Appendix

In the Appendix, we provide the following:

- comprehensive implementation details in Section A
- additional experiments, results, and discussions in Section B

A. Implementation Details

Training Setup We train FF3R on a subset of 6,500 scenes sampled from the DL3DV-10K [20] dataset. No additional 3D annotations such as depth, camera poses, or semantic labels are required; only multi-view RGB images are used as the supervision signal. We initialize the geometry transformer and depth DPT head with pretrained weights from VGGT [38], while all other modules are randomly initialized. The alternating-attention blocks are unfrozen to adapt to our downstream unified decoder structure, and we use a fixed CLIP-LSeg [15] as the semantic transformer. During training, each input image is set to 448×448, and each iteration randomly samples one scene, from which a subset of context views (up to 16 views) is further selected. The model is optimized using AdamW [22] with a cosine learning rate scheduler, a peak learning rate of 2×10^{-4}, and a warm-up phase of 1K iterations. Training is performed on 8 NVIDIA A100 GPUs for two days.

Training View Sampling Strategy To enhance the robustness of our model, careful design of the training view sampling strategy is crucial. Following Dust3r[40] and VGGT [38], we adopt a sequential sampling approach for DL3DV [20]. Specifically, we first randomly determine the temporal gap between the first and last frames. Within this interval, additional frames are randomly sampled to ensure that the total number of input views does not exceed 16. Since our framework imposes no requirement on temporal order, the sampled views are shuffled at each iteration. Finally, all input images are center-cropped and resized to 448×448 before being fed into the model.

Evaluation Dataset We evaluate our simultaneous geometry and semantic prediction on two widely used multi-view datasets: ScanNet [3] and the DL3DV-10K [20]. Following AnySplat [7], we first sample 72 views from the original video sequence based on spatial distribution, and further downsample them to 56 and 32 views. With the test interval set to 8, as in 3DGS [8], the corresponding numbers of context views become 32, 48, and 64, respectively. For the sparse-view setting, we use a test interval of 1. For datasets with semantic annotations (e.g., ScanNet[3]), we map the thousands of different labels into a set of common labels following [4]. To evaluate our model under more challenging and unconstrained scenarios, we additionally test on DL3DV-10K [20], which contains unbounded scenes, diverse environments, and varying lighting conditions. Since DL3DV does not provide semantic annotations, we follow Feature-3DGS [49] and adopt semantic masks predicted by LSeg [15] as pseudo ground truth. This allows us to evaluate how effectively our method lifts inherently inconsistent 2D semantic features into a geometrically consistent 3D representation.

Table 4. Ablation for Geometric Distillation.

Method	PSNR↑	SSIM↑	LPIPS↓
Ours w/o distill loss	9.37	0.248	0.783
Ours w/ distill loss	17.85	0.699	0.380

Table 5. Quantitative Results on Novel View Synthesis and Semantic Segmentation on ScanNet.

	2 Views						16 Views					
	PSNR↑	SSIM↑	LPIPS↓	mIoU↑	Acc.↑	Time(s)↓	PSNR↑	SSIM↑	LPIPS↓	mIoU↑	Acc.↑	Time(s)↓
Uni3R	15.57	0.649	0.448	0.354	0.716	0.626s	17.98	0.737	0.402	0.437	0.769	0.928s
Ours	22.70	0.787	0.285	0.486	0.754	0.884s	22.19	0.769	0.295	0.500	0.751	1.2s

B. Additional Experiments and Results

Additional Ablation Studies In Tab. 4. The results demonstrate that the distilled geometry is critical; without it, the model suffers from overfitting to the context views and fails to generalize to novel views. This highlights that our distillation mechanism is a necessary component for enabling robust, scalable reconstruction without human labels.

