Learning 3D Representations for Spatial Intelligence from Unposed Multi-View Images

Supplementary Material

Appendix

The appendix is structured as follows:

- §A provides additional implementation details.
- §B presents further experimental analyses. It includes:
  - §B.1 provides supplementary results on 3D vision tasks.
  - §B.2 provides additional findings on embodied AI tasks.
  - §B.3 examines geometric masking.
  - §B.4 reports extensive ablation studies.
  - §B.5 evaluates performance on downstream tasks.
- §C supplements failure cases analysis and presents more qualitative results.
- §D discusses the limitation and future work of our method.

A. More Implementation Details

A.1. 3D Vision Tasks Implementation Details

We adopt two complementary experimental settings for evaluating 3D vision tasks, each emphasizing different aspects of the unified 3D representation:

LSM-based Setting. Following [5], we train and evaluate on indoor scene datasets (ScanNet [4] and ScanNet++ [27]), where the model must jointly predicts geometry, appearance, and semantics from *unposed* multi-view images. This setting stresses *semantic consistency* across views and the ability to lift 2D features into a coherent 3D semantic field.

NoPoSplat-based Setting. Following [26], we train and evaluate on large-scale video datasets (RealEstate10K [30] and ACID [11]) for *pose-free* novel view synthesis. This setting emphasizes *geometric fidelity* and robustness to sparse, wide-baseline inputs without any pose supervision. The results are shown in B.1.

A.2. Embodied Tasks Implementation Details

To assess the effectiveness of UniSplat as a visual backbone for embodied AI, we evaluate our approach using both a synthetic reinforcement learning benchmark SPA [32] and a synthetic imitation learning benchmark RoboCasa [17].

A.2.1. Reinforcement Learning Benchmarks

The RL-based benchmark [32] spans 268 tasks across 8 simulators and covers both single-task and language-conditioned multi-task scenarios. In all experiments, the UniSplat encoder is frozen and only the downstream policy network is trained, ensuring a fair comparison of representation quality.

Single-task Benchmarks. We include three representative single-task settings:

- **VC-1** [14]: 14 tasks from four simulators — Adroit (AD) [10], Meta-World (MW) [28], DMControl (DMC) [22], and TriFinger (TF) [25]. Policies are 3-layer MLPs trained with 100 demonstrations per task (25 for MW) and evaluated over 50 rollouts using fixed seeds (100, 200, 300). The `[CLS]` token from UniSplat serves as the observation feature.
- **Franka Kitchen** [7]: a MuJoCo simulation of a Franka robot in a kitchen with 9-D joint-velocity action space. We evaluate five tasks: Sliding Door, Turn Light On, Open Door, Turn Knob, and Open Microwave. Policies are 2-layer MLPs trained on 25 demonstrations per task.
- **Meta-World [28]**: 48 diverse manipulation tasks, encompassing easy, medium, and hard levels. We adopt the Diffusion Policy [3] following [29], training with 10 demonstrations and evaluating over 20 rollouts per task.

Language-conditioned Multi-task Benchmarks. We further evaluate our method on two multi-task benchmarks featuring natural-language instructions:

- **RLBench** [9]: 71 executable tasks split into two groups according to PolarNet [2] categories (35 and 36 tasks). Each task has 100 demonstrations for training and 25 rollouts for evaluation. For each group, we train a language-conditioned multi-task agent. We use RVT-2 [6] as the policy backbone, replacing its convolutional encoder with our frozen UniSplat encoder.
- **LIBERO** [12]: 130 tasks across five suites (Spatial, Object, Goal, LIBERO-10, LIBERO-90). We train the official transformer-based language-conditioned policy with 20 demonstrations per task, no data augmentation, and pre-extracted visual features from UniSplat.

Policy Training and Evaluation. For all settings, we adhere to the training hyperparameters and evaluation protocols of the respective benchmarks to ensure comparability with prior work [32]. The frozen UniSplat encoder outputs either the `[CLS]` token (for MLP/Diffusion policies) or unpatchified feature maps (for transformer-based policies).

A.2.2. Imitation Learning Benchmarks

UniSplat can extract informative visual representations from image observations, which should benefit imitation learning. We empirically validate this by evaluating UniSplat on the RoboCasa benchmark [17]. RoboCasa comprises 24 atomic kitchen tasks with associated natural-language instructions, spanning actions such as pick-and-place, opening, and closing. For each task, we use a limited set of 50 human demonstrations and 3,000 synthetic demonstrations generated with MimicGen [16]. We train our model on these demonstra-

Table A1. **Quantitative comparison** of novel view synthesis on the RE10K dataset under various overlap settings (§B.1).

