Supplementary Material of VistaDream: Sampling multiview consistent images for single-view scene reconstruction

Haiping Wang[1] Yuan Liu[2,3,†] Ziwei Liu[3] Wenping Wang[4] Zhen Dong[1,†] Bisheng Yang[1]
[1]Wuhan University [2]Hong Kong University of Science and Technology
[3]Nanyang Technological University [4]Texas A&M University
{hpwang,dongzhenwhu,bshyang}@whu.edu.cn
yuanly@ust.hk ziwei.liu@ntu.edu.sg wenping@tamu.edu

A. Appendix

A.1. Rendering videos

We provide RGB and Depth rendering videos of the reconstructed scenes in the project page: https://vistadream-project-page.github.io/.

A.2. Implementation details of VistaDream

A.2.1. Coarse Gaussian field generation

Image description with VLM. We use LLaVA [14] to generate a detailed description for the input image. The LLaVA prompt is set as: "$\langle$*image*$\rangle$ *USER: Detailly imagine and describe the scene this image is taken from? ASSISTANT: This image is taken from a scene of*". The continuation of the LLaVA response is used as the image description and fed to inpainting models in the Coarse scene reconstruction. As shown in Fig. A.1, the inpainting results using LLaVA description is much better and detailed.

Building a 3D scaffold. The $H \times W$ input image is enlarged to $1.9H \times 1.9W$ by extending in four directions and inpainted using Fooocus [21] with LLaVA image description. Subsequently, we can recover the per-pixel depth d and image focal length f using a metric depth estimator such as Metric3Dv2 [11] or Depth-Pro [2], thereby recovering the 3D points corresponding to each pixel. We follow the default hyperparameter settings of the above models. Afterward, we follow pixelSplat [3] to construct Gaussian kernels for each pixel: the *xyz* property of the Gaussian kernels is its 3D position, the *RGB* property comes from the pixel color, the *opacity* property is set to a constant, the *rotation* property is an identity matrix, and the scale is set to $d/\sqrt{2}f$. To avoid trailing artifacts, we eliminate kernels in object boundary regions based on depth variation judgment [18] and then optimize the remaining Gaussian kernels by 100 iterations [12]. For Gaussian kernel optimization, we set the learning rate of the *xyz* property to 3e-4, *RGB* to 5e-4, *scale* to 5e-3, *opacity* to 5e-2, *rotation* to 1e-3.

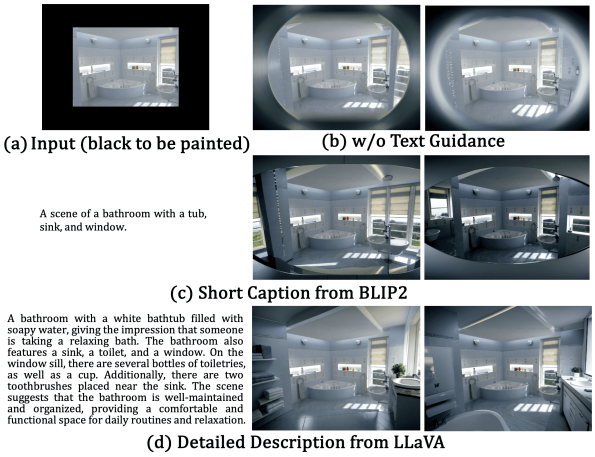

Figure A.1. *Detailed description from LLaVA is vital for inpainting.* Compared to (b) empty descriptions or (c) short captions, (d) descriptions from large Vision-Language Models are more detailed, significantly enhancing the reliability of inpainting.

Warp-and-inpaint. After scaffold initialization, we establish a spiral camera trajectory. Then we select the viewpoint with the largest missing regions to render both the partial RGB image and depth map. The RGB image is inpainted by Fooocus [21]. Taking the completed image as the condition, we use a model ϕ to estimate its depth map and optimize the depth for smoothly connecting to the existing Gaussian Field. We have two strategies for setting ϕ. The first strategy uses a diffusion model-based GeoWizard [8] to estimate depth. To ensure smooth connections, we introduce a loss between the estimated depth and the rendered one at each denoise step [20]. The second strategy employs a feedforward depth estimation model, DepthPro [2], to estimate image depth. We linearly align the estimated depth with the rendering one, and further optimize the estimation through residual smoothing [5]. The first strategy is more time-consuming but yields better results, while the second strategy is faster but may introduce distortions. In different cases, we adopt the strategy that provides better outcomes.

