Prometheus: 3D-Aware Latent Diffusion Models for Feed-Forward Text-to-3D Scene Generation

Supplementary Material

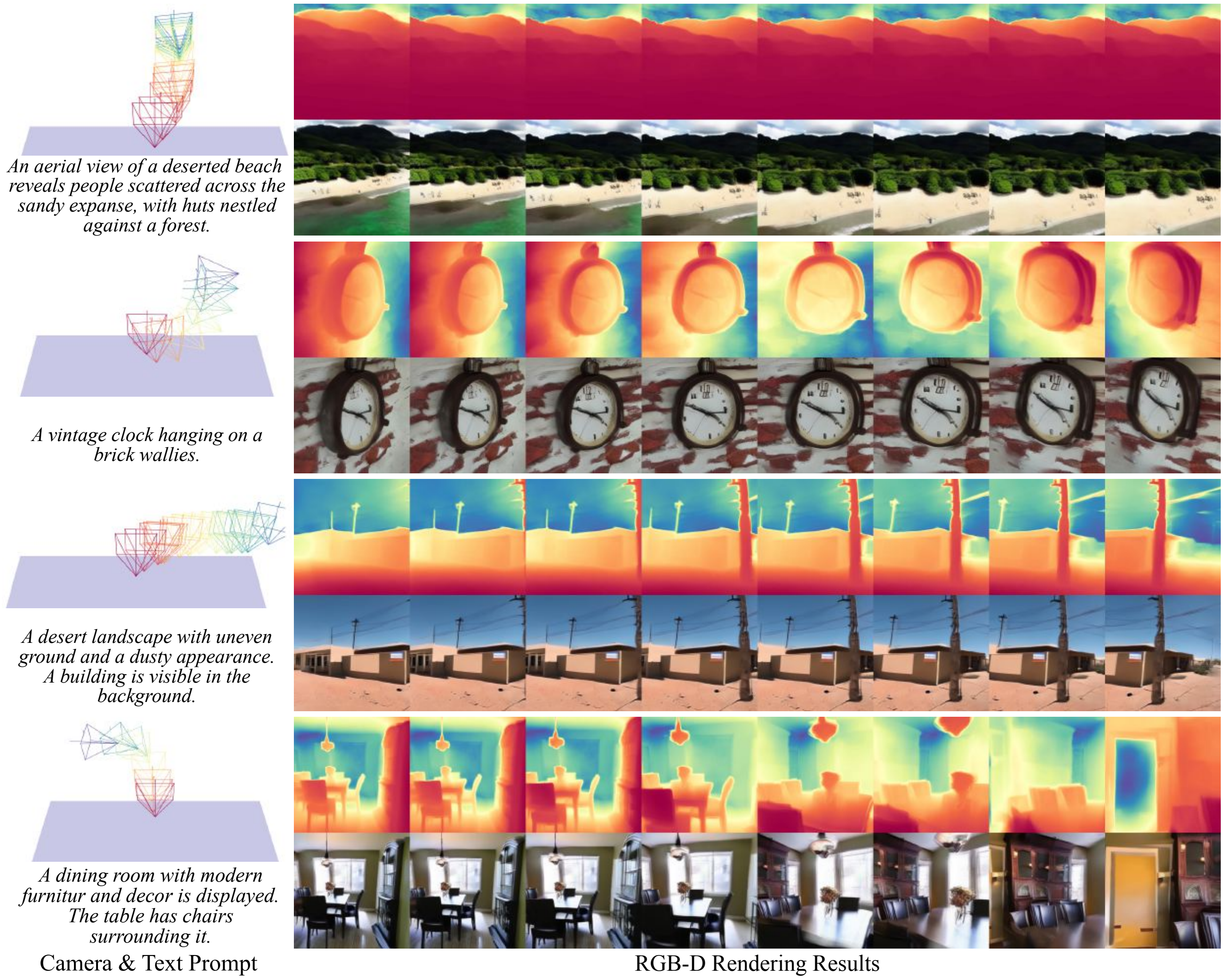

An aerial view of a deserted beach reveals people scattered across the sandy expanse, with huts nestled against a forest.
A vintage clock hanging on a brick wallies.
A desert landscape with uneven ground and a dusty appearance. A building is visible in the background.
A dining room with modern furnitur and decor is displayed. The table has chairs surrounding it.
Camera & Text Prompt
RGB-D Rendering Results

Figure 1. **More Results.** Our method can synthesize diverse results across multiple domains, taking text prompts and camera poses as input. As shown in the image, we can render diverse (indoor/outdoor/object-centric) scenes that are faithfully aligned with the given text prompt and camera trajectory, while maintaining good underlying geometry.

