InstanceTex: Instance-level Controllable Texture Synthesis for 3D Scenes via Diffusion Priors (Supplementary)

MINGXIN YANG, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
JIANWEI GUO*, MAIS, Institute of Automation, Chinese Academy of Sciences, China
YUZHI CHEN, School of Artificial Intelligence, University of Chinese Academy of Sciences, China
LAN CHEN, MAIS, Institute of Automation, Chinese Academy of Sciences, China
PU LI, MAIS, Institute of Automation, Chinese Academy of Sciences, China
ZHANGLIN CHENG*, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China
XIAOPENG ZHANG, MAIS, Institute of Automation, Chinese Academy of Sciences, China
HUI HUANG, Shenzhen University, China

ACM Reference Format:
Mingxin Yang, Jianwei Guo, Yuzhi Chen, Lan Chen, Pu Li, Zhanglin Cheng, Xiaopeng Zhang, and Hui Huang. 2024. InstanceTex: Instance-level Controllable Texture Synthesis for 3D Scenes via Diffusion Priors (Supplementary). In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3680528.3687633

In this document, we first provide more details about implementing our InstanceTex. Then we present more texturing results, ablation studies, and comparisons.

1 MORE TECHNICAL DETAILS

ControlNet Fine Tuning. To extend the original ControlNet in alignment with InstanceDiffusion, we refine the depth-based ControlNet and the lineart-based ControlNet with 10, 000 images we collected from both indoor and outdoor scenes. For the instance level captions, follow the data generation scheme proposed InstanceDiffusion [Wang et al. 2024], we utilize Grounded-SAM [Kirillov et al. 2023; Liu et al. 2023b] to generate the Bounding boxes and BLIP-V2 [Li et al. 2023] to generate distinct instance prompts.

Since 3D bounding box information is required to train the pose-aware position map ControlNet, we generate a dataset by randomly selecting and arranging multiple objects from Objaverse dataset [Deitke et al. 2023]. We generate a dataset containing 50, 000 images with ground-truth 3D position maps to train ControlNet from scratch. We set the learning rate following the official implementation of ControlNet as $1e-5$ and 50000 iterations.

*Corresponding authors: Jianwei Guo (jianwei.guo@nlpr.ia.ac.cn), Zhanglin Cheng (zl.cheng@siat.ac.cn)

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

ACM ISBN 979-8-4007-1131-2/24/12...$15.00
https://doi.org/10.1145/3680528.3687633

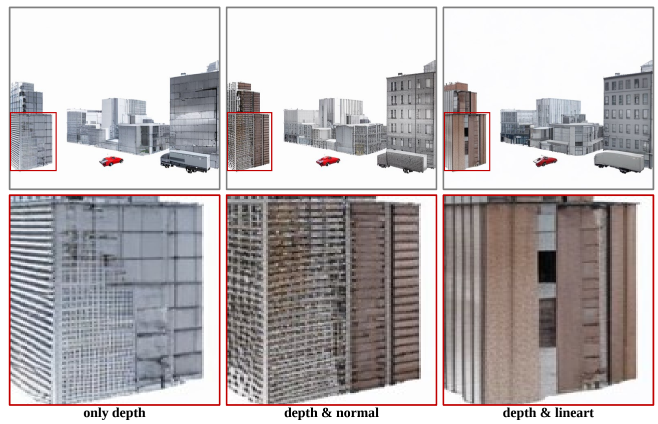


Fig. 1. Ablation study on the different geometric conditions, including depth-only, depth with normal, and depth with lineart.