Method	Small			Medium			Large			Average		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Pose-Required												
pixelSplat	20.277	0.719	0.265	23.726	0.811	0.180	27.152	0.880	0.121	23.859	0.808	0.184
MVSplat	20.371	0.725	0.250	23.808	0.814	0.172	27.466	0.885	0.115	24.012	0.812	0.175
Supervised Pose-Free												
MASt3R	16.305	0.516	0.451	18.106	0.561	0.377	17.975	0.524	0.402	17.617	0.539	0.403
CoPoNeRF	17.393	0.585	0.462	18.813	0.616	0.392	20.464	0.652	0.318	18.938	0.619	0.388
Splatt3R	17.789	0.582	0.375	18.828	0.607	0.330	19.243	0.593	0.317	18.688	0.593	0.317
NoPoSplat	22.514	0.784	0.210	24.899	0.839	0.160	27.411	0.883	0.119	25.033	0.838	0.160
Self-Supervised Pose-Free												
SelfSplat	14.828	0.543	0.469	18.857	0.679	0.328	23.338	0.798	0.208	19.152	0.680	0.328
Ours	**22.765**	**0.789**	**0.205**	**25.246**	**0.845**	**0.156**	**27.872**	**0.891**	**0.113**	**25.397**	**0.843**	**0.157**

Table A2. **Multi-Task Imitation Learning Results in Robocasa** (§B.2). Average success rates (%) of multi-task agents trained with 50 human demonstrations or 3000 generated demonstrations per task. Results are evaluated over 50 episodes.

	PnP CabToCounter		PnP CounterToCab		PnP CounterToMicrowave		PnP CounterToSink		PnP CounterToStove		PnP MicrowaveToCounter	
Method	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000
BC-transformer	2	18	6	28	2	18	2	44	2	6	2	8
GWM	18	32	4	22	14	44	20	38	2	18	20	26
Ours	**25**	**39**	**10**	**34**	**17**	**51**	**26**	**47**	**8**	**22**	**29**	**36**
	PnP SinkToCounter		PnP StoveToCounter		Open SingleDoor		Open DoubleDoor		Close DoubleDoor		Close SingleDoor	
	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000
BC-transformer	8	42	6	28	46	50	28	48	28	46	56	94
GWM	22	38	18	44	58	62	28	42	50	58	54	90
Ours	**29**	**48**	**27**	**46**	**63**	**68**	**34**	**51**	**54**	**64**	**60**	**96**
	Open Drawer		Close Drawer		TurnOn Stove		TurnOff Stove		TurnOn SinkFaucet		TurnOff SinkFaucet	
	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000
BC-transformer	42	74	80	96	32	46	4	24	38	34	50	72
GWM	56	90	80	90	46	80	22	40	52	48	44	66
Ours	**60**	**93**	**82**	**96**	**51**	**84**	**29**	**49**	**55**	**56**	**58**	**77**
	Turn SinkSpout		CoffeePress Button		TurnOn Microwave		TurnOff Microwave		CoffeeServe Mug		CoffeeSetup Mug	
	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000	H-50	G-3000
BC-transformer	54	96	48	74	62	90	70	60	22	34	0	12
GWM	72	90	76	90	64	84	70	54	36	50	16	28
Ours	**79**	**96**	**82**	**93**	**70**	**92**	**74**	**68**	**44**	**60**	**25**	**36**

tions and use UniSplat as the state encoder. We benchmark against BC-Transformer [15] and the Gaussian World Model (GWM) [13], reporting performance in terms of success rate. The results are shown in §B.2.

B. More Experimental Analysis

B.1. More Results on 3D Vision Tasks

Novel View Synthesis. As shown in Table A1, we compare our method with several methods on the RE10K dataset across different overlap settings. Our method outperforms all previous pose-free methods and even surpasses some pose-required methods, demonstrating the effectiveness of our unified 3D representation and training strategy. Qualitative results are shown in Fig. A2- A4.

B.2. More Embodied AI Results

Robocasa Multi-Task Imitation Learning. We further evaluate our UniSplat representation on the Robocasa [32] multi-task imitation learning benchmark, which includes 24 manipulation tasks in a simulated kitchen environment. As shown in Table A2, our method consistently outperforms prior visual backbones, demonstrating the effectiveness of UniSplat for spatial reasoning and task execution in embodied AI.

B.3. Analysis of Geometric Masking

Pose Estimation with Gaussian-based PnP Pose and Pose Head. To test whether the dual masking improves geometry reasoning rather than merely regularizing features, we evaluate relative pose estimation using (i) RANSAC-PnP with the predicted 3D Gaussian centers from different fields and (ii) learnable pose heads. As shown in Table A3, we observe

stable trends: appearance field $\mathcal{G}_{\text{app}}$ yields the most reliable correspondences, anchor field $\mathcal{G}_{\text{anchor}}$ follows, coarse geometry field $\mathcal{G}_{\text{geo}}$ lags; and the final pose C_{final} consistently refines the coarse estimate C_{coarse}. Because PnP relies exclusively on physically consistent 2D–3D correspondences, gains here indicate improved cross-view geometric consistency and provide empirical support that the proposed masking mechanism enhances geometry awareness.

Table A3. **Relative Pose Estimation** with Gaussian-based PnP Poses and Pose Heads (§B.3). We evaluate AUC with various thresholds on RealEstate10K [30] and ACID [11] datasets.

	RE10K			ACID		
Method	5° ↑	10° ↑	20° ↑	5° ↑	10° ↑	20° ↑
$\mathcal{G}_{\text{geo}}$	0.510	0.664	0.796	0.305	0.472	0.612
$\mathcal{G}_{\text{anchor}}$	0.578	0.714	0.822	0.336	0.495	0.648
$\mathcal{G}_{\text{app}}$	0.592	0.736	0.834	0.344	0.504	0.653
C_{coarse}	0.524	0.693	0.805	0.311	0.487	0.625
C_{final}	**0.607**	**0.748**	**0.842**	**0.354**	**0.516**	**0.661**

B.4. More Ablation Studies

Ablation on the Number of Gaussian Latent Tokens. Table A4 shows that increasing the number of Gaussian latent tokens from 64 to 128 improves mIoU, accuracy, and perceptual quality. 256 tokens give the best overall metrics with marginal gains over 128, while 512 tokens slightly degrade performance, suggesting diminishing returns and possible overfitting. An intermediate token count balances reconstruction fidelity and segmentation accuracy.

