VSC: Visual Search Compositional Text-to-Image Diffusion Model

Supplementary Material

A. Limitations

Our focus is not on object–object relationships; therefore, we do not account for interactions such as semantic relationships or shared attributes like color between objects. Given this scope, we rely only on basic dependency parsing provided by the spaCy library. However, incorporating language models in place of simple dependency parsing could be beneficial for handling such complex relationships. Another aspect of the object-object relationship is the spatial relationship. As VSC utilizes subject-driven generation, correct spatial information can be achieved by providing additional bounding boxes [41].

B. Implementation Details

VSC uses spaCy's transformer-based dependency parser to extract entity nouns that are not directly modifying other nouns by analyzing the syntactic dependency graph. We use Stable Diffusion 1.4, 2.1, and 3.5 [9, 30]. The MLP module is composed of 2 linear layers with a LayerNorm. For the image encoder, we use OpenAI's `clip-vit-large-patch14` for Stable Diffusion 1.4 and LAION's `CLIP-ViT-H-14` for Stable Diffusion 2.1 and Stable Diffusion 3.5. Note that we only optimize the last few layers in the image encoders. We set the number of training steps to $60,000$ steps, the learning rate to $1e-5$, and the batch size to 8. We use 2 NVIDIA A5000 GPUs, and the training takes approximately 45 hours on 90,000 synthetic images.

C. Ablation Studies

We evaluate the performance of VSC with and without the localization loss L_{loc} on the T2I-CompBench dataset using Stable Diffusion 3.5 in Table 6. Our results demonstrate that L_{loc} enhances the overall performance of VSC. Furthermore, the localization loss assists the separation of multiple subjects and mitigates identity blending, as observed in FastComposer [39].

Method	Color	Texture	Shape
VSC (3.5) w L_{loc}	**0.83**	**0.73**	**0.58**
VSC (3.5) w/o L_{loc}	0.79	0.70	0.53

Table 6. VSC performance with and without L_{loc}

D. Qualitative Results

We include additional visualizations of VSC's generated images in Figure 11 and Figure 12, providing a comprehensive overview of the model's capability to generate diverse and high-quality outputs under varying conditions.

In figure. 9, we provide more qualitative results on how scaling the synthetic dataset influences the model's performance.

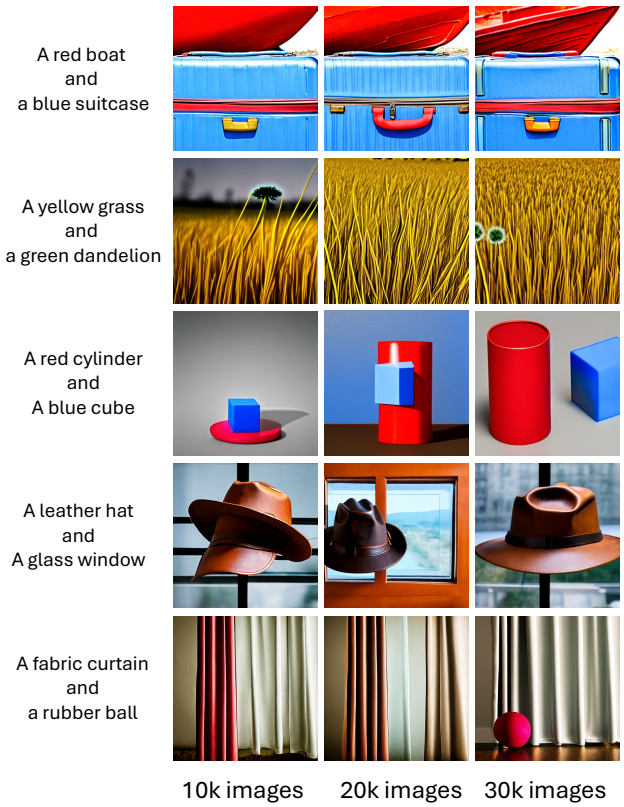

Figure 9. More qualitative results on scaling dataset

Similarly, we provide more qualitative results on VSC's performance while increasing the number of objects in the prompts in figure. 10.

E. Data Creation

We provide illustrative examples in Figure 13 to demonstrate the process of selecting generated images for inclusion in the final synthetic dataset. The selection process involves several critical steps to ensure the quality and relevance of the data. First, we utilize BLIP to perform a visual-question

