TRAINING-FREE LOOPED VISION GEOMETRY TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Can the geometric predictions of pretrained feed-forward models be improved without retraining or fine-tuning? We study training-free looped inference across VGGT, Pi3, Depth Anything 3, MapAnything, and VGGT-Ω. An architecture-aware wrapper repeats a contiguous block window with damped updates, preserving pretrained weights, native attention, and prediction heads. We compare original (*raw*) and loop-augmented (*loop*) inference using one fixed configuration per model across static and dynamic benchmarks. Looping raises VGGT's HiRoom reconstruction F1 from 53.79 to 60.43; MapAnything improves every measured reconstruction cell across five benchmarks. Benefits depend on the task and window. On HiRoom, higher raw patch-feature similarity within a window is associated with larger F1 gains for VGGT and Pi3 after accounting for window length. Consistently updating intermediate features also enables useful loops across layers supplying dense prediction heads. These results demonstrate training-free improvements to multi-view geometry and connect their benefits to the structure of pretrained representations.

1 INTRODUCTION

Multi-view geometry has increasingly become a feed-forward prediction problem. Given a set of images, pretrained transformers can jointly recover camera poses, depth, and 3D structure without optimizing a scene-specific representation (Wang et al., 2025a;b; Lin et al., 2025; Keetha et al., 2025; Wang et al., 2026). Training these reconstruction foundation models, however, requires large-scale datasets and substantial computational resources (Wang et al., 2025a; Lin et al., 2025). This investment motivates a central question: *can we improve the geometric predictions of existing feed-forward models without retraining or fine-tuning them?*

Repeating a contiguous group of pretrained transformer layers—a *loop window*—offers one route. Training-free looping improves frozen language models (Chen et al., 2026; Lys et al., 2026); here we ask *where and how existing geometric transformations can be reused without disrupting their predictions*.

Geometry models complicate this transfer: their backbones mix within- and across-view attention, sometimes using register or scene-level scale tokens. Dense heads also consume features saved at intermediate *feature-cache layers*, so the state passed to the next block is not the only output to consider. Camera poses, affine-aligned depth, and reconstructed point clouds can furthermore respond differently to the same loop.

We study training-free looping in five visual geometry models, comparing original inference (*raw*) with loop-augmented inference (*loop*) using fixed pretrained weights. Figure 1 shows selected reconstruction comparisons. Each model uses *one loop configuration across the reported suite*: synthetic and real data, indoor and outdoor scenes, object and scene scales, and static and dynamic content. Looping improves diverse backbones, but gains depend on the task and window; the selected Pi3 configuration, for example, loses reconstruction F1 on ETH3D.

On HiRoom, higher within-window patch similarity is associated with larger reconstruction F1 gains for VGGT and Pi3 after accounting for window length. Windows containing feature-cache layers can also outperform those lying between them, linking loop benefits to representation structure rather than cache locations alone.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

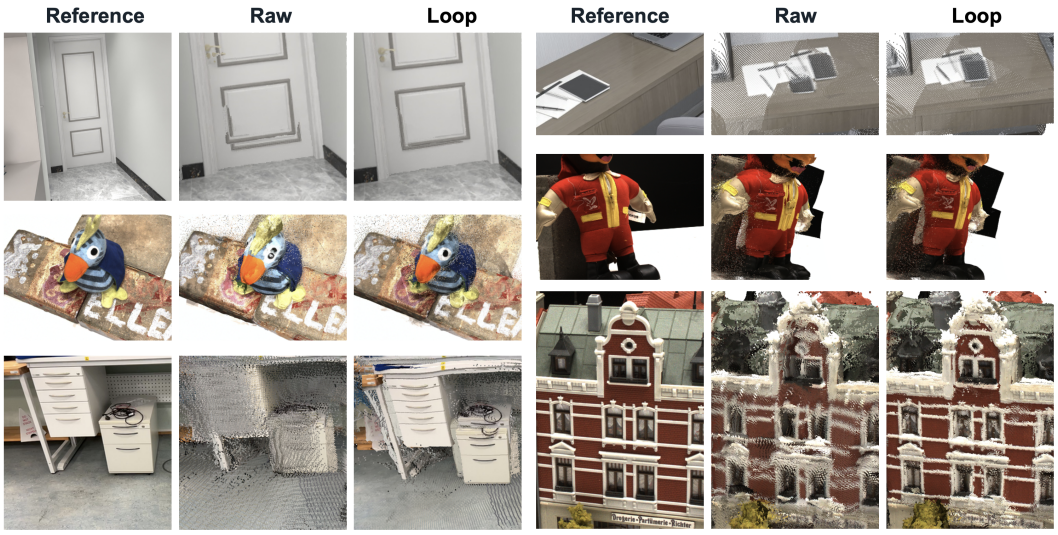


Figure 1: Qualitative reconstruction examples. From top to bottom in the left group, then top to bottom in the right group, the models are VGGT, Pi3, MapAnything, VGGT, DA3, and VGGT-Ω. Triplets show input RGB, raw, and loop, with identical observation cameras and rendering settings for each raw/loop pair. Selected structures become more coherent, although some artifacts remain. References are input photographs, not pixel-aligned novel-view ground truth.

Our contributions are threefold: (1) an inference-only looping interface respecting native attention, special tokens, and dense-head feature capture; (2) a five-model comparison with fixed per-model configurations and complete per-benchmark reporting; and (3) an empirical characterization of loop windows through representation similarity, task specialization, and feature-cache experiments.

Related work. Feed-forward geometric models provide our pretrained backbones (Wang et al., 2025a;b; Lin et al., 2025; Keetha et al., 2025; Wang et al., 2026). Training-free looping and inner-loop inference reuse frozen language-model layers (Chen et al., 2026; Lys et al., 2026). Déjà View learns shared recurrent computation for multi-view reconstruction (Burzio et al., 2026), whereas our recurrence is added after training to existing checkpoints. Easi3R also intervenes in a frozen geometry model, but targets motion through a different attention-based procedure (Chen et al., 2025). Layer-wise geometric probes (Stary et al., 2025) examine intermediate geometric representations, while block-recurrent analyses (Jacobs et al., 2025) characterize representation changes across vision-transformer layers.

2 Architecture-aware frozen computation reuse

2.1 Windowed damped updates

Let $g_{s:e} = B_e \circ \cdots \circ B_s$ be a contiguous, shape-preserving window in a pretrained backbone, with zero-based inclusive endpoints. Each B_i is a within-view/cross-view pair or a physical transformer block, depending on the architecture. Layers before and after the window form the prefix and suffix; prediction heads decode features into geometric outputs. The prefix supplies h_0; raw inference produces the endpoint $a = g_{s:e}(h_0)$. Following Chen et al. (2026), we replace this one-step endpoint with

$$h_{j+1} = h_j + \frac{g_{s:e}(h_j) - h_j}{K}, \qquad j = 0, \ldots, K-1. \tag{1}$$

Here $g_{s:e}(h_j)$ is the window proposal and h_{j+1} the updated, or accepted, state. The unchanged suffix and heads decode h_K. The normal forward supplies the first evaluation, so the wrapper adds $K-1$ replays; $K=1$ recovers the native endpoint. Window length and K are independent; we explore $K \in \{2, 3, 4\}$.