Table A4. **Ablation on the Number of Gaussian Latent Tokens** (§B.4). We evaluate the effect of varying the number of Gaussian latent tokens on reconstruction and segmentation performance.

# T_{coarse}	mIoU↑	Acc.↑	PSNR↑	SSIM↑	rel↓
64	0.5469	0.8237	25.05	0.8602	3.35
128	0.5587	0.8284	25.33	0.8684	3.23
256	**0.5625**	**0.8334**	**25.65**	**0.8782**	**3.10**
512	0.5617	0.8312	25.53	0.8756	3.11

B.5. Downstream Tasks

We further validate the utility of UniSplat as a general 3D visual backbone on EmbodiedScan [23], a large-scale, egocentric multi-modal 3D perception benchmark with oriented 3D boxes, semantic occupancy, and language prompts. We follow the official data organization, view sampling, and metric protocols to evaluate three downstream tasks: multi-view 3D detection, multi-view semantic occupancy predictions, and multi-view 3D visual grounding. To adapt UniSplat for specific tasks, we append lightweight task-specific heads to its pretrained multi-view transformer encoder. For 3D object detection, we attach a 3D detection head predicting oriented boxes (center, size, rotation) from UniSplat's fused multi-view geometric-semantic features. For semantic occupancy prediction, a 3D decoder is added, taking the voxelized dense features to predict semantic grids. For 3D visual grounding, we equip the 3D decoder with a cross-modal fusion transformer that integrates encoded language features with the 3D scene representation, followed by a grounding head sharing the detection architecture. This setup directly measures how well self-supervised 3D representations learned from unposed images generalize to complex indoor perception tasks without any depth supervision.

Multi-view 3D Object Detection. As shown in Table A5, UniSplat with RGB-only inputs consistently surpasses camera-only baselines and even strong RGB-D systems on EmbodiedScan, indicating that its unified geometric–semantic representation yields reliable oriented box estimates without depth supervision. We provide qualitative results in Fig. A8.

Table A5. **Multi-view 3D object detection** (§B.5).

Method	Input	AP_{25}	AR_{25}	AP_{50}	AR_{50}
ImVoxelNet [20]	RGB	6.15	20.39	2.41	6.31
VoteNet [18]	Depth	3.20	6.11	0.38	1.22
FCAF3D [19]	Depth	9.07	44.23	4.11	20.22
EmbodiedScan [23]	RGB-D	16.85	51.07	9.77	28.21
Ours	RGB	**28.69**	**62.24**	**15.34**	**39.57**

Multi-view Semantic Occupancy Prediction. As shown in Table A6, UniSplat delivers markedly better voxel-level semantics than prior RGB methods and is competitive with or exceeds RGB-D variants, reflecting dense, scene-consistent 3D priors learned from unposed images.

Table A6. **Multi-view Semantic Occupancy Prediction** (§B.5).

Method	Input	mIoU
OccNet [21]	RGB	8.07
SurroundOcc [24]	RGB	9.10
EmbodiedScan [23]	RGB	10.48
EmbodiedScan [23]	RGB-D	19.97
Ours	RGB	**27.45**

Multi-view 3D Visual Grounding. As shown in Table A7, UniSplat shows clear improvements over RGB-D baselines across overall, easy, and hard settings, demonstrating robust cross-modal alignment and spatial grounding from images alone. Results show UniSplat provides consistent gains over camera-only baselines across all tasks, highlighting its effectiveness as a unified RGB-only 3D backbone.

Table A7. **Multi-view 3D Visual Grounding** (§B.5).

Method	Input	Overall	Easy	Hard
ScanRefer [1]	RGB-D	12.85	13.78	9.12
BUTD-DETR [8]	RGB-D	22.14	23.12	18.23
L3Det [31]	RGB-D	23.07	24.01	18.34
EmbodiedScan [23]	RGB-D	25.72	27.11	20.12
Ours	RGB	**36.88**	**38.13**	**31.42**

C. More Visualizations

C.1. Failure Cases

Fig. A1 illustrates typical failure modes of our method, including blur and visual artifacts in occluded regions, as well as under extreme viewpoint changes.

Figure A1. **Failure cases** (§C.1).

C.2. Novel View Synthesis Visualizations

We present additional qualitative results and visualizations to further illustrate the effectiveness of UniSplat across varying scene conditions and view overlaps, as shown in Fig. A2–A5.

C.3. Semantic Segmentation Visualizations

We provide more qualitative comparisons of novel-view semantic segmentation on ScanNet, as shown in Fig. A6.

C.4. More Feature Map Visualizations

We provide additional visualizations of the learned feature maps and depth maps from UniSplat in Fig. A7, demonstrating the model's ability to capture detailed geometric and semantic information.

