ACKNOWLEDGMENT

The authors thank Minghui Xu for beneficial discussions. This work is partially supported by the AIM for Composites, an Energy Frontier Research Center funded by the U.S. Department of Energy (DOE), Office of Science, Basic Energy Sciences (BES), under Award # DE-SC0023389 and by the US National Science Foundation (NSF; Grant Number MTM2-2025541, OIA-2242812). The authors acknowledge research support from Clemson University with a generous allotment of computation time on the Palmetto cluster.

REFERENCES

Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. MultiMAE: Multi-modal multi-task masked autoencoders. 2022.

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

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment any 3d gaussians. *arXiv preprint arXiv:2312.00860*, 2023.

Agneet Chatterjee, Tejas Gokhale, Chitta Baral, and Yezhou Yang. On the robustness of language guidance for low-level vision tasks: Findings from depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2794–2803, June 2024.

Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting, 2024. URL https://arxiv.org/abs/2309.16585.

Zhiwen Fan, Peihao Wang, Xinyu Gong, Yifan Jiang, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation from complex real-world scenes. *arXiv e-prints*, pp. arXiv–2209, 2022.

Michael Grupp. evo: Python package for the evaluation of odometry and slam. https://github.com/MichaelGrupp/evo, 2017.

Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. *arXiv preprint arXiv:2403.18118*, 2024.

Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting, 2024.

Huy Ha and Shuran Song. Semantic abstraction: Open-world 3D scene understanding from 2D vision-language models. In *Proceedings of the 2022 Conference on Robot Learning*, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Ajay Jain, Ben Mildenhall, Jonathan T. Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. 2022.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *International Conference on Computer Vision (ICCV)*, 2023.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/pdf/2205.15585.pdf.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. *arXiv preprint arXiv:2403.06912*, 2024.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022.

Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. *arXiv preprint arXiv:2112.03221*, 2021.

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.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Jaeho Moon, Juan Luis Gonzalez Bello, Byeongjun Kwon, and Munchurl Kim. From-ground-to-objects: Coarse-to-fine self-supervised monocular depth estimation of dynamic objects with ground contact prior. *arXiv preprint arXiv:2312.10118*, 2023.

Jangho Park, Gihyun Kwon, and Jong Chul Ye. Ed-nerf: Efficient text-guided editing of 3d scene using latent space nerf. *arXiv preprint arXiv:2310.02712*, 2023.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. *arXiv preprint arXiv:2312.16084*, 2023.

Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 497–500, 2001.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. *arXiv preprint arXiv:2311.18482*, 2023.

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9043–9052, June 2023.

Gabriela Ben Melech Stan, Diana Wofk, Scottie Fox, Alex Redden, Will Saxton, Jean Yu, Estelle Aflalo, Shao-Yen Tseng, Fabio Nonato, Matthias Muller, et al. Ldm3d: Latent diffusion model for 3d. *arXiv preprint arXiv:2305.10853*, 2023.

Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2024.

Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. *arXiv*, 2023.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Tianyou Liang, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Mvimgnet: A large-scale dataset of multi-view images. In *CVPR*, 2023.

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19447–19456, June 2024.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2D Feature Representations by 3D-Aware Fine-Tuning. In *European Conference on Computer Vision (ECCV)*, 2024.

Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. 2021.

Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta Kadambi. Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21676–21685, 2024a.

Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Suya Bharadwaj, Tejas You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. *arXiv preprint arXiv:2404.06903*, 2024b.

Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 371–378, 2001.

A APPENDIX

A.1 DETAILS OF CALCULATING THE WEIGHT

To compute λ_{ij}, we first calculate the Absolute Pose Error (APE) for each pose pair using the formula: $E_{ij} = P_i^{-1} P_j$, where P_i and P_j are the different camera poses respectively. After obtaining E_{ij}, the APE is calculated as:

$$APE_{ij} = \|E_{ij} - I_{4\times4}\|_F, \tag{12}$$

where $I_{4\times4}$ is the identity matrix F and represents the Frobenius norm. In each iteration, the APE values are normalized across all image pairs to derive the weights λ_{ij}, as: $\lambda_{ij} = \frac{APE_{ij}}{\sum_k APE_k}$, where k represents each image pair within one iterations. This normalization ensures they reflect the relative contributions of each pose error in a consistent manner. This method is implemented based on the APE computation approach in the evo library (Grupp, 2017).

