

LangScene-X: Reconstruct Generalizable 3D Language-Embedded Scenes with TriMap Video Diffusion

Supplementary Material

1. Additional Implementation Details

Datasets. We adopt our method on LERF-OVS [4] and Scannet [1] dataset. The original annotations in the LERF-OVS dataset may not be consistent across different frames (e.g., the same object may not be annotated in all frames in which it appears), potentially leading to inaccurate evaluations. To address this issue, we refine the annotations of the LERF-OVS dataset to ensure consistency and alignment of annotations across all frames. We annotate the Scannet dataset in a manner similar to that of the LERF-OVS dataset.

Metrics. We utilize mIoU and localization mAcc as evaluation metrics for open-vocabulary semantic segmentation. To compute localization mAcc, we first smooth the relevancy map values by applying a mean convolution filter with a kernel size of 20, which mitigates the influence of outliers. Next, we locate the maximum score within the relevancy map and deem it correctly localized if its coordinates lie within the ground truth bounding box. Finally, we compute the average accuracy across all classes to determine the overall performance.

Baselines. We select one 2D language feature extraction method and three state-of-the-art 3D language-embedded scenes reconstruction methods as baselines for comparison:

- **LSeg** [5] is a language-driven 2D semantic segmentation method which extracts pixel-aligned language fetures for each input image.
- **LangSplat** [7] is a groundbreaking work that introduces language features into 3DGS [3] for the first time. It assigns compressed CLIP [8] features to each gaussian primitive, enabling the construction of a 3D language-embedded scene.
- **LangSurf** [6] is an extension of LangSplat that enhances the 3DGS reconstruction process by incorporating additional geometric constraint terms. It also jointly optimizes the language features and geometric properties of each gaussian primitive to better align the language features with object surfaces.
- **LSM** [2] is a feed-forward method that directly regress 3DGS with language features from uncalibrated and unposed images.

2. Additional Experiments and Analysis

2.1. More Comparison Results

2D Segmentation Comparison Results. Additionally, we visualize the 2D segmentation results and compare them

Table 1. **Comparison Across Varying Numbers of Views.** We report the open-vocabulary localization accuracy (%) and 2D semantic segmentation (IoU scores) on LERF-OVS (Teaime) [4]. The **bold** denotes the best results.

	2 views	3 views	4 views	6 views	8 views
mAcc	9.72	30.53	51.71	79.42	85.26
mIoU	3.08	16.27	32.36	48.02	53.14

with those of other methods [2, 5–7], demonstrating our method’s superior performance in segmenting comprehensive objects with sharper boundaries.

Mesh comparison Results. We also conduct a qualitative comparison between our method and LangSurf [6] in terms of mesh quality with sparse two views as inputs. As shown in Figure 2 and Figure 3, our method is capable of producing larger-scale scene reconstructions with fine-grained details, thereby enabling more powerful scene perception.

Comparison Results with Varying Numbers of Views. The TriMap Video Diffusion Model is capable of not only interpolating videos from two input views but also performing interpolation among multiple input views. To achieve interpolation with more than two input views, we first perform pairwise interpolation between adjacent views across all input views to generate local video segments. These local video segments are then concatenated to produce the final global video. To assess the performance of different views, we evaluate 2D semantic segmentation results for various views in the scene ”Teatime” of LERF-OVS [4] dataset. In the evaluation, we fix the first and last views within a set of views while varying the total number of views in the set. As shown in Table 1, the mAcc and mIoU increase when the number of views increase. As shown in Table 1, both mAcc and mIoU improve as the number of views increases. To fully leverage the potential of multiple views, the first and last views are positioned sufficiently far apart. Consequently, settings with fewer views(e.g. two views) may fail to generate coherent results, leading to significantly degraded performance and suboptimal numerical outcomes.