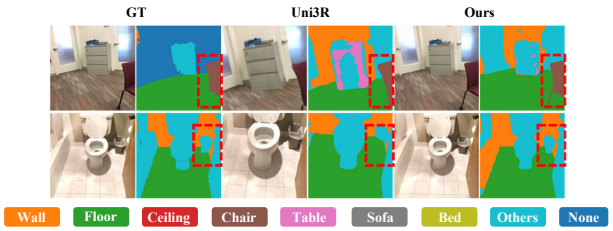

Figure 7. Qualitative Results on Novel View Synthesis and Semantic Segmentation on ScanNet.

Additional Baseline Comparisons We further evaluate our method against the state-of-the-art approach Uni3R [34]. As shown in Tab. 5 and Fig. 7, our method consistently outperforms this baseline.

Additional Qualitative Results We provide additional qualitative results of our model on simultaneous geometry and semantic reasoning in ScanNet [3] and DL3DV-10K [20] in Figs. 8, 9, and 10. As shown in Fig. 8, our

method preserves sharp boundaries between different semantic regions. Even when the 2D semantic features are inconsistent, the proposed Geometry-Guided Feature Warping effectively injects 3D awareness into the semantic features, resulting in improved generalization across challenging viewpoints.

Moreover, with the Semantic-Aware Voxelization module, our model reduces local visual artifacts by enforcing semantic consistency within each voxel, as illustrated in Fig. 9. Finally, benefiting from our fully annotation-free training strategy, the model requires no explicit semantic annotations and can be trained on large, diverse datasets such as DL3DV-10K [20]. This enables strong generalization across indoor and outdoor scenes under varying lighting conditions, as demonstrated in Fig. 10.

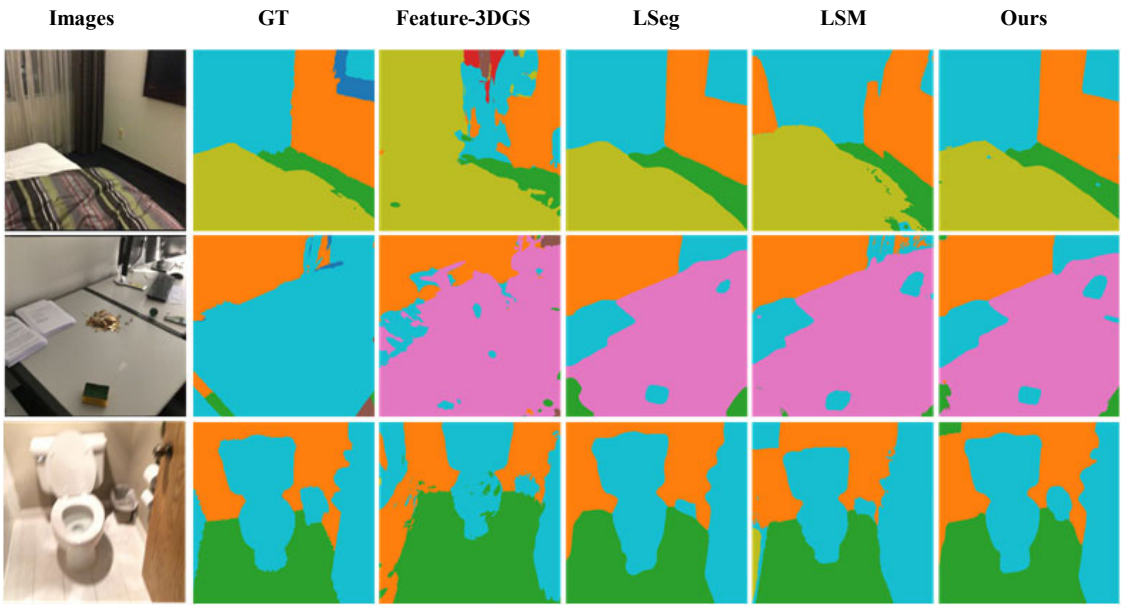

Figure 8. Qualitative results of open-volcabulory semantic segmentation.

Figure 9. Qualitative results of novel view synthesis.

