Gen3R: 3D Scene Generation Meets Feed-Forward Reconstruction

Supplementary Material

1. Implementation Details

1.1. Processing Input Conditions

We employ multiple conditions into the diffusion process, including a text prompt $\mathbf{y}$, a condition image sequence $\mathcal{I}_{cond}$ with a flexible number of available frames (missing images are set to zero), corresponding binary masks $\mathcal{M}$ and optional per-view camera conditions $\mathcal{T}_{cond}$. The condition images $\mathcal{I}_{cond}$ are encoded into appearance latents $\mathcal{A}_{cond}$ by pretrained RGB VAE $\mathcal{E}_{\mathcal{W}}$:

$$\mathcal{E}_{\mathcal{W}} : \mathcal{I}_{cond} \to \mathcal{A}_{cond} \in \mathbb{R}^{n\times h\times w\times c}, \tag{1}$$

while the masks $\mathcal{M}$ are downsampled to $\mathcal{M}_a \in \mathbb{R}^{n\times h\times w\times 4}$ to match the latent resolution. To ensure dimensional consistency with the noised latents, we initialize the geometry branch's condition latents $\mathcal{G}_{cond} \in \mathbb{R}^{n\times h\times w\times c}$ and corresponding masks $\mathcal{M}_g \in \mathbb{R}^{n\times h\times w\times 4}$ as **zeros**.

Finally, the appearance and geometry latents are fused with their respective latent masks along the channel dimension, and the two modalities are further concatenated in the width dimension to construct the unified condition latent:

$$\mathcal{Z}_{cond} = [\mathcal{A}_{cond} \oplus \mathcal{M}_a; \mathcal{G}_{cond} \oplus \mathcal{M}_g] \in \mathbb{R}^{n\times h\times 2w\times c'}, \tag{2}$$

where $(\cdot\oplus\cdot)$ means concatenation along channel dimension, and $c' = c + 4$. The input to the diffusion model is then constructed by concatenating the noised latents $\mathcal{Z}_t$ with the condition latents $\mathcal{Z}_{cond}$ along the channel dimension:

$$\mathcal{Z}_{in} = \mathcal{Z}_t \oplus \mathcal{Z}_{cond}, \tag{3}$$

$$G_\theta : \mathcal{Z}_{in} \to \hat{\mathcal{Z}}_{t-1}. \tag{4}$$

1.2. Model Architectures

Geometry Adapter. We obtain our adapter $(\mathcal{E}_{\text{adp}}, \mathcal{D}_{\text{adp}})$ by modifying Wan's causal VAE [11]. The adapter projects VGGT [12] geometry tokens $\mathcal{V} \in \mathbb{R}^{N\times L\times h_v\times w_v\times C}$ into the video diffusion model's latent space and maps them back:

$$\mathcal{E}_{\text{adp}} : \mathcal{V} \to \mathcal{G} \in \mathbb{R}^{n\times h\times w\times c}, \tag{5}$$

$$\mathcal{D}_{\text{adp}} : \mathcal{G} \to \mathcal{V} \in \mathbb{R}^{N\times L\times h_v\times w_v\times C}, \tag{6}$$

where $L = 5$, since we broadcast VGGT's camera tokens of each frame to the spatial resolution $h_v \times w_v$ ($h_v = w_v = 40$), and concatenate it with the other 4 tokens along the L dimension.

To match the VAE input format, we first reshape $\mathcal{V}$ into $\mathcal{V}' \in \mathbb{R}^{N\times h_v\times w_v\times (L\times C)}$. Accordingly, we set the adapter input dimension to $L \times C = 10240$ and use hidden dimensions $[512, 256, 128, 128]$. We then re-sample the input tokens $\mathcal{V}'$ to a spatial resolution of $h \times w = 70 \times 70$ using nearest-exact interpolation, and apply a 2D convolution to project the channels to 1024. The resulting features are processed by causal convolution layers, where we keep the spatial resolution unchanged, yielding geometry latents $\mathcal{G} \in \mathbb{R}^{n\times h\times w\times c}$. Similarly, the decoder $\mathcal{D}_{\text{adp}}$ mirrors the encoder architecture in reverse, reconstructing the geometry tokens $\mathcal{V}$ from the latents $\mathcal{G}$.

