CustomTex: High-fidelity Indoor Scene Texturing via Multi-Reference Customization

Supplementary Material

A. Implementation Details

We introduce more implementation details about our method. The image super-resolution model used in distillation has an upscale factor of 1 and operates without color alignment. For LoRA injection, we target the key (to_k), query (to_q), value (to_v) and output projection (to_out) modules within the attention mechanism of the UNet in Stable Diffusion 1.5. We use a LoRA rank and lora_alpha of 4, with dropout disabled and the bias terms of the target modules left unchanged. All other settings follow PEFT defaults. As a result, the number of trainable parameters in the LoRA modules is 797,184, accounting for only 0.09% of the total model parameters.

B. Supplementary Experiments

B.1. Ablation Study

Can we directly conduct image super-resolution to the generated textures as a post-processing step? To investigate this, we evaluate an important variant of our proposed CustomTex, which does not incorporate pixel-level distillation during optimization, but performs image super-resolution [3] on the generated textures as a post-processing step. We denote this variant as *post-SR*. The quantitative evaluation results presented in Tab. 1 indicate that this variant leads to a significant drop in two image quality assessment metrics (Q-Align IQA and IAA). This finding is further corroborated by the qualitative comparison in Fig. 1, which shows that the textures produced by *post-SR* contain obvious blurriness, noise and artifacts, whereas our method yields better sharpness and clarity. Hence, we conclude that incorporating the image super-resolution model into the distillation process is more effective for improving texture quality than using it as a post-processing step.

Fig. 2 compares the UV textures before and after the super-resolution operation, demonstrating that the process does not improve the texture quality. This is because UV textures, unlike the rendered images in Fig. 1, typically lack the high-level and middle-level semantic structures (e.g., distinct objects, object parts, or surface textures) found in natural images. Since most super-resolution models were trained on natural images, they are ill-suited for UV textures and fail to produce satisfactory results when applied directly.

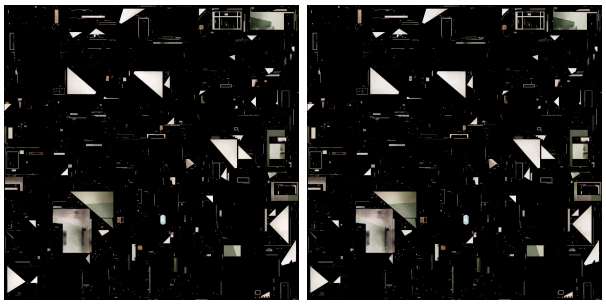
Another potential super-resolution approach is to apply the process to the rendered images rather than the textures themselves. While this can enhance the immediate visual output, it does not improve the intrinsic quality of the textures. Acquiring high-quality textures is essential, as many downstream applications rely on them as fundamental 3D assets alongside the mesh.

Method	CLIP-I↑	CLIP-FID↓	Q-Align IQA↑	Q-Align IAA↑
post-SR	0.746	114.612	2.959	2.190
full model	**0.797**	**106.229**	**4.469**	**3.629**

Table 1. Quantitative ablation study results.

Figure 1. Qualitative ablation study results.

Why can our method produce textures with less "baked-in" shading? This is primarily attributed to the multi-reference conditioning mechanism. In the main paper, we have conducted an ablation study on multi-reference input, which is denoted as *w/o multi-ref*. In this setting, we concatenate all reference images into one large composite input. As shown in Fig. 3, this ablated setting introduces obvious shading on the generated textures, such as walls and floors. The Stable Diffusion model often introduces strong, global shading to enhance perceived realism. By using instance masks, our method decomposes the global



before super-resolution after super-resolution

Figure 2. Textures before and after image super-resolution operation. These two textures have a resolution of $4,096 \times 4,096$, and correspond to the scene shown in the first row of Fig. 1. Black pixels in textures denote invalid regions where UV coordinates do not map to any vertex in the mesh.

w/o multi-ref full model

Figure 3. Qualitative ablation study results.

image generation process into the generation of local object appearances, thereby preventing the formation of overarching shading across the entire image.

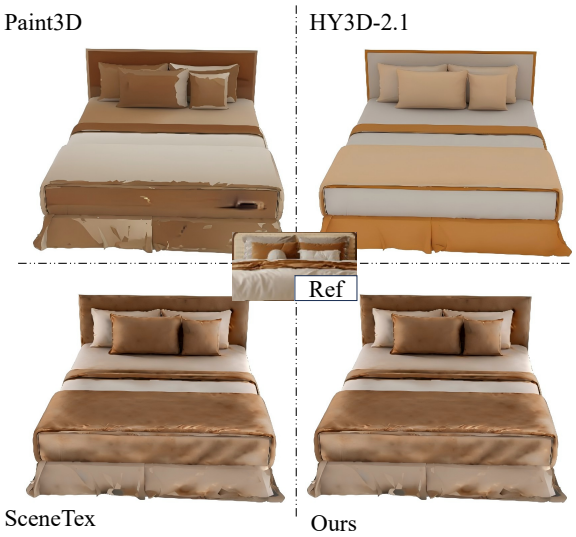


Figure 4. Single-object texture generation without image stitching.