Equation 1 can be viewed as damped forward Euler on the residual field $g(h) - h$ with step $1/K$. We also consider the raw-endpoint anchor of Chen et al. (2026), blending the loop state with the

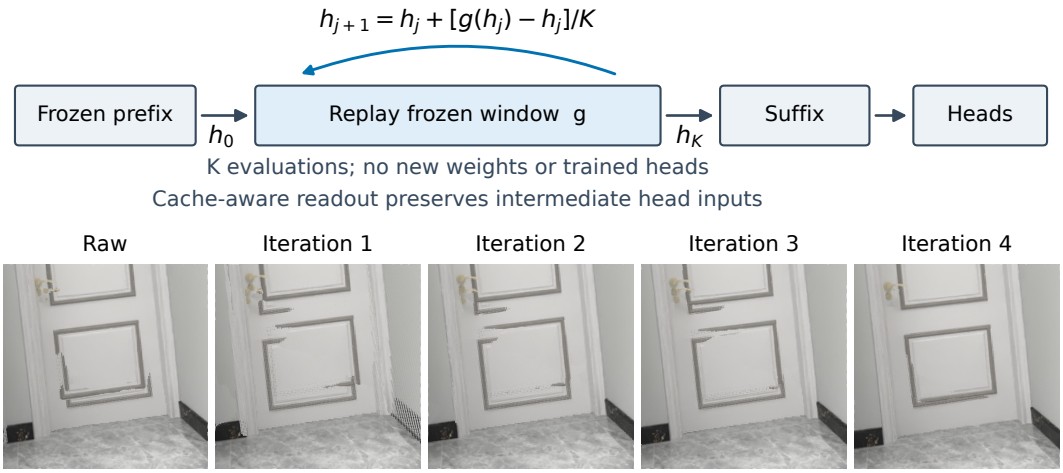


Figure 2: Inference-only computation reuse. Top: repeated window evaluation with damped state updates; prefix, suffix, and heads remain unchanged. Bottom: VGGT reconstructions from accepted states $h_1, \ldots, h_4$ of one $[5, 8]$, $K = 4$ HiRoom run, each decoded through the complete suffix and heads. Raw is a separate baseline, not h_0. After whole-scene similarity alignment, all panels share a viewing camera and rendering settings.

original one-step output:

$$h_{\text{out}} = \beta a + (1 - \beta) h_K. \tag{2}$$

The blend weight is $\beta \in [0, 1]$, and the suffix then receives h_{out}. Our shared configurations use $\beta = 0$, except for the VGGT-Ω setting in Table 1.

2.2 Respecting the native computation graph

VGGT and Pi3 replay complete within-view/global pairs in their principal searches. VGGT-Ω uses frame/inter-frame pairs, including communication through special tokens. For DA3, we search physical blocks in its initial single-view stage separately from complete pairs in its later multi-view stage. MapAnything's global/frame fusion pairs update the scene-level scale token together with the per-view state.

VGGT-Ω also varies which tokens accept updates and whether repetition applies to a whole window, each pair, or one attention sublayer (Appendix H). Token-restricted updates preserve unselected tokens at their raw endpoint values without removing them from native attention.

2.3 Intermediate features for dense prediction

Dense prediction heads such as DPT (Ranftl et al., 2021) read multiple intermediate states. We call the backbone layers whose outputs are saved for these heads *feature-cache layers*. When such a layer lies inside a replayed window, its saved feature can come from a replay forward pass without receiving the corresponding damped update. The base VGGT and VGGT-Ω searches exclude windows containing feature-cache layers. MapAnything permits a feature-cache layer at the window endpoint by committing the complete updated state before native feature capture.

For VGGT, we separately maintain both the frame and global cached features during looping. Let (F_i, G_i) be the actual frame and global outputs of pair i. VGGT's DPT head reads their concatenation $[F_c, G_c]$ at each feature-cache index c, so both components require an updated state after looping. For a window starting at $s > 0$, the backbone starts from $h_0 = G_{s-1}$. For every feature-cache index c inside the window, auxiliary states begin at $(f_{c,0}, q_{c,0}) = (F_{s-1}, G_{s-1})$ and evolve as

$$\begin{aligned} f_{c,j+1} &= f_{c,j} + \big(F_c(h_j) - f_{c,j}\big)/K, \\ q_{c,j+1} &= q_{c,j} + \big(G_c(h_j) - q_{c,j}\big)/K. \end{aligned} \tag{3}$$

Here $F_c(h_j)$ and $G_c(h_j)$ denote the intermediate frame and global features computed while evaluating the window from h_j. The head reads $[f_{c,K}, q_{c,K}]$; these auxiliary states never feed back into attention. When the feature-cache layer is the window endpoint $c = e$, $q_{c,K} = h_K$.

3 MULTI-MODEL GEOMETRIC EVALUATION

3.1 SETUP AND REPORTING

We evaluate official pretrained VGGT-1B, original Pi3, DA3-Giant-1.1, MapAnything, and VGGT-Ω-1B/512 checkpoints. Inference is RGB-only; GT geometry, poses, depths, and masks are used only for evaluation. Raw and loop share checkpoints, native preprocessing and mixed precision, input order, sampling, and evaluators. Checkpoint versions are detailed in Appendix A.

The suite combines benchmarks used in recent geometry evaluations: HiRoom (SVLightVerse; Lin et al., 2025), ETH3D (Schöps et al., 2017), DTU (Aanæs et al., 2016), 7Scenes (Shotton et al., 2013), NRGBD (Azinović et al., 2022), ScanNet++ (Yeshwanth et al., 2023), DyCheck (Gao et al., 2022), Sintel (Butler et al., 2012), and TUM-Dynamic (Sturm et al., 2012). We report camera AUC@3/AUC@30, reconstruction F1 and distance, and depth $\delta_{1.25}$/AbsRel wherever the corresponding task is defined. Because native resolutions and reconstruction representations differ, we compare each model with its own raw inference. Protocols follow DA3 and Pi3/VGGT-Ω evaluations (Lin et al., 2025; Wang et al., 2025b; 2026), with model-specific details in Appendix B.

Configuration selection. We explore window endpoints and K, selecting one configuration per model using cross-benchmark aggregate metrics (Appendix D). Its window, K, token scope, and anchor remain fixed across all benchmarks, with no per-dataset routing. We call one complete model inference with repetition confined to its window *single-pass inference*. Separately labeled *task-specialist configurations* target camera pose, reconstruction, or depth with the same frozen weights. Unsupported tasks have dashes or omitted rows.

3.2 ONE CONFIGURATION, MULTIPLE BENCHMARKS AND TASKS

Table 1 lists the shared configurations and their aggregate changes from raw. VGGT, Pi3, and DA3 improve all six aggregate metrics, and VGGT-Ω improves its four applicable metrics. MapAnything improves five of the six aggregate metrics, with a small decrease in mean depth accuracy. Per-benchmark results, including regressions, appear in Tables 2, 3, 4, 5, and 6.

Table 1: Shared configurations and cross-benchmark changes. AUC, F1, and depth accuracy changes are percentage points; distance/AbsRel reductions are mean per-benchmark relative percentages. Positive is better throughout. Zero-based inclusive indices count native attention pairs for VGGT, Pi3, and VGGT-Ω, and physical blocks otherwise. All tokens are updated without anchoring, except VGGT-Ω (patch-only, $\beta = 0.125$); DA3 uses a single-view block, and MapAnything includes its scale token.

		Camera		Reconstruction		Depth	
Model	Window/K	ΔAUC@3	ΔAUC@30	ΔF1	Dist. red. (%)	$\Delta\delta_{1.25}$	AbsRel red. (%)
VGGT	$[5, 8]/4$	+0.762	+0.127	+2.283	+2.56	+0.226	+1.85
Pi3	$[4, 8]/3$	+0.206	+0.182	+0.334	+2.01	+0.088	+1.35
DA3	$[6, 6]/3$	+0.319	+0.113	+0.263	+0.54	+0.059	+0.06
MapAnything	$[2, 3]/2$	+1.322	+0.590	+2.217	+3.49	−0.060	+1.18
VGGT-Ω	$[8, 9]/4$	+0.336	+0.152	–	–	+0.094	+0.49

Substantial geometric improvements are possible. On HiRoom, VGGT improves AUC@3 from 48.19 to 52.20, AUC@30 from 88.02 to 89.54, F1 from 53.79 to 60.43, and Overall from 10.33 to 8.88 cm. MapAnything improves all four F1 and all five distance cells in the reconstruction suite: for example, ETH3D F1 increases from 57.76 to 61.13, 7Scenes F1 from 52.67 to 55.88, and DTU distance falls from 7.4345 to 6.8887 mm.

Table 2: Static-scene camera pose. The same model-specific shared loop configuration is used in all main tables. VGGT-Ω uses 10-frame sampling on ETH3D and 7Scenes, unlike the other models' corresponding main-protocol cells. DTU* denotes the released mixed-view artifact audit, not a nominal DTU-64 reproduction. In Tables 2–6, small parentheses show $\Delta = \text{loop} - \text{raw}$ before score rounding: green indicates improvement, red regression, and black a delta rounding to zero.

