Code Implementation

We have included our code implementation in the supplementary zip file, and also released it in the following anonymous repository: https://anonymous.4open.science/r/siu3r-BF1E.

Appendix

In this appendix, we provide additional content to complement the main manuscript:

- Appendix A: Additional Implementation Details
- Appendix B: Comparisons with Per-Scene Optimization Methods
- Appendix C: Additional Visualizations

A Additional Implementation Details

A.1 Data Preprocessing

As described in Sec.4.1 of our main manuscript, we utilize ScanNet[1] for training and validation. We adopt the official training and validation dataset splitting of ScanNet, and then resize and crop original images to centered images at 256×256 resolution. The camera's intrinsic parameters have also been adjusted accordingly. We followed [2]'s camera conventions, where intrinsics are normalized and extrinsic parameters are OpenCV-style camera-to-world matrices.

Our data samples are obtained by randomly sampling context image pairs with certain overlaps. The overlap is determined by a pair-wise Intersection over Union (IoU) metric as shown in Fig.I. During training, we constrain the IoU to $[0.3, 0.8]$ to randomly select our training samples from scenes. Specifically, the IoU metric can be calculated as follows:

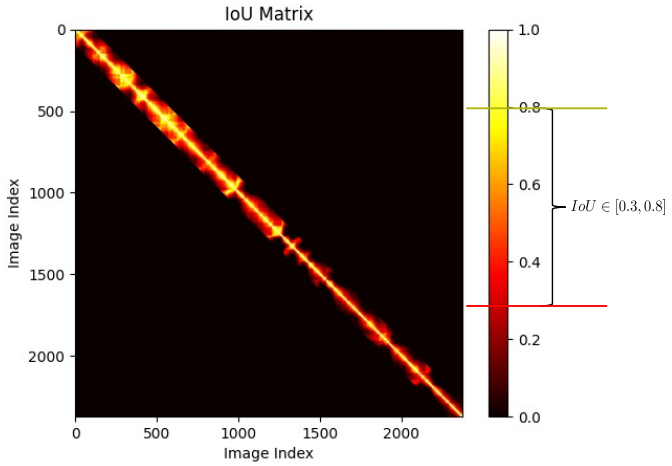

Figure I: **IoU matrix of ScanNet scene0011_00**

1. For a pair of images $\boldsymbol{I}_1, \boldsymbol{I}_2$, obtain their depths $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, poses $\boldsymbol{P}_1$, $\boldsymbol{P}_2$ and intrinsic K.
2. Unproject $\boldsymbol{D}_1$ into world coordinates and project them to $\boldsymbol{D}_2$'s camera to obtain $\boldsymbol{D}_1'$. Only depths satisfying $|\boldsymbol{D}_1' - \boldsymbol{D}_2| < 0.1$ are considered as valid.
3. Calculate the intersection over union ratio as:

$$IoU_{i \to j} = \frac{\text{\#valid projected depths}}{\text{\#total depths}}$$

4. Calculate $IoU_{1 \to 2}$ and $IoU_{2 \to 1}$.

5. Define the final IoU as:

$$IoU = \frac{IoU_{1\to 2} + IoU_{2\to 1}}{2}$$

The same IoU-based sampling strategy is also adopted in our evaluation, where we select 1,860 context image pairs to formulate the validation set. The curated evaluation benchmark and its processing scripts will be made publicly available for reproducing our results.

A.2 Network Architecture and Hyperparameters

In Table I (a), the order from top to bottom are the network details of Image Encoder, Gaussian Decoder, Unified Query Decoder, respectively. In Table I (b), we specify loss weights for Eq.5 in our main manuscript, which is followed by parameters used in our training phase. To enable the Unified Query Decoder to leverage MASt3R features for scene understanding, we pre-trained the decoder on COCO dataset [3] while keeping the Image Encoder's weights frozen. The pre-trained weights will be publicly released to facilitate further research and development.

