Scal3R: Scalable Test-Time Training for Large-Scale 3D Reconstruction

Supplementary Material

A. Model Details

We provide the detailed model architecture in this section.

Overall architecture. As stated in Section 4.1, we build our model up as a large transformer, as VGGT [22]. We then attach a Global Context Memory (GCM) module after 4 specific global attention layers, namely 4th, 11th, 17th, and 24th, whose outputs are used as input features of the two DPT [12] decoders to predict the depth maps and point clouds. The total number parameters of the newly added GCM module is 75.55M, namely 0.076B.

Global Context Memory module. The GCM module consists of three components: a query-key-value projection layer, three compact MLP networks W_1, W_2, W_3 serving as the Adaptive Memory Units (AMUs), and an output projection layer. The forward pass of the GCM module is performed as follows:

$$f_W(x) = W_2\big(\mathrm{SiLU}(W_1 x) \circ (W_3 x)\big), \tag{12}$$

where $\circ$ denotes the element-wise product, after we update the AMUs W_1, W_2, W_3 in the inner loop using K, V as detailed in the Section 4.2, we can use the updated AMUs to compute the GCM output $f_W(Q)$. The query-key-value projection layer is the standard linear projection layer, which projects the upstream feature $x \in \mathbb{R}^{M\times d}$ into multi-head query $Q \in \mathbb{R}^{M\times nh\times hd}$, key $K \in \mathbb{R}^{M\times nh\times hd}$, and value $V \in \mathbb{R}^{M\times nh\times hd}$, where nh is the number of heads and hd is the dimension of each head, where $nh \times hd = d$. Define the hidden dimension of the AMUs as $hd \times k$, with k being a scaling factor, then $W_1, W_3 \in \mathbb{R}^{hd\times hd\times k}$ and $W_2 \in \mathbb{R}^{hd\times k\times hd}$. The total state size of the GCM module is calculated as:

$$\text{state size} = nh \times hd \times hd \times k = \frac{d^2}{nh} \times k. \tag{13}$$

Specifically, we set the number of heads nh to 1 to maximize the state size for larger memory capacity, and set the scaling factor k to 4 to balance the memory capacity and computational efficiency.

B. Evaluation Details

B.1. Dataset Details

Our benchmarks are built on four datasets: Virtual KITTI [1], KITTI Odometry [4], Oxford Spires [19], and ETH3D [15]. These datasets feature long, large-scale sequences with diverse weather and lighting conditions, urban driving scenarios, and indoor and outdoor scenes, respectively. We present more details about the datasets in the following.

Virtual KITTI [1] is a synthetic dataset comprising 50 outdoor street-scene sequences spanning diverse weather and lighting conditions (e.g., fog, morning, overcast, overcast, rain and sunset). Sequence lengths range from 223–837 frames, with path lengths spanning 52–711 meters.

KITTI Odometry [4] is a real-world benchmark of 11 sequences collected from urban driving scenarios with varied lengths and street layouts. Sequence lengths range from 271–4,661 frames, covering 0.39–5.07 km of travel, and pose challenging long-sequence tracking conditions.

Oxford Spires [19] is a real-world dataset with 6 sequences, `2024-03-12-keble-college-02`, `2024-03-12-keble-college-03`, `2024-03-12-keble-college-04`, `2024-03-12-keble-college-05`, `2024-03-13-observatory-quarter-01`, `2024-03-13-observatory-quarter-02`, featuring challenging loop closures and extreme view sparsity across indoor and outdoor scenes. Sequence lengths range from 351–787 frames, covering 280–773 meters. To ensure reliable supervision and fair evaluation, we filtered out views with large LiDAR–RGB timestamp discrepancies and removed scenes that consequently contained fewer than 50 frames despite spanning several hundred meters.

ETH3D [15] provides high-resolution indoor and outdoor images with ground-truth depth from laser sensors. We select 11 scenes: `courtyard`, `electro`, `kicker`, `pipes`, `relief`, `delivery area`, `facade`, `office`, `playground`, `relief 2`, `terrains`, for the benchmark. The number of frames in each scene ranges from 14 to 76.

B.2. Evaluation Details

We provide the detailed evaluation in this section.