Diffusion Transformer. We adapt the DiT architecture from VideoX-Fun's Wan2.1 [11] to accommodate our joint appearance-geometry latents. Specifically, we set the input channel dimension to $c + c' = 36$. To support width-wise concatenation of appearance and geometry latents, we modify the positional embeddings so that corresponding pixels in the left and right halves of the latents share identical RoPE embeddings.

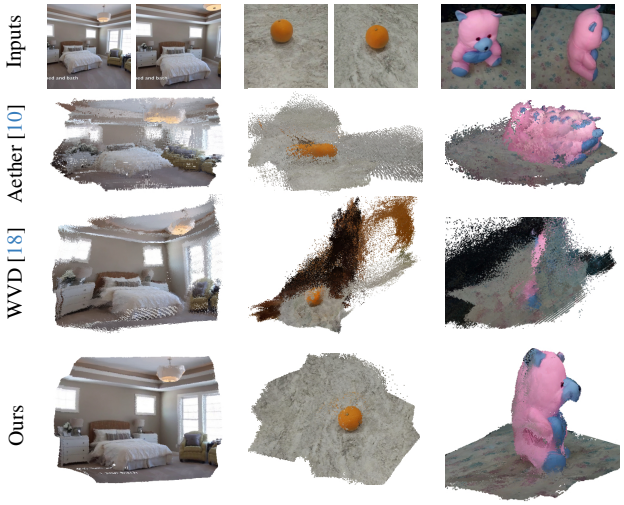


Figure 1. **Qualitative Comparison of Geometry Generation** in the 2-view based setting.