Model	Method	HiRoom		ETH3D		DTU*		7Scenes		ScanNet++	
		AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑
VGGT	raw	48.19	88.02	29.42	81.68	79.66	85.26	23.11	84.73	59.55	94.37
VGGT	loop	52.20 (+4.02)	89.54 (+1.52)	29.19 (-0.23)	82.87 (+1.18)	79.85 (+0.19)	85.33 (+0.06)	23.09 (-0.03)	84.77 (+0.03)	60.29 (+0.73)	94.54 (+0.17)
Pi3	raw	66.80	94.78	31.95	85.11	62.36	94.79	25.52	85.93	54.11	93.49
Pi3	loop	66.94 (+0.14)	94.79 (+0.00)	32.22 (+0.28)	85.80 (+0.69)	62.42 (+0.06)	94.90 (+0.10)	26.26 (+0.74)	86.06 (+0.14)	54.15 (+0.03)	93.80 (+0.32)
DA3	raw	80.15	95.90	48.46	91.41	93.99	99.38	27.50	86.69	84.66	98.27
DA3	loop	80.20 (+0.05)	95.90 (-0.00)	48.54 (+0.08)	91.68 (+0.27)	94.21 (+0.21)	99.40 (+0.02)	27.51 (+0.01)	86.69 (+0.01)	84.77 (+0.11)	98.27 (+0.00)
MapAnything	raw	30.51	87.88	24.05	82.65	10.37	81.96	16.89	82.58	26.56	89.66
MapAnything	loop	33.50 (+2.99)	89.01 (+1.12)	29.87 (+5.82)	85.08 (+2.43)	10.33 (-0.05)	81.85 (-0.11)	17.97 (+1.08)	82.71 (+0.13)	28.86 (+2.30)	90.13 (+0.47)
VGGT-Ω	raw	–	–	44.18	91.44	–	–	54.81	94.25	–	–
VGGT-Ω	loop	–	–	44.04 (-0.13)	91.51 (+0.07)	–	–	55.87 (+1.06)	94.39 (+0.14)	–	–

Table 3: Camera pose on NRGBD (static) and three dynamic benchmarks. All use the frozen 10-frame sampling protocol.

Model	Method	NRGBD		DyCheck		Sintel		TUM-Dynamic	
		AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑
VGGT	raw	70.04	95.90	21.16	77.43	18.36	66.06	24.81	75.38
VGGT	loop	70.04 (-0.00)	95.97 (+0.07)	22.33 (+1.16)	76.73 (-0.70)	18.99 (+0.63)	64.03 (-2.03)	25.19 (+0.37)	76.21 (+0.83)
Pi3	raw	76.38	96.96	27.72	81.01	16.93	59.34	20.65	67.43
Pi3	loop	75.88 (-0.49)	97.05 (+0.09)	28.89 (+1.16)	81.58 (+0.58)	16.40 (-0.53)	59.20 (-0.14)	21.11 (+0.46)	67.29 (-0.14)
DA3	raw	93.50	99.19	50.63	89.83	21.75	62.27	27.04	68.56
DA3	loop	93.99 (+0.49)	99.27 (+0.08)	51.16 (+0.53)	90.01 (+0.17)	22.86 (+1.11)	62.70 (+0.43)	27.31 (+0.28)	68.58 (+0.03)
MapAnything	raw	35.14	87.13	4.71	61.85	6.67	38.01	6.39	47.17
MapAnything	loop	33.99 (-1.15)	87.43 (+0.30)	4.76 (+0.05)	62.04 (+0.19)	7.14 (+0.48)	38.00 (-0.01)	6.76 (+0.37)	47.95 (+0.79)
VGGT-Ω	raw	87.49	98.12	52.12	91.07	51.59	87.58	40.56	85.93
VGGT-Ω	loop	88.15 (+0.66)	98.21 (+0.09)	52.91 (+0.79)	91.28 (+0.20)	51.69 (+0.11)	88.06 (+0.48)	40.09 (-0.46)	85.86 (-0.06)

The effect depends on the backbone and metric. For Pi3, the shared setting improves HiRoom F1 from 80.41 to 81.70 and DTU distance from 2.9933 to 2.7703 mm, but reduces ETH3D F1 from 73.41 to 71.36. DA3's shared local-block setting improves HiRoom F1 from 85.20 to 86.46, yet many depth changes are small. VGGT-Ω's aggregate gains are also modest; its token-restricted setting reduces some adverse interactions but does not improve every camera or depth cell.

3.3 Task specialization and output consistency

Different tasks favor different interventions: VGGT's camera specialist improves pose but worsens depth, whereas its depth specialist improves both depth metrics (Appendix Table 9). DA3's camera specialist improves all 18 pose cells (Appendix Tables 10 and 11). Appendix E gives configurations and complete results, including effects outside the target task.

Routing complete task outputs from separate specialists improves VGGT's aggregate camera/depth results while retaining its reconstruction specialist's scores. Two mixtures of cameras and depths from different passes instead worsen mean relative reconstruction distance by 0.23% and 0.46% from raw, versus a 2.56% reduction with the unified reconstruction setting (Appendix F). Better camera AUC and affine-aligned depth therefore need not yield compatible cross-pass geometry.

Table 4: Unposed reconstruction. F1 is a percentage; distance is Overall in cm for HiRoom, metres for ETH3D/7Scenes/ScanNet++, and Chamfer distance in mm for DTU (22 scans). ETH3D uses complete GT-depth support. A dash is not a zero.

Model	Method	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
		F1↑	Dist.↓	F1↑	Dist.↓	F1↑	Dist.↓	F1↑	Dist.↓	F1↑	Dist.↓
VGGT	raw	53.79	10.33	62.86	0.5748	–	1.9405	48.16	0.1401	67.40	0.0791
VGGT	loop	60.43 (+6.64)	8.88 (-1.44)	64.52 (+1.66)	0.5759 (+0.0010)	–	1.9414 (+0.0009)	48.12 (-0.04)	0.1425 (+0.0025)	68.27 (+0.87)	0.0785 (-0.0007)
Pi3	raw	80.41	3.69	73.41	0.4483	–	2.9933	51.80	0.0869	65.75	0.0701
Pi3	loop	81.70 (+1.29)	3.68 (-0.01)	71.36 (-2.05)	0.4538 (+0.0055)	–	2.7703 (-0.2230)	53.19 (+1.39)	0.0842 (-0.0026)	66.46 (+0.70)	0.0698 (-0.0003)
DA3	raw	85.20	4.63	79.60	0.4456	–	1.7832	50.55	0.1229	78.33	0.0675
DA3	loop	86.46 (+1.26)	4.60 (-0.02)	78.51 (-1.10)	0.4436 (-0.0020)	–	1.7646 (-0.0186)	51.49 (+0.94)	0.1220 (-0.0009)	78.27 (-0.06)	0.0675 (+0.0000)
MapAnything	raw	45.01	8.48	57.76	0.5929	–	7.4345	52.67	0.0982	36.43	0.1110
MapAnything	loop	45.99 (+0.98)	8.11 (-0.37)	61.13 (+3.37)	0.5915 (-0.0014)	–	6.8887 (-0.5458)	55.88 (+3.21)	0.0947 (-0.0035)	37.74 (+1.32)	0.1088 (-0.0022)

Table 5: Static video-depth evaluation. Each sequence has a single L1 scale-and-shift fit; accuracy is shown in percent.

Model	Method	ETH3D		7Scenes		NRGBD	
		$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓
VGGT	raw	98.35	0.0385	92.71	0.0656	98.96	0.0155
VGGT	loop	98.40 (+0.05)	0.0370 (-0.0015)	92.76 (+0.05)	0.0656 (+0.0000)	98.97 (+0.01)	0.0156 (+0.0001)
Pi3	raw	99.68	0.0272	93.59	0.0629	98.97	0.0132
Pi3	loop	99.71 (+0.03)	0.0262 (-0.0010)	93.62 (+0.03)	0.0626 (-0.0002)	98.88 (-0.08)	0.0137 (+0.0005)
DA3	raw	99.50	0.0217	93.48	0.0610	99.15	0.0090
DA3	loop	99.49 (-0.01)	0.0218 (+0.0001)	93.48 (+0.01)	0.0610 (-0.0000)	99.15 (-0.00)	0.0089 (-0.0000)
MapAnything	raw	98.32	0.0353	93.78	0.0694	98.75	0.0289
MapAnything	loop	98.48 (+0.15)	0.0348 (-0.0005)	93.80 (+0.01)	0.0690 (-0.0005)	98.77 (+0.02)	0.0291 (+0.0002)
VGGT-Ω	raw	99.91	0.0138	94.65	0.0539	99.67	0.0098
VGGT-Ω	loop	99.91 (+0.00)	0.0138 (-0.0000)	94.63 (-0.02)	0.0538 (-0.0000)	99.67 (+0.00)	0.0098 (-0.0000)

3.4 Qualitative reconstruction examples

Figure 1 compares dense RGB point clouds from raw and loop inference on two HiRoom scenes, three DTU scans, and one ScanNet++ scene. With the observation camera fixed, loop reduces repeated contours around the notebook, door moulding, bird, and zipper, and improves cabinet and facade coherence. DA3 uses its reconstruction specialist; MapAnything and VGGT-Ω use additional configurations, not their shared main-table settings. Input counts appear in Appendix H.