A.2 DETAILS OF DATASET

We create a correspondence pair dataset based on the training set of DL3DV-10K (Ling et al., 2024) dataset to fine-tune our VAE encoder. We randomly sample 784 scenes and extract correspondence pairs from the multi-view images by using COLMAP. The correspondence points for each scene will be pre-computed before the model fine-tuning process. We use a sequential matcher with the number of overlapping images set to 10 and the number of quadratic overlaps set to 1. Such overlapping searching strategy ensures our model not only learns from easy and dense correspondence, but also from challenging cases among far-view image pairs, adding great robustness for our model. The ability to remain consistency in large view difference is particularly necessary for the outdoor unbounded reconstruction. Moreover, we set the minimum number of inliers and minimum ratio of inliers to 15 and 0.25 with the loop detection to make sure the extracted correspondence is accurate enough. We also train the same number of latent 3D Gaussian splatting scenes from the DL3DV-10K datasets to create a paired dataset of images and rendered latents, which are used for Stage-III decoder fine-tuning.

A.3 IMPLEMENTATION DETAILS

For Stage-I, we employ the pre-trained VAE model ($f = 8$, KL), from LDM model zoo as the backbone VAE model. We fine-tune the VAE on 2 NVIDIA A100-80GB GPUs for around one day, by using the correspondence pair dataset with an image resolution of 512×512, the base learning rate of $4.5e-06$, and the default optimizer. For Stage-III, we fine-tune the decoder on the image-latent dataset with 2 NVIDIA A100-80GB GPUs for around one day.

In the implementation of LRF, we normalize the latent input to the radiance field using the scale of all input views to stabilize radiance field optimization, and apply denormalization during rendering. During the VAE encoding stage, we start the discriminator at step 501 for better image quality, and we set $KL_{\text{weight}} = 1.0 \times 10^{-6}$, and $\mathcal{D}_{\text{weight}} = 0.5$. For the decoder training, we use the same configuration as the original VAE, except $KL_{\text{weight}} = 0$ to ensure only the decoder was optimized.

A.4 IMAGE RECONSTRUCTION PERFORMANCE

To verify that our approach does not degrade the performance on downstream tasks, we evaluate the image reconstruction performance of our fine-tuned VAE by calculating PSNR between the original images and the reconstructed images. As shown in Table 4, adding the correspondence consistency constraint to inject 3D awareness and applying a regularization loss to keep the latent space close to the original latent space perform minimal impact on the VAE's reconstruction performance. This ensures that our VAE model can still be effectively used in conjunction with other pre-trained models, such as the Stable Diffusion model, without any fine-tuning.

Table 4: Evaluation of PSNR for images reconstructed by VAEs on NeRF-LLFF, DL3DV-10K, and Mip-NeRF360 datasets.

Method	Metric	NeRF-LLFF	DL3DV-10K	Mip-NeRF360
VAE	PSNR ↑	23.47	24.59	24.54
Our-VAE	PSNR ↑	23.59	23.25	24.24

A.5 More Image Generation Results

Fig. 7 demonstrates that our VAE model can generate 3D objects guided by text prompts without any fine-tuning of the diffusion model. Moreover, Fig. 8 shows that our VAE can also improve the GSGEN (Chen et al., 2024) to achieve better 3D generations with complicated text prompts.

A.6 Efficiency Analysis

Table 5 demonstrates that our method reduces input resolutions, model storage space, and GPU usage for photorealistic NVS, which is particularly useful in cases with limited communication bandwidth and storage. For instance, some individuals may not have GPUs with large memories, where our method is an efficient solution for them to run photorealistic NVS algorithms.

Table 5: Efficiency comparison of different image-space and latent-space NVS methods.