2.2. More Visual Results

We present the outputs of the TriMap Video Diffusion Model in Figure 6 and Figure 7. It can be noticed that the visual content within each video output is consistent, and the RGB, normals and semantic maps video share a consistent camera trajectory.

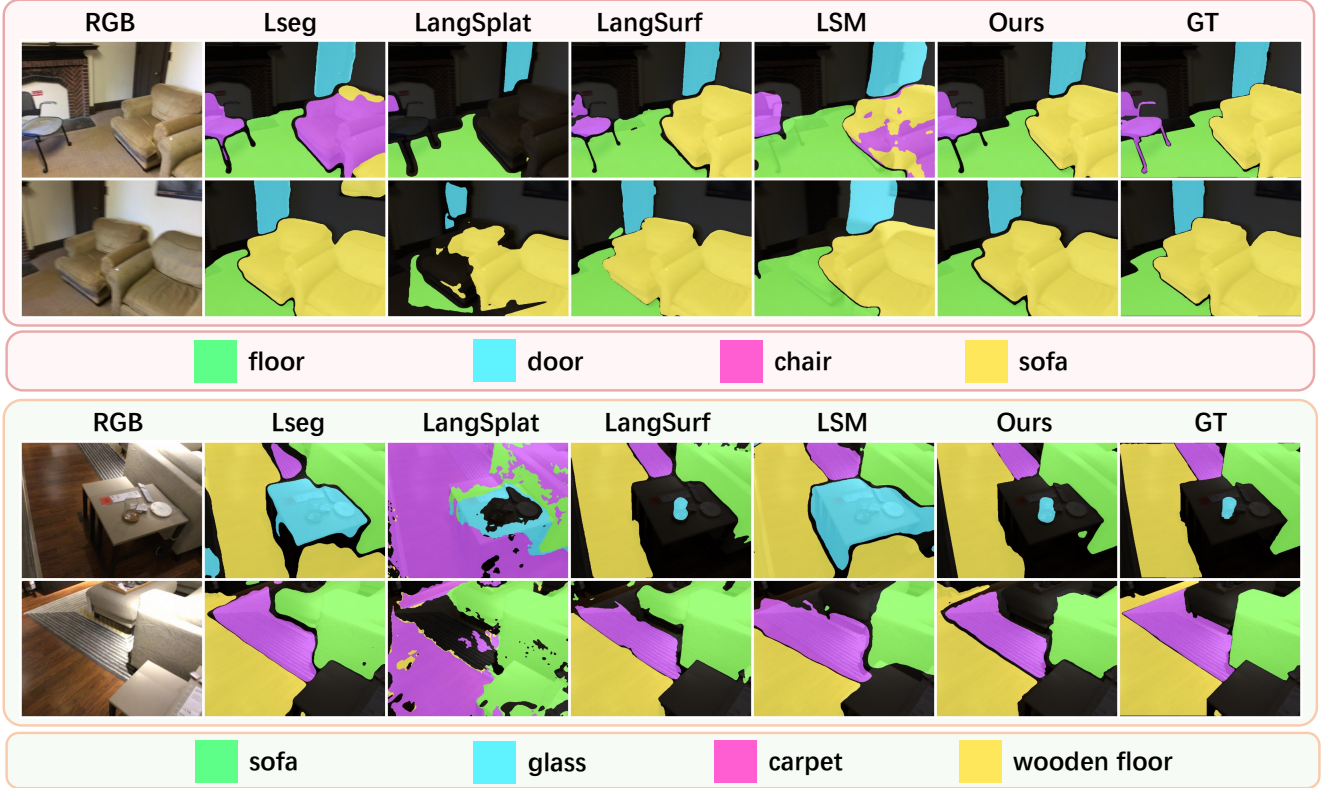


Figure 1. Qualitative comparison between our method and other methods in terms of 2D open-vocabulary semantic segmentation on Scannet [1].