Handling distant objects and occlusions. When texturing scene meshes, we encountered significant occlusions during the texturing process, resulting in disordered and unaesthetic textures. Additionally, when viewed from certain angles, distant objects will cause textures that are not faithful to the mesh geometry. Inspired by the two-stage inpainting paradigm, we have devised an effective technique to address these challenges. For distant or occluded objects viewed from a single viewpoint, we choose not to unwrap the object's texture into an RGB UV texture, but instead utilize only the latent UV texture. This approach is motivated by the assumption that textures generated for distant objects tend to be inaccurate and, as such, are not suitable to be unwrapped to the RGB texture. On the other hand, the inaccurate latent UV texture, which represents the latent of the early stage of denoising, has a reduced impact on the generated RGB texture. Therefore, it still can be effectively leveraged for inpainting other viewpoints. Benefiting from this scheme, our approach is further relieved from the disordered texture caused by distant objects and occlusions.

In our implementation, we establish an object projected area threshold of 80×80 pixels within a 512×512 image to determine whether an object is distant/occluded.

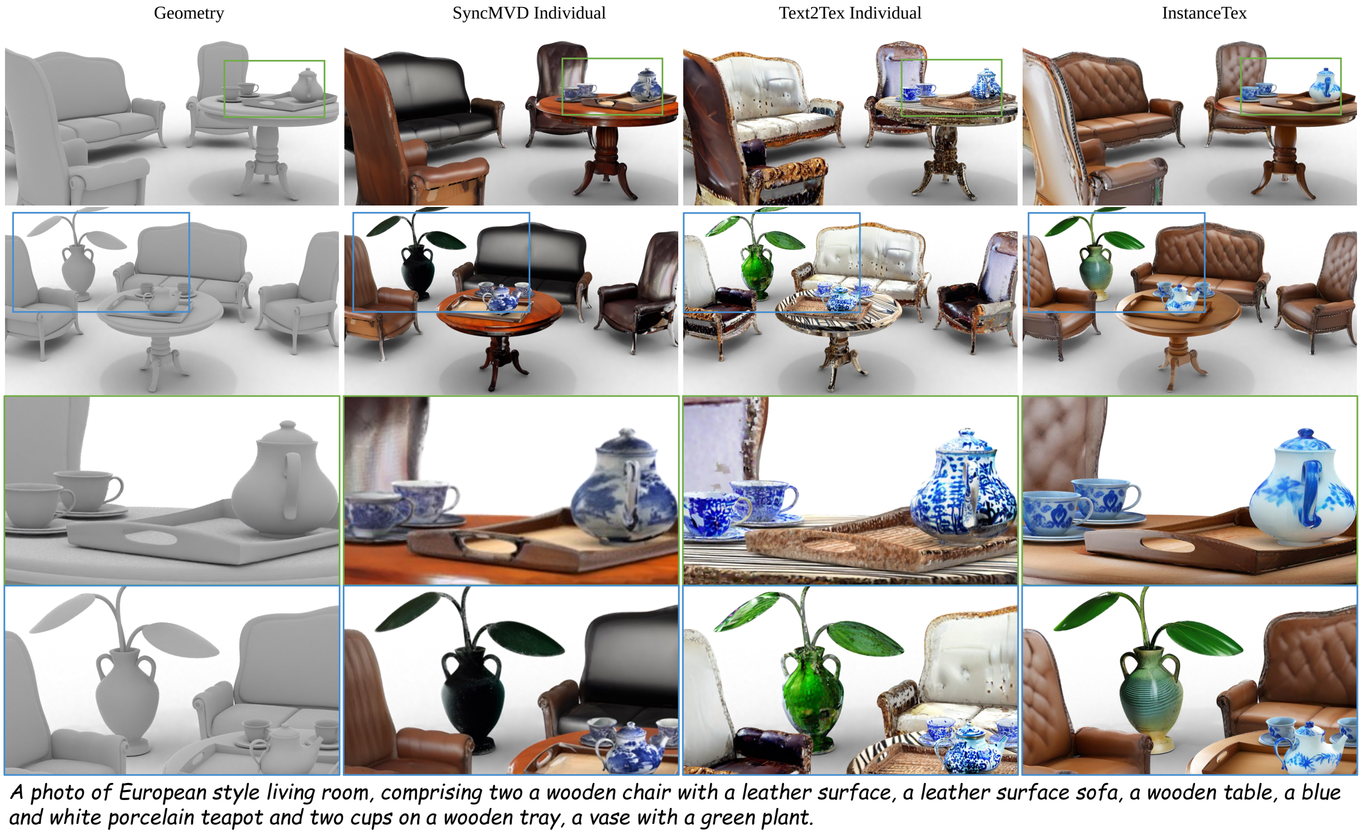


Fig. 2. Full visual comparison of scene texture generation on an indoor scene (*Furniture-1*) with two individual texturing baselines SyncMVD [Liu et al. 2023a] and Text2Tex [Chen et al. 2023].

