

Supplementary Materials of Scene Diffusion: Text-driven Scene Image Synthesis Conditioning on a Single 3D Model

Anonymous Authors

1 FURTHER DISCUSSION ON THE SHADING ADAPTIVE TRANSFORMATION \mathcal{T}

In the section of methodology, we outline how the shading adaptive transformation \mathcal{T} functions within the SACA to reconcile the contradiction between two training objectives and facilitate network’s learning on the prior of global shading coherence. In support of this argument, we offer an intuitive comparison in Figure 1. The training hyperparameters are the same for both setups, saving for the exclusion of \mathcal{T} in the first one. As can be seen, not utilizing \mathcal{T} results in the decreased shading diversity for the object across the various scenes. The object may seem dark even in the scene with bright lighting. As a contrast, the object’s shading exhibits more dynamic changes when \mathcal{T} is employed.

We also include a brief analysis in the main text to confirm that the utilized \mathcal{T} can effectively handle the variations in the ambient light color of the object. Figure 2 displays visual illustrations to support this assertion. Take case 1-3 as the examples, when object is in the scene with relatively uniform lighting, \mathcal{T} can largely eliminate the shading difference between c and x_0 . However, since the current \mathcal{T} does not take diffuse reflection and specular highlight colors into account, the shading variances they initiated are still hard to eliminate, especially in the samples as case 4. This is also a topic that we plan to explore in the future research. Overall, the current form of \mathcal{T} is concise and theoretically rational. The application of it has been shown to yield beneficial results for the network’s learning.

2 VISUAL DEMONSTRATION OF THE FPTs’S EFFECT

In the main text, the EoG indicator is employed to quantitatively verify that FPTs can efficiently reduce the high-frequency signals in the object area. Figure 3 further provides the visual evidences to uphold this assertion. The top and bottom rows exhibit the sequences of predicted \hat{x}_0 during the denoising process, without or with FPTs, respectively. In the absence of FPTs, the high-frequency components in the object area will grow quickly. Though there is a decline in the subsequently, they still be unreasonably high in the end. The introduction of FPTs alleviates this problem. As depicted in the figure, it softens the growth of high frequency components in object area, regulating them to the reasonable extent eventually.

3 FURTHER INTRODUCTION TO THE DATASET

The dataset used in this work is constructed based on 3D-FUTURE. It consists of 20240 high-quality interior design rendering images and the textured 3D models of included furniture. Condition images are create based on Blender 3.6. We setup the object and camera positions according to the annotation, place a daylight source above them, and then execute the single-model rendering. Since a scene

image usually contains multiple furniture, it will correspond to multiple condition images. The text prompt corresponding to each scene image is obtained based on the annotation and LLaVA-1.5. After leaving out the test objects and the scene images containing them, there are 19127 different scene images and 49,963 condition-text-output triples involved in training. The condition images can be divided into five categories according to the furniture they contain: bed (4148), sofa (12259), table (4129), chair (5195) and shelf (24232). Figure 4 shows some examples of the training data.

4 IMPLEMENTATION DETAILS OF THE COMPARISONS

Four alternative methods are selected for comparison in this paper. This part provides the implementation details of them. All these methods are diffusion-based. Unless otherwise specified, all methods use the same sampling setup as the proposed method, where 100-step DDIM sampler and the classifier-free guidance scale of 7 are used.

When performing **BLIP-Diffusion**, to achieve the control over the pose of objects, the recommended paradigm¹ that integrating with ControlNet is used. Concretely, we use the original condition image as the input to subject encoder, and the canny map of it as the input to ControlNet.

For **SD-Inpainting**, we use the implementation provided by Stable Diffusion WebUI². Masked content is set to *latent nothing*, inpaint area is set to *whole picture*, and denoising strength is 1.

For **InstructPix2Pix**, we use the online demo³ provided by the original authors. Text CFG and image CFG are set to 7.5 and 1.5 according to the recommendation. The editing instruction is given as “Change the background to...”.

For **ControlNet**, we use our dataset to train it based on its original codebase⁴. The base network is Stable Diffusion V2.1. Except for the utilization of \mathcal{L}_{saca} , all the training hyperparameters are same as the proposed method.

5 MORE VISUAL RESULTS

This part provides more visual illustrations about the experiments. Figure 5,6 show more results about the application effect of scene diffusion. Figure 7 shows more results about the comparison with existing alternatives. Figure 8 shows the original images of the results exhibited in the ablation study of main text. Figure 9,10 show more results about the expanded applications of the proposed method.