1. More Generation Results

We present additional generation results of multi-view images and depth maps across diverse text prompts and camera trajectories in Fig. 1. These results underscore the robustness of our approach in managing both object-level and scene-level prompts for 3D scene generation. Then we present more scene-level generation comparison results with Director3D, the concurrent scene-level feedforward text-to-3D method, as shown in Fig. 2.

2. More Evaluation Metrics

We adopt BRISQUE and NIQE, as used by Director3D [6], for a fair comparison. These metrics are chosen because the text prompts for evaluation are generated by LLMs with no reference images. To enable a more comprehensive comparison with no-background methods, we randomly se-

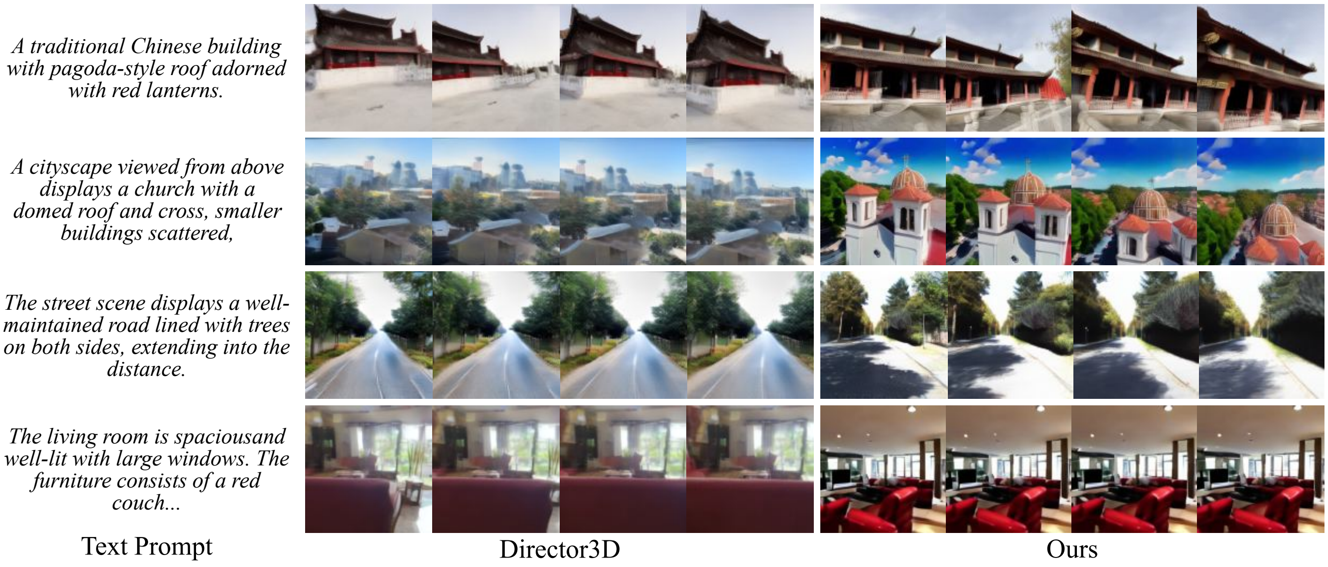

A traditional Chinese building with pagoda-style roof adorned with red lanterns.
A cityscape viewed from above displays a church with a domed roof and cross, smaller buildings scattered,
The street scene displays a well-maintained road lined with trees on both sides, extending into the distance.
The living room is spaciousand well-lit with large windows. The furniture consists of a red couch...
Text Prompt
Director3D
Ours

Figure 2. **Qualitative comparison with Director3D.** We compare `Prometheus` against baselines under varying difficulty settings. As overlap gradually decreases, the advantages of our method continue to grow. Moreover, as shown in the depth map, our method exhibits superior geometry quality across all settings.