Method	Input resolution	Training Time ↓	GPU Usage ↓	Storage ↓	Rendering FPS ↑	Decoding FPS ↑	PSNR ↑	SSIM ↑	LPIPS ↓
3DGS	*512×512*	*5.9 min*	*3 GB*	*200.41 MB*	*100*	-	*26.17*	*0.778*	*0.009*
3DGS/8	64×64	**3.1 min**	1 GB	**59.15 MB**	**200**	-	14.03	0.352	0.541
3DGS-VAE	64×64	4.8 min	2 GB	250.97 MB	80	20	20.57	0.595	0.346
Latent-NeRF	64×64	27.2 min	10 GB	350.50 MB	0.09	20	18.16	0.530	0.432
Ours	64×64	3.9 min	**1 GB**	96.42 MB	180	20	**22.45**	**0.667**	**0.197**

A.7 More Experimental Results

To demonstrate the effectiveness and generalizability of our method for 3D latent reconstruction, we show more NVS and 3D generation results on four datasets covering indoor scenes, outdoor scenes, and object-level scenes. As shown in Fig. 9, 10, 11 and 12, our method yields a significant improvement in image quality.

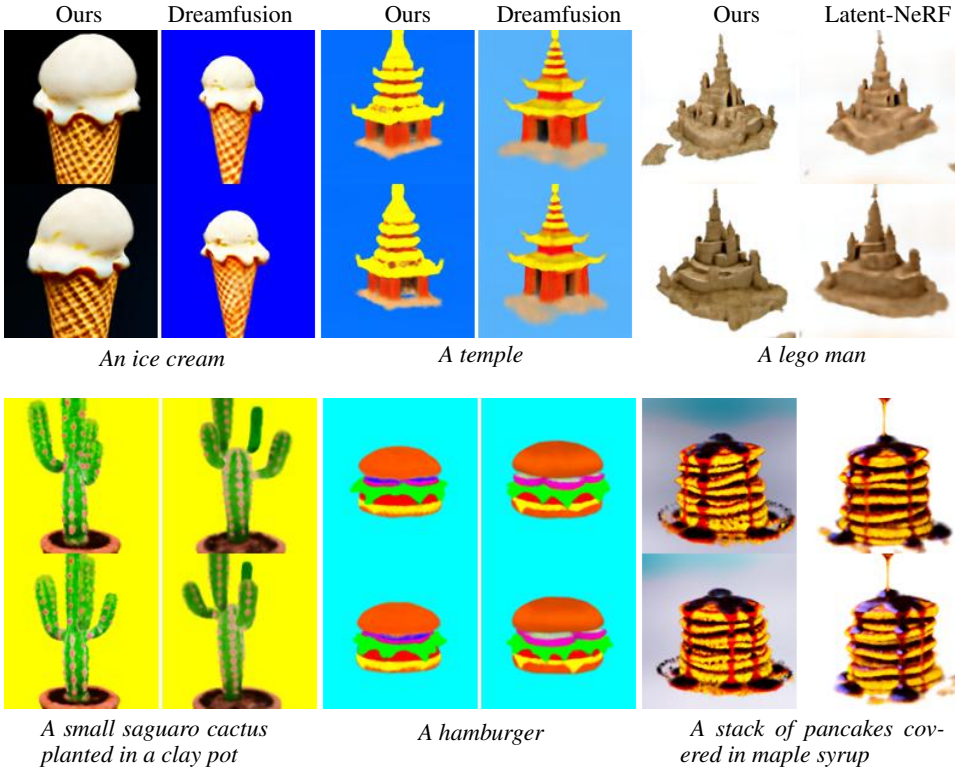

Figure 7: Samples for text-to-3D generation on the image and latent space.

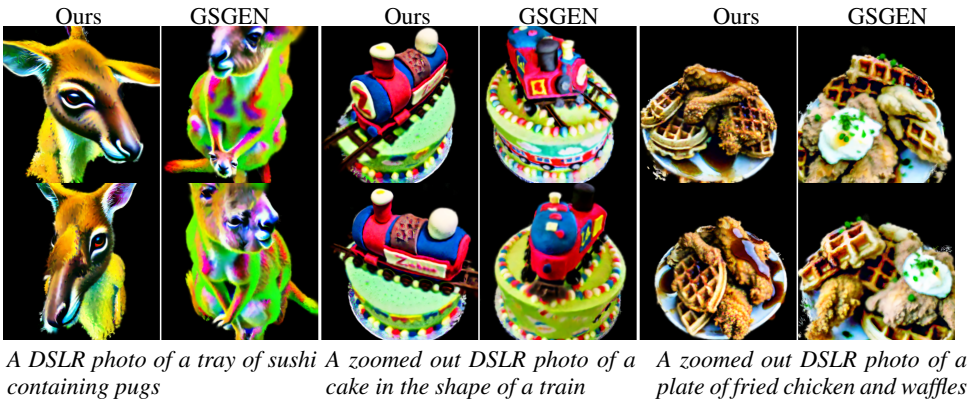

Figure 8: More samples for text-to-3D generation on the image space.