Pose metrics. For pose accuracy evaluation, we follow the protocol introduced in [3, 6] , and report results using the Absolute Trajectory Error (ATE), Relative Rotation Error (RRE in °/100m), and Relative Translation Error (RTE in m/100m), providing a comprehensive assessment of both translation and rotation accuracy. All metrics are calculated after Sim(3) alignment of predicted pose trajectories with the ground truth.

Reconstruction metrics. We evaluate 3D reconstruction with Chamfer Distance (CD) and F1-score. Let $\mathcal{G}$ be the ground-truth point cloud and $\mathcal{P}$ the predicted point cloud after Sim(3) alignment with the ground-truth using the Umeyama algorithm [21]. Denote by $\mathrm{dist}(A \to B)$ the average nearest-neighbour distance from each point in $\mathcal{A}$ to $\mathcal{B}$. We define accuracy as $\mathrm{dist}(\mathcal{P} \to \mathcal{G})$ and completeness as

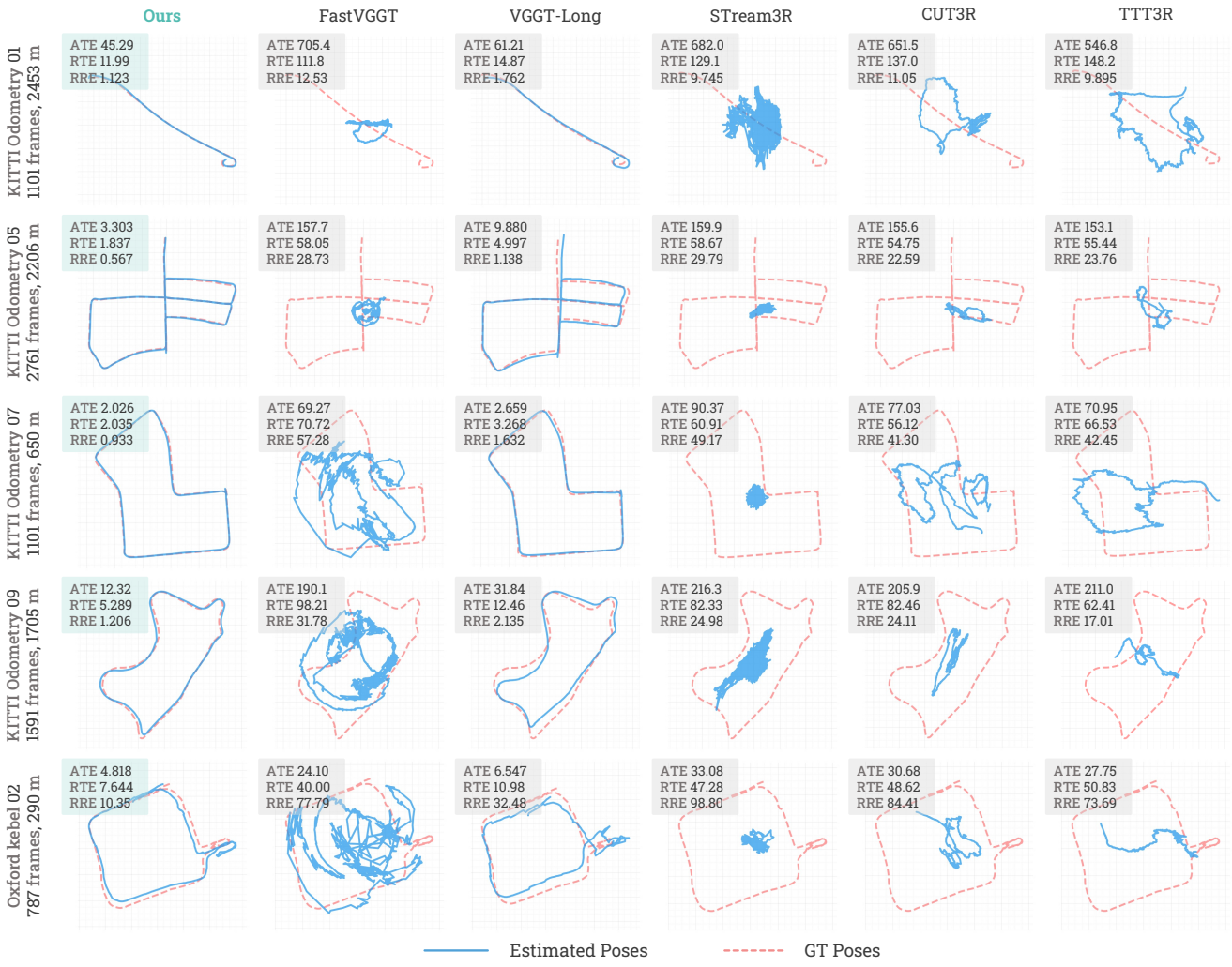

Figure 5. **Camera trajectory comparison.** Scal3R preserves global structure with substantially lower drift, whereas baselines frequently lose tracking or diverge, demonstrating our capability of reconstructing large-scale scenarios with high accuracy.

$\text{dist}(\mathcal{G} \to \mathcal{P})$, then the Chamfer Distance (CD) is defined as the average of accuracy and completeness. Given a distance threshold d, we define the precision and recall as:

$$\text{precision} = \frac{1}{|\mathcal{P}|} \sum_i [\text{dist}(\mathcal{P}_i \to \mathcal{G}) < d], \tag{14}$$

$$\text{recall} = \frac{1}{|\mathcal{G}|} \sum_i [\text{dist}(\mathcal{G}_i \to \mathcal{P}) < d], \tag{15}$$

where $[\cdot]$ denotes the Iverson bracket [5]. Then, the F1-score is computed as:

$$\text{F1} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{16}$$

Evaluation details. All evaluations use the full set of frames per sequence. We set chunk size to 60 and overlap to 30 for both Scal3R and VGGT-Long [3] across all datasets, and we follow official evaluation protocols for the remaining baselines [2, 6, 9, 10, 16, 23, 26]. Pose metrics are directly evaluated on Virtual KITTI [1], KITTI Odometry [4], and Oxford Spires [19] datasets with no extra hyperparameters. Reconstruction metrics are evaluated on ETH3D [15], Virtual KITTI [1], and Oxford Spires [19] datasets, using dataset-specific distance thresholds (ETH3D: 0.25, Virtual KITTI: 1.0, Oxford Spires: 4.0) to reflect differences in scale and sparsity. For baselines that fail to produce valid camera trajectories or reconstructions on a scene, we assign the worst valid score among the compared methods on that scene when computing dataset averages in Sections 5.1 and 5.2.

Ablation details. To simplify training while preserving the generality of our conclusions, we train all ablation models on a subset of the datasets listed in