Texture resolution	Inference time (s)
$1,024 \times 1,024$	0.15
$2,048 \times 2,048$	0.54
$3,072 \times 3,072$	1.19
$4,096 \times 4,096$	2.41
$8,192 \times 8,192$	9.69
$12,288 \times 12,288$	21.69

Table 2. Inference time across different texture resolutions

B.2. Additional Comparisons

Comparison on single-object texturing. In Fig. 4, we present a comparison on single-object texture generation (without image stitching) task. The style of the texture generated by our method is still more consistent with the reference. Paint3D [6] and HY3D-2.1 [4] perform well in single-object texturing but struggles with scene texturing.

Comparison with closed-source methods. In the main paper, we have compared our method against existing texture synthesis methods for which code or pre-trained models are publicly available. We do not include two most recent works, InstanceTex [5] and RoomPainter [2], as their implementations have not been released. To enable visual comparison, we reproduce their results using the images presented in their original papers and retrieve the most visually similar object images from online repositories to serve as reference images for our method. As shown in Fig. 5, the textures generated by our method exhibit greater visual richness and realism than those produced by the two competitors.

Figure 5. Visual comparison with InstanceTex [5] and RoomPainter [2].

Figure 6. Qualitative results on challenging scenes with diverse room types and complex layouts.

B.3. Inference Time

By leveraging a multi-resolution hash grid for implicit texture representation, our method generates texture maps at arbitrary resolutions. We further enhance inference efficiency through a patch-based texture generation strategy. As a result, our method produces textures up to 4K within seconds and maintains high efficiency even at 12K resolution, requiring only about 20 seconds. Complete inference speeds across various resolutions are detailed in Tab. 2.

Method	Visual Quality↑	Prompts Consistency↑
Paint3D [6]	2.792	2.817
HY3D-2.1 [4]	2.525	2.619
SceneTex-IPA [1]	3.842	3.617
CustomTex (Ours)	**4.008**	**4.125**

Table 3. User study results on visual quality and prompt consistency, averaged over 60 participants on a 1–5 scale. Our method achieves the highest scores across both criteria.

B.4. User Study.

We conducted a user study to further evaluate CustomTex in terms of perceived visual quality and consistency. A total of 60 participants rated each generated texture on visual quality, including clarity, color and composition, as well as the consistency between the texture content and the image prompt. Ratings were provided on a 1-to-5 scale, where 1 indicates "very poor" or "completely inconsistent", and 5 indicates "very good" or "highly consistent". The results reported in Tab. 3 indicate that our method is consistently preferred.

C. More Visual Results

Fig. 6 presents visual results produced by our method on two challenging scenes, a kitchen and an office, which feature diverse layouts and materials. These results indicate that our method performs well in complex scene texturing. Fig. 7-10 show the $2,000 \times 2,000$ resolution images rendered from our generated textures in two scenes "living room" and "bedroom". Our method produces visually compelling textures with significantly reduced blurriness, artifacts and "baked-in" shading. Fig. 11 shows that our method successfully generalizes to various exaggerated styles beyond furniture. The results, generated from reference images including "dark", "Van Gogh" and "Cyberpunk" styles, confirm its broad applicability. More visual results are presented in the supplementary video.

References

[1] Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Scenetex: High-quality texture synthesis for indoor scenes via diffusion priors. In *Proc. of CVPR*, pages 21081–21091, 2024. 4

[2] Zhipeng Huang, Wangbo Yu, Xinhua Cheng, ChengShu Zhao, Yunyang Ge, Mingyi Guo, Li Yuan, and Yonghong Tian. Roompainter: View-integrated diffusion for consistent indoor scene texturing. In *Proc. of CVPR*, pages 574–584, 2025. 2, 3

[3] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proc. of CVPR*, pages 2333–2343, 2025. 1

[4] Tencent Hunyuan3D Team. Hunyuan3d 2.1: From images to high-fidelity 3d assets with production-ready pbr material. *arXiv preprint arXiv:2506.15442*, 2025. 2, 4

[5] Mingxin Yang, Jianwei Guo, Yuzhi Chen, Lan Chen, Pu Li, Zhanglin Cheng, Xiaopeng Zhang, and Hui Huang. Instancetex: Instance-level controllable texture synthesis for 3d scenes via diffusion priors. In *Proc. of SIGGRAPH Asia*, pages 59:1–59:11, 2024. 2, 3

[6] Xianfang Zeng, Xin Chen, Zhongqi Qi, Wen Liu, Zibo Zhao, Zhibin Wang, Bin Fu, Yong Liu, and Gang Yu. Paint3d: Paint anything 3d with lighting-less texture diffusion models. In *Proc. of CVPR*, pages 4252–4262, 2024. 2, 4

Figure 7. The “living room" texture generated by our method is rendered into $2,000 \times 2,000$ resolution image.