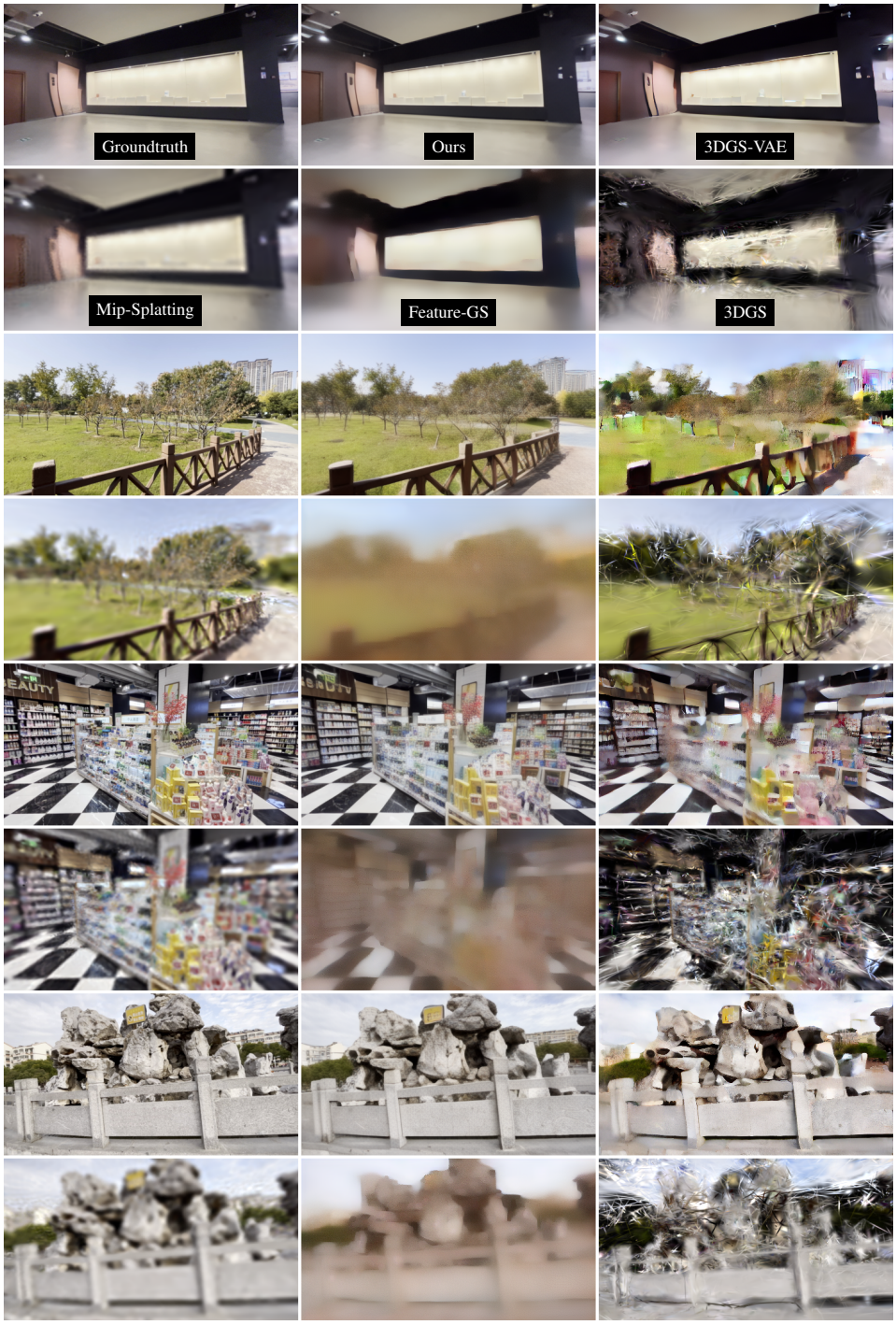

Figure 9: More NVS results on the **DL3DV-10K** dataset.

