BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion

Jinheng Xie[1] Yuexiang Li[2*] Yawen Huang[2] Haozhe Liu[2,3] Wentian Zhang[2]
Yefeng Zheng[2] Mike Zheng Shou[1*]
[1] Show Lab, National University of Singapore [2] Jarvis Lab, Tencent
[3] AI Initiative, King Abdullah University of Science and Technology
{sierkinhane,mike.zheng.shou}@gmail.com

1. Limitations

We present some synthetic results with unusual prompts in Fig. 1. Here we found some scenarios in which BoxDiff may fail to synthesize realistic images: i) **combinations of objects that infrequently co-occur**; ii) **uncommon locations as spatial conditions for objects**. For example in Fig. 1, when "car" and "basin", which infrequently co-occur, are in a sentence as the text prompt and the uncommon bounding boxes are as the conditions, the synthetic results will be unrealistic and not adhering to the spatial conditions. Besides, the proposed BoxDiff cannot synthesize realistic images when given some uncommon scenes like "a giraffe flying in the sky" or "a mountain underneath the water".

2. Scribble as Conditions

Simply, scribble can be transformed into bounding boxes, which seamlessly fit the proposed three constraints. In addition, an objectness constraint can be additionally added to further control an object's content or direction. Given a set of scribble $\mathcal{C} = \{\mathbf{c}_i\}$ with each $\mathbf{c}_i$ containing Q points $\{(x_1^i, y_1^i), (x_2^i, y_2^i), \cdots, (x_Q^i, y_Q^i)\}$, the objectness constraint can be formulated as:

$$\mathcal{L}_{\mathbf{s}_i}^5 = 1 - \frac{1}{Q}\sum \textbf{select}(\mathbf{A}_i^t, \mathbf{c}_i), \tag{1}$$

$$\mathcal{L}_{OC} = \sum_{s_i \in \mathcal{S}} \mathcal{L}_{\mathbf{s}_i}^5, \tag{2}$$

where **select**$(\cdot)$ selects corresponding elements using $\mathbf{c}_i$ from $\mathbf{A}_i^t$. $\mathcal{L}_{OC}$ can be added to the overall constraints when the scribble is given.

3. Implementation Details

All experimental results are obtained using the official Stable Diffusion v1.4 text-to-image synthesis model. The number of denoising steps is set as 50 with a fixed guidance scale of 7.5, and the synthetic images are in a resolution of 512×512. We use a Gaussian kernel with a size of 3×3 and a standard deviation $\sigma = 0.5$. P in **topk**$(\cdot)$ is set as 80% of the number of the mask regions $\mathbf{M}_i$ and $(1 - \mathbf{M}_i)$ so that P is adaptively set according to the size

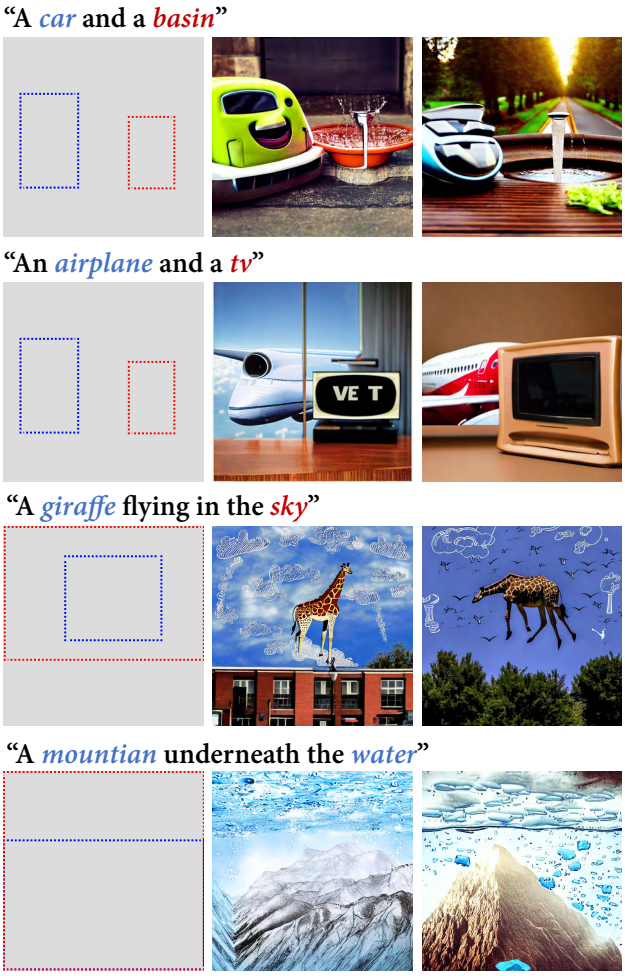


Figure 1: Image synthesis with unusual prompts and locations.