Figure 8. The “living room" texture generated by our method is rendered into $2,000 \times 2,000$ resolution image.

Figure 9. The “bedroom" texture generated by our method is rendered into $2,000 \times 2,000$ resolution image.

Figure 10. The “bedroom" texture generated by our method is rendered into $2,000 \times 2,000$ resolution image.

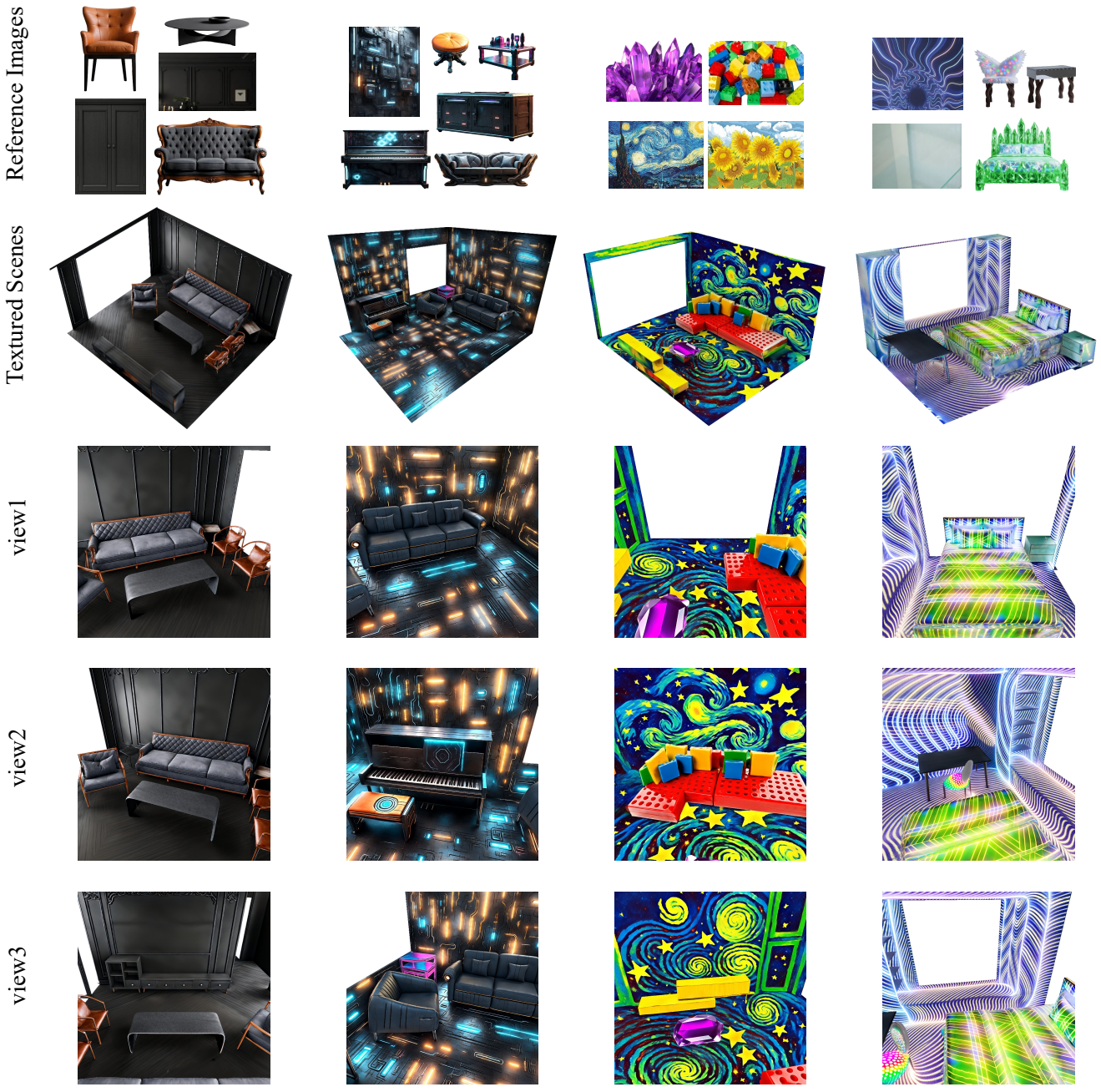


Figure 11. Using images with more diverse styles as the reference images.