C.5. More 3D Object Detection Visualizations

We include more qualitative results of 3D object detection on EmbodiedScan in Fig. A8, showcasing the accurate oriented bounding box predictions achieved by UniSplat.

D. Discussion and Limitations

Limitations. Although UniSplat achieves strong performance across diverse 3D vision and embodied AI tasks, several limitations remain. First, the framework still relies on pseudo-supervision from large pre-trained teacher models for geometry and semantics, which may propagate biases and inaccuracies from these teachers into the learned representation. Second, while our geometry-aware masking and coarse-to-fine splatting improve robustness to sparse unposed views, performance degrades in extremely limited or highly textureless scenes, indicating that geometry induction could be further strengthened. Third, our experiments are primarily conducted on indoor datasets; scaling to large-scale outdoor or highly dynamic environments may require additional adaptations, such as motion modeling or more robust pose estimation. Finally, although the hierarchical Gaussian representation improves efficiency compared to dense splatting, rendering and training remain computationally intensive relative to purely latent approaches, which may limit deployment in resource-constrained settings.

Future Work. Future research could address these limitations in several ways. One direction is to develop geometry and semantic priors that are learned in a fully self-supervised manner, reducing dependence on external teacher models. Another is to design adaptive masking and rendering strategies that adjust to scene complexity and viewpoint coverage, further improving robustness in sparse or degenerate cases. Extending UniSplat to handle dynamic, open-world environments and outdoor scenes would broaden its applicability, potentially requiring integration of temporal modeling and more generalizable camera estimation. Finally, incorporating large-scale language-scene interaction and multi-modal grounding could enable richer spatial reasoning and task understanding for embodied agents.

References

[1] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, 2020. 3

[2] Shizhe Chen, Ricardo Garcia, Cordelia Schmid, and Ivan Laptev. Polarnet: 3d point clouds for language-guided robotic manipulation. *arXiv preprint arXiv:2309.15596*, 2023. 1

[3] Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 2023. 1

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 1

[5] 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 *NeurIPS*, 2024. 1

[6] Ankit Goyal, Valts Blukis, Jie Xu, Yijie Guo, Yu-Wei Chao, and Dieter Fox. Rvt-2: Learning precise manipulation from few demonstrations. *arXiv preprint arXiv:2406.08545*, 2024. 1