(a) Network Architecture		
Image Encoder	architecture	ViT encoder with Adapter[4]
	initialization	MASt3R[5]
	# depth of ViT encoder	24
	# embed dim of ViT encoder	1024
	# attn heads of ViT encoder	16
	positional embedding	RoPE
	# patchsize	16
	# interaction blocks of adapter	[5, 11, 17, 23]
	attention of adapter	MSDeformAttn
	# attention heads of adapter	16
	# inplanes of adapter spatial prior module	64
	# embed dim of adapter spatial prior module	1024
	# ref points	4
	# deform ratio	0.5
Gaussian Decoder	architecture	ViT decoder with DPT head[6]
	initialization	MASt3R[5]
	# depth of ViT decoder	12
	# embed dim of ViT decoder	768
	# attn heads of ViT decoder	12
	# channels of DPT head	83
	# sh degree	4
	# min gaussian scale	0.5
	# max gaussian scale	15.0
Unified Query Decoder	architecture	mask decoder[7, 8]
	# queries	100
	# probability score threshold of queries τ_c	0.5
	# probability score threshold of pixels τ	0.3
	# attn layers for text refer segmentation	6
(b) Hyperparameters		
Loss Weights	# $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$	1.0, 0.5, 0.05, 0.05, 1
Training Details	learning rate scheduler	Cosine
	# epochs	100
	# learning rate	1e-4
	# batch size on each device	3
	training devices	8 * RTX 4090
	optimizer	AdamW[9]
	# beta1, beta2	0.9, 0.95
	# weight decay	0.05
	# warm-up epochs	3
	# gradient clip	1.0

Table I: Details of network architecture and hyperparameters. In the table, "#" denotes numerical parameters. We present parameters that specify our network architecture, and parameters used in our loss functions and training phase, in (a) and (b), respectively.

A.3 Implementation Details about Versatile 3D Editing

As shown in Fig.1 of main manuscript, our simultaneous modeling of scene understanding and 3D reconstruction enables diverse 3D scene manipulations through unified pixel-aligned representations, including instance removal, replacement, relocation, and recoloring.

Here we take instance removal as an example to derive the implementation of such 3D editing:

1. Conduct inference to obtain SIU3R field with pixel-aligned 3D masks $\boldsymbol{M}$ and Gaussians $\boldsymbol{\mathcal{G}}$.
2. Remove Gaussians for a specified instance (ID $= ins\_id$):

$$\boldsymbol{\mathcal{G}'} = \boldsymbol{\mathcal{G}} \setminus \{\boldsymbol{g_v^{ij}} | \boldsymbol{M}_{ins}^{v,ij} = ins\_id\}$$

3. The modified Gaussians $\boldsymbol{\mathcal{G}'}$ are rendered into original context views to obtain images $\boldsymbol{\mathcal{I}'}$, with an off-the-shelf diffusion-based inpainting model [10] applied to fill the removed regions while ensuring visual coherence.
4. Conduct inference once again and rebuild SIU3R field from $\boldsymbol{\mathcal{I}'}$.

For other 3D editing tasks (i.e. instance replacement, relocation and recoloring), we adopt a similar approach powered by different diffusion models [11–13].

B Comparisons with Per-Scene Optimization Methods

We also compare our approaches to methods (i.e., Feature-3DGS[14] and NeRF-DFF[15]) that require dense view capturing and per-scene optimization. Both of the two per-scene optimization methods follow a feature alignment paradigm similar to the feed-forward method LSM[16], where their 3D understanding capabilities are powered by off-the-shelf 2D vision language models that can only support language-guided segmentation. To enable the training of Feature-3DGS and NeRF-DFF, we uniformly select dense views (i.e., $\sim$100 images) as input and conduct per-scene optimization for each scene to align 3DGS or NeRF field with 2D features via rasterization. As shown in Table II, our method surpasses all of the feature alignment-based approaches by a large margin in the task of scene understanding, no matter they perform reconstruction in per-scene optimization (Feature-3DGS and NeRF-DFF) or feed-forward (LSM) manner. Besides, benefiting from our align-free framework, our method can further enable instance-level understanding tasks such as instance and panoptic segmentation. Furthermore, our method is the fastest in reconstruction speed, significantly surpassing Feature-3DGS and NeRF-DFF, and leading ahead of LSM. Considering that Feature-3DGS and NeRF-DFF use much more training images than our method, our performance in novel view synthesis is acceptable while achieving the best depth accuracy owing to our mask-guided geometry refinement module. As shown in Fig.II, we also make qualitative comparisons with these feature alignment-based methods, where our method achieves superior mask quality and semantic coherence.