¹https://huggingface.co/docs/diffusers/main/en/api/pipelines/blip_diffusion

²<https://github.com/AUTOMATIC1111/stable-diffusion-webui>

³<https://huggingface.co/spaces/timbrooks/instruct-pix2pix>

⁴<https://github.com/llyasviel/ControlNet>

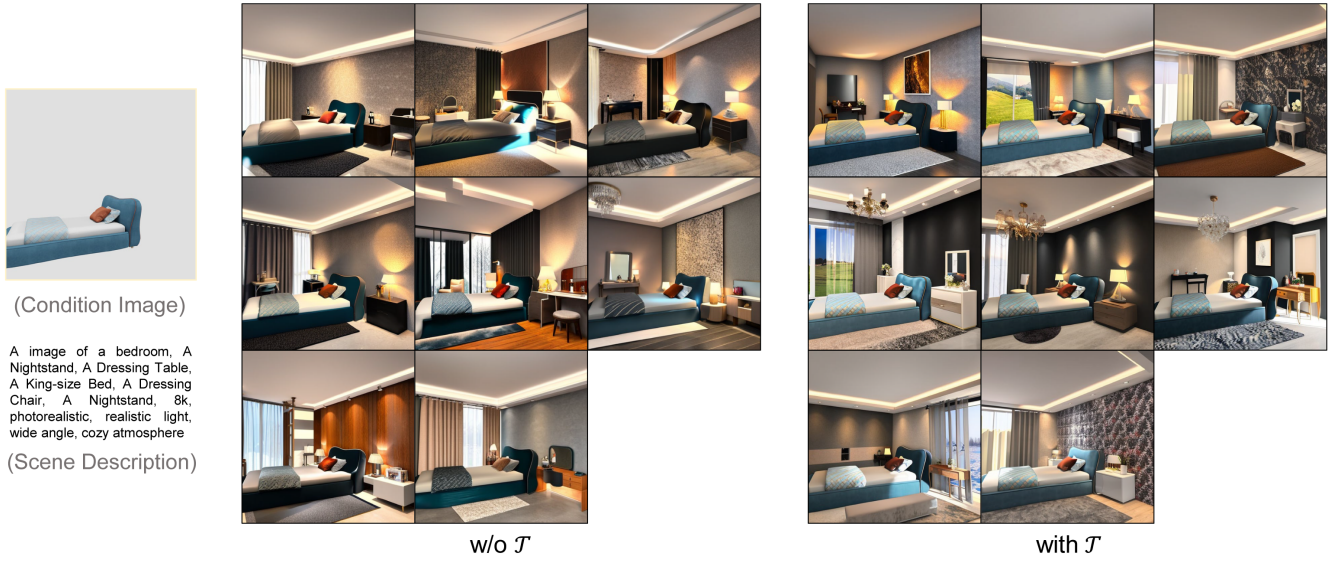


Figure 1: The comparison between the setups with or without the shading adaptive transformation \mathcal{T} . Each group of images are produced in the same mini-batch.

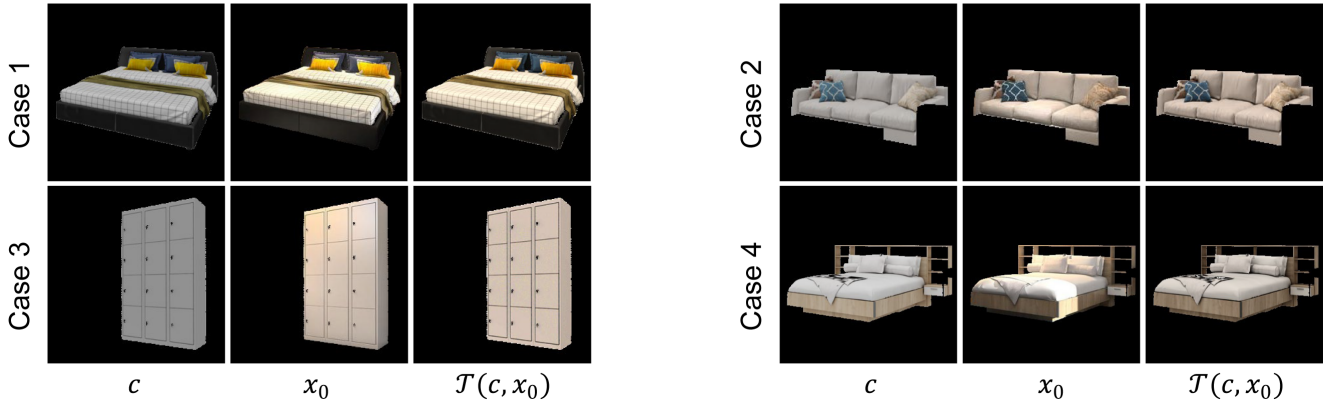


Figure 2: The effect of shading adaptive transformation \mathcal{T} . Since \mathcal{T} is solely conducted on the object areas, the background areas in these images are omitted.