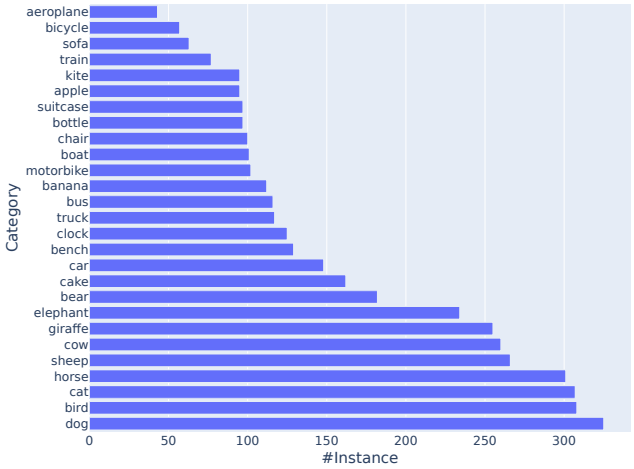


Figure 2: The number of instances of each category in the spatial conditions for zero-shot performance comparison. Here, an instance means a corresponding bounding box.

of the mask. L in **sample**$(\cdot)$ is set as 6, which means that 3 error terms around the given two coordinates are selected, respectively. All experiments are conducted on the NVIDIA TESLA V100 GPU with 32 GB memory.

As fully-supervised layout-to-image methods are restricted to a limited scope of categories, we select 9 common animals and 18 common objects from the detection results as the candidate classes. In total, there are 4,274 valid bounding boxes to be the spatial layout conditions. For a fair comparison, we propose to compare the performance of conditional image synthesis on the newly collected layout (no intersection with COCO and VG). The collected candidate categories (9 animals and 18 objects) for spatially conditional text-to-image synthesis are presented below:

{ Animals: ['bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'giraffe'],
Objects: ['bicycle', 'car', 'motorbike', 'aeroplane', 'bus', 'train', 'truck', 'boat', 'bench', 'suitcase', 'kite', 'bottle', 'banana', 'apple', 'cake', 'chair', 'sofa', 'clock']}

Fig. 2 presents the bar chart for the number of instances of each candidate category. An instance means a corresponding bounding box for conditional image synthesis. It can be seen that there is a maximum number of instances up to 300 and a minimum one around 50. In total, there are 4,274 valid instances (bounding boxes) for image synthesis.

Selection of Target Tokens: Typically, given a text prompt such as "a rabbit and a balloon", if we are interested in controlling the synthesis of the rabbit and balloon, each single target token or word, e.g., "rabbit" and "balloon", is enough to extract the corresponding cross-attentions for box-constrained diffusion. However, sometimes we are interested in controlling the objects in the form of compound nouns. For example, given a text prompt such as "Palm trees sway in the gentle breeze", if we aim to control the synthesis of the palm trees, how to perform BoxDiff with two cross-attention maps for a single semantic target? In our experiments, we found that a single token almost dominates the cross-attention for the target semantic. For example, as shown in Fig. 3, to control the synthesis of palm trees, the cross-attention of "trees" is enough for BoxDiff to limit the palm trees within the given conditional box while retaining the correct semantics of "palm trees".

Table 1: Ablation studies on various **topk**$(\cdot)$.

topk (largest)	topk (smallest)	random	T2I-Sim	AP(↑)
✓			0.3513	22.3
	✓		0.3206	12.8
		✓	0.3491	21.4

Table 2: Ablation studies on various P in **topk**$(\cdot)$.

20%	40%	80%	100%	T2I-Sim	AP(↑)
✓				0.3523	13.2
	✓			0.3516	18.5
		✓		0.3513	22.3
			✓	0.3489	24.8

4. More Ablation Studies

In this section, we provide more ablation studies to validate the effectiveness and necessity of **topk**$(\cdot)$ and **sample**$(\cdot)$ in Eqs. (4), (6), (10), and (13).

Table 1 presents the influence of various sampling methods on the quality and precision of synthetic images. One can observe from the table that selecting P elements with the largest value in **topk**$(\cdot)$ obtains the best T2I-Sim and AP. We argue that a pixel within the given box in a higher response represents a higher probability that the object will appear or be synthesized in the pixel. Therefore, sampled elements with a high response obey the prior where the object will appear in the spatial dimension. By contrast, selecting P elements with the smallest value or P random elements in **topk**$(\cdot)$ achieves a lower T2I-Sim and AP. The comparison further validates the effectiveness and necessity of **topk**$(\cdot)$ (sampling of P elements with the largest value).