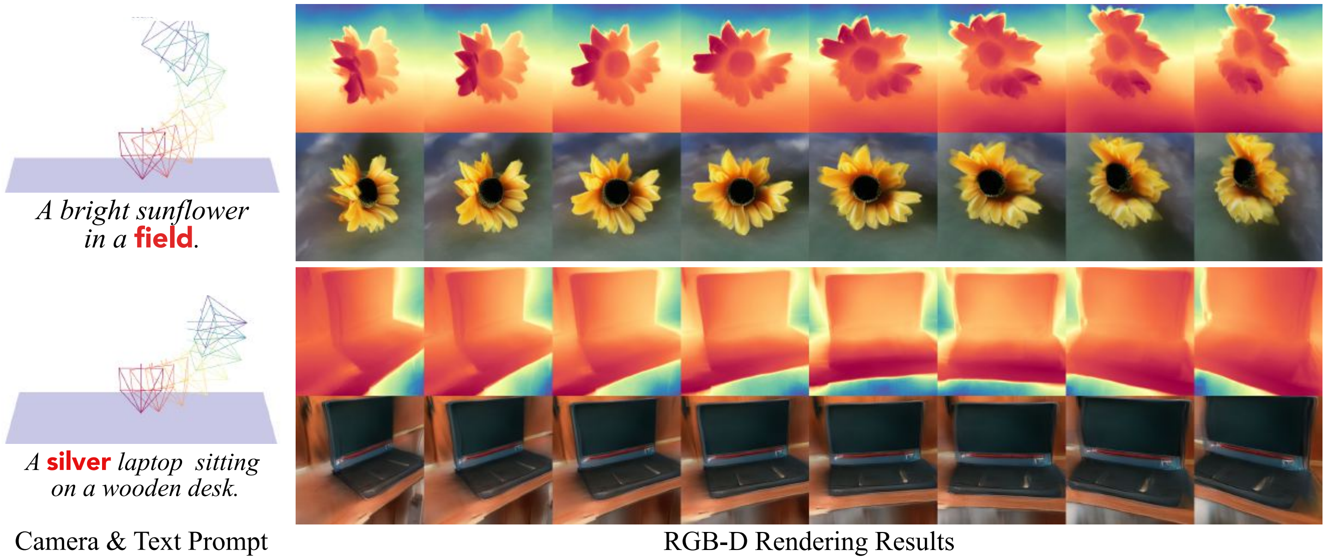

A bright sunflower in a field.
A silver laptop sitting on a wooden desk.
Camera & Text Prompt
RGB-D Rendering Results

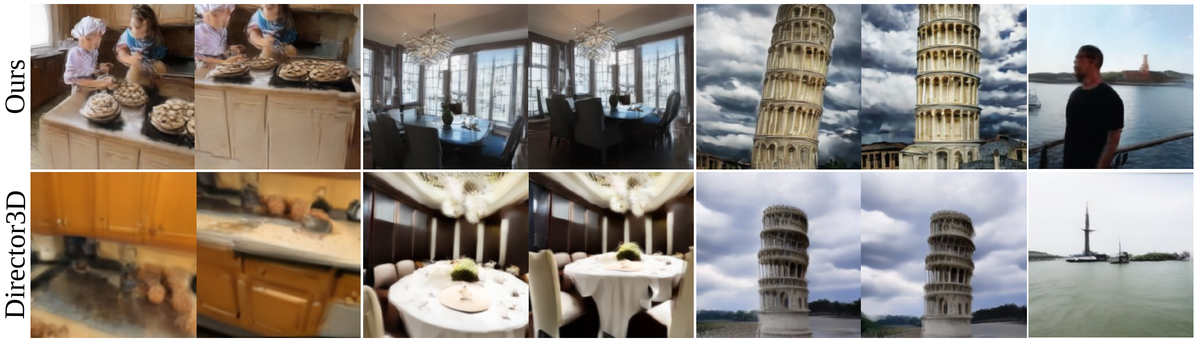
Figure 3. **Multiview-inconsistency cases.** We show Multiview-inconsistency, the main factor contributing to the failure cases of our method. As shown in the images, due to the lack of explicit 3D representation during multiview generation in latent space, `Prometheus` will encounter view inconsistency under large rotations or extreme viewpoints.