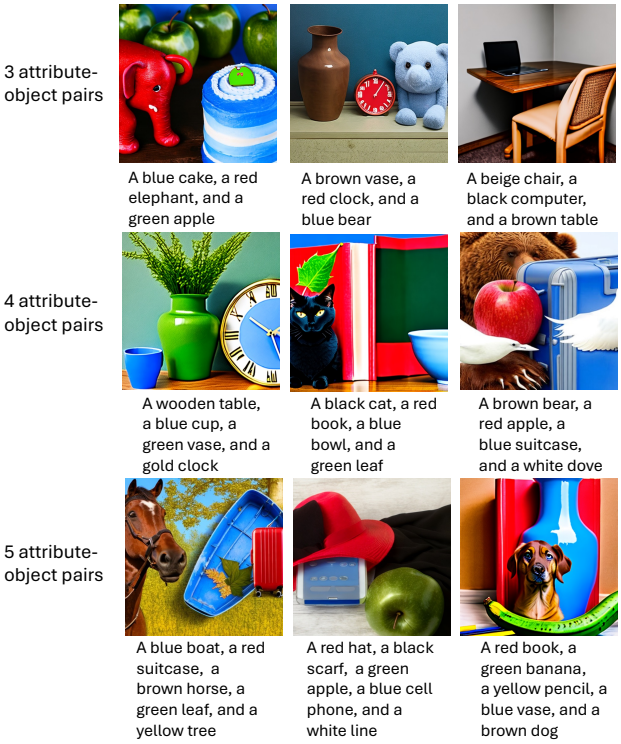

Figure 10. More qualitative results on scaling the number of attribute-object pairs in the prompts.

answering (VQA) task, which helps assess the semantic coherence between the image and its textual description. This step ensures the generated content aligns with the intended prompt or context.

Next, we extract the instance segmentation of the generated image to identify distinct objects or regions within the scene. These segments are then approximated to bounding boxes, which provide a simplified representation of object locations Finally, we crop the bounding boxes and compute the text-image alignment score using CLIP.

In our selection process, the generated images undergo strict filtering based on predefined thresholds to ensure the quality and relevance of the synthetic data. Mostly, images with a BLIP-VQA score lower than 0.8 or a CLIP alignment score below 0.7 are typically discarded. These thresholds are carefully chosen to maintain a high textual and visual coherence standard within the dataset.

E.1. Data Curation

Remove	Color	Texture	Shape	HM
CLIP	0.81	0.71	0.53	0.672
BLIP-vqa	0.79	0.68	0.53	0.649

Table 7. Ablation study on the data curation process with base model as Stable Diffusion 3.5

We conducted additional training in 2 different scenarios: Removing the CLIP score and removing the BLIP-vqa scores in the data curation process. The results are depicted in Table 7, showing that BLIP-vqa significantly curates high-quality data while CLIP is less influential.