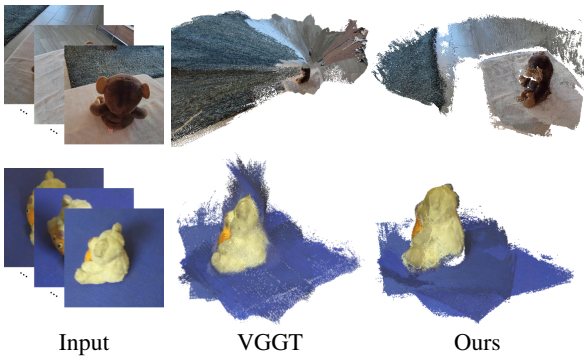


Figure 2. **Qualitative Comparison of Geometry Reconstruction.**

Cond.	Co3Dv2						WildRGB-D						TartanAir					
Method	PSNR ↑	SSIM ↑	LPIPS ↓	I2V Subj. ↑	I2V BG ↑	I.Q. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	I2V Subj. ↑	I2V BG ↑	I.Q. ↑	PSNR ↑	SSIM ↑	LPIPS ↓	I2V Subj. ↑	I2V BG ↑	I.Q. ↑
1-view																		
LVSM [4]	14.08	0.5623	0.5698	0.9482	0.9581	0.3579	13.9483	0.5239	0.5195	0.9692	0.9713	0.4004	14.44	0.5044	0.5210	0.9325	0.9540	0.3542
Gen3C [8]	15.82	0.5666	0.5095	0.9134	0.9355	0.4335	14.60	0.5463	0.4513	0.9622	0.9646	0.4629	13.95	0.4731	0.5385	0.9142	0.9403	0.3713
GF [14]	10.25	0.3150	0.6761	0.7933	0.8193	0.5320	11.8944	0.4147	0.5940	0.9215	0.9214	0.5310	10.21	0.3249	0.6249	0.7447	0.7864	0.4379
Aether [10]	12.78	0.5106	0.6052	0.9229	0.9395	0.4411	11.87	0.4289	0.5973	0.9595	0.9614	0.4786	12.88	0.4585	0.5645	0.9295	0.9480	0.4303
WVD [18]	13.35	0.4733	0.5765	0.9339	0.9484	0.5355	12.95	0.4522	0.5362	0.9669	0.9671	0.5513	12.77	0.4513	0.5652	0.9271	0.9473	0.4571
Ours	**16.09**	**0.5754**	**0.4997**	**0.9535**	**0.9588**	**0.5383**	**14.73**	**0.5501**	**0.4398**	**0.9715**	**0.9716**	**0.5609**	**15.04**	**0.5069**	**0.5073**	**0.9350**	**0.9546**	**0.4620**
2-view																		
DepthSplat [16]	10.45	0.3262	0.6167	0.8314	0.8585	0.2992	16.22	0.5382	0.4518	0.9012	0.9067	0.3779	13.87	0.4585	0.5195	0.8073	0.8474	0.3301
LVSM [4]	17.87	0.5986	0.4534	0.9467	0.9519	0.4064	**19.13**	**0.6789**	**0.3134**	**0.9747**	0.9730	0.4555	**17.79**	**0.5685**	**0.4265**	**0.9415**	**0.9569**	0.3628
Gen3C [8]	17.16	0.5927	0.4776	0.9149	0.9361	0.4263	17.81	0.6307	0.3882	0.9636	0.9651	0.4634	15.24	0.5055	0.5318	0.9119	0.9376	0.3668
GF [14]	12.67	0.3855	0.5998	0.7645	0.7925	0.4969	13.51	0.3991	0.4609	0.8785	0.8852	0.5374	12.06	0.3670	0.5666	0.7447	0.7946	0.4589
Aether [10]	14.28	0.5405	0.5498	0.9322	0.9426	0.4647	13.79	0.4884	0.5161	0.9491	0.9512	0.4685	14.53	0.4989	0.5153	0.9294	0.9496	0.4267
WVD [18]	14.66	0.5101	0.5334	0.9246	0.9409	0.5306	16.27	0.5421	0.4098	0.9631	0.9646	0.5627	14.22	0.4605	0.5266	0.9116	0.9371	0.4680
Ours	**18.01**	**0.6085**	**0.4371**	**0.9547**	**0.9597**	**0.5405**	18.88	0.6448	0.3256	0.9746	**0.9755**	**0.5685**	17.34	0.5581	0.4416	0.9385	0.9559	**0.4748**

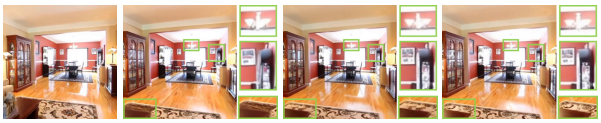
Table 1. **Quantitative Comparison of Appearance Generation.** We compare both 1-view and 2-view settings with camera conditions.

Cond.	Method	RealEstate10K					DL3DV-10K				
		I2V Subj. ↑	I2V BG ↑	Aes.Q. ↑	I.Q. ↑	M.S. ↑	I2V Subj. ↑	I2V BG ↑	Aes.Q. ↑	I.Q. ↑	M.S. ↑
1-view	Aether [10]	0.9743	0.9770	0.5118	0.5060	0.9885	0.9377	0.9501	0.4704	0.4872	0.9600
	WVD [18]	0.9815	0.9843	0.5125	0.5653	0.9895	0.9274	0.9412	0.4555	0.4916	0.9542
	Ours	**0.9879**	**0.9890**	**0.5291**	**0.5761**	**0.9929**	**0.9461**	**0.9561**	**0.4727**	**0.5187**	**0.9701**
2-view	Aether [10]	0.9852	0.9843	0.5278	0.5187	0.9923	0.9485	0.9521	0.4846	0.5026	0.9685
	WVD [18]	0.9929	0.9923	0.5336	0.5973	0.9938	0.9403	0.9518	0.4760	0.5338	0.9685
	Ours	**0.9949**	**0.9947**	**0.5369**	**0.6009**	**0.9947**	**0.9549**	**0.9576**	**0.4881**	**0.5357**	**0.9719**

Table 2. **Quantitative Comparison of Appearance Generation** without camera conditions.