[7] Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019. 1

[8] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *ECCV*, 2022. 3

[9] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020. 1

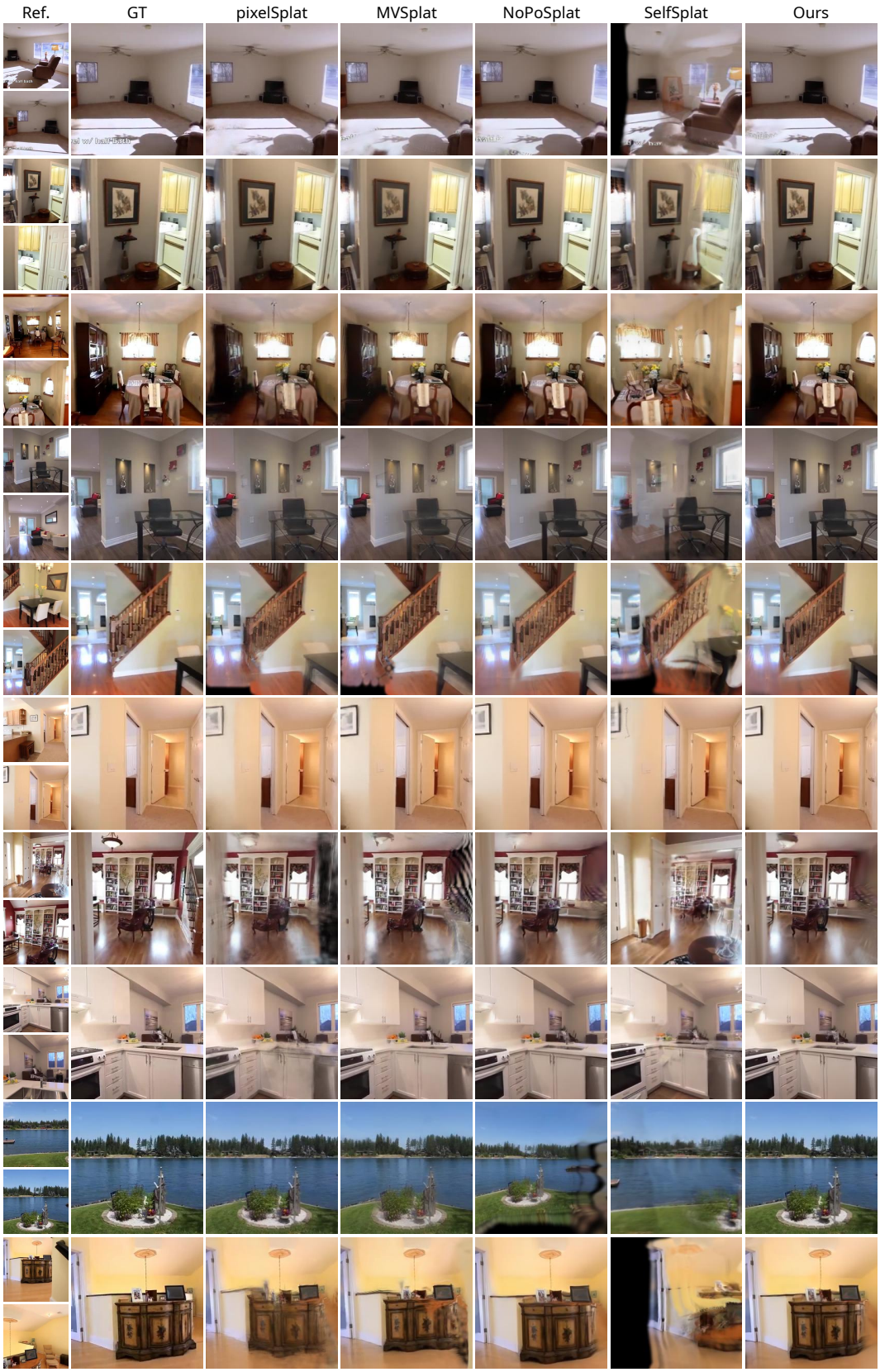

Figure A2. **More qualitative comparisons on RE10K with small image overlap** (§C.2).

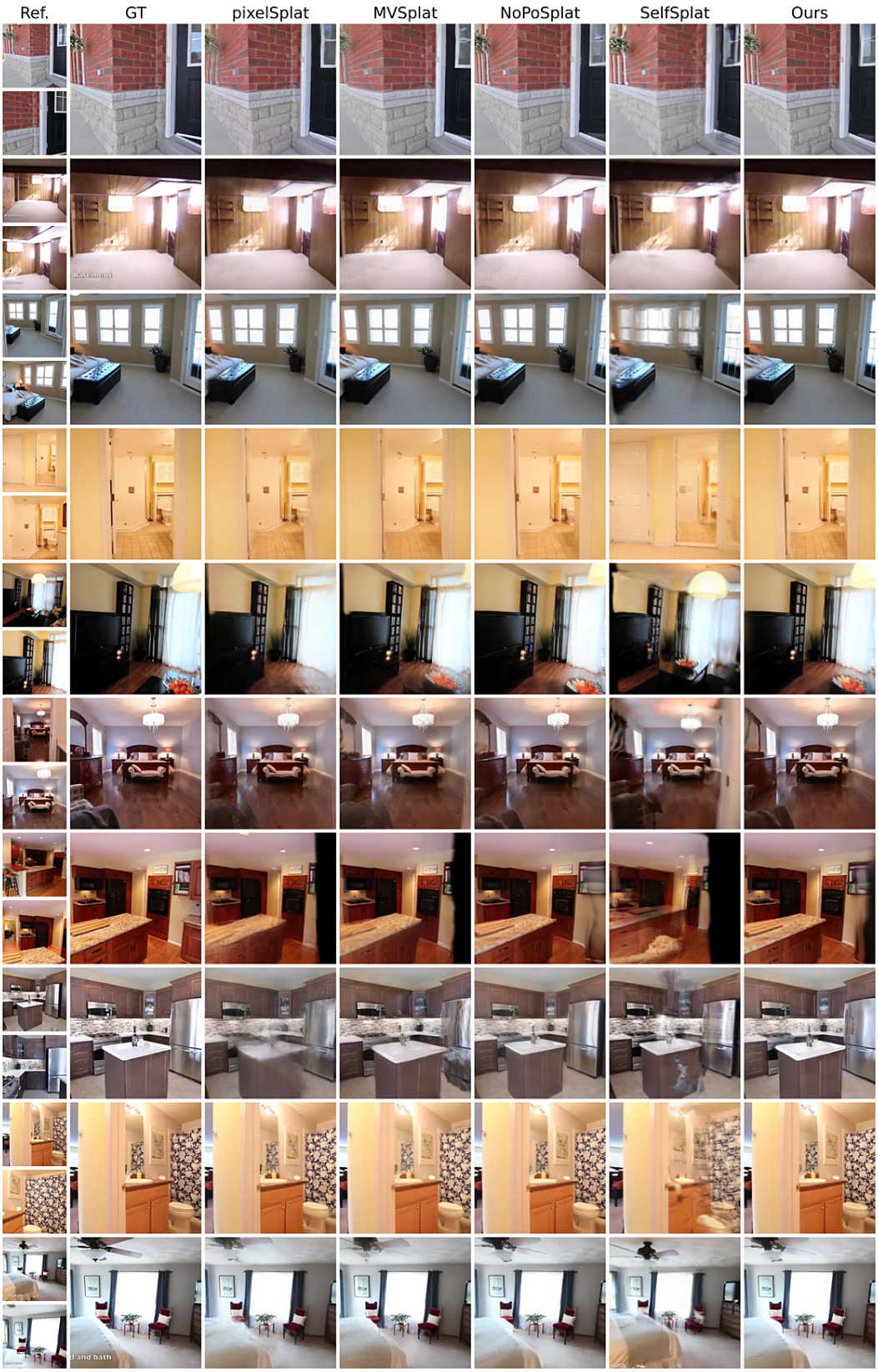

Figure A3. **More qualitative comparisons on RE10K with medium image overlap** (§C.2).