Table II: **Quantitative Comparisons with Per-Scene Optimization Methods.**

	Depth Estimation		**Novel View Synthesis**			**Scene Understanding**			**Efficiency**
	AbsRel ↓	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	mIoU ↑	mAP ↑	PQ ↑	Reconstruction Time ↓
Feature-3DGS[14]	0.1546	0.3585	**28.69**	**0.8893**	0.2171	0.3965	-	-	145.30min
NeRF-DFF[15]	0.1846	0.4151	20.12	0.6252	0.5136	0.3410	-	-	2.71min
LSM[16]	0.07468	0.2190	21.88	0.7336	0.3035	0.2745	-	-	0.24s
Ours	**0.07421**	**0.2081**	25.96	0.8220	**0.1841**	**0.5922**	**0.2817**	**0.6612**	**0.13s**

C Additional Visualizations

C.1 Instance, Panoptic and Text-Referred Segmentation

In our main manuscript, we have included qualitative comparisons with other methods and demonstrated our superiority in semantic segmentation. Here we provide additional qualitative results of instance and panoptic segmentation for further demonstration. As shown in Fig.III, compared to 2D-based method Mask2Former[8] that leads to noisy mask boundaries, our method exhibits significantly higher mask quality. As shown in Fig.III, when performing panoptic segmentation, our method exhibits excellent mask consistency across different views and significantly outperforms other methods in mask quality. Similar effects can also be observed in text-referred segmentation results as shown in Fig.V. We attribute the superiority of our method to the simultaneous modeling of scene understanding and 3D reconstruction, which effectively leverages 3D geometric clues to aggregate semantic information from different views, and propagates them back to the original views to ensure cross-view consistency and improve segmentation accuracy.