2.3. Applications

By constraining Gaussian primitives to align with object surfaces, our method enables the representation of semantic features through individual surface-bound Gaussian primitives. This property facilitates direct semantic applications of 3D point clouds, eliminating the need for intermediate 2D rendering. To validate the effectiveness of our language-embedded scene reconstruction, we apply our method to tasks such as 3D semantic segmentation and 3D object removal.

3D Semantic Segmentation. Figure 4 demonstrates the results of 3D semantic segmentation. For each 3D point, we query it using a text prompt and highlight the activated points or enclose them within a bounding box.

3D Object Removal. Figure 5 demonstrates the results of 3D object removal.

References

- [1] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 2, 4, 7
- [2] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, et al. Large spatial model: End-to-end unposed images to semantic 3d. *Advances in Neural Information Processing Systems*, 37:40212–40229, 2025. 1
- [3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 1
- [4] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lurf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 1, 4, 6
- [5] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 1
- [6] Hao Li, Roy Qin, Zhengyu Zou, Diqi He, Bohan Li, Bingquan Dai, Dingwen Zhang, and Junwei Han. Langsurf: Language-embedded surface gaussians for 3d scene understanding. *arXiv preprint arXiv:2412.17635*, 2024. 1, 4
- [7] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 1
- [8] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh,

Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. [1](#)

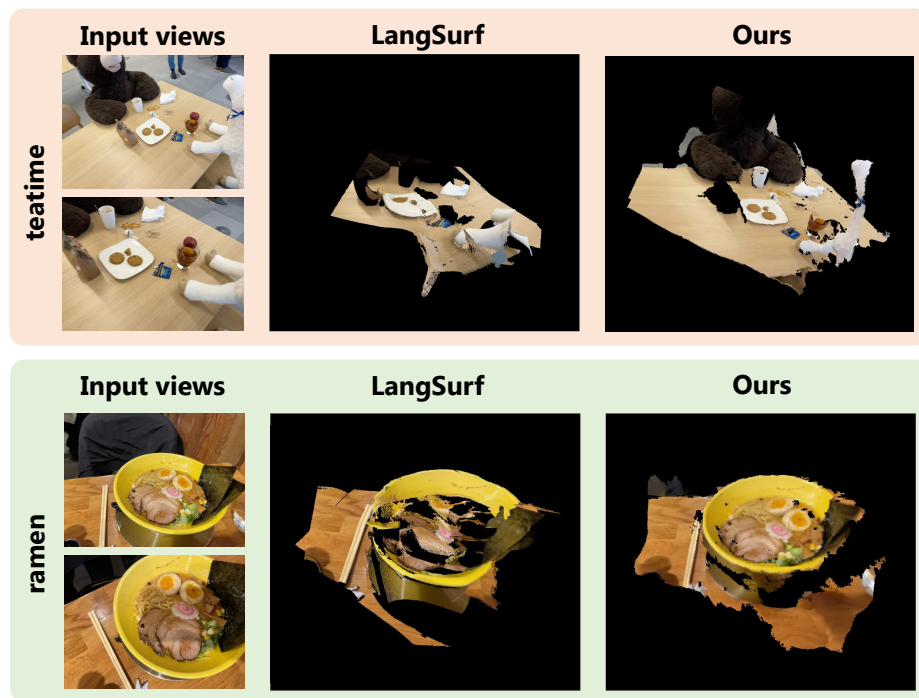


Figure 2. Qualitative Results of mesh quality between our method and LangSurf [6] on LERF-OVS [4].