Computation. A loop adds $K - 1$ evaluations of its selected window, not $K - 1$ complete model passes. There are no new learned parameters. We distinguish this inference overhead from evaluation-pipeline throughput in Appendix H.

4 Where does repeated computation help?

The same loop update can improve one window and degrade another. We investigate this difference on HiRoom by asking three questions: whether similar layers make better loop windows, how

Table 6: Dynamic video-depth evaluation, using the same frozen frames and evaluator for raw and loop.

Model	Method	DyCheck		Sintel		TUM-Dynamic	
		$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓	$\delta_{1.25}$ ↑	AbsRel↓
VGGT	raw	93.41	0.0640	62.86	0.2808	92.84	0.0803
VGGT	loop	94.65 (+1.24)	0.0601 (-0.0039)	62.84 (-0.02)	0.2715 (-0.0093)	92.87 (+0.03)	0.0813 (+0.0011)
Pi3	raw	97.34	0.0475	68.88	0.5116	96.70	0.0474
Pi3	loop	97.50 (+0.16)	0.0463 (-0.0012)	69.26 (+0.39)	0.4936 (-0.0181)	96.70 (+0.00)	0.0467 (-0.0008)
DA3	raw	95.71	0.0536	67.01	0.2591	94.49	0.0830
DA3	loop	95.73 (+0.02)	0.0536 (-0.0000)	67.34 (+0.33)	0.2585 (-0.0006)	94.50 (+0.01)	0.0828 (-0.0002)
MapAnything	raw	93.91	0.0773	61.36	0.4286	94.40	0.0930
MapAnything	loop	93.33 (-0.58)	0.0776 (+0.0003)	61.36 (+0.00)	0.4086 (-0.0200)	94.43 (+0.04)	0.0917 (-0.0013)
VGGT-Ω	raw	98.60	0.0279	92.36	0.0831	97.39	0.0350
VGGT-Ω	loop	98.60 (+0.01)	0.0278 (-0.0001)	92.91 (+0.55)	0.0815 (-0.0016)	97.41 (+0.02)	0.0350 (-0.0001)

loop-induced changes propagate through later layers, and whether layers supplying features to DPT constrain window placement.

4.1 Do similar layers make better loop windows?

A window may be easier to reuse if its layers already produce similar representations: repeating such layers could be less disruptive than repeating layers whose features differ sharply. We examine this possibility by relating raw feature similarity to reconstruction gains on HiRoom, and then asking whether windows in earlier and later layers behave differently.

We measure layer similarity as the cosine between feature vectors for corresponding image patches in raw inference, averaging over patches and views within each scene and then weighting scenes equally. For each window, we average over all distinct layer pairs and compare this value with the F1 change from raw, averaged over the tested iteration counts. Higher similarity means that the layers produce more closely aligned feature directions.

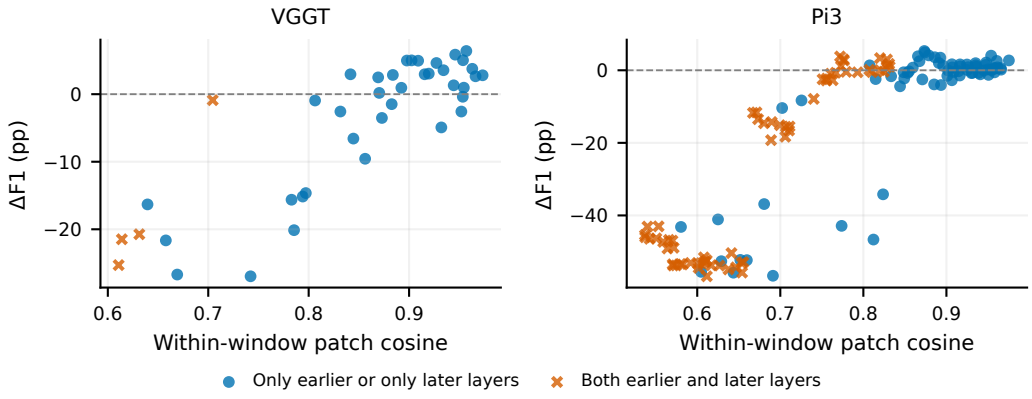


Figure 3: Raw feature similarity and loop gains on HiRoom. Each point is a window, with F1 change averaged over the tested K. Blue circles use only earlier or only later layers; orange crosses span both. Positive changes improve over raw. Reported rank associations control for window length and replay strategy.

We first inspect how these similarities vary through the backbone. In Figure 5(a), VGGT's two relatively warm squares along the diagonal indicate that earlier layers resemble one another, as do later layers, while similarity between the two parts is lower. This divides VGGT into earlier layers $[0, 12]$ and later layers $[13, 23]$; Pi3 exhibits a similar division. These are divisions identified from feature similarity, not DPT feature-cache locations.

Figure 3 tests whether this feature structure corresponds to loop gains. Each point represents one window: blue circles use only earlier or only later layers, and orange crosses contain both. For VGGT and Pi3, higher similarity is associated with better F1 changes; after accounting for window length and replay strategy, the rank associations are 0.717 and 0.675. Windows containing both earlier and later layers have lower gains on average than matched windows confined to one side, although some Pi3 windows still improve, as the orange crosses above zero show. This positive similarity–gain relationship is weaker for DA3 and absent for MapAnything in the corresponding controlled comparison (Appendix G).

To examine the role of window position more directly, Figure 5(b) moves a fixed-length VGGT window through the backbone while holding the iteration count fixed. The upper curve shows within-window similarity and the lower curve shows F1 change. A region of early windows combines high similarity with positive gains; both curves fall around the division between earlier and later layers. Their Spearman correlation is 0.850. Farther into the later layers, similarity partly recovers but F1 remains below raw, showing that similarity and position must be considered together. The broader VGGT comparison is consistent with this pattern: 16 of 21 earlier-only windows improve F1, versus four of 16 later-only windows, with median changes of $+2.95$ and -12.10 points.

These observations favor internally similar windows in the earlier part of VGGT and show a positive similarity–gain association in Pi3. Similarity therefore provides a useful clue to window placement, but the gains also depend on where those layers lie in the backbone. The partitioning procedure, exact configurations, and statistical controls are given in Appendix G.

4.2 What happens during looping and in later layers?

A possible explanation for useful looping is that successive updates settle toward a stable representation. Alternatively, the important distinction may emerge after the loop, when later layers process the changed features. We compare windows that improve and worsen reconstruction to examine both possibilities. Inside the loop, we follow the updated states h_j that are actually passed onward. After the loop, we compare loop and raw features at the same layer. In both comparisons, we measure $1 - \cos$ between corresponding patch-token features; smaller values indicate more closely aligned directions.

Figure 4 shows a clearer contrast in the later layers than in step shrinkage. On the left, the directional change between consecutive loop states decreases for both beneficial and harmful windows. On the right, the beneficial $[5, 8]$ window remains close to the raw feature directions, whereas the harmful $[12, 15]$ and $[13, 16]$ windows produce larger deviations that grow again in later layers. For these measured windows, reconstruction quality is therefore associated with how the changed features evolve through the rest of the backbone, not simply with whether successive loop steps become smaller. Exact configurations, additional windows, and uncertainty estimates are provided in Appendix G.