2 ABLATION STUDY OF LINEART CONDITION

Lineart-based ControlNet is a robust and widely used tool in 2D image generation, particularly for outdoor buildings and indoor room scenes, which represent most cases in our experiments. We found that using only the depth map as the geometric condition for the scene mesh often results in images that fail to capture precise local geometric details, leading to inconsistent and disordered textures. Therefore, we integrate lineart as an additional condition in our pipeline to provide accurate local geometric cues for texture generation.

As illustrated in Fig. 1, incorporating lineart as an additional geometric condition yields better performance than using depth condition alone. As referenced in the zoom-in figure (indicated by the red boxes), our depth with lineart conditioned model demonstrates the most consistency compared to other conditions. We also evaluate the generated textures of commonly used normal maps as a condition for comparison, which also demonstrate inferior performance compared to our approach.

3 USER STUDY

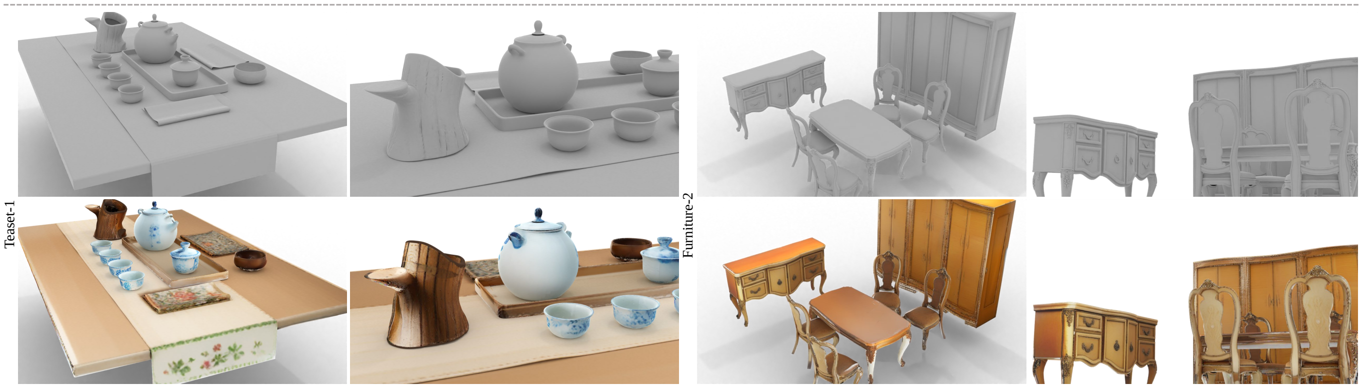
We conducted a user study to evaluate InstantTex against other compared methods, collecting feedback from 23 participants using the same set of questionnaires. To comprehensively evaluate texturing results, following TEXTure [Yu et al. 2023] and SceneTex [Chen et al. 2024], we designed 5 questions (listed in Table 1) to assess the texturing in terms of aesthetics, harmony, realism, and prompt-fidelity, respectively. Specifically, we designed two questionnaires on consistency: one to evaluate texture consistency within individual objects,

Table 1. The questionnaires of the scene texturing evaluation assess the aspects of aesthetics, harmony, realism, and prompt fidelity, respectively.