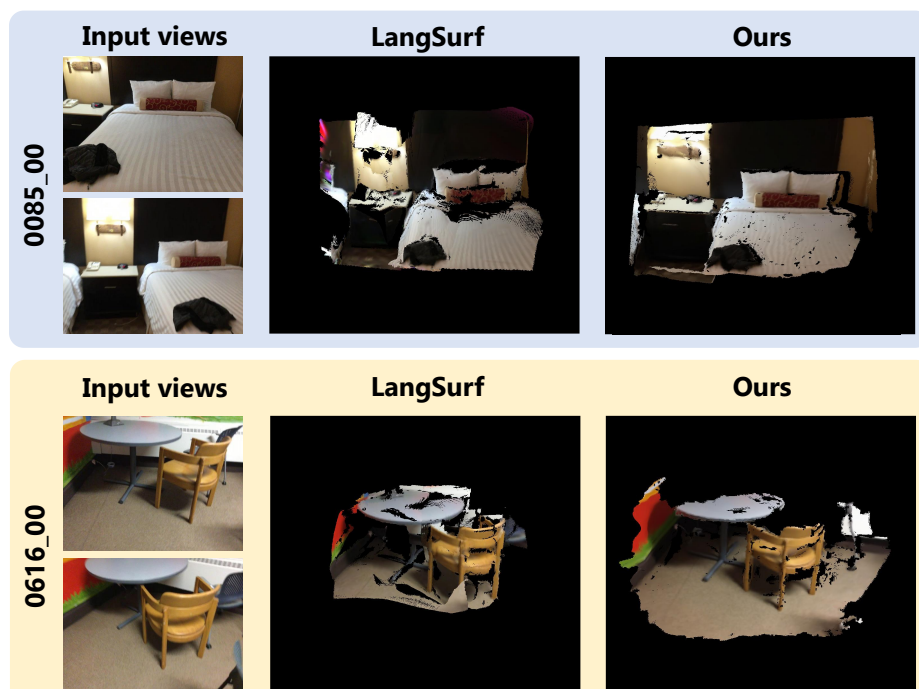


Figure 3. Qualitative Results of mesh quality between our method and LangSurf [6] on Scannet [1].