Then, we construct a set of Gaussian kernels on the completed RGB-D regions as above. We filter them with two additional checks: 1) *Occlusion avoidance*: We project the Gaussian center onto already processed viewpoints, and if its depth is less than the original depth at any viewpoint, it is discarded. 2) *Boundary exclusion*: we remove the kernels on the object boundaries as mentioned above. The remaining kernels are integrated into the Gaussian field. This is followed by a 256-step scene optimization process. The above "warp-and-inpaint" process is iteratively executed several times to obtain the coarse Gaussian field.

A.2.2. Multiview Consistency Sampling for Refinement

Multi-view Consistency Sampling. In our implementation, we uniformly sample $N = 8$ views along the spiral trajectory, with an image resolution of 512×512. Afterward, we encode and add $T = 10$ steps of noise to each view by a 50-step DDPM sampler [9]. We use the Latent Consistency Model of Stable Diffusion (LCM-SD) [15] for noise prediction for its strong performance following DreamLCM [22]. We remove Classifier Free Guidance (CFG) in LCM and find better results without it. We perform weighted rectification and std-alignement of Eq. 3 of Sec.3.2.1 of main paper on the predicted and rectified noise map $\hat{\epsilon}$ and $\bar{\epsilon}$ for direct operations in latent space. ϵ has a linear relationship with μ according to Eq.1 of Sec.3.2.1. In each sampling step of MCS, we use the denoising multi-view images to optimize a copy of the coarse Gaussian field by 2560 steps to enforce consistency, where we set a smaller learning rate of *xyz* in Gaussian kernels, specifically 1e-4, to avoid geometry distortions.

Gaussian field refinement. In our implementation, we optimize the coarse Gaussian Field by 2560 steps with the refined multi-view images and enlarged input image.

To run VistaDream within a 24GB VRAM limit, we need to allocate some time for model swapping. Specifically, we transfer only the currently active model to the GPU while keeping the others in CPU memory. This ensures efficient memory usage to maintain the overall workflow's integrity.

A.3. LVM-IQA metrics

Typical perception metrics such as PSNR and SSIM need pixel-wise ground truth (GT) of novel views, which is hard to acquire for our task. Instead, we use VLM-based metrics, as VLMs can understand and evaluate images well without requiring GT as proven by LLaVA [14], InternVL [4], and Qwen2.5-VL [1].

Given a set of rendered images, we perform the Image Quality Assessment using the above LVMs, called LVM-IQA. The prompt is designed as: "$\langle$*image*$\rangle$ *USER:* $\langle$*question*$\rangle$*, just answer with yes or no? ASSISTANT:*". The $\langle$*question*$\rangle$ placeholder is replaced according to different evaluation purposes as follows:

Coarse GS reconstruction (s)			MCS Refine (s per step)			Overall (min)
LLaVA Describe	Zoom-out &Inpaint	Warp&Inpaint (per step)	$\hat{\mu}$ Sample	3D GS Optimize	Stepwise Denoise	
2.00	36.87	11.69	0.36	11.05	0.41	6.86

Table A.1. *Average time consumption of VistaDream to reconstruct a scene*. In our default setting, we conduct 10 warp&inpaint steps in coarse GS reconstruction and 10 denosing steps in MCS. "Overall" includes the time spent on model loading and I/O operations.

- For noise level (**Noise-Free**): "*Is the image free of noise or distortion*"
- For sharp edge (**Edge**): "*Does the image show clear objects and sharp edges*"
- For scene structure (**Structure**): "*Is the overall scene coherent and realistic in terms of layout and proportions in this image*"
- For image details (**Detail**): "*Does this image show detailed textures and materials*"
- For image quality (**Quality**): "*Is this image overall a high-quality image with clear objects, sharp edges, nice color, good overall structure, and good visual quality*"

We then calculate the proportion of "yes" responses and report the average result of the five aspects.