Please review the images then answer the following questions to rate texture quality on a scale of 1 to 5.	
(Q1)	Please review the generated images about the overall aesthetic appeal.
(Q2)	Please review the generated images about harmonious textures for each single object.
(Q3)	Please review the realism of the generated images.
(Q4)	Please review how the generated images match the description text.
(Q5)	Please review the generated images about harmonious textures for the whole scene.

(a). A photo of traditional style bedroom, comprising a bed with a floral pattern on the bedding, a dark wooden wardrobe with visible wood grain texture, two wooden bedsides, a chair upholstered in velvet fabric.

(b). A photo of a modern children's bedroom, comprising a single bed with a bright and colorful bedspread, two modern cylindrical bedside tables, a stylish cabinet, a modern closet filled with a variety of cloth in different colors, a cozy chair.



(c). A photo of a Chinese tea set table, comprising a set of porcelain tea cups with blue and white pattern , a porcelain tea kettle, a wooden table, a tablecloth with classical pattern, two traditional Chinese handkerchief, a wooden container and a wooden tea bowl, a wooden tray.

(d). A photo of a European style room, comprising four a wooden chair in Baroque style, a a wooden dinning table, a wooden low cabinet, a wooden wardrobe.

Fig. 3. Evaluation on the texture generation for challenging complex cases, where the scenes *EchoScene-1* and *EchoScene-2* are generated by EchoScene [Zhai et al. 2024], and *Teaset-1* and *Furniture-2* are randomly selected scenes with either intricate geometry or complex text prompts.

Table 2. We report User Study results for quantitative comparisons, including Visual Quality (VQ), which summarizes the first three question scores, Prompt Fidelity (PF) and Scene Consistency (SC) which reflects the last two question scores.

Method	VQ↑	PF↑	SC↑
TEXTure	1.39	1.19	2.19
Text2tex	2.45	1.12	3.88
SyncMVD	2.17	1.57	3.91
SceneTex (only indoors)	**4.37**	2.93	4.17
InstanceTex (Ours)	4.35	**4.54**	**4.35**

and another to assess scene texture consistency, which is often overlooked by approaches that focus solely on individual objects. Then we summarized the first three questions into the Visual Quality (VQ) criterion by averaging the score, and the last two questions into Prompt Fidelity (PF) and Scene Consistency (SC) criterion.

As shown in Table 2, benefiting from instance layout guidance, InstantTex achieves textures with high fidelity to the given prompt without losing the scene consistency, as reflected in the highest PF and SC scores. Our performance on VQ score is also significantly superior to the common baselines TEXTure [Yu et al. 2023], Text2tex [Chen et al. 2023], SyncMVD [Liu et al. 2023a] and achieves comparable results with SceneTex [Chen et al. 2024], an optimization-based approach which consumes much more time to converge.

4 TEXTURING COMPLEX SCENES

To validate InstanceTex's robustness and generality, we conduct a stress test on several challenging scenes: noisy meshes with irregular scene geometry, scene meshes comprising sets of objects with repeated elements, and complex text prompts. Initially, we utilize EchoScene [Zhai et al. 2024], a scene layout-based 3D indoor scene mesh generation approach, to produce the noisy meshes. Then we randomly select several generated scene meshes and manually obtain the corresponding scene layout. As illustrated in Fig. 3, the texturing performance of IntanceTex on noisy meshes remains stable, comparable to its performance on manually-created meshes with fine geometry. It verifies that InstanceTex does not impose any special requirements on input meshes, as depth and position maps are independent of vertex/face order. Furthermore, integrated with layout-guided scene generation approaches, InstanceTex can generate textured 3D assets in an end-to-end manner.