4.3 Can loop windows include DPT feature-cache layers?

DPT reconstructs dense geometry from features saved at several intermediate backbone layers. One may therefore expect a loop window containing these feature-cache layers to disrupt the head inputs: a saved feature from a replay forward pass can differ from the updated state used by the rest of the model. This motivates testing whether a window can include these layers and still improve predictions when its cached features receive consistent loop updates.

We use the paired feature updates in Equation 3 and move a fixed-length window through VGGT, including positions that contain DPT feature-cache layers. Figure 5 displays this scan. Window $[3, 6]$, which contains DPT feature-cache layer 4 internally, reaches F1 61.11 versus raw 53.79, the highest F1 in the scan. Its paired scene-bootstrap 95% improvement interval is $[1.59, 13.63]$ points. With consistent feature-cache updates, these layers therefore need not be excluded from useful loop

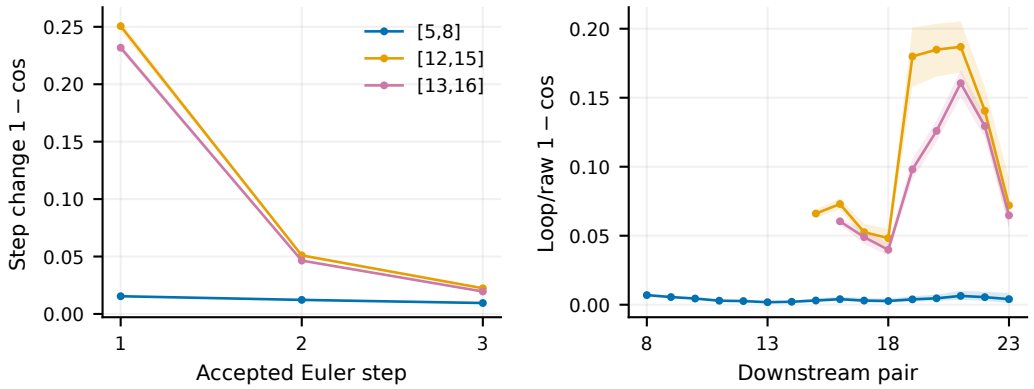


Figure 4: VGGT patch-feature directions on HiRoom. Left: $1 - \cos$ between consecutive updated loop states. Right: $1 - \cos$ between loop and raw features at each subsequent pair. Shading shows pointwise 95% scene-bootstrap intervals.

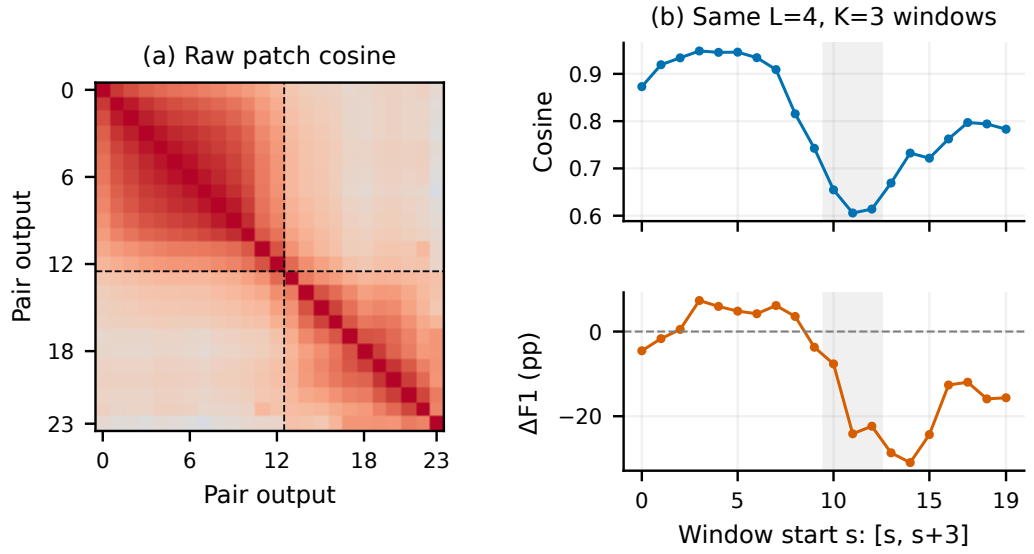


Figure 5: VGGT similarity and F1 gain on HiRoom. (a) Raw patch-feature cosine between frame/global pair outputs; dashed lines separate similarity-based earlier/later layers. (b) Within-window similarity (top) and F1 change (bottom) for windows $[s, s+3]$, $K = 3$. Gray shading marks windows spanning the earlier/later division, not DPT feature-cache locations.

windows. Placement still matters: later windows in the same scan substantially reduce F1. The complete scan settings and comparisons controlling for window position are given in Appendix G.

5 Conclusion

We demonstrate training-free improvements across five geometric backbones using one configuration per model. Gains vary with window placement and task. HiRoom analyses associate VGGT/Pi3 gains with feature similarity and show that consistent intermediate-feature updates enable useful windows containing DPT feature-cache layers.

AI USE STATEMENT

Generative AI tools assisted with experimental planning, software development, workflow management, exploratory result analysis, literature organization, manuscript drafting, and editing. Quantitative claims in this manuscript are tied to recorded experiment outputs. The authors are responsible for the final verification of the methods, results, citations, and text.

REFERENCES

Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120 (2):153–168, 2016. doi: 10.1007/s11263-016-0902-9. URL https://doi.org/10.1007/s11263-016-0902-9.

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6290–6301, 2022. doi: 10.1109/CVPR52688.2022.00619. URL https://openaccess.thecvf.com/content/CVPR2022/html/Azinovic_Neural_RGB-D_Surface_Reconstruction_CVPR_2022_paper.html.

Alessandro Burzio, Tobias Fischer, Sven Elflein, Qunjie Zhou, Riccardo de Lutio, Jiawei Ren, Jiahui Huang, Shengyu Huang, Marc Pollefeys, Laura Leal-Taixé, Zan Gojcic, and Haithem Turki. Déjà View: Looping transformers for multi-view 3D reconstruction. *arXiv preprint arXiv:2605.30215*, 2026. URL https://arxiv.org/abs/2605.30215.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision – ECCV 2012*, pp. 611–625, 2012. doi: 10.1007/978-3-642-33783-3_44. URL https://link.springer.com/chapter/10.1007/978-3-642-33783-3_44.

Lizhang Chen, Jonathan Li, Chen Liang, Ni Lao, and Qiang Liu. Training-free looped transformers. *arXiv preprint arXiv:2605.23872*, 2026. URL https://arxiv.org/abs/2605.23872.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3R: Estimating disentangled motion from DUSt3R without training. *arXiv preprint arXiv:2503.24391*, 2025. URL https://arxiv.org/abs/2503.24391.

Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33768–33780, 2022. doi: 10.52202/068431-2447. URL https://papers.neurips.cc/paper_files/paper/2022/hash/dab5a29f6614ec47ea0ca85c140226fd-Abstract-Conference.html.

Mozes Jacobs, Thomas Fel, Richard Hakim, Alessandra Brondetta, Demba Ba, and T. Andy Keller. Block-recurrent dynamics in vision transformers. *arXiv preprint arXiv:2512.19941*, 2025. URL https://arxiv.org/abs/2512.19941.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3D reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. URL https://arxiv.org/abs/2509.13414.

Haotong Lin, Sili Chen, Junhao 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. URL https://arxiv.org/abs/2511.10647.

Jonathan Lys, Vincent Gripon, Bastien Pasdeloup, Axel Marmoret, Lukas Mauch, Fabien Cardinaux, and Ghouthi Boukli Hacene. Inner loop inference for pretrained transformers: Unlocking latent capabilities without training. *arXiv preprint arXiv:2602.14759*, 2026. URL https://arxiv.org/abs/2602.14759.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *arXiv preprint arXiv:2103.13413*, 2021. URL https://arxiv.org/abs/2103.13413.

Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3260–3269, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Schops_A_Multi-View_Stereo_CVPR_2017_paper.html.

Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGB-D images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2930–2937, 2013. URL https://openaccess.thecvf.com/content_cvpr_2013/html/Shotton_Scene_Coordinate_Regression_2013_CVPR_paper.html.

Michal Stary, Julien Gaubil, Ayush Tewari, and Vincent Sitzmann. Understanding multi-view transformers. *arXiv preprint arXiv:2510.24907*, 2025. URL https://arxiv.org/abs/2510.24907.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 573–580, 2012. doi: 10.1109/IROS.2012.6385773. URL https://cvg.cit.tum.de/data/datasets/rgbd-dataset.