We use 11 scenes from RealmDreamer [17] for quantitative assessment, including *bathroom*, *bear*, *bedroom*, *bust*, *kitchen*, *living-room*, *car*, *lavender*, *piano*, *victorian*, and *steampunk*. For each scene, we sample 50 viewpoints along the reconstruction trajectory for rendering and evaluation.

A.4. More analysis

Detailed time consumption of VistaDream. The time consumption details of VistaDream to reconstruct a scene is shown in Table A.1, where we conduct 10 warp&inpaint steps in coarse GS reconstruction and 10 denosing steps in MCS. It can be seen that VistaDream requires $\sim$ 7 minutes to reconstruct a scene, which is much faster than optimization-based RealmDreamer (2.5 hours) [17].

Choice of w_t in Eq. 3 in Sec.3.2.1 of main paper. In Fig. A.2, we show qualitative results using different w_t. When $w_t = 0$, the multi-view images are optimized independently to obtain high-quality but inconsistent images, yielding noisy and chaotic details after optimizing the scene. As the value of w_t increases, the consistency guidance is strengthened, leading to more accurate scene optimization. However, some finer details may be lost in this process to satisfy consistency. Empirically, we found that setting w_t between 0.3 and 0.8 achieves optimal results, striking a balance between detail enhancement and overall coherence. In this section, as well as in the "Compare MCS with SDS refinement" section in main text, we did not optimize the scene with the input image and outpaint image, in order to more accurately reflect effects of SDS and MCS.

VistaDream with sparse view inputs. VistaDream also supports sparse-view inputs. As shown in Fig, A.3, given

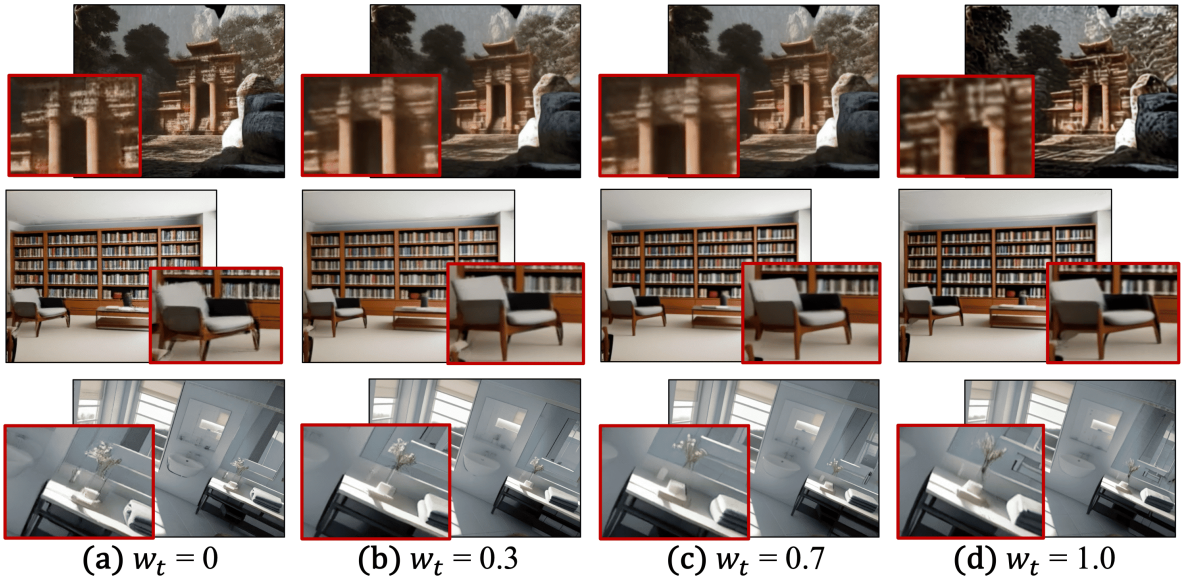

Figure A.2. *Set different w_t in Eq. 3 in Sec.3.2.1 of main paper.* When w_t is set to 0, the optimization of the Gaussian scene lacks multi-view consistency, leading to chaotic reconstructions and noisy details. As w_t increases, multi-view consistency improves, facilitating a more accurate optimization of the Gaussian field but slightly loses some details.