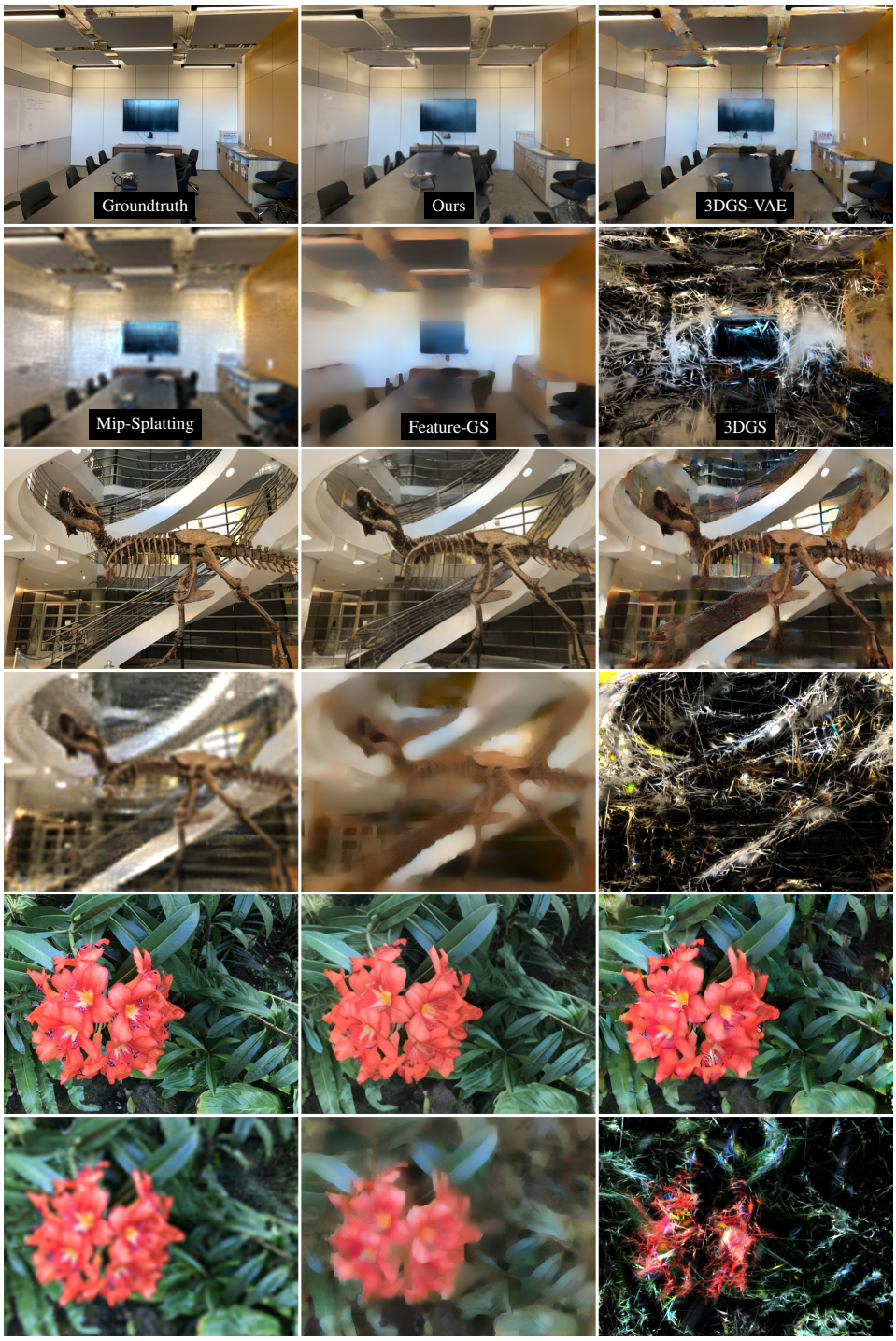

Figure 10: More NVS results on the **NeRF-LLFF** dataset.