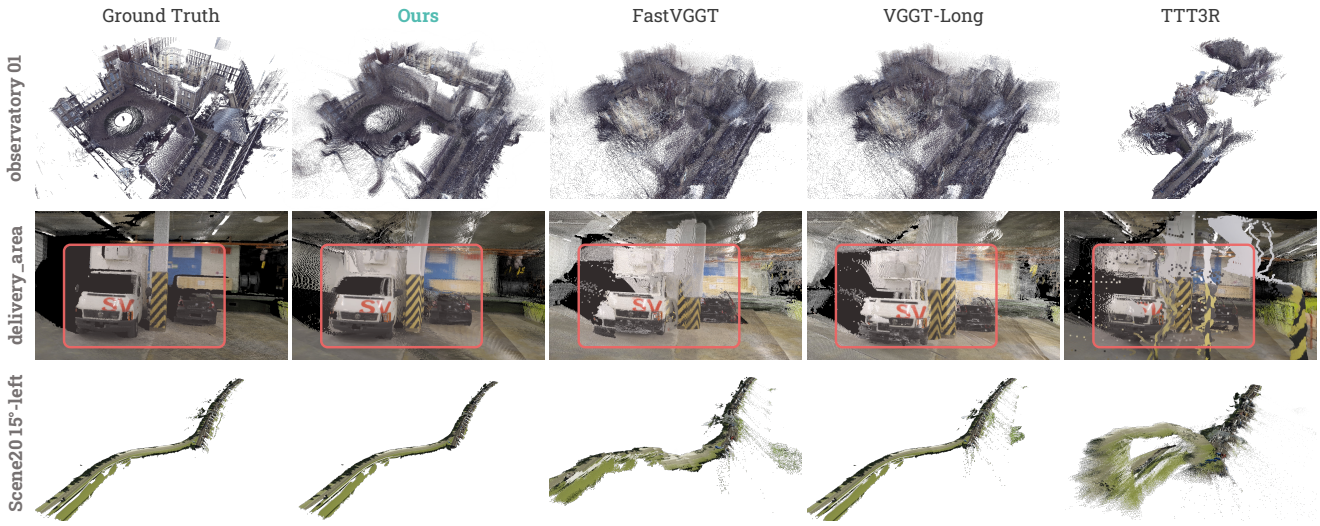

Figure 6. **Point-cloud reconstruction comparison.** Scal3R produces more accurate large-scale reconstructions for large-scale outdoor environments where baselines often fail, and achieves higher local geometric accuracy and consistency in indoor scenes.

Section 4.3, excluding the object-centric datasets WildRGB [24], Co3Dv2 [13], and Aria Digital Twin [11]. State size ablations are trained on 16 NVIDIA A800 GPUs for 60k iterations for a fair comparison. We randomly select 7 sequences from Virtual KITTI [1] (`Scene01 15-deg-left`, `Scene02 30-deg-left`, `Scene06 clone`, `Scene18 morning`, `Scene20 rain`), Oxford Spires [19] (`2024-03-12-keble-college-04`, `2024-03-13-observatory-quarter-01`) for evaluation, covering diverse weather and lighting conditions, urban driving scenarios, and indoor and outdoor scenes. Global context ablations are trained on 8 NVIDIA A800 GPUs for 60k iterations for a fair comparison. We select KITTI Odometry [4] sequences `01`, `03`, `04`, `10`, and Virtual KITTI [1] sequences `Scene20` for evaluation, featuring challenging long-sequence tracking conditions.

C. Additional Results

We provide additional long-sequence camera trajectory results in Figure 5. As demonstrated in Figure 5, our method is capable of reconstructing extremely large-scale long sequences with small drift, whereas baselines frequently lose tracking or diverge significantly, showcasing the effectiveness of our proposed global context representation and aggregation mechanism. We provide additional long-sequence reconstruction results in Figure 6, the results illustrate the improvement of our method over baselines on both large-scale accurate reconstruction and local geometric consistency.

C.1. Additional Benchmark Comparisons

We further evaluate pose accuracy on three additional benchmarks in Table 4. For ScanNet++ [25], we use five sequences: `419cbe7c11`, `98b4ec142f`, `bb87c292ad`, `c24f94007b`, and `ebc200e928`. For TUM-RGBD [17], we evaluate all scenes. For Waymo [18], we follow VGGT-Long [3] and use the same nine test scenes. Compared with the main paper benchmarks, these datasets are denser video regimes with stronger short-range overlap, so they are useful for checking whether our gains persist when recent streaming and video-based baselines are relatively better matched to the evaluation setting. We follow the same evaluation protocol as in the main paper and report ATE after Sim(3) alignment with the ground truth. We set the chunk size and overlap to 120 and 60, respectively, for both Scal3R and VGGT-Long [3] on ScanNet++ [25] and TUM-RGBD [17], and to 60 and 30, respectively, on Waymo [18]. As shown in Table 4, Scal3R achieves the best ATE on ScanNet++ (0.08) and TUM-RGBD (0.07), with clear margins over strong video-based baselines such as STream3R and TTT3R. This shows that the proposed global context mechanism is not only helpful for the large-scale sparse settings emphasized in the main paper, but also remains effective on denser long-video benchmarks. On Waymo, Scal3R remains competitive on long driving sequences, indicating good transfer across different video regimes.

C.2. Runtime Scaling with Sequence Length

We further analyze runtime scaling with sequence length in Table 5. As the sequence length increases from 150 to 990 frames, the total runtime grows approximately linearly, while throughput remains stable at around 2.6–2.9 FPS. Meanwhile, the relative pose error remains within 0.07–0.08 m, indicating that Scal3R maintains stable pose accuracy as the sequence length increases.

Table 4. **Additional pose benchmark comparisons.** We report ATE (m, lower is better) on three supplementary benchmarks. The best results are in **bold**, and the second best are underlined.