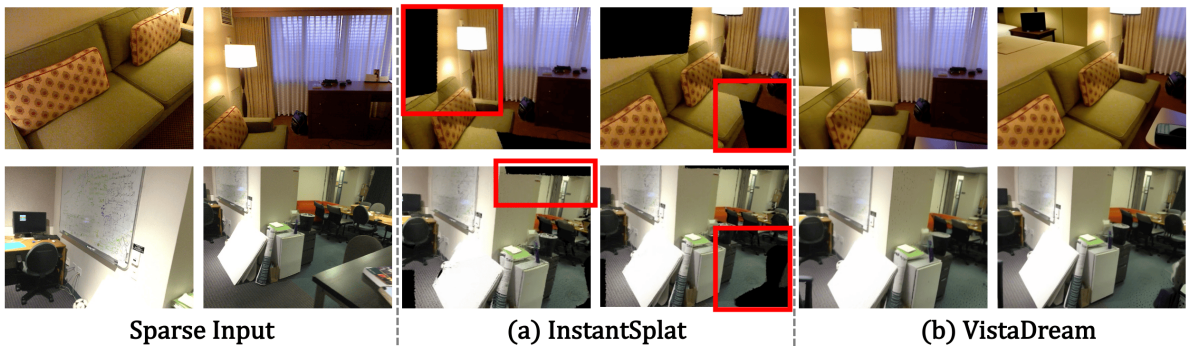

Figure A.3. *VistaDream with sparse inputs.* (a) Input sparse views (2 images). (b) Novel view renderings from the SoTA sparse-view reconstruction method, InstantSplat, leave gaps in unseen regions. (c) VistaDream effectively reconstructs a complete scene using warp-and-inpaint and MCS.

Methods	NF	Edge	Struc.	Detail	Quality	Avg.
InstantSplat	0.39	0.09	0.34	0.83	0.41	0.41
NVComposer	<u>0.75</u>	<u>0.26</u>	0.52	**0.96**	<u>0.68</u>	<u>0.63</u>
Ours-Coarse	0.66	0.23	<u>0.56</u>	0.89	0.62	0.59
Ours	**0.80**	**0.45**	**0.75**	<u>0.94</u>	**0.76**	**0.74**

Table A.2. *LLaVA-IQA results with sparse-view inputs.*

unposed sparse views, existing sparse reconstruction methods are unable to reconstruct unseen regions, resulting in substantial gaps. In contrast, VistaDream can reconstruct a complete scene. Specifically, we use Dust3r [19] to recover the relative poses and build a 3D Scaffold with the sparse inputs, then apply the warp-and-inpaint strategy with LLaVA prompt guidance to fill in the missing regions. MCS is used for scene optimization based on this foundation.

In Table A.2, we quantitatively compare the scene reconstruction performance on 7 scenes between VistaDream, InstantSplat [7], and concurrent NVComposer [13], given 2 input images. We modify LLaVA questions to add "regardless of large black regions" to ensure a meaningful evaluation of InstantSplat scenes with missing regions (zeros or else). Our method achieves an 11% average improvement. On these scenes, MCS achieves 15% average improvements, which further verifies the effectiveness of the proposed MCS on scene optimization.

A.5. More ablation studies

Ablating scaffold construction in VistaDream. In Table A.3, we ablate our core designs in the first stage of VistaDream.

a) *Global Scaffold by Zoom-out&Inpaint*. We propose to build a 3D scaffold for better inpainting connections by zoom-out&inpaint operation. This brings a 5% average improvement for VistaDream by comparing model-a and model-b in Table A.3. It is because the scaffold provides a reliable initialization and constraint for most regions of the