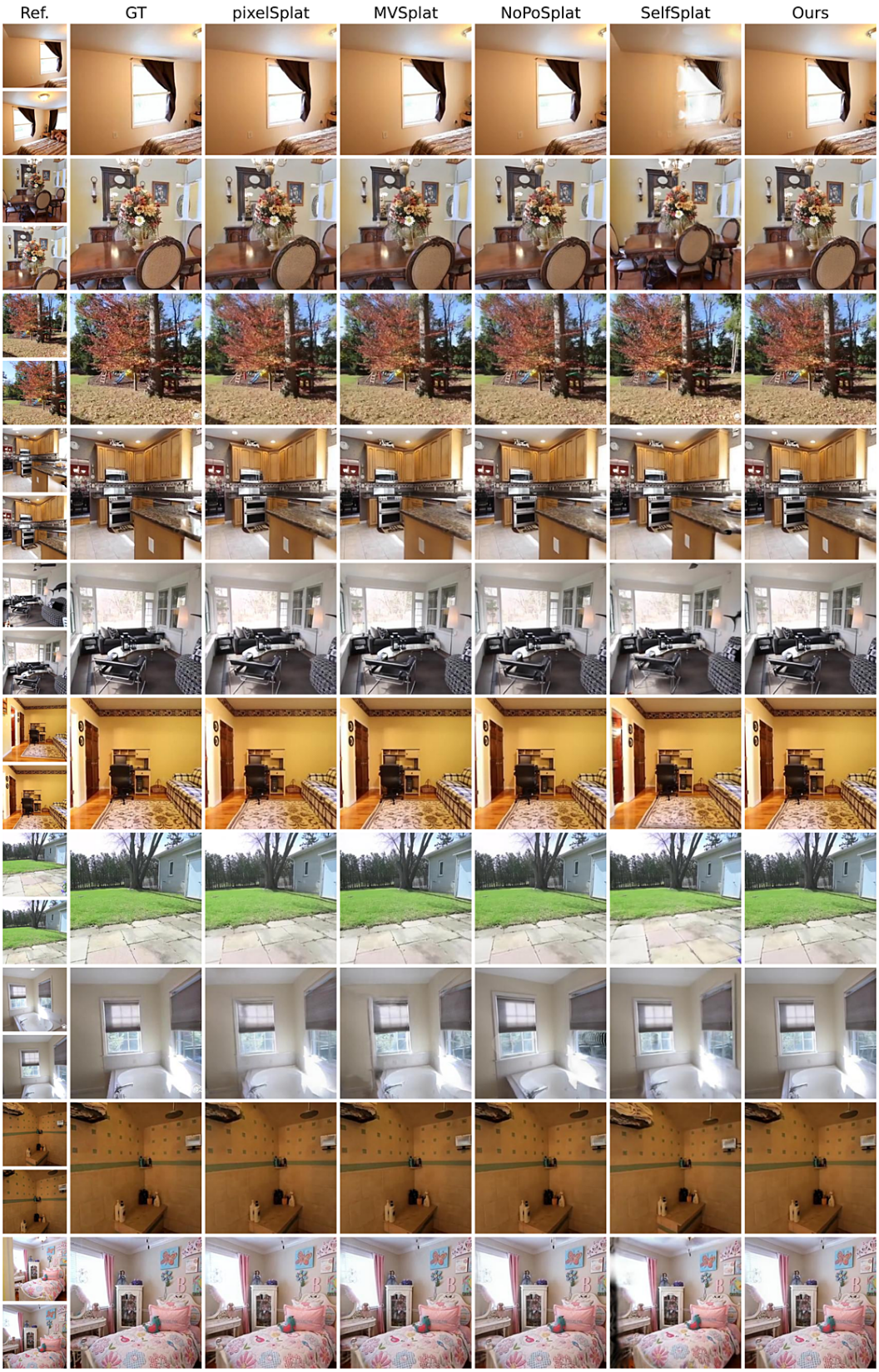

Figure A4. **More qualitative comparisons on RE10K with large image overlap** (§C.2).