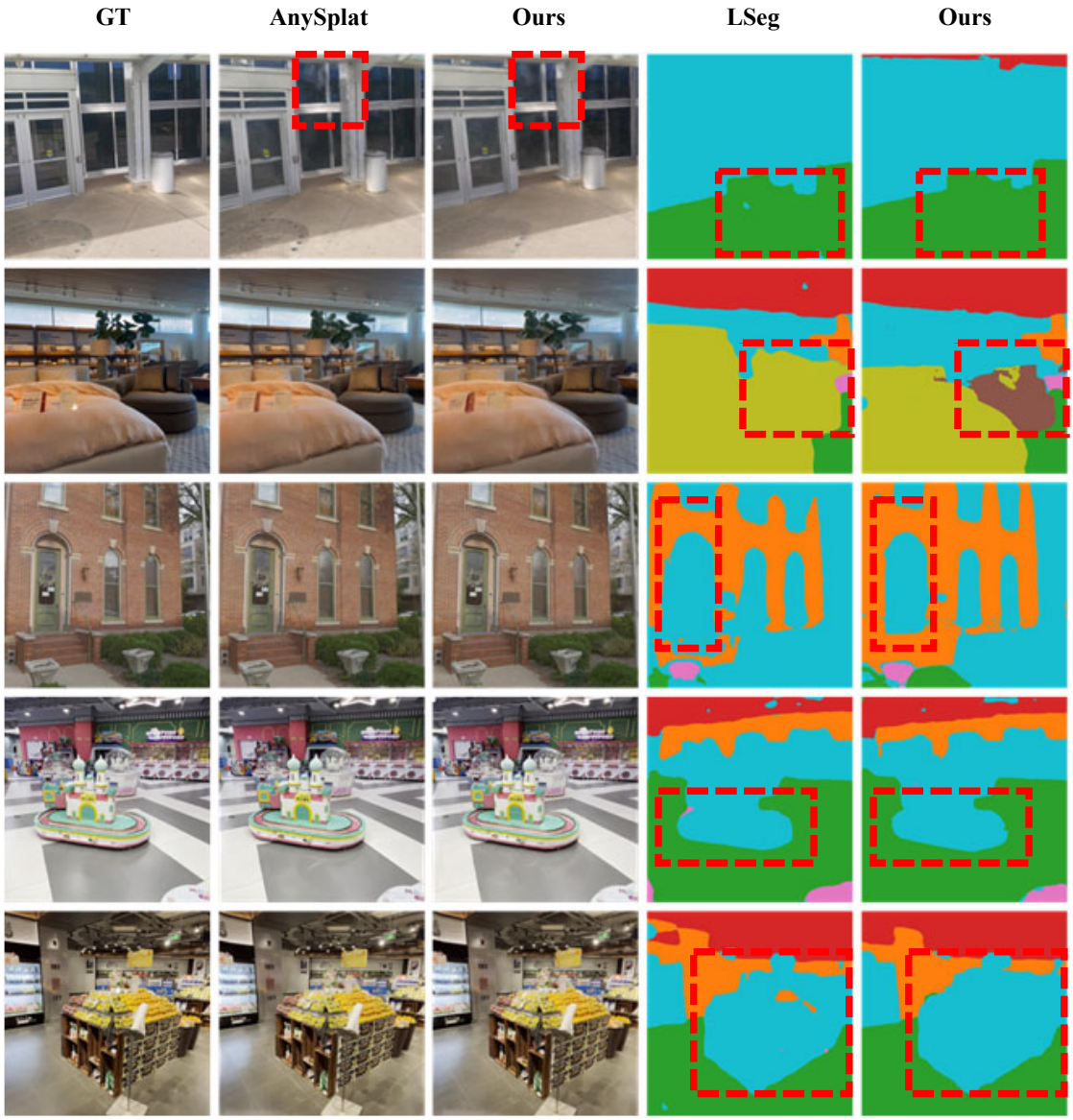

Figure 10. Qualitative results on DL3DV-10K [20] demonstrating generalization across diverse indoor and outdoor scenes.