Cond.	LLFF		Mip-NeRF 360		ScanNet++	
Method	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
1-view						
LVSM [4]	12.39	0.5742	12.71	0.6328	15.25	0.4531
SEVA [19]	11.43	0.6562	12.77	0.5898	13.97	0.4688
Aether [10]	11.01	0.6133	11.06	0.6640	12.47	0.5351
Ours	**13.19**	**0.5078**	**13.17**	**0.5703**	**15.42**	**0.4420**
2-view						
LVSM [4]	**18.39**	**0.3266**	**15.56**	0.5039	**21.31**	**0.2754**
SEVA [19]	16.86	0.3769	14.86	0.5080	18.07	0.3438
Aether [10]	13.66	0.4589	11.43	0.6289	17.14	0.4018
Ours	17.66	0.3496	15.32	**0.4941**	20.11	0.2964

Table 3. **Quantitative Comparison of Appearance Generation on Out-of-Distribution Datasets.** We compare both 1-view and 2-view settings with camera conditions.



Input Bolt3D Ours GT

Figure 3. **Quali. Comparison with Bolt3D** in 1-view setting.

	Method	PSNR↑	SSIM↑	LPIPS↓
1-view RE10K	Bolt3D	21.54	0.747	0.234
	Ours	**22.36**	**0.762**	**0.204**
2-view RE10K	Bolt3D	23.13	0.806	0.166
	Ours	**26.65**	**0.851**	**0.151**
2-view Co3D hyd.	Bolt3D	17.38	0.437	0.390
	Ours	**17.84**	**0.478**	**0.372**
2-view Co3D ted.	Bolt3D	18.94	0.605	0.393
	Ours	**19.81**	**0.722**	**0.390**

Table 4. **Quantitative Comparison with Bolt3D.**

2. Additional Comparison Results

2.1. 3D Generation

Comparison on 3D Generation with Camera Conditions. We provide the full appearance evaluation results on Co3Dv2 [7], WildRGB-D [15] and TartanAir [13] datasets in Tab. 1. Gen3R consistently surpassing existing methods across all metrics and datasets in the 1-view setting, and achieves leading performance in the 2-view setting. Additional qualitative comparisons of 3D generation are shown in Fig. 4 and Fig. 1. As observed, LVSM [4], Aether [10] and WVD [18] fail to synthesize images from novel viewpoint in 1-view setting, primarily due to poor camera controllability. While Gen3C [8] can generate plausible contents, it exhibits notable shifts caused by inaccurate depth estimation. In contrast, our methed produces high-fidelity results that adhere closely to the camera conditions and maintain better 3D structure, as shown in Fig. 1.

Comparison on 3D Generation without Camera Conditions. We further demonstrate our capability to generate 3D scenes from images without camera conditions. To assess this, we report the VBench Score [2, 3], focusing on I2V Subject (I2V Subj.), I2V Background (I2V BG), Aesthetic Quality (Aes.Q.), Imaging Quality (I.Q.) and Motion Smoothness (M.S.) on RealEstate10K [20] and DL3DV-10K [5] datasets. As shown in Tab. 2, our method clearly outperforms Aether [10] and WVD [18], illustrating its superior ability in generating high-quality 3D scenes.