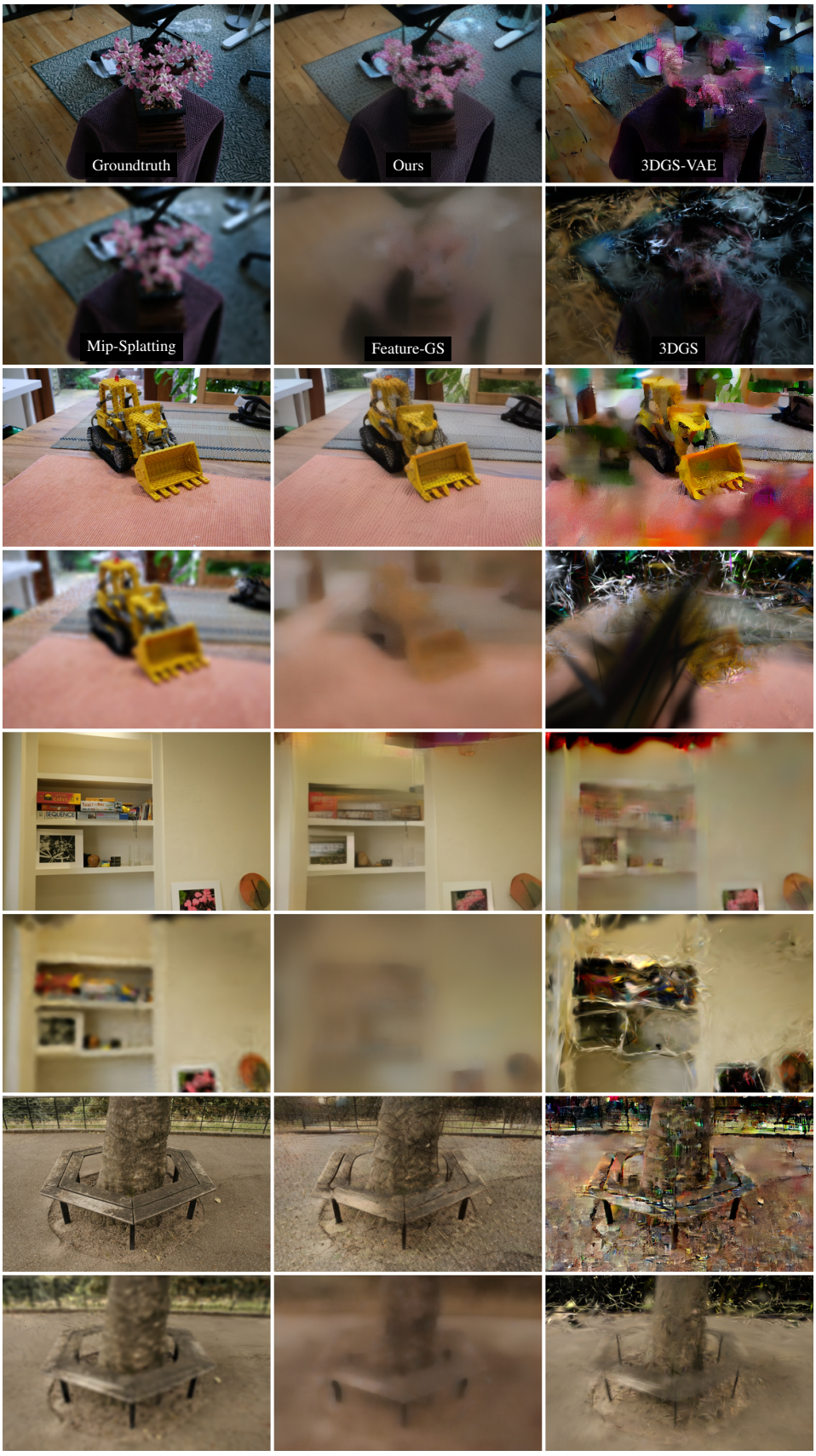

Figure 11: More NVS results on the **Mip-NeRF360** dataset.

Figure 12: More NVS results on the **MVImgNet** dataset.