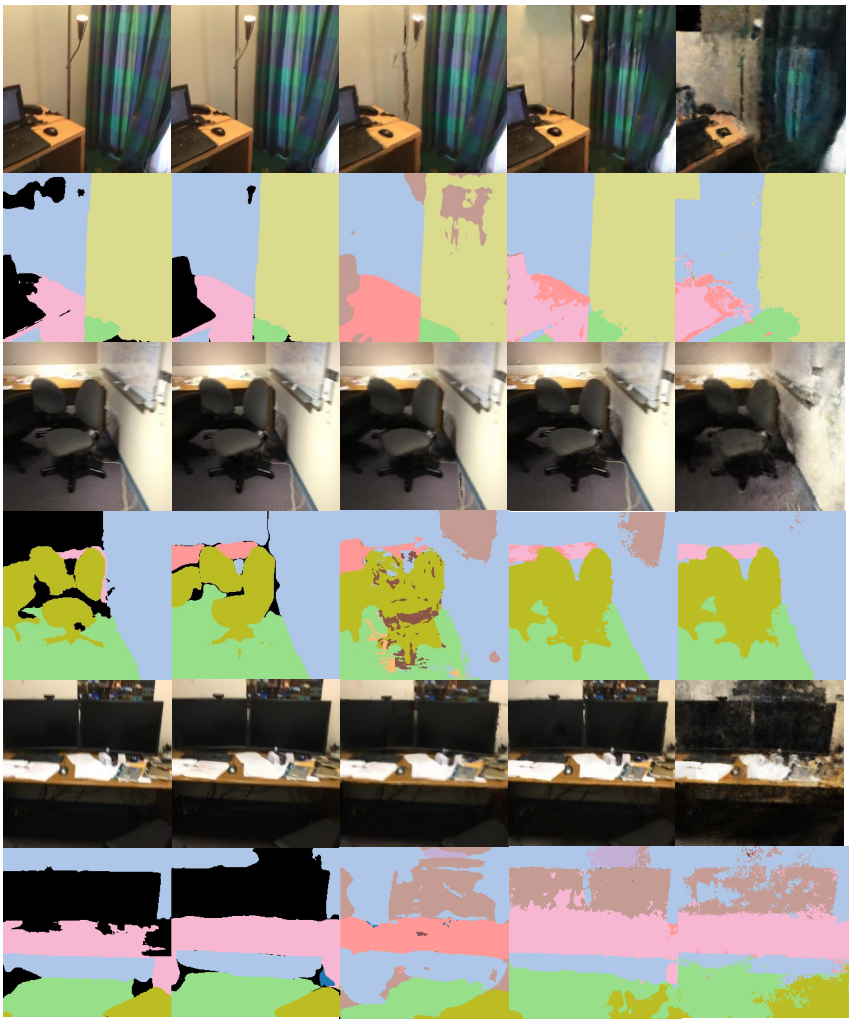

Figure II: **Qualitative Comparisons with Per-Scene Optimization Methods.** Note that the per-scene optimization methods Feature-3DGS and NeRF-DFF require about 100 images for each scene during training, while our method and LSM perform with only 2 context images in a feed-forward manner.

C.2 Depth Estimation

As illustrated in Fig.VI, compared to other feed-forward reconstruction methods, our approach achieves significantly superior depth quality with less artifacts and better coherence. We attribute this to our mask-guided geometry refinement module, which ensures geometry consistency within the same object instances under semantic guidance of 2D masks, and thus reduces erroneous depth variations that typically observed in other approaches.

C.3 Versatile 3D Editing

As shown in Fig.VII, we present a comprehensive set of versatile 3D editing results, demonstrating SIU3R's potential for diverse 3D manipulation applications. Furthermore, this capability establishes an effective baseline that bridges geometric reconstruction, scene understanding and manipulation in 3D environments.