In Table 2, we provide the performance of BoxDiff using various P in **topk**$(\cdot)$ and P is adaptively set according to the percentage of the number of elements in $\mathbf{M}_i$ and $(1-\mathbf{M}_i)$. When 80% elements of $\mathbf{M}_i$ and $(1-\mathbf{M}_i)$ are selected in the Box-Constrained Diffusion, respectively, the BoxDiff can obtain the best T2I-Sim and a relatively higher AP.

Table 3: Ablation studies on various L in **sample**$(\cdot)$.

6	10	14	T2I-Sim	AP(↑)
✓			0.3513	22.3
	✓		0.3486	22.0
		✓	0.3489	22.1

We conduct an ablation study to determine the value of L used in Eqs. (10) and (13) and the results are presented in Table 3. When $L = 6$, the best T2I-Sim and AP are

achieved. When L varies from 6 to 14, though AP is relatively stable, T2I-Sim is accordingly decreased. This validates that more constraints on the cross-attentions will affect the quality of the synthetic images, and representative sampling is sufficient for image synthesis obeying the spatial conditions while retaining higher image quality.

5. More Visualization Results

Scribble as Conditions: As discussed in the paper, the proposed BoxDiff can also interact with other types of spatial conditions such as scribble. Here, we provide more synthetic samples using scribble conditions in Fig. 3. Beyond object-bounding boxes, scribble provides more pixel information about the object content. This motivates the proposed objectness constraints $\mathcal{L}_{OC}$, which can further control the object's content or direction. For example, as shown in the second row of Fig. 3, the content of the sailboat and palm trees are relatively consistent with the scribble conditions, *i.e.,* the blue and orange line.

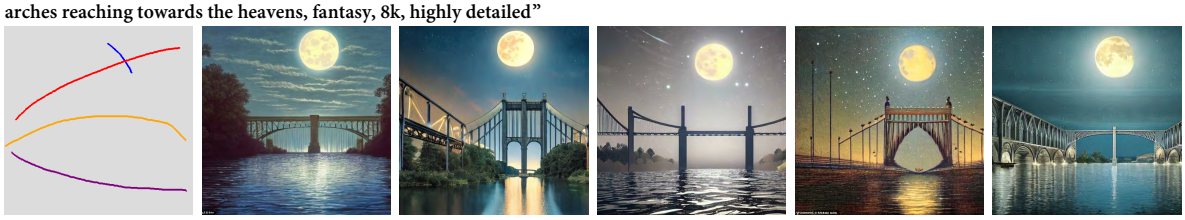
More Visual Comparisons: In Fig. 4, we provide more visual comparison among Stable Diffusion, Structure Diffusion, and the proposed BoxDiff. It can be seen that contents such as tie and hat occasionally are missed in the samples synthesized by Stable Diffusion and Structure Diffusion. In contrast, samples generated by the proposed BoxDiff are relatively consistent with the spatial conditions. Besides, each target object is correctly presented in the resulting images.

More visualization results are shown in Fig. 6, 8, and 9.

"A *rabbit* and a blue *balloon*"

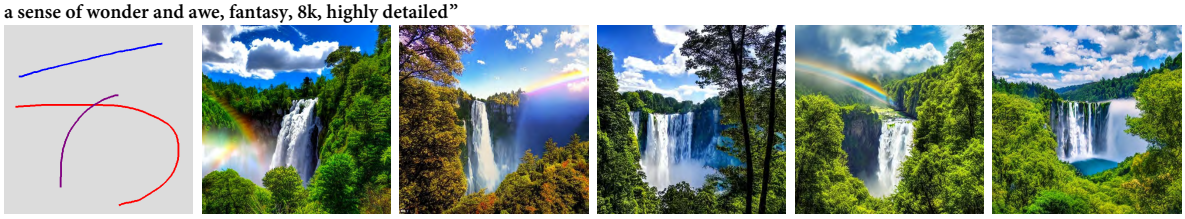
"An ancient *sailboat* glides across the *ocean*, accompanied by the radiance of the *moonlight*, fantasy, 8k, highly detailed"

"The *moon* hangs high in the *sky*, casting a shimmering reflection on the calm *river* below. A grand *bridge* spans the width of the waterway, its arches reaching towards the heavens, fantasy, 8k, highly detailed"