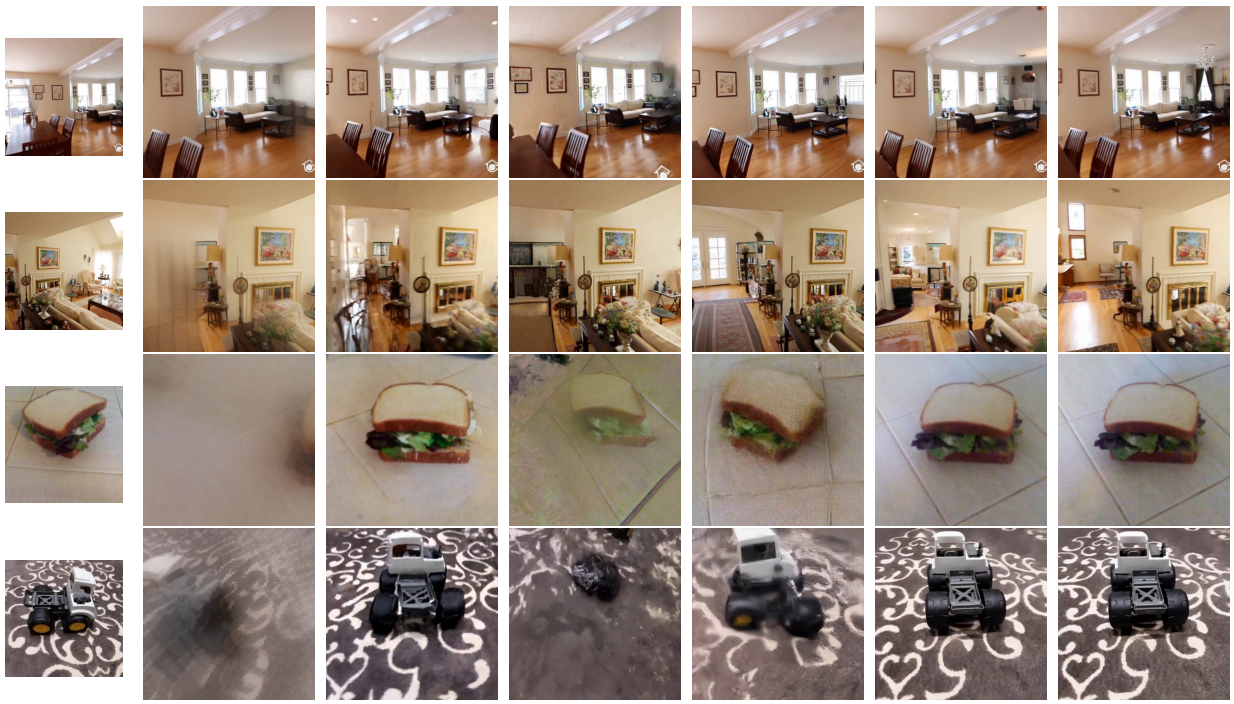


Figure 4. **Qualitative Comparison of Novel View Synthesis** in 1-view setting with camera conditions.