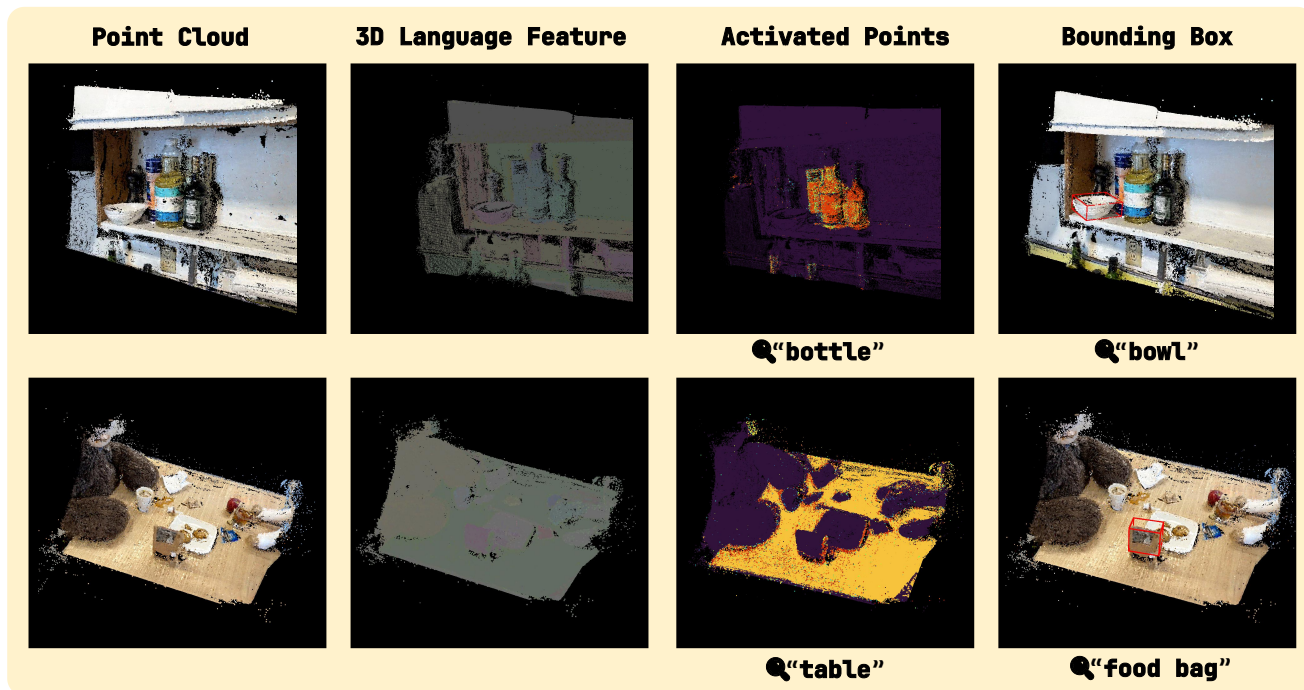


Figure 4. Visual Results on 3D semantic segmentation.

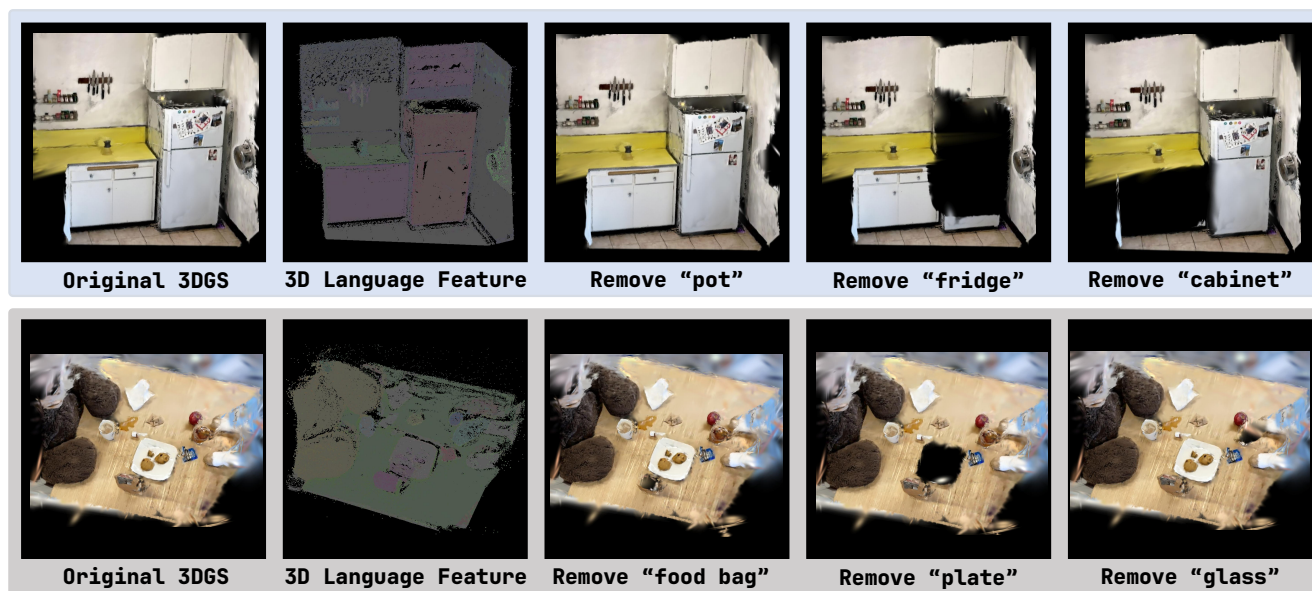


Figure 5. Visual Results on 3D semantic segmentation.