Method	ScanNet++ avg. 924 fr.	TUM-RGBD avg. 926 fr.	Waymo avg. 198 fr.
MASt3R-SLAM [10]	0.47	0.08	7.63
VGGT-SLAM [9]	0.29	0.12	7.43
StreamVGGT [26]	1.70	0.63	45.10
STream3R [6]	1.75	0.63	42.20
CUT3R [23]	1.27	0.54	9.40
TTT3R [2]	0.55	0.31	3.49
FastVGGT [16]	1.56	0.42	**1.28**
VGGT-Long [3]	0.13	0.08	1.78
COLMAP [14]	GT	0.19	25.63
MASt3R-SfM [7]	1.50	0.39	3.95
DROID-SLAM [20]	0.97	0.11	6.67
DPVO++ [8]	0.91	0.10	1.35
Ours	**0.08**	**0.07**	1.52

Table 5. **Runtime scaling with sequence length.** Using the same single-GPU evaluation setting as the main-paper resource comparison, we report relative pose error (RPE, m), total inference time, and FPS as the input length increases. Runtime grows smoothly with sequence length while RPE remains stable.

Frames	150	270	510	990
RPE (m) ↓	0.08	0.08	0.07	0.08
Time (s) ↓	51.19	98.81	195.24	382.80
FPS ↑	2.93	2.73	2.61	2.59

C.3. Failure Cases

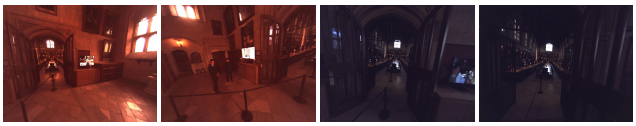

Figure 7. **Failure case under abrupt illumination changes.** Large appearance shifts within a sequence weaken cross-chunk correspondences and can lead to inaccurate global alignment.

We further summarize representative failure modes of Scal3R. The first arises from severe appearance inconsistency within a sequence (e.g., abrupt illumination or color shifts), as illustrated in Figure 7. In such cases, the appearance gap across chunks weakens the reliability of cross-chunk correspondences. The second occurs under extreme view sparsity, for example when only tens of images cover scenes spanning hundreds of meters or even kilometers. In such extreme cases, even local predictions can fail due to the lack of sufficient geometric constraints.

References

[1] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 2, 6, 7, 8, 1, 3

[2] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 3, 6, 7, 2, 4

[3] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it–pushing vggt's limits on kilometer-scale long rgb sequences. *arXiv preprint arXiv:2507.16443*, 2025. 2, 4, 6, 7, 1, 3

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 1, 2, 6, 7, 3

[5] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4): 1–13, 2017. 2

[6] Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Shuai Yang, Bo Dai, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. *arXiv preprint arXiv:2508.10893*, 2025. 3, 6, 7, 1, 2, 4

[7] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 2, 6, 7, 4

[8] Lahav Lipson, Zachary Teed, and Jia Deng. Deep patch visual slam. In *European Conference on Computer Vision*, pages 424–440. Springer, 2024. 3, 6, 7, 4

[9] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 6, 7, 2, 4

[10] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025. 3, 6, 7, 2, 4

[11] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 6, 3

[12] 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. 1

[13] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 6, 3

[14] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE confer-*

ence on computer vision and pattern recognition, pages 4104–4113, 2016. 2, 6, 7, 4

[15] Thomas Schöps, Torsten Sattler, and Marc Pollefeys. BAD SLAM: Bundle adjusted direct RGB-D SLAM. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7, 8, 1, 2

[16] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 2, 6, 7, 4

[17] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 7, 3

[18] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 7, 3

[19] Yifu Tao, Miguel Ángel Muñoz-Bañón, Lintong Zhang, Jiahao Wang, Lanke Frank Tarimo Fu, and Maurice Fallon. The oxford spires dataset: Benchmarking large-scale lidar-visual localisation, reconstruction and radiance field methods. *International Journal of Robotics Research*, 2025. 1, 2, 6, 7, 8, 3

[20] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 2, 3, 6, 7, 4

[21] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002. 8, 1

[22] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3, 4, 5, 6, 1

[23] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 2, 3, 6, 7, 4

[24] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024. 6, 3

[25] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 6, 7, 3

[26] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 3, 6, 7, 2, 4