For more intricate scenes, we evaluate InstantTex on two additional cases: a tea set scene featuring repetitive elements like tea

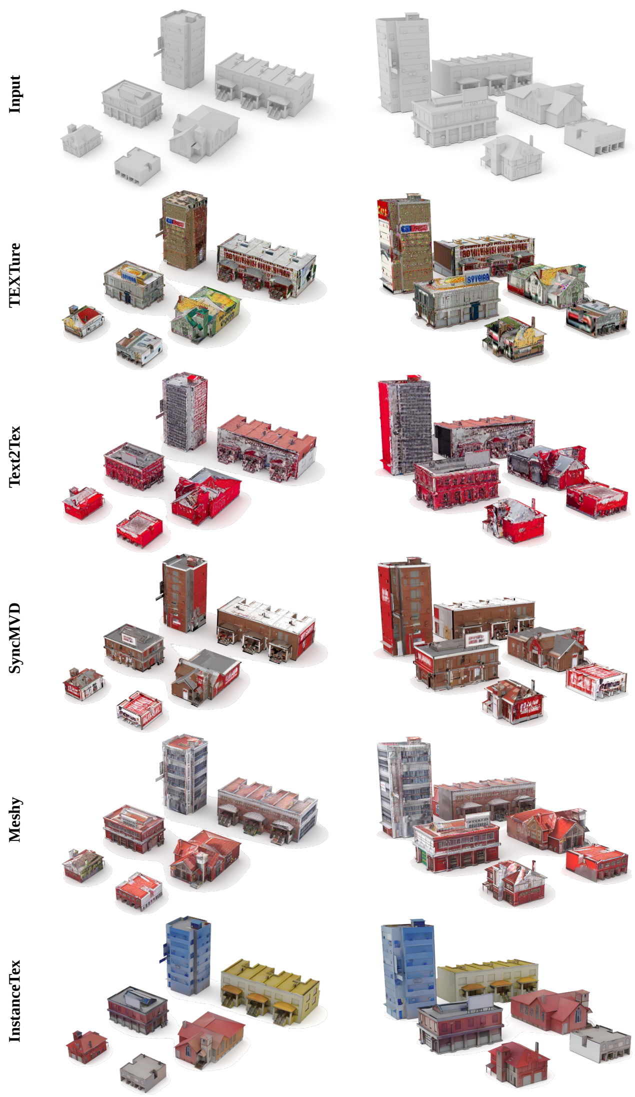


Fig. 4. Comparisons of texture generation for an outdoor scene, *Block-1*.

Fig. 5. Comparisons of texture generation for an outdoor scene, *Block-2*.

cups with the complex text prompt "blue-and-white porcelain", and a furniture set scene with multiple chairs and closets. As illustrated in Fig. 3 (c) and (d), InstantTex not only performs harmonious mesh texturing across distinct objects, but also produces textures with complex patterns that faithfully reflect input text prompts.

5 MORE VISUAL COMPARISON

For a detailed comparison, we show all of the outdoor and indoor results compared with TEXTure [Richardson et al. 2023], Text2tex [Chen et al. 2023], SyncMVD [Liu et al. 2023a], and Meshy, each of which exhibits two different views for better visualization, as well as the rendered input meshes. Fig. 4, Fig. 5, Fig. 6 and Fig. 7 are four cases of outdoor scenarios, named Block-1, Block-2, Block-3, and Garden. Those methods show multi-view inconsistency results (*e.g.,* the roof of the wooden cabin in Fig. 7) and over-fragmentation (*e.g.,* the exterior facades in Fig. 4). Besides, they are difficult to create and interpret prompts for instance control (*e.g.,* three villas with red roof with a white stone wall and a white warehouse in Fig. 5). In contrast, our InstanceTex achieves higher-quality texture, seamless, and instance-aware texturing with accurate semantic alignment to input prompts.

In addition, we show two multi-view renderings of generated texturing rooms from 3D-FRONT [Fu et al. 2021] in Fig. 8, as the indoor comparison. We add the results of SceneTex [Chen et al. 2024], a specially designed method for texting indoor scenes. From these comparisons, we can conclude that SceneTex is able to synthesize high-quality textures with overall coherent styles. However, it fails to match the input prompt correctly because it cannot precisely control the appearance of a target object. It is also worth noticing that SceneTex is an entirely optimization-based pipeline and takes more than 25 hours to coverage, whereas ours can synthesize textures with comparable quality with greater fidelity to the text prompt in around half an hour.