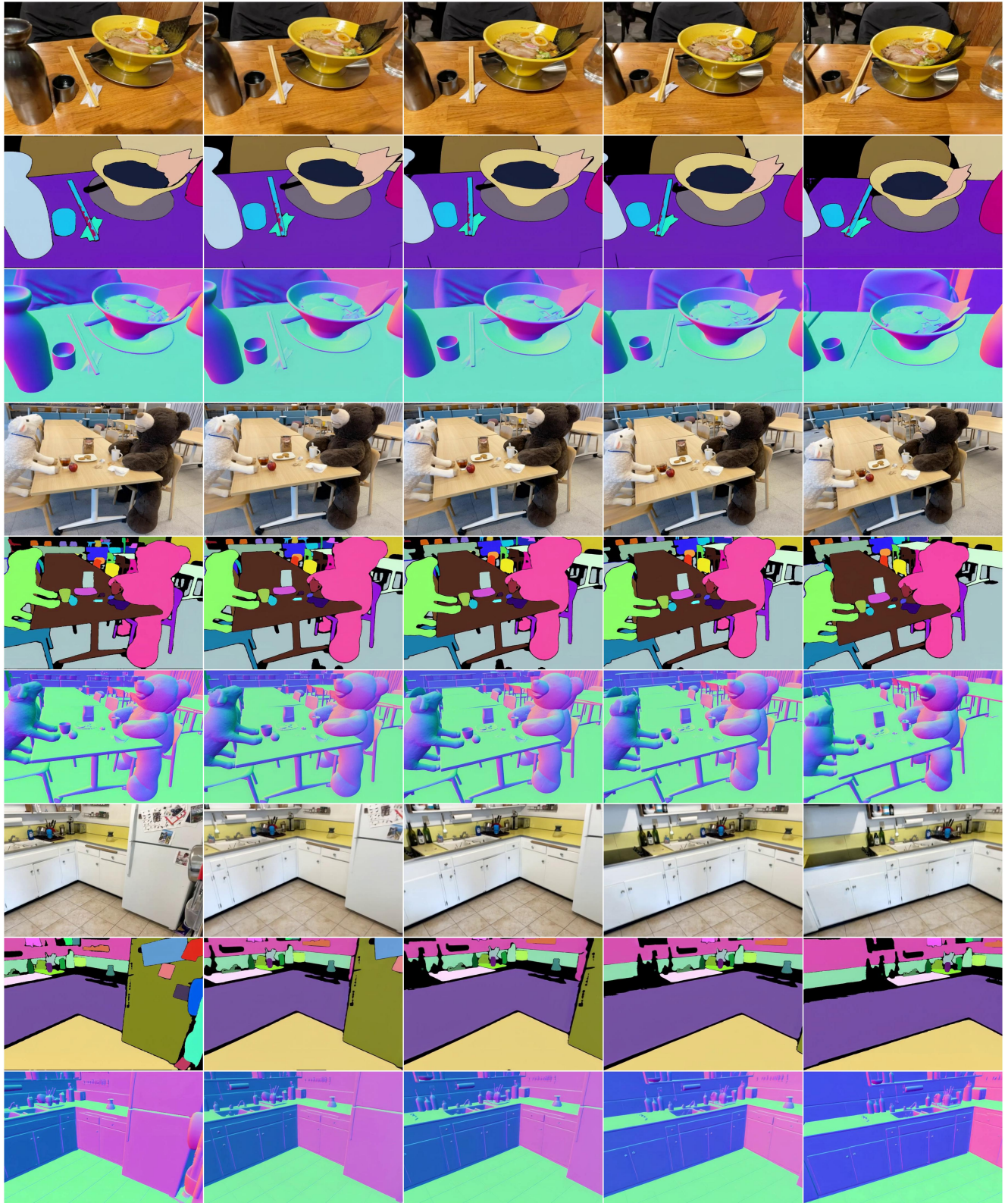


Figure 6. **Visual Results on LERF-OVS [4] Dataset.** The first and last columns represent the input views, while the intermediate columns depict the generated views.