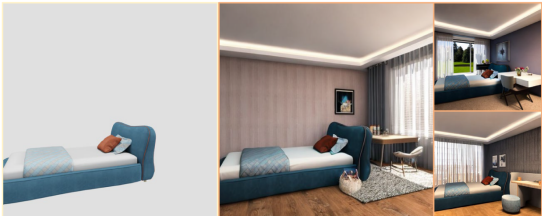


Figure 3: The comparison between the sequences of predicted \hat{x}_0 with or without FPTs.

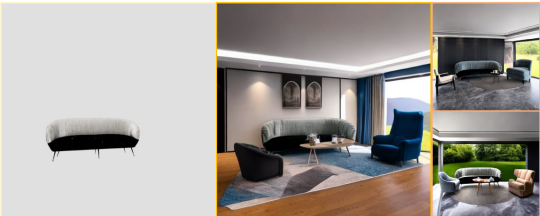


Figure 4: The examples of constructed condition-text-output data.

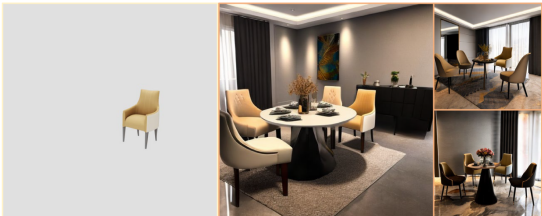
Single-object



A detailed depiction of a bedroom scene, showcasing a single bed, a desk, and an office chair, each positioned to complement the room's overall ambience.



A living room composition, showcasing a loveseat sofa and an armchair, arranged to create a harmonious and inviting space.



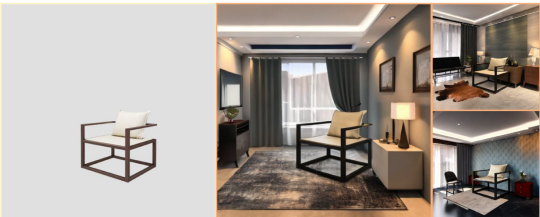
A living room scene, featuring dining chairs and an accompanying dining table, arranged to evoke a welcoming atmosphere.



A living room visualized with a table, a corner table, a loveseat sofa, and a sofastool, each element thoughtfully positioned to craft a cohesive and inviting atmosphere.



A living room scene, elegantly set with a dining table, dining chairs, and a sideboard, curated to blend seamless functionality with sophisticated style.

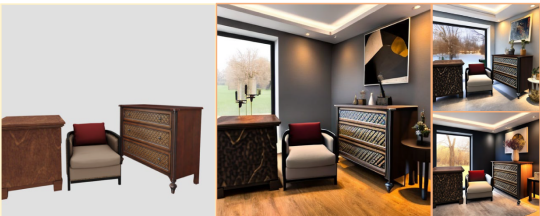


A depiction of a living room furnished with an office chair, a TV stand, a nightstand, and a corner table, arranged to balance functionality with aesthetic appeal.

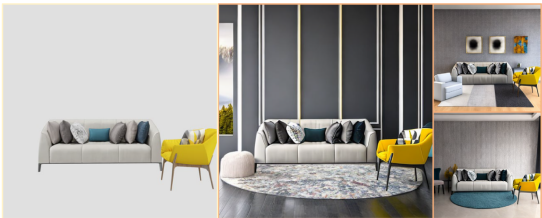
Multi-object



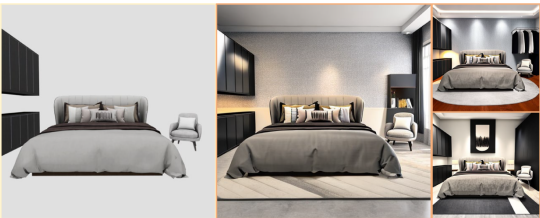
A living room setting with a king-size bed as the focal point, accompanied by a TV stand and a nightstand.



A living room setting, highlighted by an armchair, a corner cabinet and a corner table, meticulously placed to enhance the space's comfort and aesthetics.



A living room featuring a multi-seat sofa and an armchair, arranged for cozy gatherings.



A bedroom with a king-size bed, wardrobe, office chair, and white walls, arranged for elegance and comfort.