"A hot air *balloon* hovers over rolling hills, while *lotus* leaves floats in a tranquil *pond*. In the distance, majestic *mountains* rise up amongst the greenery. The scene is serene and picturesque, capturing the beauty of nature in a single glance, fantasy, 8k, highly detailed"

"A canvas of blue *sky*, dotted with fluffy white clouds, stretches over a *valley* below. A majestic *waterfall* cascades down, its mist creating rainbows in the sun. The trees sway gently in the breeze, their leaves rustling a soothing melody. The scenery is serene and captivating, evoking a sense of wonder and awe, fantasy, 8k, highly detailed"



"A canopy of white clouds stretches across a blue *sky*, as crystal clear *waters* meet the sandy shores of a tranquil *beach*. Palm *trees* sway in the gentle breeze, their fronds rustling like whispers in the wind. And in the shade of a solitary coconut tree, a lone figure gazes out at the serene expanse before them, feeling at peace amidst the beauty of nature, fantasy, 8k, highly detailed"

Figure 3: Synthetic samples using scribble spatial conditions.

Figure 4: More visual comparison among Stable Diffusion, Structure Diffusion, and the proposed BoxDiff.

“A *bear* wearing *sunglasses* looks very proud”

“A *giraffe* wearing *sunglasses* looks very proud”

“A *rabbit* wearing *sunglasses* looks very proud”

“A *duck* wearing *sunglasses* looks very proud”

“A colorful *parrot* and a red *hat*”

“A *cat* and a red *hat*”

“A *rabbit* and a red *hat*”

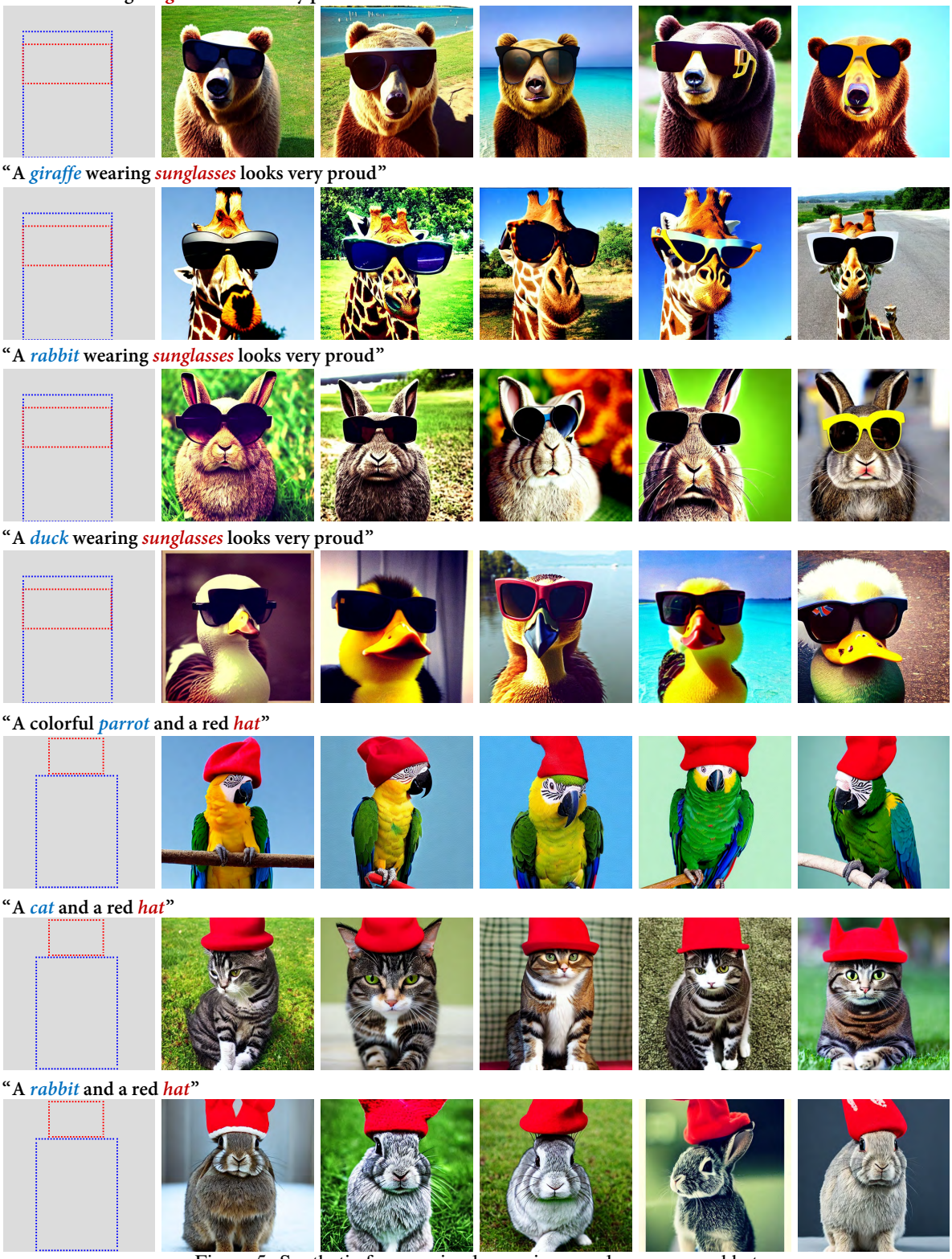
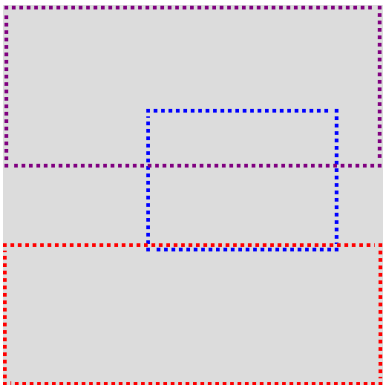
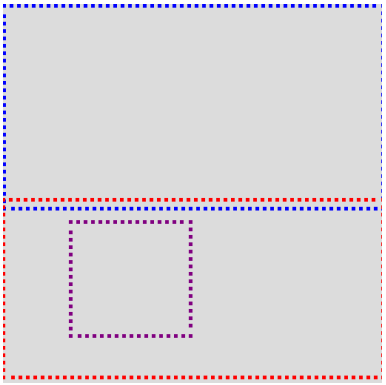