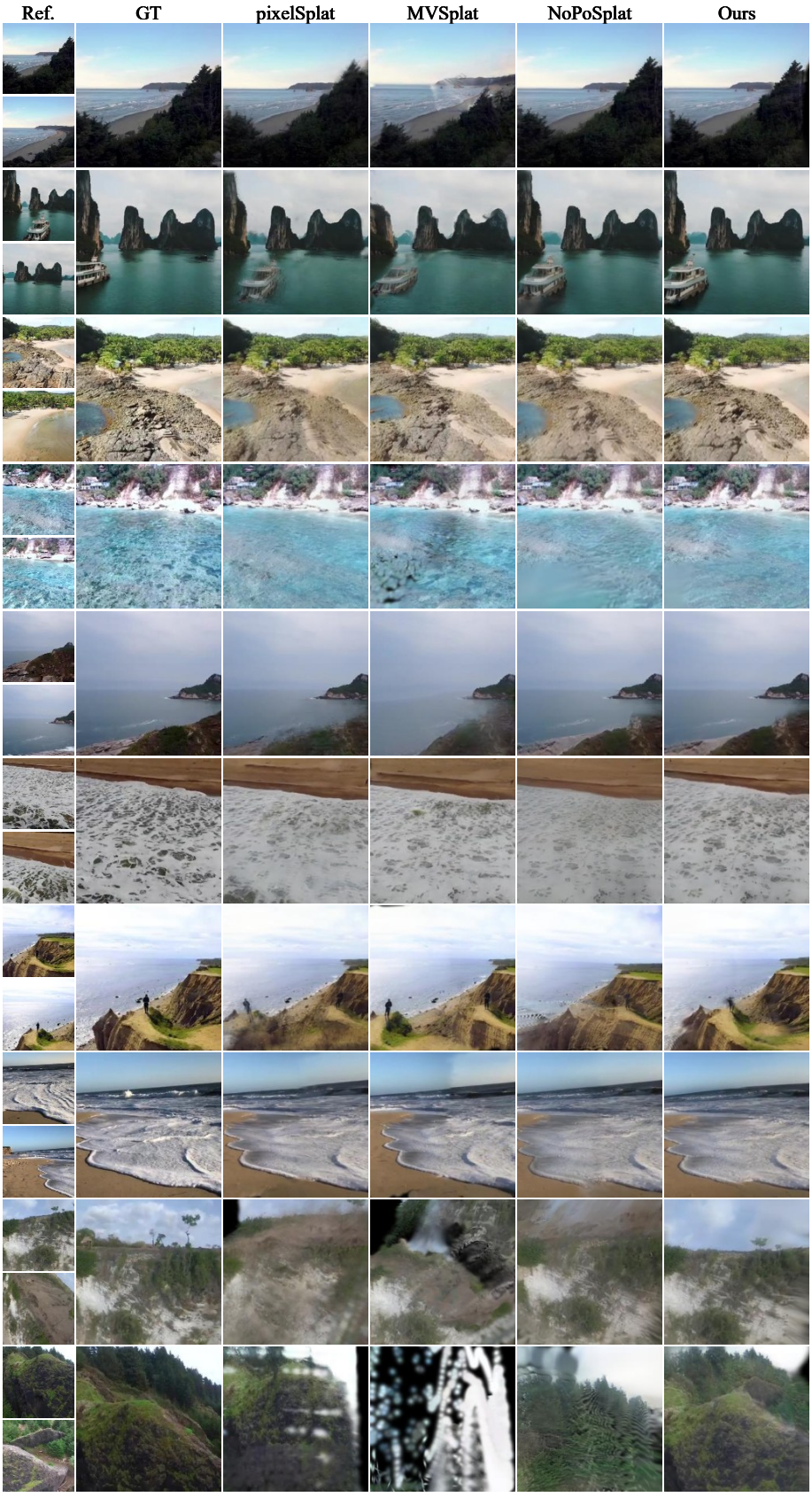

Figure A5. **More qualitative comparisons on ACID** (§C.2).

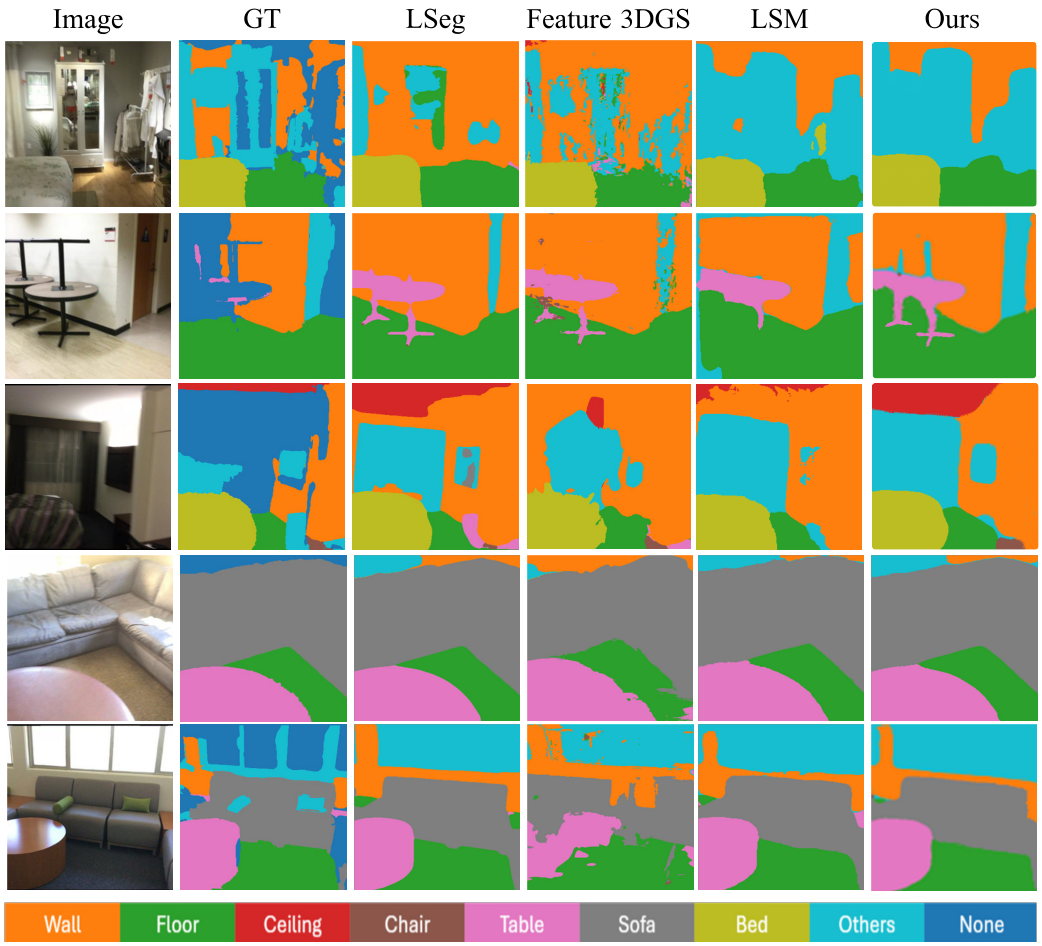

Figure A6. **More qualitative comparison of novel-view segmentation on ScanNet** (§C.3).