SVLightVerse. SVLightVerse: Large-scale photorealistic indoor dataset with spatially-varying HDRI lighting, 2025. URL https://jerrypiglet.github.io/SVLightVerse/. Project page linked as the HiRoom source by Depth Anything 3; accessed September 26, 2026.

VGGT-Omega Authors. VGGT-Ω reproduction. https://github.com/facebookresearch/vggt-omega/blob/main/reproduction.md, 2026. Official checkpoint and benchmarking advisory; accessed September 23, 2026.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025a. URL https://arxiv.org/abs/2503.11651.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. *arXiv preprint arXiv:2605.15195*, 2026. URL https://arxiv.org/abs/2605.15195.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Pi3: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025b. URL https://arxiv.org/abs/2507.13347.

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 (ICCV)*, pp. 12–22, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Yeshwanth_ScanNet_A_High-Fidelity_Dataset_of_3D_Indoor_Scenes_ICCV_2023_paper.html.

A CHECKPOINTS AND NATIVE GEOMETRY

We use VGGT-1B, original Pi3 (not Pi3X), DA3-Giant-1.1, the frozen released MapAnything checkpoint, and VGGT-Ω-1B/512. Raw/loop comparisons preserve each model's native precision and preprocessing. VGGT uses a BF16 backbone with FP32 geometry heads; Pi3 raw/loop evaluations both use Torch 2.4.1. DA3 uses its native 504 upper-bound preprocessing and first-reference camera-head path. MapAnything uses native 518 resizing/cropping, image-only inference, and its scene-level scale token. Depth along the optical axis is distinguished from ray distance during output adaptation. Point-map reconstruction is used for Pi3 and MapAnything; VGGT and DA3 use predicted cameras and depths with the corresponding fusion evaluator.

VGGT-Ω checkpoint version. Our VGGT-Ω study retains the originally released 1B/512 checkpoint for paired raw/loop comparisons. The authors subsequently reported a potential training-data contamination risk whose impact on that checkpoint could not be conclusively established, and recommended an August 2026 retrained checkpoint at resolution 416 for future benchmarking (VGGT-Omega Authors, 2026). Our results characterize the historical 512 checkpoint, not the recommended retrained model. The two versions are not silently mixed, and transfer of the selected loop settings to the latter is not measured here.

B EXACT EVALUATION SCOPE

Table 7: Benchmark subsets. Scene counts refer to the actual evaluated subsets, not the size of each original dataset. Static describes scene content, not whether the camera moves.

Dataset	Scenes	Origin/content	Tasks
HiRoom	30	Synthetic indoor, static	pose, reconstruction
ETH3D	11	Real indoor/outdoor, static	pose, reconstruction, depth
DTU pose artifact	13	Real objects, static	pose
DTU reconstruction	22	Real objects, static	reconstruction
7Scenes	7	Real indoor, static	pose, reconstruction, depth
NRGBD	9	Synthetic indoor, static	pose, depth
ScanNet++	20	Real indoor, static	pose, reconstruction
DyCheck	14	Real, dynamic	pose, depth
Sintel	14	Synthetic, dynamic	pose, depth
TUM-Dynamic	8	Real indoor, dynamic	pose, depth

HiRoom uses all 400 released images (10–23 per scene); its 30-scene list is retained without exclusions. ETH3D main pose/reconstruction uses all 377 filtered views. ScanNet++ uses exactly 100 seed-42 selected frames per scene. DTU reconstruction uses all 49 views in each of the 22 scans, reference index 33, and the released DA3 evaluation settings; its depth-based and native-point-map construction paths remain distinct. The released pose artifact named DTU-64 is internally mixed: seven scenes have 49 images and six have 64. We retain its cells as an artifact audit rather than claiming faithful reproduction of a nominal 64-view protocol.

Video sampling and camera evaluation. NRGBD and the dynamic datasets use Pi3-style random-order ten-frame samples. A NumPy RandomState initialized once per dataset at seed 42 samples without replacement in scene-list order, retaining the sampled frame order. For ETH3D depth with VGGT, Pi3, DA3, and MapAnything, we shuffle each scene's frame indices with Python's seed-42 generator, take the first ten, and restore their original order. VGGT-Ω instead uses the Pi3-style random-order sample on all six of its benchmarks, including ETH3D depth. Camera metrics use relative rotation and translation-direction angular errors over camera pairs and average AUC over scenes. VGGT-Ω's random 10-frame sample is not published in its paper; our deterministic surrogate is therefore not an exact replay of its reported sample. The first four models' main 7Scenes camera entries use DA3-BENCH, whereas VGGT-Ω's entries use the ten-frame VGGT evaluator. The DA3-BENCH camera evaluator composes world-to-camera matrices as $E_i^{-1}E_j$, whereas the

ten-frame evaluator follows VGGT's released $E_i E_j^{-1}$ convention. Each raw/loop pair uses the same evaluator; these two protocols are not treated as interchangeable cross-model comparisons.

Depth. A single affine transform $s\hat{d} + t$ is fitted by L1 loss for each sequence, jointly across its selected views, rather than independently per image. AbsRel and $\delta_{1.25}$ are weighted by valid-pixel count. Depth ranges in metres are ETH3D $(0, 80)$; 7Scenes/NRGBD $(0, 10)$; DyCheck $(0, 3)$; Sintel/TUM $(0, 70)$. Native image crop/resize mappings and excluded out-of-crop regions are retained. These aligned metrics do not measure unaligned metric-scale accuracy.

Reconstruction. Unposed reconstructions are aligned using the predicted and GT trajectories through the frozen Umeyama–RANSAC path (seed 42). VGGT/DA3 use scaled depths and TSDF outside DTU, and the released valid-depth backprojection path on DTU; Pi3/MapAnything align native world-point maps on all reconstruction datasets. The F1 distance threshold is 0.05 m on HiRoom, 7Scenes, and ScanNet++, and 0.25 m on ETH3D. Native point maps use the corresponding point-map evaluator without inventing intrinsics. DTU reports the mean of public accuracy/completeness distances in millimetres. Overall distances in different units are never averaged directly; aggregate improvements first use each benchmark's relative change.

Complete ETH3D support. The initial minimal download contained GT depths for only 177 of the 377 reconstruction views. The released loader admitted invalid regions when missing depths fell back to all-ones support. We repaired the support to finite, positive GT depth intersected with the released image-mask rule for every view, leaving inference, trajectory alignment, reference geometry, and thresholds unchanged. Both raw and loop rows use this repaired support. Four raw-model F1 scores changed from 37.16/58.11/44.95/46.00 to 62.86/73.41/79.60/57.76 for VGGT/Pi3/DA3/MapAnything. A GT-camera/GT-depth courtyard control reaches F1 99.9654; this is an evaluator diagnostic, not a model score. The repair does not change video-depth or pose evaluations.

C Per-benchmark result conventions

Each full-suite model has 39 metric cells (18 camera, 9 reconstruction, 12 depth); VGGT-Ω has 24 camera/depth cells across six benchmarks. Tables 2–6 report raw and the same shared loop from Table 1. Specialists are not substituted into these rows. Blank task coverage is shown with dashes or omitted rows, never zeros. Values are snapshots of validated result records, rounded only for presentation.

D Search coverage and selection details

The native index axes are 24 attention pairs for VGGT and VGGT-Ω, 18 attention pairs for Pi3, 40 physical blocks for DA3, and 16 fusion blocks for MapAnything. Window endpoints are zero-based and inclusive. Across-benchmark selection uses mean AUC@3, AUC@30, F1, relative reconstruction-distance reduction, depth accuracy, and relative AbsRel reduction; VGGT-Ω uses its four applicable metrics. Mean ranks balance these metrics, with improved-cell counts as a complementary criterion; model-specific choices are described below.