Figure 5: More results about the application effect of Scene Diffusion in Single-object and Multi-object scenarios.

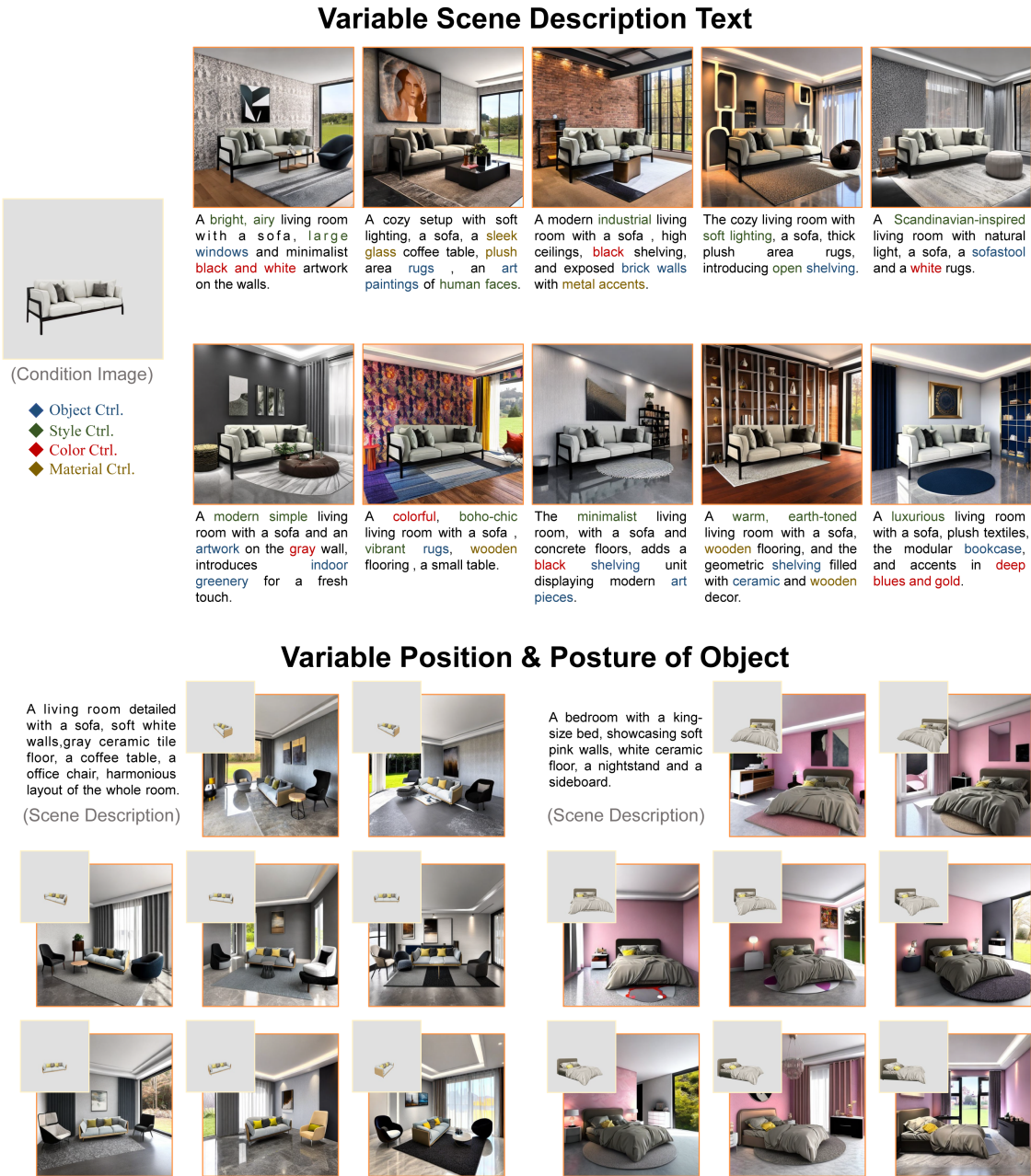


Figure 6: More results about the application effect of Scene Diffusion in Variable Scene Description Text and Variable Position & Pose of Object scenarios.