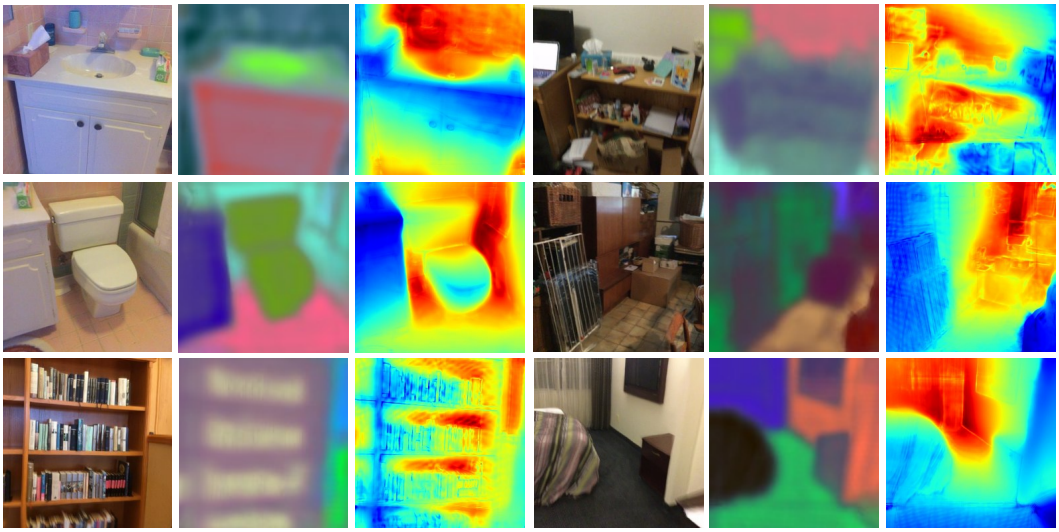

Figure A7. **More visualizations of feature maps and depth maps** (§C.4).

Figure A8. **More qualitative results of 3D object detection** (§C.5).

[10] Vikash Kumar. *Manipulators and Manipulation in high dimensional spaces*. PhD thesis, 2016. 1

[11] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *ICCV*, 2021. 1, 3

[12] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *NeurIPS*, 2023. 1

[13] Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. Gwm: Towards scalable gaussian world models for robotic manipulation. In *ICCV*, 2025. 2

[14] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an Artificial Visual Cortex for Embodied Intelligence? In *NeurIPS*, 2023. 1

[15] Ajay Mandlekar, Danfei Xu, Josiah Wong, Soroush Nasiriany, Chen Wang, Rohun Kulkarni, Li Fei-Fei, Silvio Savarese, Yuke Zhu, and Roberto Martín-Martín. What matters in learning from offline human demonstrations for robot manipulation. In *CoRL*, 2022. 2

[16] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *CoRL*, 2023. 1

[17] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024. 1

[18] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *ICCV*, 2019. 3

[19] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *ECCV*, 2022. 3

[20] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *CVPR*, 2022. 3

[21] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, 2023. 3

[22] Saran Tunyasuvunakool, Alistair Muldal, Yotam Doron, Siqi Liu, Steven Bohez, Josh Merel, Tom Erez, Timothy Lillicrap, Nicolas Heess, and Yuval Tassa. dm_control: Software and tasks for continuous control. *Software Impacts*, 2020. 1

[23] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 3

[24] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, 2023. 3

[25] Manuel Wüthrich, Felix Widmaier, Felix Grimminger, Joel Akpo, Shruti Joshi, Vaibhav Agrawal, Bilal Hammoud, Majid Khadiv, Miroslav Bogdanovic, Vincent Berenz, et al. Trifinger: An open-source robot for learning dexterity. *arXiv preprint arXiv:2008.03596*, 2020. 1

[26] Botao Ye, Sifei Liu, Songyou Peng, Haofei Xu, Xueting Li, Ming-Hsuan Yang, and Marc Pollefeys. No Pose, No Problem: Surprisingly Simple 3d Gaussian Splats from Sparse Unposed Images. In *ICLR*, 2025. 1

[27] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023. 1

[28] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020. 1

[29] Yanjie Ze, Gu Zhang, Kangning Zhang, Chenyuan Hu, Muhan Wang, and Huazhe Xu. 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. *arXiv preprint arXiv:2403.03954*, 2024. 1

[30] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 1, 3

[31] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection. *arXiv preprint arXiv:2309.09456*, 2023. 3

[32] Haoyi Zhu, Honghui Yang, Yating Wang, Jiange Yang, Limin Wang, and Tong He. Spa: 3d Spatial-Awareness Enables Effective Embodied Representation. In *ICLR*, 2025. 1, 2