Fig. 6. Comparisons of texture generation for an outdoor scene, *Block-3*.

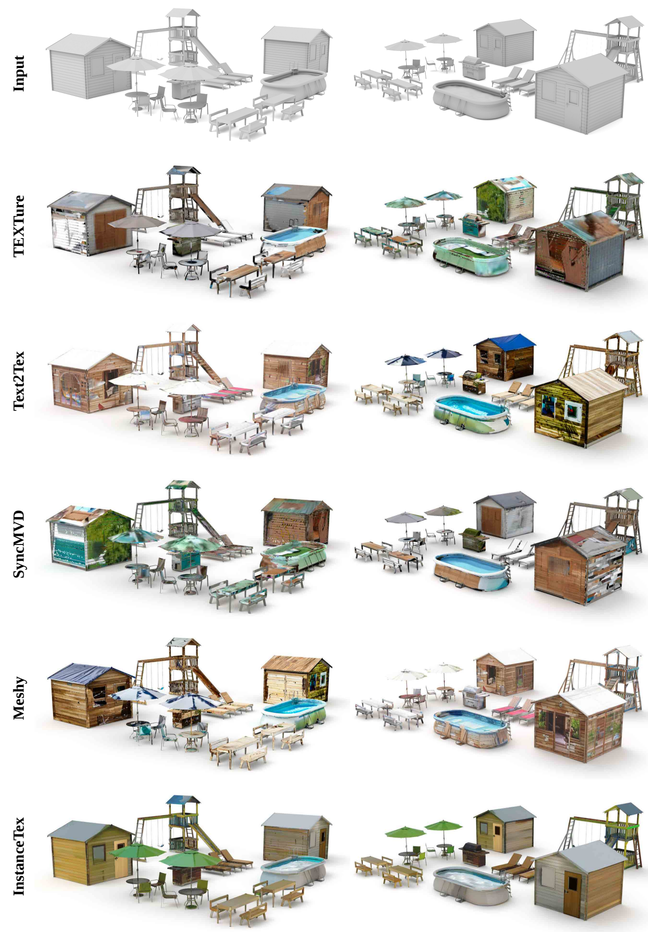


Fig. 7. Comparisons of texture generation for an outdoor scene, *Garden*.

6 QUANTITATIVE COMPARISON

We show the detailed quantitative comparison in Table 3 case by case in our dataset. The number of instances in each scene is also presented to indicate the complexity of each scene, ranging from 6 to 12. We first calculate the Fréchet Inception Distance (FID) [Heusel et al. 2017] and Kernel Inception Distance (KID) [Bińkowski et al. 2018], which measure the difference between the output distribution of the generated images of ControlNet and our textured objects under specified viewpoints. Then we utilize CLIP-Score [Hessel et al. 2021] to validate the congruence between our generated texture and the provided text prompts.

Our method outperforms prior methods on all metrics in each case by a significant margin. When the number of instances is similar, the FID and KID metrics of compared methods for outdoor scenes are higher than those for indoor scenes (*e.g.,* , Block-1/Block-2 vs. Room 6). However, our method maintains relatively stable metric values for both indoor and outdoor scenes. This demonstrates the superior capability of InstanceTex in generating realistic and high-fidelity textures across diverse scenes with different categories and styles. In particular, InstanceTex achieves nearly 13% and 31% improvements in CLIP-Score on indoor and outdoor scenes respectively, indicating our model's superiority in semantic-aligned texture generation.

REFERENCES

Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations.*

Dave Zhenyu Chen, Haoxuan Li, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2024. SceneTex: High-Quality Texture Synthesis for Indoor Scenes via Diffusion Priors. In *IEEE Computer Vision and Pattern Recognition (CVPR).*

Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. 2023. Text2tex: Text-driven texture synthesis via diffusion models. In *IEEE International Conference on Computer Vision (ICCV).*

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. 2023. Objaverse: A universe of annotated 3d objects. In *IEEE Computer Vision and Pattern Recognition (CVPR).* 13142–13153.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *IEEE International Conference on Computer Vision (ICCV).* 10933–10942.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP.*

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

Fig. 8. Viusal comparison of texture generation for two indoor scenes, *Room3* and *Room6*.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *IEEE International Conference on Computer Vision (ICCV)*. 4015–4026.