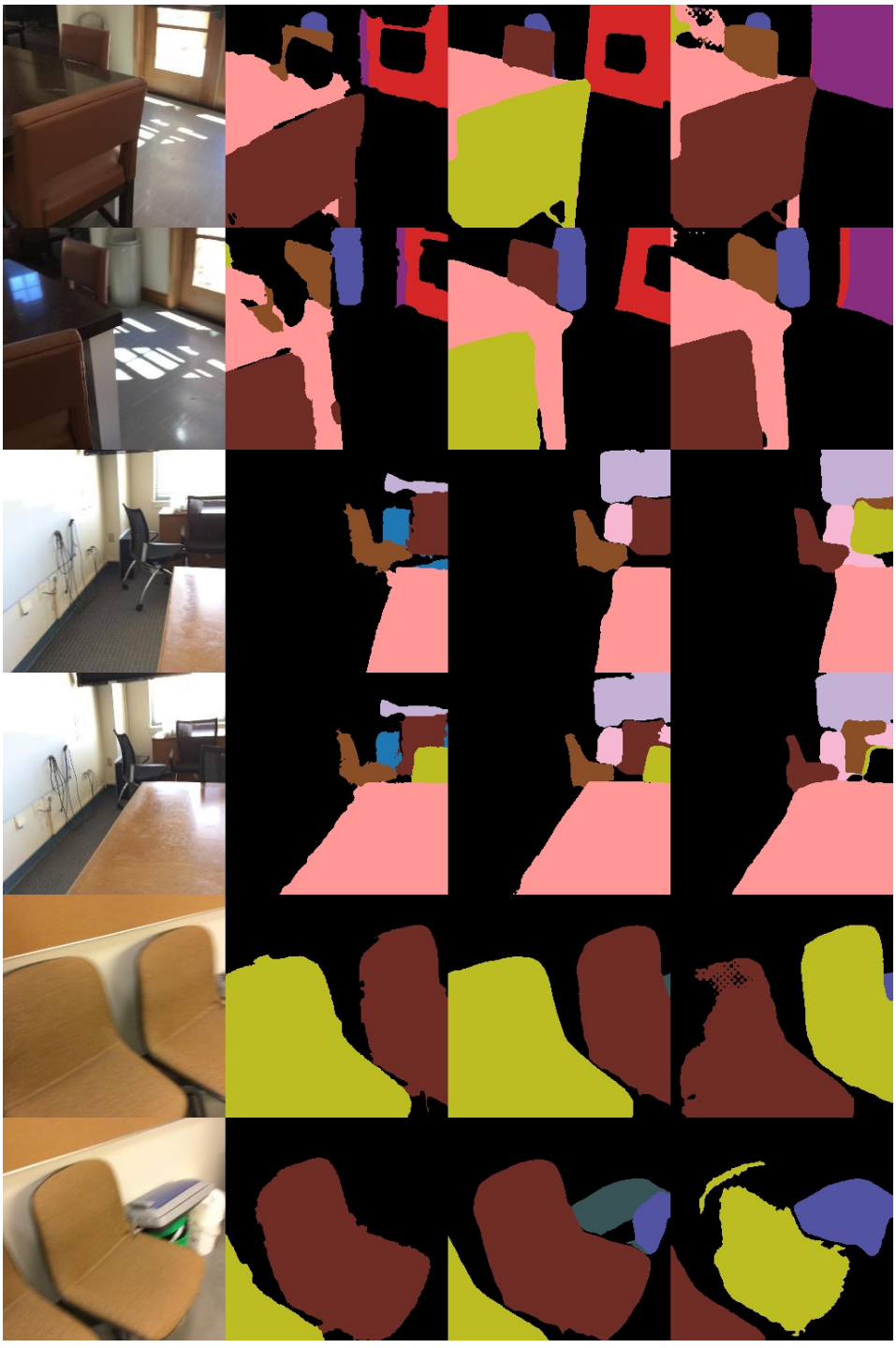

Figure III: **Qualitative Results of Instance Segmentation.**

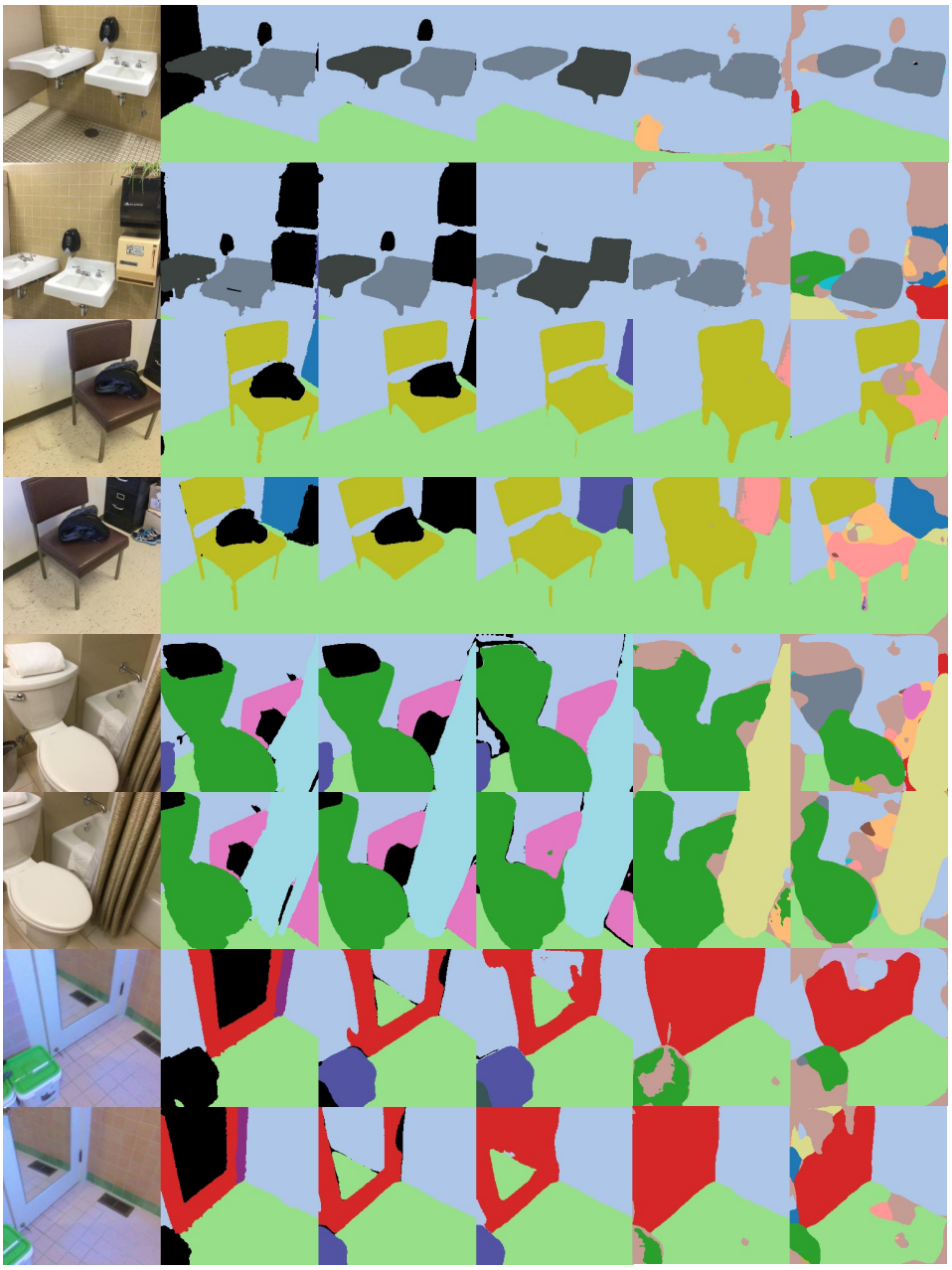

Figure IV: **Qualitative Results of Panoptic Segmentation.**

Figure V: **Qualitative Results of Text-Referred Segmentation.**

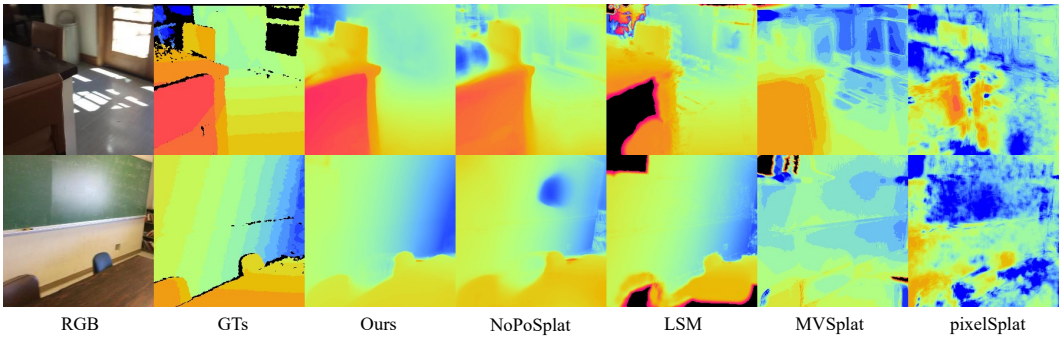

Figure VI: **Qualitative Results of Depth Estimation.**

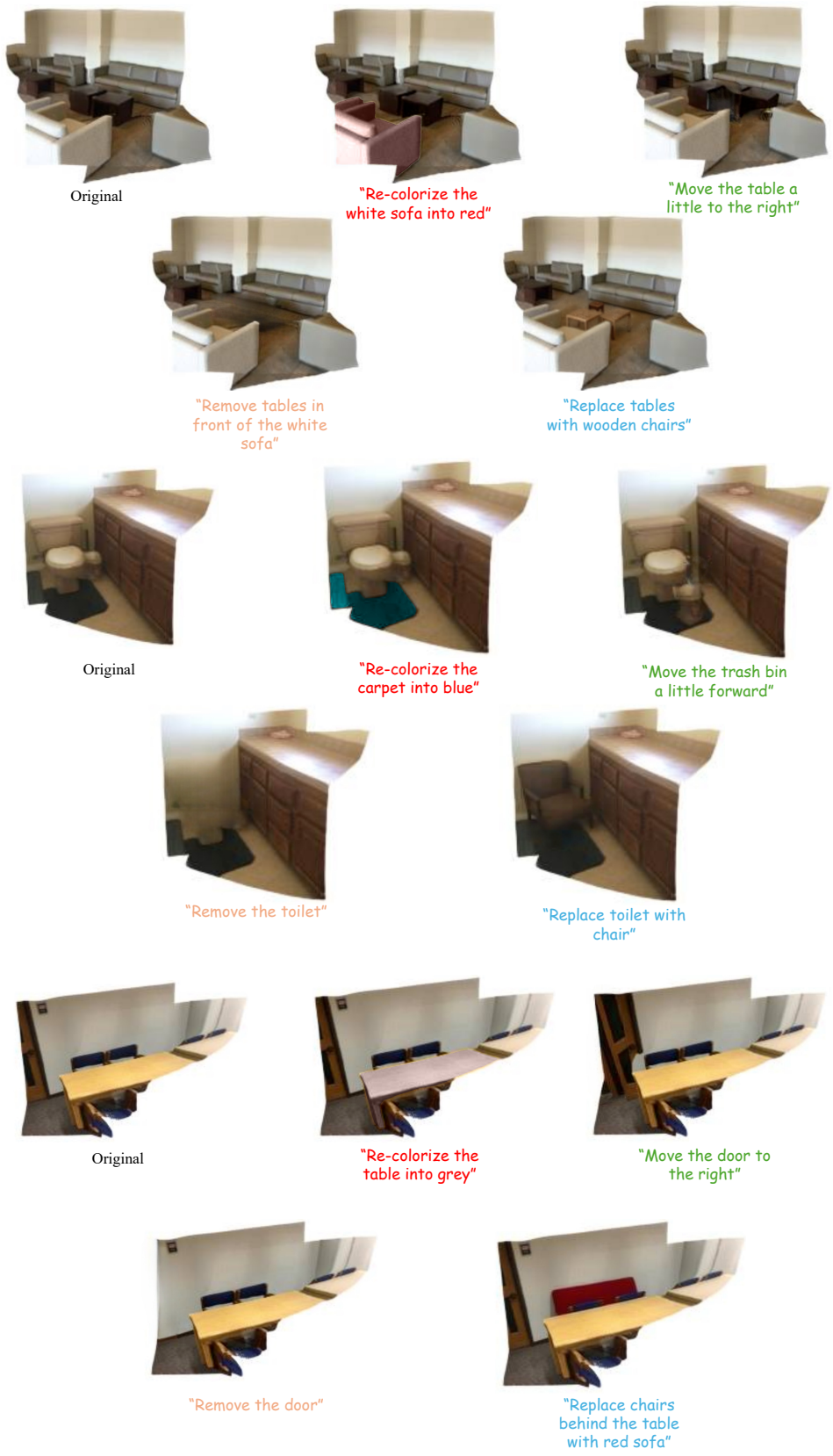

Figure VII: **Qualitative Results of Versatile 3D Editing.**

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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443, July 2017.

[2] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. No Pose, No Problem: Surprisingly Simple 3D Gaussian Splats from Sparse Unposed Images. In *The Thirteenth International Conference on Learning Representations*, October 2024.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[4] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision Transformer Adapter for Dense Predictions. In *The Eleventh International Conference on Learning Representations*, September 2022.

[5] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R, June 2024.

[6] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021.

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-Pixel Classification is Not All You Need for Semantic Segmentation. In *Advances in Neural Information Processing Systems*, volume 34, pages 17864–17875. Curran Associates, Inc., 2021.

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-Attention Mask Transformer for Universal Image Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[9] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[10] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023.

[11] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[12] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2149–2159, 2022.

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[14] 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*, pages 21676–21685, 2024.

[15] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022.

[16] Zhiwen Fan, Jian Zhang, Wenyan Cong, Peihao Wang, Renjie Li, Kairun Wen, Shijie Zhou, Achuta Kadambi, Zhangyang Wang, Danfei Xu, Boris Ivanovic, Marco Pavone, and Yue Wang. Large Spatial Model: End-to-end Unposed Images to Semantic 3D. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024.