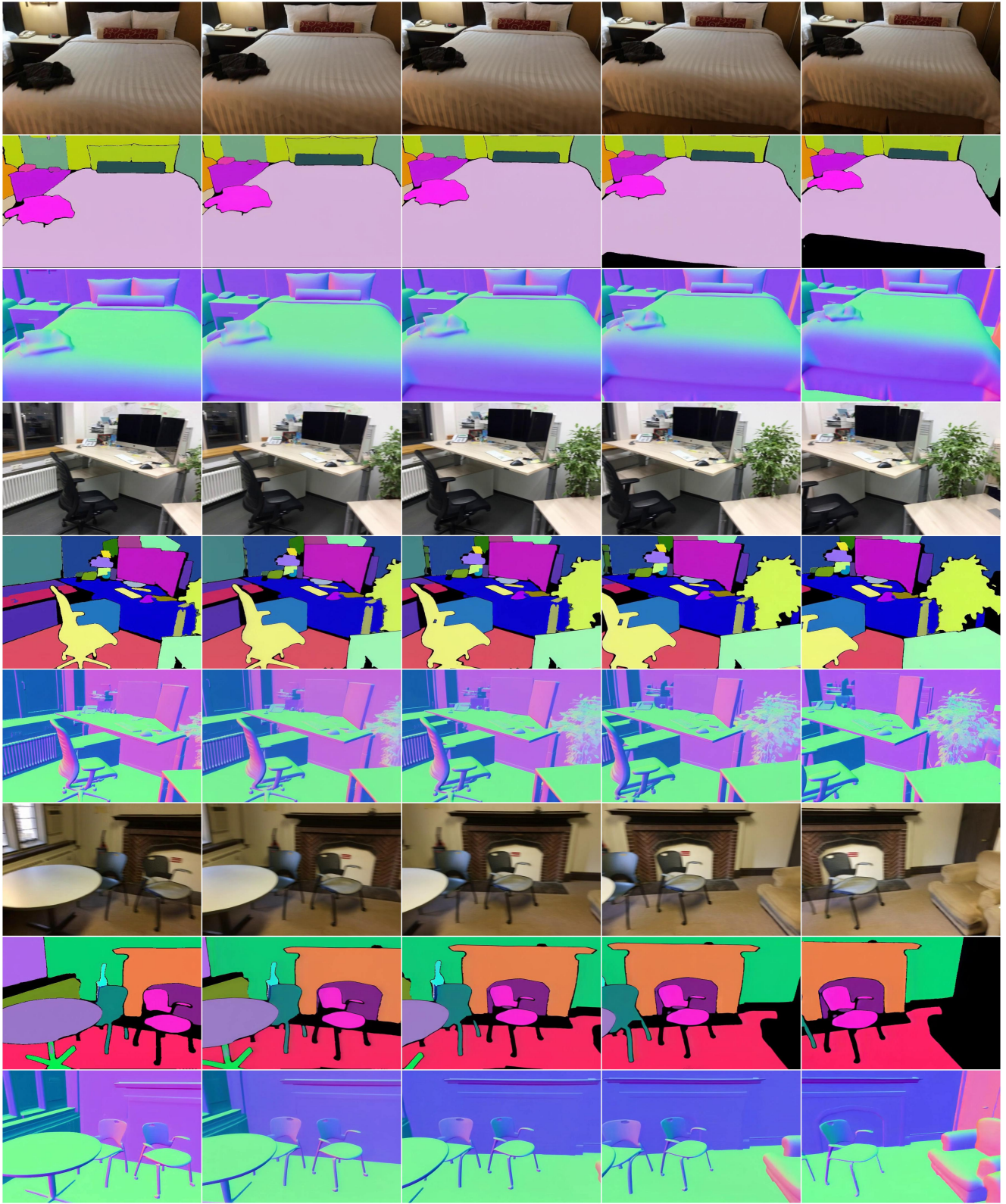


Figure 7. **Visual Results on ScanNet [1] Dataset.** The first and last columns represent the input views, while the intermediate columns depict the generated views.