Seven Pi3 configurations are explicitly discarded after non-finite predictions or empty valid-point outputs; 452 complete the full suite. Their separately valid HiRoom outputs remain in the HiRoom-only analysis rather than being deleted because another benchmark failed. DA3's 318 configurations consist of 45 multi-view and 273 initial single-view configurations. MapAnything's 48 configurations include windows ending at a feature-cache layer; the wrapper updates those features before the head reads them. For VGGT-Ω, merging the one-factor study with earlier combined-strategy searches gives 2,333 distinct strategies. For each of the 183 window/K settings, the one-factor study varies replay mode, anchor strength, or token scope one at a time from pair-based whole-window looping with all-token updates and $\beta = 0$. The 2,133-study itself has 183 configurations for each of base, layer replay, five anchor strengths, and four nontrivial token scopes, plus 60 frame-only and 60 inter-only controls.

Table 8: Completed search spaces. Completeness refers to these explicitly defined spaces, not every mathematically possible recurrence design. $K \in \{2, 3, 4\}$ throughout.

Model	Admissible windows/design	Configurations
VGGT	61 pair windows excluding DPT feature-cache indices 4, 11, 17, 23	183
Pi3	All contiguous pair windows in 0–16; retain pair 17	459
DA3	15 later pair windows and all 91 single-view windows in physical blocks 0–12	318
MapAnything	16 global/frame pair windows; feature-cache layers may end a window but cannot lie strictly inside it	48
VGGT-Ω	Global one-factor design across window, K, anchor, scope, and replay unit	2,133

Shared settings and specialists. For VGGT, the mean-rank choice with improvements in all six aggregate metrics is $[5, 8]/4$; $[5, 8]/3$ improves more individual cells, and $[5, 8]/2$ uses less added computation. Pi3 uses the fixed $[4, 8]/3$ shared configuration, which improves 31 of 39 cells at unrounded precision. DA3 $[6, 6]/3$ is the only configuration in its completed grid that improves all six aggregate metrics. MapAnything $[2, 3]/2$ is selected by displayed improved-cell count (31 improved, seven regressed, one tied), not by improving every aggregate metric. This setting updates all tokens, including its scene-level scale token. VGGT-Ω uses patch-only $[8, 9]/4$, $\beta = 0.125$, selected by balanced aggregate rank across the combined strategies. The patch-plus-anchor setting is not part of the global one-factor patch branch, which has $\beta = 0$.

Historical ETH3D re-evaluation. The repaired-support history covers the historically top-ranked third of configurations: 61 of 183 VGGT and 151 of 452 valid Pi3 configurations, plus raw references. Candidates were fixed by historical ranks across the six aggregate metrics before reading repaired scores. This budgeted re-evaluation does not establish a new exhaustive ranking under the corrected protocol. Within those corrected candidates, Pi3 $[5, 10]/3$ has stronger balanced aggregate changes than the fixed main-table choice but fewer individually improved cells; it is not silently substituted into the main rows. VGGT's balanced choice remains $[5, 8]/4$. DA3/MapAnything sweeps already used complete support, and VGGT-Ω has no reconstruction cell in this suite.

Failures and computation. Long windows can produce severe degradation, but length alone is not a valid rejection rule: Pi3's corrected-candidate leaders include length-six and length-seven windows. Likewise, increasing K can improve one task while worsening another. These comparisons hold the damped Euler update fixed while changing the window and K.

E SINGLE-PASS TASK-SPECIALIST RESULTS

We report task-specialist configurations for VGGT, DA3, and VGGT-Ω alongside each model's shared setting. Each row uses one complete model inference, with the same frozen weights as raw inference; all evaluated outputs come from that configuration. Thus these tables expose both target-task gains and changes in the other tasks, before any multi-pass routing or mixing. VGGT's shared setting is also its reconstruction specialist, so they occupy one row labeled shared/recon.

Configurations. Table 9 specifies every window and K, with zero-based inclusive indices. VGGT windows use frame/global pairs and VGGT-Ω windows use frame/inter-frame pairs; all DA3 settings here use physical blocks in its single-view stage. VGGT and DA3 update all tokens without an anchor. For VGGT-Ω, the shared setting updates patches with $\beta = 0.125$, the camera specialist updates patches and registers with $\beta = 0$, and the depth specialist updates patches and camera tokens with $\beta = 0$. These scopes restrict the loop update, not participation in the native attention operations.

Metric conventions. The six aggregate changes use equal weights across benchmarks: nine for each camera metric, four for F1, five for reconstruction distance, and six for each depth metric. VGGT-Ω uses six benchmarks for each of its four applicable metrics. Distance and AbsRel changes

are computed as relative error reductions within each benchmark before averaging. Tables 10, 11, 12, 13, and 14 give the corresponding absolute scores and changes from raw. The raw references and evaluation protocols are identical to the main tables, including complete GT-depth support for ETH3D reconstruction. Small parentheses show differences from raw computed before rounding; green denotes improvement, red regression, and black a difference rounding to zero. Dashes denote unsupported tasks, not zero scores.

Table 9: Single-pass shared and task-specialist configurations: aggregate changes from each model's raw inference. AUC, F1 and depth accuracy changes are percentage points; Dist./AbsRel are the mean per-benchmark relative error reductions (percent). Positive values indicate improvement in every column. All supported tasks are shown for every configuration, not only its selection task. VGGT's shared and reconstruction-specialist configuration is identical.

Model	Method	Window/K	ΔAUC3	ΔAUC30	ΔF1	Dist. red.	$\Delta\delta$	AbsRel red.
VGGT	shared/recon.	$[5, 8]/4$	+0.762	+0.127	+2.283	+2.56	+0.226	+1.85
VGGT	camera	$[12, 12]/3$	+1.125	+0.710	+1.913	+0.14	-0.164	-2.68
VGGT	depth	$[5, 9]/4$	+0.614	+0.040	+1.752	+0.73	+0.385	+2.50
DA3	shared	$[6, 6]/3$	+0.319	+0.113	+0.263	+0.54	+0.059	+0.06
DA3	camera	$[2, 10]/2$	+1.381	+0.672	+0.105	+0.20	-0.385	-0.91
DA3	recon.	$[0, 1]/4$	-0.016	-0.239	+0.404	+2.34	-0.683	-0.52
DA3	depth	$[2, 12]/2$	-0.041	+0.135	-0.947	-3.72	+0.089	+0.70
VGGT-Ω	shared	$[8, 9]/4$	+0.336	+0.152	–	–	+0.094	+0.49
VGGT-Ω	camera	$[8, 9]/3$	+0.590	+0.225	–	–	+0.066	-0.07
VGGT-Ω	depth	$[8, 9]/4$	-0.070	+0.015	–	–	+0.100	+0.57

Table 10: Single-pass specialists: static-scene camera pose. Configurations are defined in Table 9; raw references and protocols match Table 2. Each row uses complete outputs from one model inference. DTU* retains the mixed-view artifact protocol; VGGT-Ω uses 10-frame sampling.

Model	Method	HiRoom		ETH3D		DTU*		7Scenes		ScanNet++	
		AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑
VGGT	raw	48.19	88.02	29.42	81.68	79.66	85.26	23.11	84.73	59.55	94.37
VGGT	shared/recon.	52.20 (+4.02)	89.54 (+1.52)	29.19 (-0.23)	82.87 (+1.18)	79.85 (+0.19)	85.33 (+0.06)	23.09 (-0.03)	84.77 (+0.03)	60.29 (+0.73)	94.54 (+0.17)
VGGT	camera	52.95 (+4.76)	89.56 (+1.55)	29.15 (-0.27)	84.49 (+2.80)	80.27 (+0.60)	85.86 (+0.60)	23.12 (+0.00)	84.73 (-0.00)	60.80 (+1.25)	94.59 (+0.22)
VGGT	depth	51.14 (+2.95)	89.01 (+0.99)	29.88 (+0.46)	83.21 (+1.52)	79.80 (+0.14)	85.27 (+0.01)	22.84 (-0.28)	84.69 (-0.04)	61.17 (+1.62)	94.74 (+0.37)
DA3	raw	80.15	95.90	48.46	91.41	93.99	99.38	27.50	86.69	84.66	98.27
DA3	shared	80.20 (+0.05)	95.90 (-0.00)	48.54 (+0.08)	91.68 (+0.27)	94.21 (+0.21)	99.40 (+0.02)	27.51 (+0.01)	86.69 (+0.01)	84.77 (+0.11)	98.27 (+0.00)
DA3	camera	83.21 (+3.06)	97.05 (+1.15)	48.62 (+0.16)	91.75 (+0.33)	95.06 (+1.07)	99.50 (+0.12)	27.93 (+0.43)	86.75 (+0.07)	85.14 (+0.48)	98.33 (+0.06)
DA3	recon.	80.39 (+0.24)	95.78 (-0.12)	48.16 (-0.30)	91.50 (+0.08)	92.32 (-1.67)	99.22 (-0.16)	27.42 (-0.08)	86.69 (-0.00)	84.66 (-0.00)	98.25 (-0.02)
DA3	depth	79.69 (-0.46)	96.22 (+0.32)	47.50 (-0.96)	91.44 (+0.03)	93.46 (-0.54)	99.33 (-0.05)	28.12 (+0.62)	86.80 (+0.11)	84.94 (+0.28)	98.30 (+0.03)
VGGT-Ω	raw	–	–	44.18	91.44	–	–	54.81	94.25	–	–
VGGT-Ω	shared	–	–	44.04 (-0.13)	91.51 (+0.07)	–	–	55.87 (+1.06)	94.39 (+0.14)	–	–
VGGT-Ω	camera	–	–	43.03 (-1.14)	91.34 (-0.10)	–	–	56.72 (+1.90)	94.46 (+0.20)	–	–
VGGT-Ω	depth	–	–	43.97 (-0.20)	91.39 (-0.05)	–	–	56.30 (+1.48)	94.38 (+0.13)	–	–