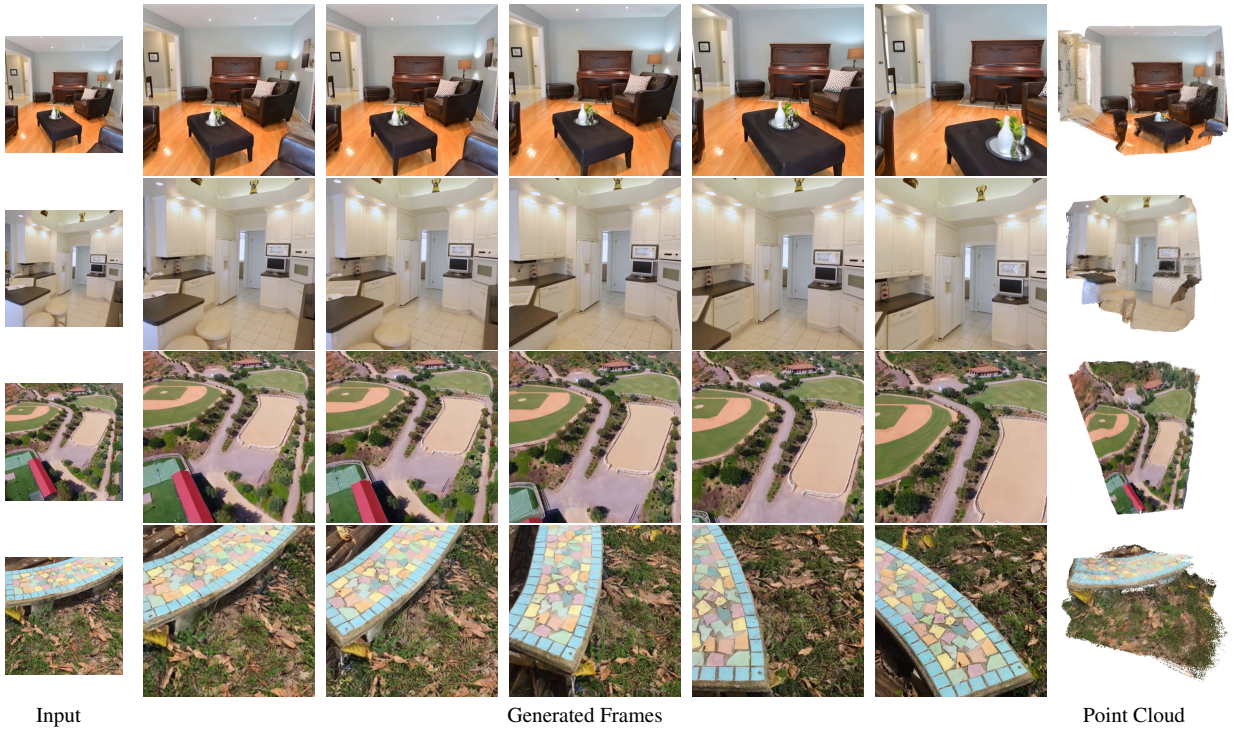


Figure 5. **More Qualitative Results** in 1-view setting with camera conditions.

Comparison on 3D Generation on Out-of-Distribution Datasets. We report appearance generation results in 1-view and 2-view settings on LLFF [6], Mip-NeRF 360 [1] and ScanNet++ [17] test sets, which are excluded from the training datasets. As shown in Tab. 3, our method handles these unseen scenes well, yields conclusions consistent with the main paper, and demonstrates superior performance over SEVA [19].

Figure 6. **More Qualitative Results** in 2-view setting with camera conditions.