Figure 11. More qualitative results of VSC

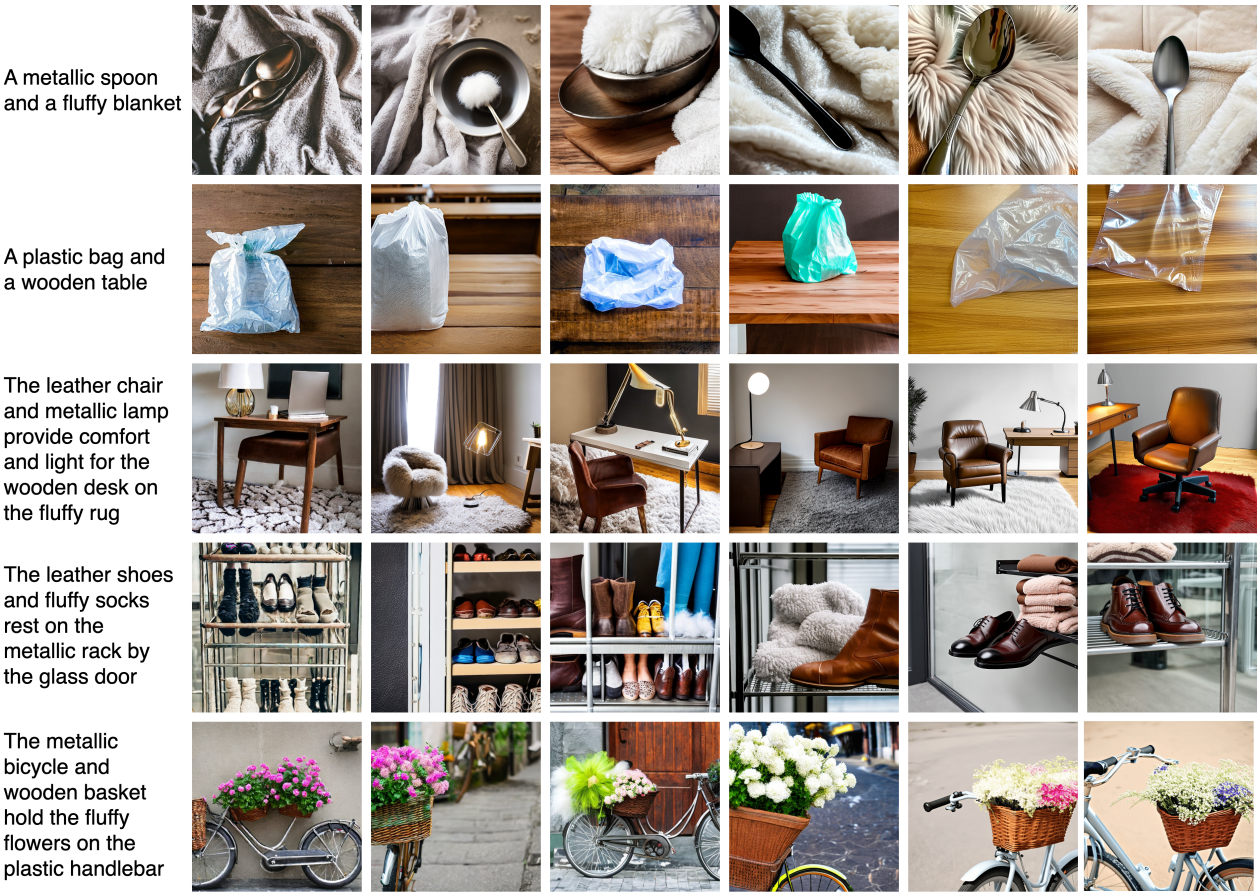

Figure 12. More qualitative results of VSC

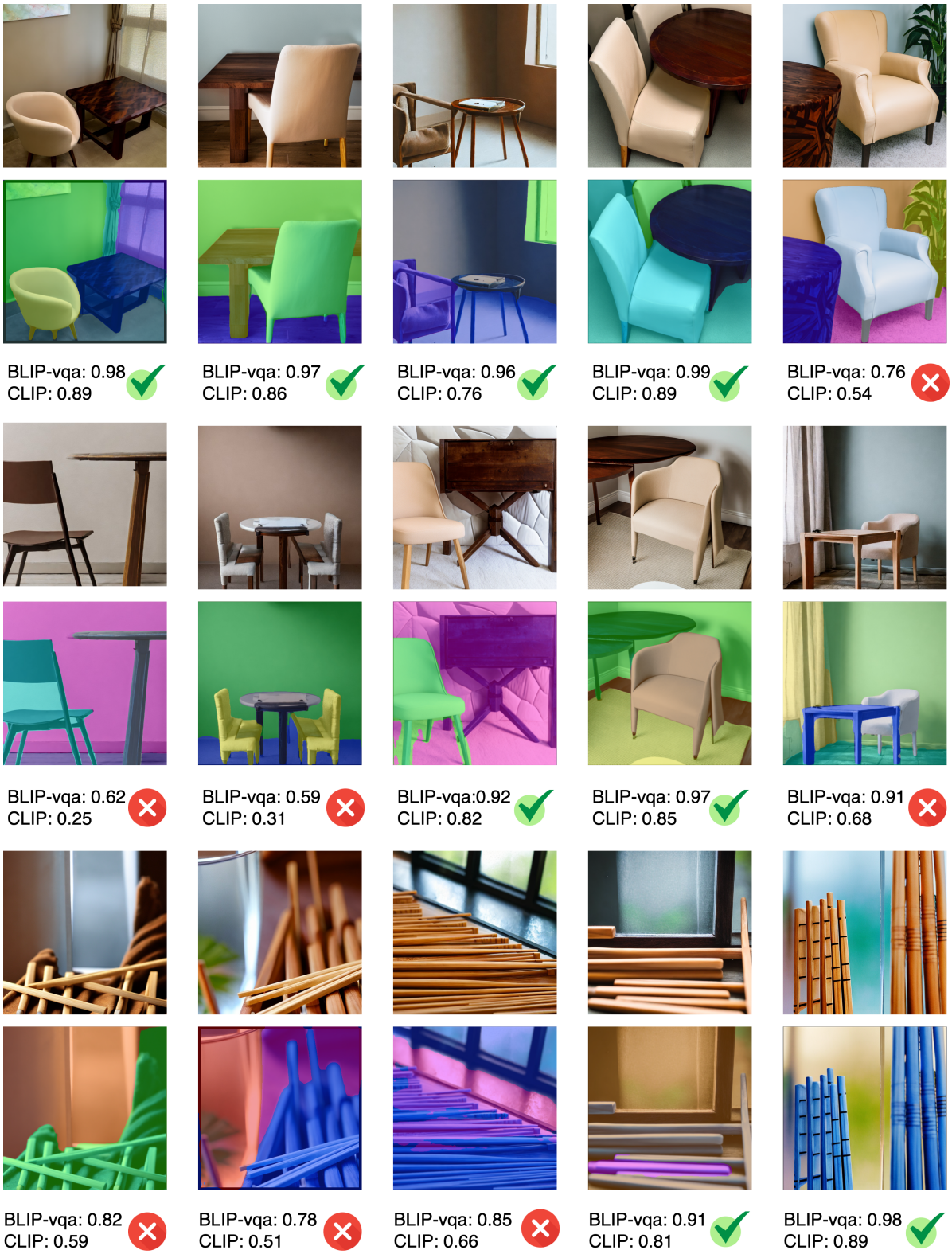

Figure 13. Visualization of our synthetic data creation pipeline. Each generated image is evaluated with the BLIP-vqa metrics and CLIP score between the text and its highest aligned mask. Mostly, the image with BLIP-vqa lower than 0.8 or CHLIP score lower than 0.7 would be discarded.