Figure 5: Synthetic funny animals wearing sunglasses or a red hat.



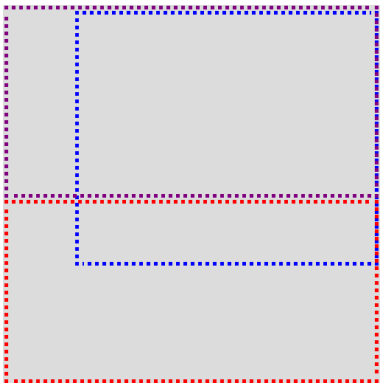
“A magnificent *castle* stands atop a hill, overlooking a serene *lake*, The crystal-clear water reflects the majestic beauty of the castle and the surrounding forests, The *sky* above is a stunning blue, dotted with fluffy white clouds, The landscape seems to be from another world, like a mystical fairyland waiting to be explored, fantasy, 8k, highly detailed”

Figure 6: Synthetic castles, lakes, and sky with the same spatial conditions.



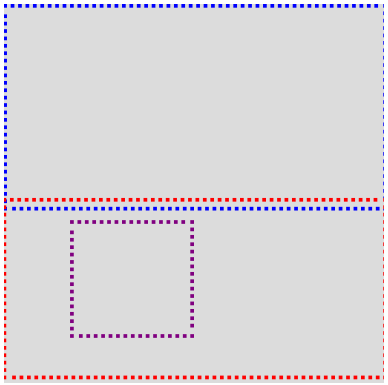
“The *Aurora* Borealis illuminates the night sky above a serene *lake*, while a cozy *tent* is pitched on the shore. The colors of the aurora dance and swirl, casting a magical glow over the water. It's a breathtaking scene, like something out of a fairytale, where one can escape from the worries of the world and lose themselves in the beauty of the moment, fantasy, 8k, highly detailed”

Figure 7: Synthetic aurora, lakes, and tents with the same spatial conditions.



“Higly detailed, majestic royal tall *ship* on a calm *sea*,realistic painting, by Charles Gregory Artstation and Antonio Jacobsen and Edward Moran, (long shot), clear blue *sky*, intricated details, 4k”

Figure 8: Synthetic ship, sea, and sky with the same spatial conditions.



"The *sky* is blanketed with a twinkling carpet of stars, casting a serene glow over a calm *lake*. A cozy *tent* is pitched beside the water, providing the perfect vantage point to take in the breathtaking celestial display. The stars reflect on the surface of the water, creating a mesmerizing scene. It's a peaceful and beautiful moment, where one can bask in the wonder of the starry sky and the tranquility of the natural world, fantasy, 8k, highly detailed"

Figure 9: Synthetic starry sky, lakes, and tents with the same spatial conditions.