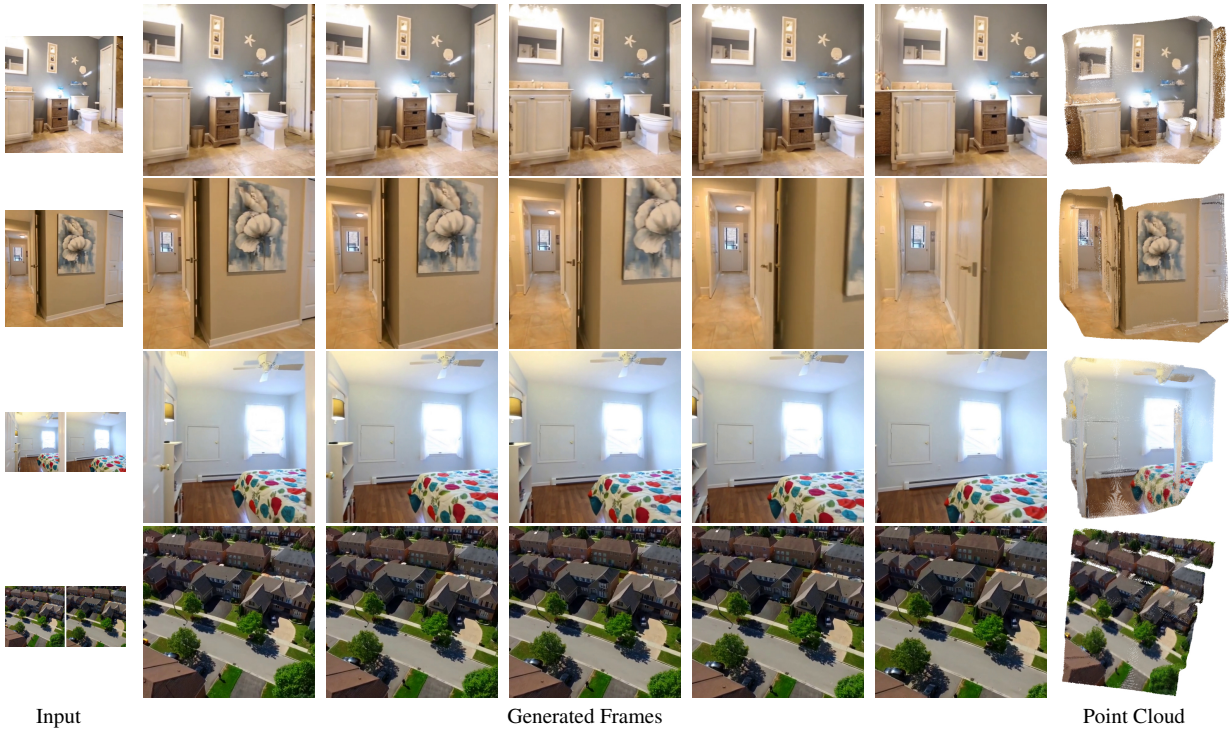


Figure 7. **More Qualitative Results** in 1-view and 2-view settings *without* camera conditions.

Comparison with Bolt3D. Rendering our point cloud requires 3DGS-head, which we consider promising future work. Currently, we compare our generated multi-view appearance with Bolt3D [9] renderings using their official test split in Tab. 4 and Fig. 3. Although this setup is not perfectly fair, our superior results highlight Gen3R's strength in producing high-fidelity appearance, validating the RGB quality needed to lift high-quality appearance into 3D representations.

2.2. Feed-forward 3D Reconstruction

Comparison on Camera Pose Estimation. We evaluate our method on RealEstate10K and WildRGB-D datasets for camera pose estimation, as reported in Tab. 5. Our approach achieves competitive results compared to VGGT, while notably surpassing Aether, showing the versatility and robust-

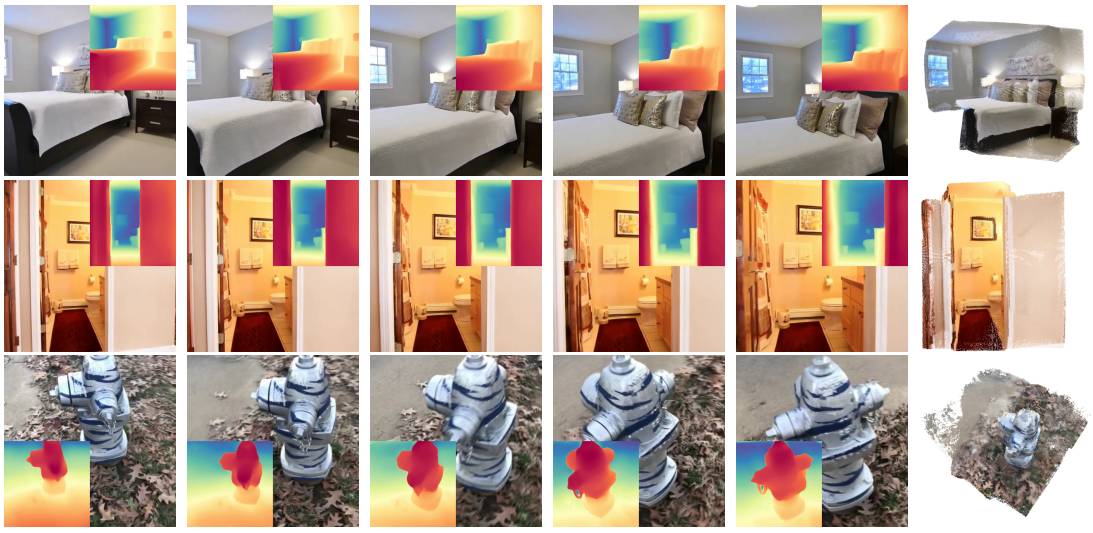


Figure 8. **More Qualitative Results** of feed-forward reconstruction.

Method	RealEstate10K	WildRGB-D
	AUC@30 ↑	AUC@30 ↑
Aether	0.7291	0.7303
VGGT	**0.8387**	**0.8406**
Ours	0.8265	0.8391

Table 5. **Quantitative Comparison of Camera Pose Estimation** in feed-forward 3D reconstruction.

Method	ScanNet++		
	Accuracy ↓	Completeness ↓	CD ↓
Aether	0.3187	0.3022	0.3105
VGGT	**0.1396**	0.1162	0.1279
Ours	0.1455	**0.0963**	**0.1209**

Table 6. **Quantitative Comparison of Zero-shot Geometry Reconstruction** in ScanNet++ dataset.