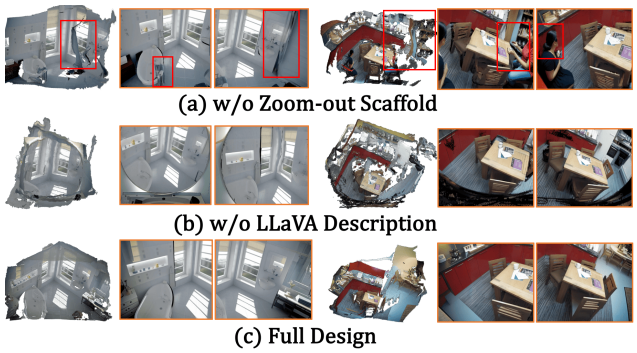

Figure A.4. *Ablating 3D global scaffold in coarse scene reconstruction*. (a) Replacing our zoom-out scaffold with moving-based scaffold of RealmDreamer yields distorted scenes and unwanted human regions. (b) Reconstructing with short captions of BLIP2 as inpaint prompt yields telescope-like or mirror-like images. (c) Using full designs improves the scaffold and scene quality.

Id	Scaffold	Prompt	NF	Edge	Sturc.	Detail	Quality	Avg.
(a)	None	BLIP2	0.85	0.13	0.47	0.93	0.54	0.58
(b)	Ours	BLIP2	0.89	0.21	0.50	0.92	0.63	0.63
(c)	Ours	LLaVA	**0.91**	**0.29**	**0.54**	**0.97**	**0.71**	**0.68**
(d)	Moving	LLaVA	0.89	0.23	0.52	0.93	0.69	0.65

Table A.3. *Ablating the coarse stage of VistaDream*. For 3D scaffold construction, "Ours" means our zoom-out&inpaint strategy, "None" means we directly conduct warp-and-inpaint without building a scaffold, "Moving" denotes the camera moving-based scaffold construction from RealmDreamer [17], specifically largely outpainting the scene scope on multiple neighboring views using advanced Depth-Pro for depth estimation and Fooocus for inpainting. For the inpaint prompt construction, "BLIP2" means the short description from BLIP2, "LLaVA" denotes using the detailed descriptions from LLaVA as inpainting instruction. We use LLaVA for Image Quality Assessment here.

scene, as also proven by previous methods [10, 17]. Without this operation, the scene may exhibit severe geometric distortions caused by the unstable out-/in-painting[17].

b) *Prompt inpainting with LVM*. Using the zoom-out operation is non-trivial. Leveraging rich textual prompts from large vision models (LVMs) for inpainting demonstrates significant enhancements in the quality of warp&inpainted regions. Compared to the short descriptions generated by BLIP2, this strategy achieves a 5% average quality improvement by comparing model-b (Fig. A.4b) and model-c (Fig. A.4c) in Table A.3, greatly avoiding unwanted telescope-like or mirror-like artifacts. Note that the quantitative results gains 10% in the coarse stage by eliminating large scene distortion with our designs. Removing large distortions is easier than MCS which improves the quality and details of the un-distorted scene. Though difficult, our MCS achieves better quality and gains a further 4% improvement.

Replace zoom-out with typical moving-based scaffold. Previous methods [10, 16, 17] construct moving-based scaffolds by moving cameras to neighboring viewpoints to largely outpaint the 3D scene on them They suffer from inconsistent geometric connections across the multiple outpainted views, yielding distorted scaffold and final scene.

In VistaDream, we find that a single-step zoom-out operation, though simple, suffices to initialize a more reliable scaffold, which not only ensures texture consistency but also significantly enhances the connectivity accuracy of subsequently inpainted regions. In Table A.3, by compare model-c (Fig. A.4c) and model-d (Fig. A.4a), our scaffold construction achieves 3% average improvement. Note that we adopt advanced Depth-Pro for monocular depth estimation, Fooocus for inpainting, and our other designs in moving-based scaffold for a fair comparison with our zoom-out strategy.

Improve other reconstruction methods with MCS. In Table 2 of main paper, we apply MCS to other reconstruction methods, specifically single-view based Realm-DP and sparse-view based InstantSplat. We re-implement RealmDreamer [17] and use an advanced monocular depth estimation method Depth-Pro [2], denoted by Realm-DP, all other settings are the same as the main experiment. For InstantSplat, we further introduce additional boundary exclusion for improvement and better visual quality. All other settings are the same as Table A.2.

A.6. Additional qualitative results

Given various input images, the results in Fig. A.6 and Fig. A.7 demonstrate that VistaDream produces clear, accurate, and highly consistent 3D scenes. In Fig. A.8, VistaDream achieves scene reconstruction from text inputs by incorporating a text-to-image generation model [6]. Moreover, in Fig. A.9, our method can generate different plausible scenes using different random seeds.

