Bounded Guidance loss encourages the Bounded Attention map of each key corresponding to subject s_i, to be within the b_i bounding box. To this end, for each subject key we consider the ratio between the attention within the corresponding bounding box, to the entire Bounded Attention map (see Figure 6). Formally, we aggregate the following loss on the different subjects:

$$\mathcal{L}_i = 1 - \frac{\sum_{\mathbf{x}\in b_i,\mathbf{c}\in C_i} \hat{\mathbf{A}}\left[\mathbf{x},\mathbf{c}\right]}{\sum_{\mathbf{x}\in b_i,\mathbf{c}\in C_i} \hat{\mathbf{A}}\left[\mathbf{x},\mathbf{c}\right] + \alpha \sum_{\mathbf{x}\notin b_i,\mathbf{c}\in C_i} \hat{\mathbf{A}}\left[\mathbf{x},\mathbf{c}\right]}, \tag{3}$$

where i denotes the index of subject s_i, $\hat{\mathbf{A}}$ is the mean Bounded Attention map, averaged across heads and layers, and α is a hyperparameter that magnifies the significance of disregarding attention towards the background, as we explain later. Similarly to the above, When indexing $\hat{\mathbf{A}}$, pixel coordinates $\mathbf{x}$ represent rows, and attention-type-specific context vectors $\mathbf{c}$ represent columns. We designate C_i as the set of all s_i-related context vectors, i.e., pixel coordinates in b_i for self-attention layers, and the token identifiers of s_i in cross-attention layers. Additionally, for cross-attention layers, we include the first padding token [EoT] in C_i to enhance layout alignment [**?**].

For each subject s_i, the mask $\mathbf{M}_t$ should block the influence of opposing keys (tokens in s_j and pixels in b_j for $j \neq i$), to avoid the artifacts illustrated in Figure 3. This fosters that the queries of different subjects, including similar ones, are not erroneously forced to be far apart. Utilizing this loss, our Bounded Guidance step is defined as $z_t^{\text{opt}} = z_t - \beta \nabla_{z_t} \sum_i \mathcal{L}_i^2$.

Integrating this loss within the cross-attention layers encourages the localization of each subject's semantics within its bounding boxes [**?**]. However, as cross-attention responses tend to peak around more typical patches associated with the subject's semantics (e.g., the face of a human, the legs of a crab, etc.), it may lack control over the subject's boundaries. By applying the loss within the self-attention layers, we encourage each subject to establish its own boundaries close to its bounding box, thereby discouraging subject fusion (see Figure 2).

In the computation of the loss, we also introduce a hyperparameter α to reinforce attention to the background. This adjustment aids in preventing subject amalgamation, where a redundant subject is realized from different subject semantics in the background. To preserve image quality, we limit the application of this mode to an initial time interval $[T, T_{guidance}]$, following similar works [**?**,**?**].

5.2 Bounded Denoising

In Bounded Denoising, we compute the model's output and use it as the next latent in the series. Here, the masks aim to reduce semantic leakage between subjects, as detailed in Section 4, and to prevent unintended semantics from leaking to the background. Unlike Bounded Guidance and typical attention-based guidance approaches, Bounded Denoising can be applied throughout all time steps to mitigate leaks in fine details, which emerge only in later stages [**?**].

However, coarse masking in later stages may degrade image quality and result in noticeable stitching. To address this, after the optimization phase, for $t \in [T_{guidance}, 0]$, we replace each bounding box with a fine segmentation mask obtained by clustering the self-attention maps [**?**]. Since the subject outlines are roughly determined in the initial time steps and evolve gradually thereafter [**?**], we refine the masks periodically.

Notably, this mechanism also addresses imperfect alignments between subjects and bounding boxes after the guidance phase, which are more common

Fig. 7: Our results with SDXL. See the supplemental for more examples.

when generating numerous subjects. Thus, employing this technique enhances the robustness of our method to seed selection, ensuring proper semantics even when subjects extend beyond their initial confines (see Figure 1,7). Compared to methods that necessitate strict input masks yet remain susceptible to leakage [?, ?, ?], our method offers greater user control with simpler inputs and more satisfactory outcomes.

Method Details. Further details on the adaptation of Bounded Attention to the cross- and self-attention layers, along with a description of the subject mask refinement process, are provided in the supplementary materials.

6 Experiments

In this section, we conduct both qualitative and quantitative experiments to assess the effectiveness of our Bounded Attention method. We compare our approach with three training-free baseline methods: Layout-guidance (LG) [?], BoxDiff (BD) [?], and MultiDiffusion (MD) [?]. Additionally, we include comparisons with GLIGEN [?] and ReCo [?], which necessitate training. Since Attention-refocusing (AR) [?] is based on GLIGEN, we categorize it as a trained method for the purpose of our evaluation. For fairness, when comparing our method with other methods, we use Stable Diffusion.

6.1 Qualitative Results

SDXL results. We begin our experiments by demonstrating the efficacy of our method in challenging scenarios, particularly when tasked with generating multiple semantically similar subjects using SDXL. As can be seen in Figure 7, Vanilla SDXL fails to follow the prompt due to semantic leakage. It generates inaccurate number of dogs and mixes the features of each breed. In comparison, our approach adeptly generates each dog with its unique characteristics.

Fig. 8: Comparison of the first six images generated from the seed 0.

Non-curated results. Next, we conduct a non-curated comparison with the training-free baseline methods and present the results in Figure 8. We showcase the initial six images sampled from seed 0 for each method. We anticipate that in the absence of Bounded Attention, semantic leakage may freely blend subject features, hindering the intended layout's formation.