lected 200 sample prompts from Objaverse [3] to generate background-free scenes and sampled 8 views per scene. We evaluated our performance using FID [5] and compared it with MVDream+LGM [8, 9]. As shown in Table 1, our method achieves superior image fidelity in terms of FID.

Due to our work focusing more on in-the-wild scene-level generation, thus we also computed FID on out-of-domain prompts using 600 scenes from the LSUN [11] dataset. Tab. 1 and Fig. 4 show that our approach outperforms Director3D in generalization and image fidelity, benefiting from the larger scale of our training dataset. Our method is nearly 3x more efficient during inference due to its simple and intuitive design.

3. Number of Views During Inference

For GS-VAE, to save memory, we increase inference views via cross-view attention, validated in PixelSplat [1]. For MV-LDM, $N = 8$ is kept for training and inference. We



A mother and daughter are baking cookies together

A contemporary interior with a centrally placed table surrounded by four chairs

The Leaning Tower of Pisa, Italy is seen against a cloudy sky

A man in a black shirt stands near the water

Figure 4. **Qualitative comparison on LSUN** Benefiting from the larger scale of training dataset, `Prometheus` achieves superior generalization on the out-of-domain LSUN dataset and higher image fidelity compared to Director3D [6].

	MV. + LGM	Director3D	Ours
Objaverse	51.77	-	**36.45**
LSUN	-	38.32	**29.73**

Table 1. **FID evaluation.** Director3D cannot generate objects without backgrounds, and MVDream is not for scene-level generation. To ensure a fair comparison, both MVDream and our method are trained on Objaverse, while LSUN remains unseen for both our method and Director3D. GaussianDreamer is excluded due to its high computational cost.

will update the camera trajectory visualization in the supp. mat. to avoid confusion.

4. Get Camera Trajectory from Text

We employ the TrajDiT from Director3D [6] for trajectory generation. Recent works like ChatCam [7] also present text-to-trajectory generation, allowing for obtaining well-aligned trajectories. Moreover, as our model is jointly conditioned on camera poses and text during training, it can adaptively generate scenes that adapt to the input trajectory.

5. Reconstruction Performance

Since our work focuses on 3D generation, the GS-VAE is designed for wide-baseline generalizable reconstruction, allowing us to achieve SOTA performance in the "hard mode" case (Tab. 2 in the main paper). We attribute pixelSplat's [1] superior performance in easier modes to its explicit epipolar constraints, which likely improve per-pixel accuracy. In contrast, our method—while excelling in geometry and overall subjective quality (see LPIPS and Fig. 3 in the main paper)—may experience slight pixel-level misalignments, leading to a reduction in PSNR and SSIM.

6. Why RGB-D Joint Prediction?

In the context of per-pixel GS generation, accurate geometry reconstruction and high-quality depth estimation are intrinsically linked. Ideally, each GS should be precisely positioned at its correct depth. DepthSplat [10] demonstrates a synergy between monocular depth and per-pixel GS. Moreover, as shown in like DS-NeRF [4], accurate depth recovery is pivotal for sparse view reconstruction.

Regarding multi-view generation, modeling the joint distribution of multi-view RGB-D data ensures depth consistency. Meanwhile, obtaining depth from MV image generator introduces additional steps compared to RGD-D latent prediction: decoding latent images to RGB (425 ms), depth estimation (1242 ms), and re-encoding depth into latent space (355 ms). These increase the denoising computational time by 50.2% (original: 4.02 s) and complicate the pipeline.