Table 3. Quantitative evaluation and comparison on texture generation on seven synthetic datasets. For the comparison of texture quality, Fréchet Inception Distance (FID) [Heusel et al. 2017] and Kernel Inception Distance (KID) [Bińkowski et al. 2018] are recorded. CLIP-Score [Hessel et al. 2021] is calculated to validate the alignment between the generated texture and the provided text prompts. Statistics including the number of instances (#Instances) are shown for every scene example. The first- and second-place performances are highlighted using bold and italic fonts, respectively.

Scene	#Instances	Method	FID↓	KID↓	CLIP Score↑
Room-3	10	TEXTure	95.61	6.89	18.23
		Text2tex	94.37	7.98	21.62
		SyncMVD	89.32	6.89	*23.38*
		Meshy	99.8	8.47	22.61
		SceneTex	*88.76*	*6.38*	23.18
		Ours	**86.51**	**6.53**	**26.47**
Room-6	7	TEXTure	113.57	9.48	20.64
		Text2tex	98.73	7.25	23.64
		SyncMVD	92.15	8.07	21.57
		Meshy	*88.31*	6.75	23.45
		SceneTex	94.35	*6.74*	*24.19*
		Ours	**83.47**	**5.95**	**26.93**
Garden	12	TEXTure	111.45	9.18	16.42
		Text2tex	108.49	9.56	19.71
		SyncMVD	128.45	10.32	*22.15*
		Meshy	*85.23*	*6.57*	20.58
		SceneTex	/	/	/
		Ours	**88.18**	**6.47**	**28.15**
Block-1	6	TEXTure	137.24	10.79	18.22
		Text2tex	107.17	9.47	18.27
		SyncMVD	103.75	9.12	*21.57*
		Meshy	*93.25*	*8.19*	21.07
		SceneTex	/	/	/
		Ours	**76.35**	**5.39**	**28.19**
Block-2	8	TEXTure	119.57	9.74	21.17
		Text2tex	117.96	9.59	18.35
		SyncMVD	96.35	9.75	21.57
		Meshy	*89.45*	*7.92*	*22.19*
		SceneTex	/	/	/
		Ours	**84.21**	**6.04**	**27.37**
Block-3	7	TEXTure	137.82	11.25	19.62
		Text2tex	105.27	9.42	22.15
		SyncMVD	93.13	8.04	21.17
		Meshy	*88.19*	*6.45*	*23.21*
		SceneTex	/	/	/
		Ours	**83.15**	**5.89**	**28.19**

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

Yuxin Liu, Minshan Xie, Hanyuan Liu, and Tien-Tsin Wong. 2023a. Text-Guided Texturing by Synchronized Multi-View Diffusion. *arXiv preprint arXiv:2311.12891* (2023).

Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. 2023. TEXTure: Text-Guided Texturing of 3D Shapes. In *ACM SIGGRAPH 2023 Conference*

Proceedings. Article 54, 11 pages.
Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. 2024. InstanceDiffusion: Instance-level Control for Image Generation. In *IEEE Computer Vision and Pattern Recognition (CVPR)*.
Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Zhengzhe Liu, and Xiaojuan Qi. 2023. Texture Generation on 3D Meshes with Point-UV Diffusion. In *IEEE International Conference on Computer Vision (ICCV)*. 4206–4216.
Guangyao Zhai, Evin Pınar Örnek, Dave Zhenyu Chen, Ruotong Liao, Yan Di, Nassir Navab, Federico Tombari, and Benjamin Busam. 2024. EchoScene: Indoor Scene Generation via Information Echo over Scene Graph Diffusion. *ECCV* (2024).