A.7. Limitations

MCS Refinement might lose details. Enforcing multi-view consistency might lose some details. Typical SDS smooth scenes for detail conflicts across denoised multi-views in different iterations. Comparing with SDS, the proposed MCS mitigates detail conflicts by enforcing consistency in multiview denoising and preserves more scene details. However, the details cannot be fully preserved. In future work, we will explore adaptive optimization region selection in MCS to further enhance multi-view clarity and consistency for better details.

Backside generation limitation. Generating backside regions for corners or 360° paths remains hard for VistaDream. Such areas are outside scaffold coverage and rely solely on warp&inpaint. However, when the camera moves back, front Gaussians may be warped there and corrupt inpainting (Fig. A.5a). NVS methods like SEVA can hallucinate backsides during multi-view synthesis, but may suffer from noise or blur for multi-view inconsistency. MCS can also improve them (Fig. A.5b-c).

Figure A.5. *Generate occluded backside regions.*

Holes and floating artifices. In the warp&inpaint stage, inaccurate depth estimation for edge regions or small objects leads to holes or floating artifacts when changing viewpoints. Although MCS reduces distortions and noise to some extent as shown in Fig.7 of the main paper, we acknowledge that it does not fully address severe missing regions or significant distortions, as demonstrated in Fig. A.10. Improving the quality of depth estimation may solve these issues, which we leave as future work.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[2] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv:2410.02073*, 2024. 1, 4

[3] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pages 19457–19467, 2024. 1

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[5] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 1

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 4

[7] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 3

[8] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. *arXiv preprint arXiv:2403.12013*, 2024. 1, 9

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 2

[10] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, pages 7909–7920, 2023. 4

[11] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv:2404.15506*, 2024. 1

[12] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[13] Lingen Li, Zhaoyang Zhang, Yaowei Li, Jiale Xu, Xiaoyu Li, Wenbo Hu, Weihao Cheng, Jinwei Gu, Tianfan Xue, and Ying Shan. Nvcomposer: Boosting generative novel view synthesis with multiple sparse and unposed images. *arXiv preprint arXiv:2412.03517*, 2024. 3

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024. 1, 2

[15] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[16] Chris Rockwell, David F Fouhey, and Justin Johnson. Pixelsynth: Generating a 3d-consistent experience from a single image. In *ICCV*, pages 14104–14113, 2021. 4

[17] Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024. 2, 4

[18] Haiping Wang, Yuan Liu, WANG Bing, YUJING SUN, Zhen Dong, Wenping Wang, and Bisheng Yang. Freereg: Image-to-point cloud registration leveraging pretrained diffusion models and monocular depth estimators. In *ICLR*, 2024. 1

[19] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[20] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In *CVPR*, pages 6658–6667, 2024. 1

[21] Lvming Zhang. Fooocus. https://github.com/lllyasviel/Fooocus, 2023. 1

[22] Yiming Zhong, Xiaolin Zhang, Yao Zhao, and Yunchao Wei. Dreamlcm: Towards high-quality text-to-3d generation via latent consistency model. *arXiv preprint arXiv:2408.02993*, 2024. 2

Figure A.6. *Image-to-3D scenes.* In each example, VistaDream generates a 3D Gaussian field based on the input image (red box), which is capable of rendering novel view images (orange box).

Figure A.7. *Image-to-3D scenes.* In each example, VistaDream generates a 3D Gaussian field based on the input image (red box), which is capable of rendering novel view images (orange box)

Figure A.8. *Text-to-3D scenes.* In each example, we use Stable Diffusion 3 to generate an image based on the input text (marked in yellow). Subsequently, VistaDream generates a 3D Gaussian field from the input image (red box), which can be used to render novel view images (orange box).