It is evident from the results that none of the competing methods is able to consistently construct the input layout. Layout Guidance [**?**] frequently neglects one of the subjects, and even when it generates three subjects, it struggles to avoid leakage, resulting in puppies with kitten-like features or incorrect color assignments. BoxDiff [**?**] often generates the correct number of subjects but suffers from artifacts in the form of blobs. Similar to Layout Guidance, it encounters difficulties in properly constructing the puppy. Surprisingly, even MultiDiffusion [**?**], which generates the subjects separately, faces challenges in generating them all, with some disappearing or merging together in its bootstrapping phase.

In contrast, our method consistently outperforms these approaches, producing three subjects that align with the both prompt and layout in all six images.

Comparisons with baselines. We present a qualitative comparison in Figure 9. All competing methods, including those trained specifically for the layout-to-image task, exhibit significant visual and semantic leakage. The training-free methods perform the worst: MultiDiffusion produces disharmonious, low-quality images, while optimization-based methods often result in object fusion, combining different semantics without adhering to the layout.

The training-based approaches closely follow the layout but fail to convey the correct semantics. In the first row, they neglect the corduroy jacket, leaking the denim texture into the other jacket, or even fusing them together. In the second row, the elephant's features leak into the rhino.

In comparison, our method generates images that align with the input layout and prompt, ensuring each subject retains its unique attributes, semantics, and appearance.

Fig. 9: Qualitative comparison of our method with baseline methods: Above each row, we display the input prompt, where each subject's color matches the color of its corresponding bounding box. We compare with both training-free methods (2nd-4th columns) and trained models (5th-7th columns). See supplemental for more results.

6.2 Quantitative Results

Method	Counting			Spatial
	Precision	Recall	F1	Accuracy
Stable Diffusion [?]	0.74	0.78	0.73	0.19
Layout-guidance [?]	0.72	0.78	0.72	0.35
BoxDiff [?]	**0.81**	0.78	0.76	0.28
MultiDiffusion [?]	0.70	0.55	0.57	0.15
Ours	**0.81**	**0.91**	**0.82**	**0.43**

Table 1: Evaluation on the DrawBench dataset.

We evaluate our method's effectiveness using the DrawBench dataset [?], known for its challenging prompts designed to test a model's ability to compose multiple subjects with specific quantities and relations. We use the evaluation procedure from previous work [?, ?]. Our results, alongside those of other training-free methods, are summarized in Table 1. Unlike other approaches that do not account for semantic leakage, our method demonstrates notable improvements in both the counting and the spatial categories. Notably, other methods struggle to surpass the counting recall rates of vanilla SD, highlighting the effectiveness of Bounded Attention in generating multiple subjects.

We further conduct a quantitative comparison of semantic leakage in training-free methods through a user study. Further details are in the supplemental.

6.3 Ablation Studies

To assess the significance of each component, we conduct an ablation study where we systematically vary our method's configurations by omitting one component in each setting. We show the results in Figure 10.

Guidance is crucial for aligning the latent signal with the intended layout. Neglecting it can lead to subject generation in the background, as demonstrated

“A *red lizard* and a *turtle* on the grass.”

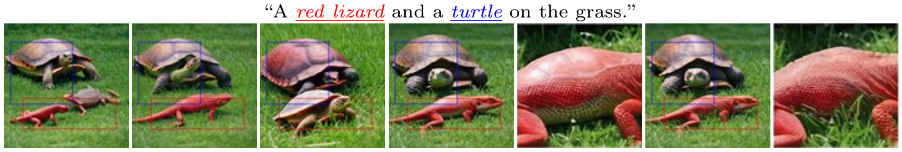

Fig. 10: Qualitative ablation. We ablate our method by skipping the guidance step (G), performing guidance without Bounded Guidance (BG), not applying Bounded Denoising (BD), and not performing Mask Refinement (MR). We show zoomed-in images of the two rightmost configurations. More examples can be found in the supplemental.

by the emergence of a turtle-lizard hybrid. However, attempting to guide the latent signal without our Bounded Guidance mechanism may cause distortions in subject appearances, such as the turtle's neck growing out of its leg and inconsistent artifacts appearing on the lizard's head and back.

Meanwhile, forgoing Bounded Denoising results in noticeable semantic leakage, with the lizard being replaced by a turtle, and the “red” attribute erroneously leaking to the wrong subject.

Lastly, incorporating mask refinement in the later stages prevents fine-details from leaking. Without mask refinement, the lizard exhibits shell-like contours on its back.

7 Conclusions

We introduce Bounded Attention, a technique designed to regulate the accurate generation of multiple subjects within an image. This approach encourages each subject to “be yourself”, emphasizing the importance of preserving individuality and uniqueness without being excessively influenced by other subjects present in the image. Our development of the Bounded Attention technique stemmed from an in-depth analysis of the root causes behind the misalignment observed between the provided prompt and the resulting generated image. Our investigation revealed that this misalignment primarily arises due to semantic leakage among the generated subjects, a phenomenon observed in both the cross and self-attention layers.

While Bounded Attention effectively mitigates a significant portion of semantic leakage, it does not entirely eliminate it. Our findings demonstrate a marked improvement in performance compared to other methods that seek to achieve semantic alignment. However, residual leakage persists, which we attribute to imperfect optimization during the guidance mode and inaccurate segmentation of the subject prior to the second phase.

While Bounded Attention excels in generating multiple subjects with plausible semantic alignment, its performance may vary across different layouts. Achieving success with Bounded Attention hinges on a strong match between the seed and the layout. Moving forward, we aim to explore methods for generating well-suited seeds tailored to specific layouts. One potential avenue is to introduce noise to the layout image, thereby creating a seed that aligns more closely with the desired outcomes.

Acknowledgement

We thank Guy Tevet for fruitful discussions and useful suggestions. This work is partly supported by a research gift from Snap Inc.

References