7. Limitations and Future Works

We then visualize the failure cases of our method in Figs. 3 and 5. Firstly, as shown in Fig. 3, despite specific designs during training and sampling aimed at mitigating 3D inconsistencies, `Prometheus` still encounters inaccuracies in rendering high-frequency structures. Secondly, as shown in Fig. 5, our method occasionally exhibits text misalignment issues. The primary cause is the joint training of single-view and multi-view models, which disrupts the original text embedding layer of the pre-trained image generation model. Designing a specialized architecture to preserve the text alignment capability of the pre-trained image generation model will address this issue.

Compared to existing methods primarily limited to object-centric generation, we pioneer scene-level generation. Indeed, `Prometheus` has room for improvement regarding scene scale, this stems from a key bottleneck in current multiview-based 3D generation: frame quantity limitations. However, recent advances in video diffusion, e.g., Diffusion Forcing [2], which progressively generates long

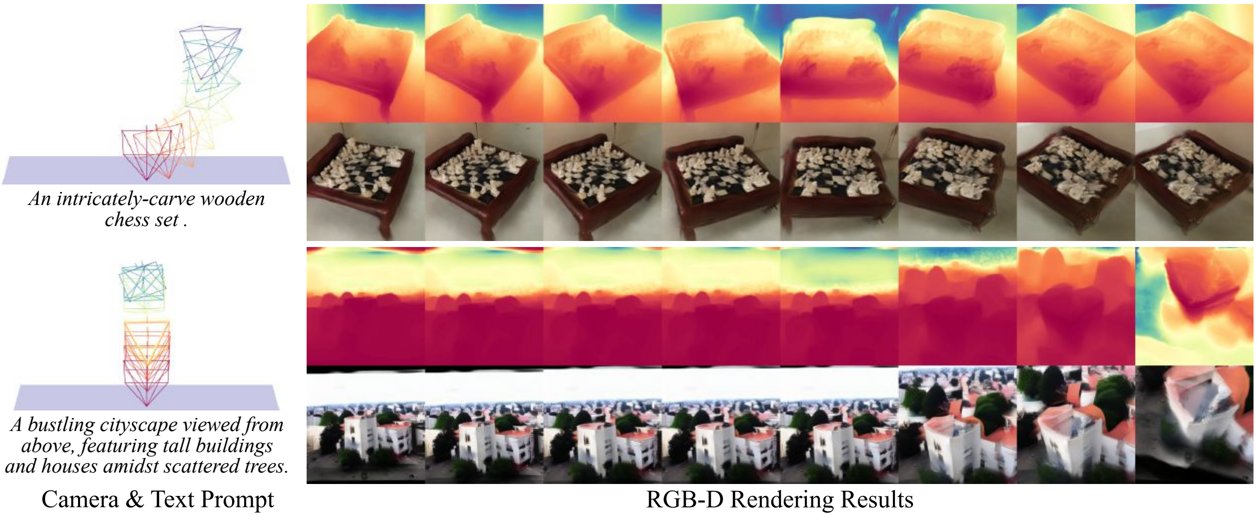

An intricately-carve wooden chess set .
A bustling cityscape viewed from above, featuring tall buildings and houses amidst scattered trees.
Camera & Text Prompt
RGB-D Rendering Results

Figure 5. **Text-misalignment cases.** We then show Text-misalignment, the second factor contributing to the failure cases of our method. As shown in the images, `Prometheus` synthesizes a black laptop instead of following the prompt, which should be silver.

sequences, hold promise for overcoming this limitation in the future.

References

[1] David Charatan, Sizhe Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[2] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *arXiv:2407.01392*, 2024. 3

[3] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2

[4] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[5] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[6] Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3d: Real-world camera trajectory and 3d scene generation from text. *arXiv.org*, 2406.17601, 2024. 1, 3

[7] Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. Chatcam: Empowering camera control through conversational ai. *arXiv preprint arXiv:2409.17331*, 2024. 3

[8] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv:2308.16512*, 2023. 2

[9] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. *arXiv.org*, 2402.05054, 2024. 2

[10] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv preprint arXiv:2410.13862*, 2024. 3

[11] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2