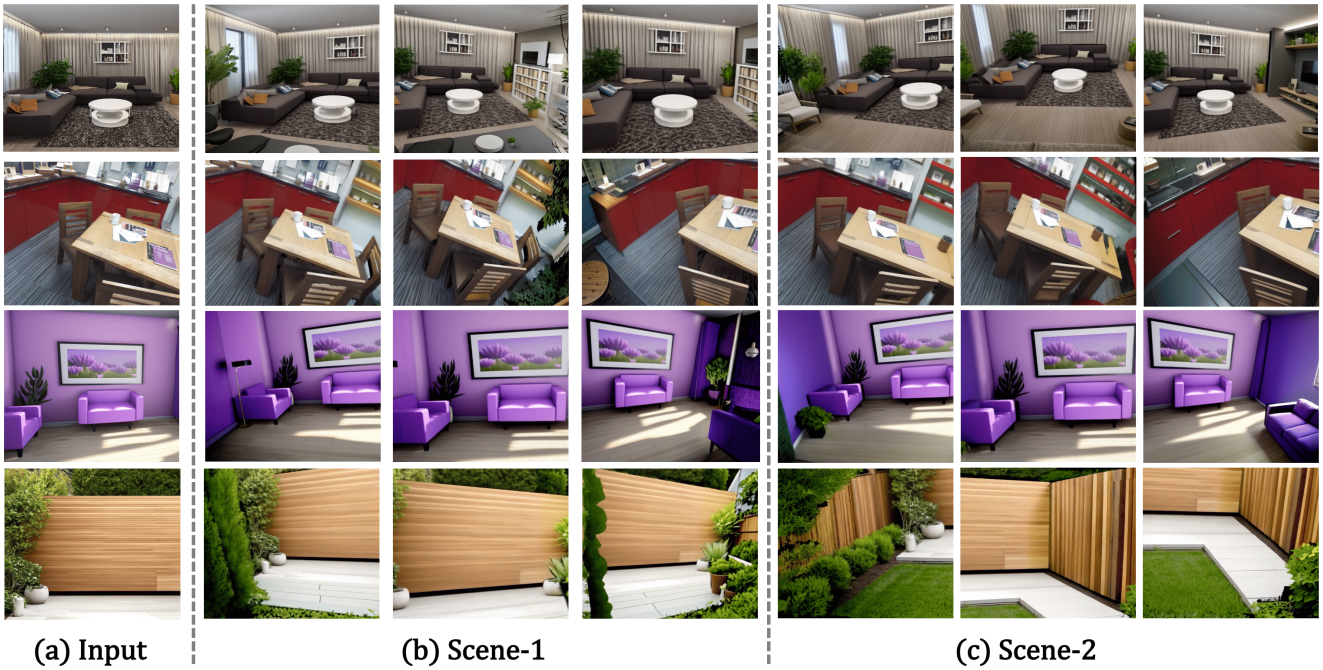

Figure A.9. *Different plausible scenes generated by VistaDream from the same input image.*

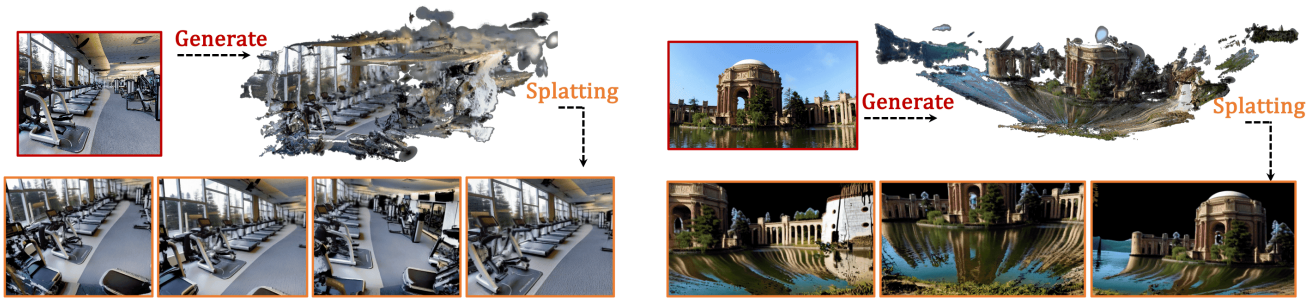

Figure A.10. *Typical failure cases.* Significant distortion (holes and floating artifacts) emerges due to the inaccurate depth estimation of GeoWizard [8].