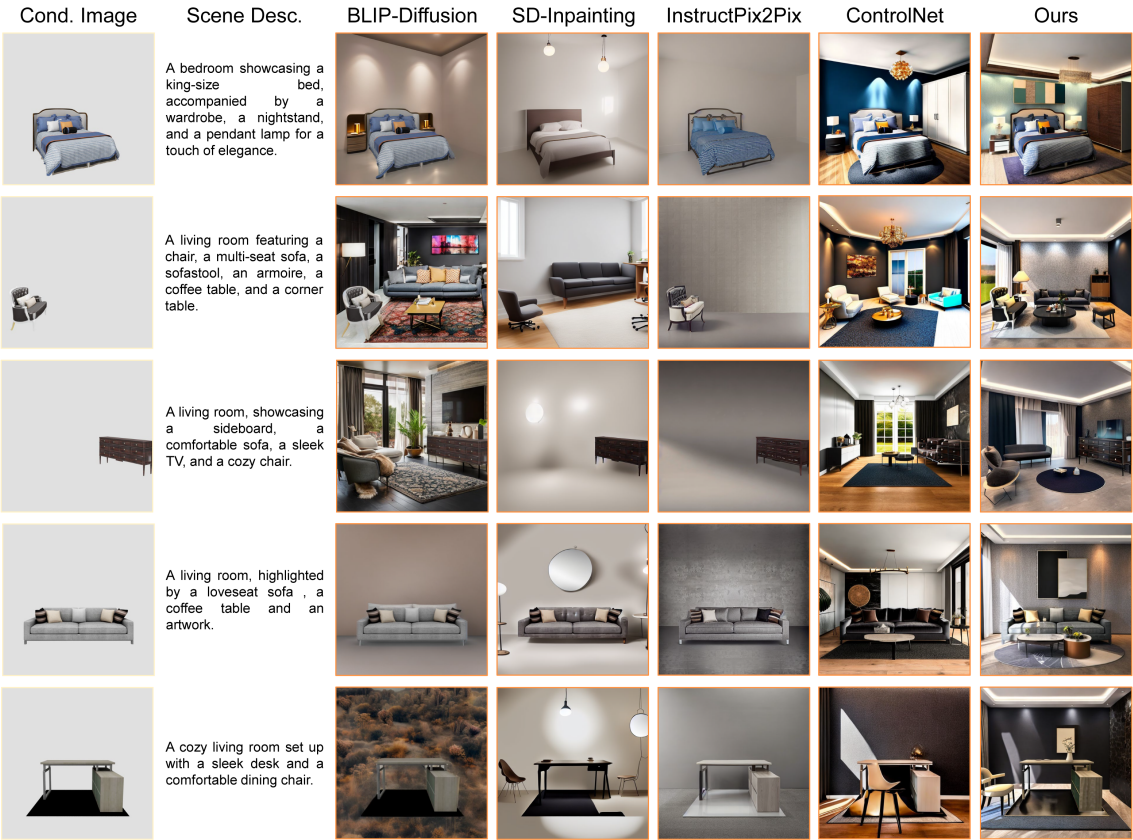


Figure 7: More results about the comparison with the existing alternatives.

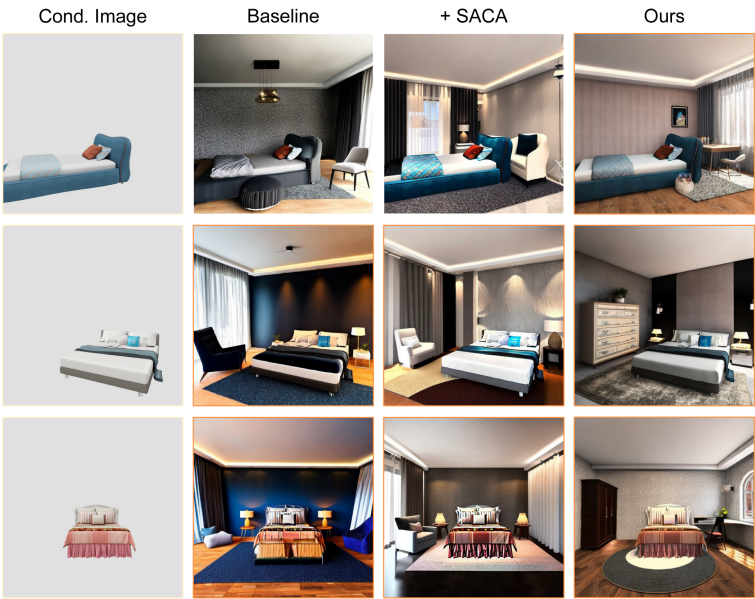


Figure 8: The original images of the results exhibited in the ablation study of main text.



Figure 9: More results about the expanded application of Scene Diffusion in Integrating with Existing ControlNet.



Figure 10: More results about the expanded application of Scene Diffusion in Generalizing to Real Image Fragment.