KV-Tracker: Real-Time Pose Tracking with Transformers

Supplementary Material

1. Applying KV-Tracker on Depth Anything V3

To demonstrate the applicability of our proposed approach in making off-the-shelf reconstruction methods efficient for online use, we show our tracking via caching method on Depth Anything 3 [3]. Despite being a very recent model, it still uses self-attention blocks for exchanging information between the input frames, so our method can be applied on it. We show evaluation results of running it on the 7-Scenes and TUM RGB-D datasets.

Table 1. Absolute Trajectory Error (ATE) RMSE in meters on 7-Scenes dataset for Depth Anything V3, 1.15B model, running at 18 FPS.

	DepthAnything
chess	0.098
fire	0.044
heads	0.055
office	0.106
pumpkin	0.143
redkitchen	0.083
stairs	0.299
Average	0.118

Table 2. Absolute Trajectory Error (ATE) RMSE in meters on TUM RGB-D dataset for Depth Anything V3, 1.15B model, running at 18 FPS.

	DepthAnything
360	0.19
desk	0.173
desk2	0.261
plant	0.096
room	0.525
rpy	0.059
teddy	0.102
xyz	0.024
Average	0.179

2. Additional results

We provide additional results in Table 3 on our baselines in the scene-level camera tracking evaluation for CUT3R [4] and TTT3R [2] with different input image resolutions for completeness. At the 224×224 resolution, their frame rate increases to 22 FPS, which is still lower than our 27 FPS.

3. Scene representation stability

We handle π^3's coordinate invariance through consistent frame anchoring. For each KV cache set, a transformation is computed to place the first keyframe at identity, while also being applied to all keyframes and tracked frames. Consecutive submaps are aligned via $\mathrm{Sim}(3)$ Umeyama alignment on overlapping keyframes, ensuring global consistency.

4. Comparison with StreamingVGGT

We provide a comparison against StreamVGGT [5] in Table 4, with their default 518 resolution. Out of the box, it runs out of memory since every frame's KV cache is saved, so we run it with keyframes similar to ours. Despite being training free, ours is more accurate, showing that full bidirectional attention between keyframes yields better scene representation for tracking.

5. KV-cache Ablation

5.1. Pose quality

We provide additional experiments with pose evaluation with and without KV caching in Table 5. As expected, full π^3 slightly outperforms the version using cached KVs.

5.2. Geometry quality

We provide a geometry evaluation by ablating the effects of caching vs full bidirectional attention in Table 6. Since we are approximating the attention pattern, we observe a slight drop in accuracy compared to running π^3 with full bidirectional attention. We use a single scale per scene to align the prediction to the GT, so these results demonstrate temporal consistency.

6. Complex camera motion

We provide evaluation results on Sintel [1] in Table 7. We demonstrate higher accuracy compared to TTT3R. We follow the test split used by π^3 on Sintel.

References

[1] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conf. on Computer Vision (ECCV)*, pages 611–625. Springer-Verlag, 2012. 1

[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. 1

Resolution	CUT3R 224	TTT3R 224	CUT3R 512	TTT3R 512	Ours 224	Ours 308
chess	0.439	0.395	0.297	0.154	0.054	**0.039**
fire	0.454	0.256	0.218	0.124	0.045	**0.039**
heads	0.225	0.121	0.115	0.097	0.040	**0.025**
office	0.383	0.321	0.356	0.196	0.075	**0.068**
pumpkin	0.391	0.185	0.249	0.228	0.145	**0.142**
redkitchen	0.322	0.132	0.118	0.136	0.050	**0.035**
stairs	0.263	0.128	0.079	0.063	**0.052**	0.063
Average	0.354	0.220	0.205	0.143	0.066	**0.059**

Table 3. Absolute Trajectory Error (ATE) RMSE in meters on 7-Scenes dataset with different input image resolutions.

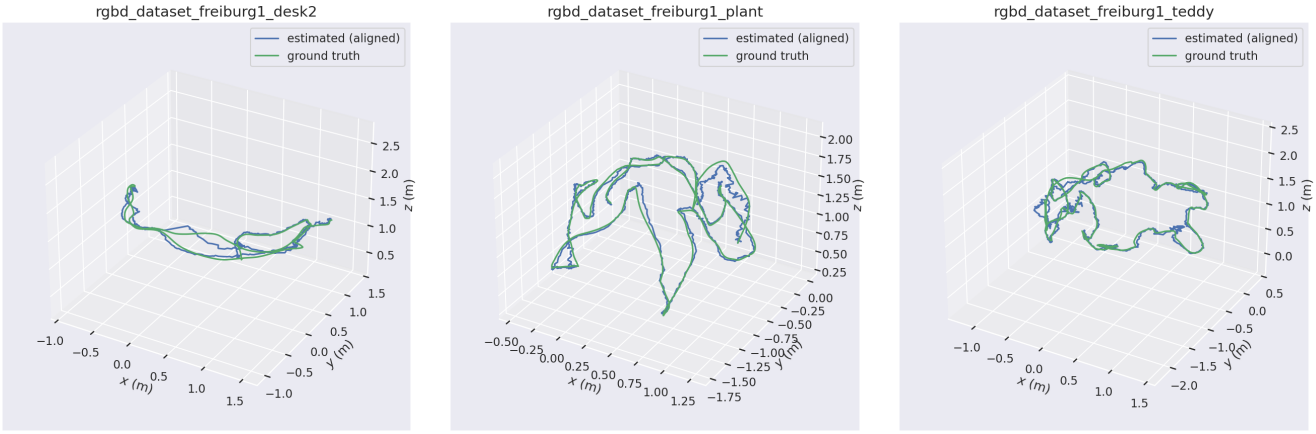

Figure 1. Qualitative trajectory visualizations on TUM RGB-D. Our estimated trajectories (blue) closely follow the ground truth paths (green), demonstrating accurate tracking across diverse indoor sequences.

StreamingVGGT	StreamingVGGT w KF	Ours
OOM	0.090	**0.059**

Table 4. Absolute Trajectory Error (ATE) RMSE in meters on 7-Scenes.

	π^3 w KV	π^3
chess	0.039	**0.035**
fire	0.039	**0.038**
heads	0.025	**0.021**
office	**0.068**	0.073
pumpkin	0.142	**0.140**
redkitchen	0.035	**0.034**
stairs	0.063	**0.062**
Average	0.059	**0.058**

Table 5. Pose estimation quality with and without KV caching. Absolute Trajectory Error (ATE) RMSE in meters on 7-Scenes.

π^3 w KV		π^3	
Acc. ↓	Comp. ↓	Acc. ↓	Comp. ↓
0.037	0.027	**0.034**	0.027

Table 6. Geometry ablation π^3 w/w.o.t KV caching on 7-Scenes.

TTT3R	Ours
0.208	**0.118**

Table 7. Absolute Trajectory Error (ATE) RMSE in meters on Sintel.

[3] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 1

[4] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 1

[5] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 1