Method	RealEstate10K			DL3DV-10K		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
VGGT* [12]	23.3927	0.8346	0.2341	22.6958	0.7557	0.2910
RGB VAE [11]	**37.5770**	**0.9819**	**0.0288**	**32.7673**	**0.9057**	**0.1031**

Table 7. **Quantitative Comparison for RGB Reconstruction.** We train an RGB head for VGGT to reconstruct images from geometry tokens. * indicates our implementation.

ness of our model.

Comparison on Geometry Reconstruction. We provide additional qualitative results of feed-forward 3D reconstruction compared with VGGT [12] in Fig. 2. It can be observed that VGGT produces noticeable floaters in the reconstructed point clouds, while our method generates significantly cleaner geometry.

In addition, to evaluate zero-shot generalization, we compare Gen3R with VGGT on ScanNet++ [17] test split. As shown in Tab. 6, VGGT slightly outperforms Gen3R in accuracy; however, our method achieves better completeness and better chamfer distance, indicating that our method generalizes reasonably to unseen scenes.

2.3. Ablation Study

RGB Head for VGGT. To validate the effectiveness of our joint latents design, we train an RGB head for VGGT to enable direct RGB reconstruction from its geometry tokens $\mathcal{V}$. We then compare its RGB reconstruction quality with that of Wan's RGB VAE [11]. The results are presented in Tab. 7. RGB VAE significantly outperforms VGGT*, as VGGT is designed primarily for geometry modeling and lacks sufficient capacity for RGB feature extraction and high-fidelity appearance reconstruction. This observation also motivates our choice to decode appearance and geometry separately. By combining the strengths of both pre-trained models, we achieve photorealistic video generation together with high-quality 3D structure.

3. More Results of Gen3R

We present additional qualitative results for both 3D generation and feed-forward 3D reconstruction in this section, including: *1) 3D Generation with Camera Conditions* (see Fig. 5 and Fig. 6); *2) 3D Generation without Camera Conditions* (see Fig. 7); and *3) Feed-Forward 3D Reconstruction* (see Fig. 8). We visualize the generated frames, depth maps of the sequences, and the global point clouds of the scenes.

Our method synthesizes globally consistent and photorealistic 3D scenes under diverse input conditions and effectively handles a wide range of scenarios, including indoor

scenes, outdoor environments, and object-centric cases. Thanks to our design, the model exhibits strong camera controllability under conditioned settings, while also enabling free scene navigation in the absence of camera inputs. Combined with support for multiple output modalities, Gen3R provides fine-grained and coherent 3D scene generation across both constrained and unconstrained regimes.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 3

[2] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[3] Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *arXiv preprint arXiv:2411.13503*, 2024. 2

[4] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias, 2025. 2, 3

[5] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, Xuanmao Li, Xingpeng Sun, Rohan Ashok, Aniruddha Mukherjee, Hao Kang, Xiangrui Kong, Gang Hua, Tianyi Zhang, Bedrich Benes, and Aniket Bera. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision, 2023. 2

[6] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 3

[7] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 2

[8] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control, 2025. 2, 3

[9] Stanislaw Szymanowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds, 2025. 4

[10] Aether Team, Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 1, 2, 3

[11] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 5

[12] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 5

[13] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. 2020. 2

[14] Haoyu Wu, Diankun Wu, Tianyu He, Junliang Guo, Yang Ye, Yueqi Duan, and Jiang Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025. 2

[15] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos, 2024. 2

[16] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth, 2025. 2

[17] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3, 5

[18] Qihang Zhang, Shuangfei Zhai, Miguel Angel Bautista Martin, Kevin Miao, Alexander Toshev, Joshua Susskind, and Jiatao Gu. World-consistent video diffusion with explicit 3d modeling. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21685–21695, 2025. 1, 2, 3

[19] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. *arXiv preprint arXiv:2503.14489*, 2025. 2, 3

[20] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images, 2018. 2