Table 11: Single-pass specialists: camera pose on NRGBD and dynamic scenes. Configurations are defined in Table 9; raw references and protocols match Table 3. Each row uses complete outputs from one model inference.

Model	Method	NRGBD		DyCheck		Sintel		TUM-Dyn.	
		AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑	AUC@3↑	AUC@30↑
VGGT	raw	70.04	95.90	21.16	77.43	18.36	66.06	24.81	75.38
VGGT	shared/recon.	70.04 (-0.00)	95.97 (+0.07)	22.33 (+1.16)	76.73 (-0.70)	18.99 (+0.63)	64.03 (-2.03)	25.19 (+0.37)	76.21 (+0.83)
VGGT	camera	69.88 (-0.16)	95.88 (-0.02)	22.38 (+1.22)	77.78 (+0.35)	18.68 (+0.32)	65.60 (-0.46)	27.22 (+2.41)	76.73 (+1.35)
VGGT	depth	70.70 (+0.66)	95.97 (+0.07)	22.59 (+1.43)	76.22 (-1.21)	17.20 (-1.16)	64.07 (-1.98)	24.54 (-0.28)	76.02 (+0.64)
DA3	raw	93.50	99.19	50.63	89.83	21.75	62.27	27.04	68.56
DA3	shared	93.99 (+0.49)	99.27 (+0.08)	51.16 (+0.53)	90.01 (+0.17)	22.86 (+1.11)	62.70 (+0.43)	27.31 (+0.28)	68.58 (+0.03)
DA3	camera	93.91 (+0.41)	99.23 (+0.05)	53.44 (+2.80)	91.05 (+1.22)	23.17 (+1.43)	62.79 (+0.52)	29.63 (+2.59)	71.08 (+2.53)
DA3	recon.	93.83 (+0.33)	99.27 (+0.08)	51.06 (+0.42)	89.45 (-0.38)	20.90 (-0.85)	60.95 (-1.32)	28.80 (+1.76)	68.24 (-0.31)
DA3	depth	92.67 (-0.82)	99.13 (-0.06)	52.43 (+1.80)	90.51 (+0.68)	19.42 (-2.33)	60.08 (-2.19)	29.07 (+2.04)	70.90 (+2.34)
VGGT-Ω	raw	87.49	98.12	52.12	91.07	51.59	87.58	40.56	85.93
VGGT-Ω	shared	88.15 (+0.66)	98.21 (+0.09)	52.91 (+0.79)	91.28 (+0.20)	51.69 (+0.11)	88.06 (+0.48)	40.09 (-0.46)	85.86 (-0.06)
VGGT-Ω	camera	87.90 (+0.41)	98.15 (+0.03)	52.38 (+0.26)	91.23 (+0.15)	53.97 (+2.38)	88.77 (+1.20)	40.28 (-0.28)	85.80 (-0.13)
VGGT-Ω	depth	87.33 (-0.16)	98.12 (+0.01)	51.59 (-0.53)	91.07 (-0.00)	51.69 (+0.11)	88.10 (+0.52)	39.44 (-1.11)	85.41 (-0.52)

Table 12: Single-pass specialists: unposed reconstruction. Configurations are defined in Table 9; raw references and protocols match Table 4. Each row uses complete outputs from one model inference. F1 is in percent; distances are cm on HiRoom, mm on DTU, and metres elsewhere. ETH3D uses complete GT-depth support.

Model	Method	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
		F1↑	Dist.↓	F1↑	Dist.↓	F1↑	Dist.↓	F1↑	Dist.↓	F1↑	Dist.↓
VGGT	raw	53.79	10.33	62.86	0.5748	–	1.9405	48.16	0.1401	67.40	0.0791
VGGT	shared/recon.	60.43 (+6.64)	8.88 (-1.44)	64.52 (+1.66)	0.5759 (+0.0010)	–	1.9414 (+0.0009)	48.12 (-0.04)	0.1425 (+0.0025)	68.27 (+0.87)	0.0785 (-0.0007)
VGGT	camera	61.52 (+7.73)	9.59 (-0.73)	65.56 (+2.70)	0.5767 (+0.0018)	–	2.0099 (+0.0694)	45.32 (-2.84)	0.1431 (+0.0031)	67.46 (+0.06)	0.0794 (+0.0002)
VGGT	depth	58.34 (+4.55)	9.91 (-0.41)	63.92 (+1.07)	0.5736 (-0.0012)	–	1.9841 (+0.0436)	47.74 (-0.42)	0.1416 (+0.0015)	69.21 (+1.80)	0.0770 (-0.0022)
DA3	raw	85.20	4.63	79.60	0.4456	–	1.7832	50.55	0.1229	78.33	0.0675
DA3	shared	86.46 (+1.26)	4.60 (-0.02)	78.51 (-1.10)	0.4436 (-0.0020)	–	1.7646 (-0.0186)	51.49 (+0.94)	0.1220 (-0.0009)	78.27 (-0.06)	0.0675 (+0.0000)
DA3	camera	87.41 (+2.21)	4.26 (-0.36)	78.56 (-1.04)	0.4473 (+0.0017)	–	1.8466 (+0.0634)	50.23 (-0.32)	0.1259 (+0.0030)	77.89 (-0.43)	0.0678 (+0.0003)
DA3	recon.	86.42 (+1.22)	4.59 (-0.03)	80.14 (+0.54)	0.4423 (-0.0033)	–	1.5837 (-0.1995)	50.35 (-0.20)	0.1238 (+0.0009)	78.38 (+0.06)	0.0677 (+0.0002)
DA3	depth	84.55 (-0.65)	4.67 (+0.05)	77.02 (-2.58)	0.4601 (+0.0145)	–	2.0014 (+0.2182)	50.56 (+0.01)	0.1242 (+0.0013)	77.76 (-0.57)	0.0682 (+0.0007)

Table 13: Single-pass specialists: static-scene depth. Configurations are defined in Table 9; raw references and protocols match Table 5. Each row uses complete outputs from one model inference. $\delta_{1.25}$ is in percent; AbsRel is unscaled. Depth uses one scale-and-shift alignment per sequence.

Model	Method	ETH3D		7Scenes		NRGBD	
		$\delta_{1.25}\uparrow$	AbsRel↓	$\delta_{1.25}\uparrow$	AbsRel↓	$\delta_{1.25}\uparrow$	AbsRel↓
VGGT	raw	98.35	0.0385	92.71	0.0656	98.96	0.0155
VGGT	shared/recon.	98.40 (+0.05)	0.0370 (-0.0015)	92.76 (+0.05)	0.0656 (+0.0000)	98.97 (+0.01)	0.0156 (+0.0001)
VGGT	camera	98.20 (-0.15)	0.0390 (+0.0005)	92.76 (+0.05)	0.0654 (-0.0002)	98.95 (-0.01)	0.0166 (+0.0012)
VGGT	depth	98.07 (-0.28)	0.0381 (-0.0004)	92.90 (+0.19)	0.0648 (-0.0008)	98.94 (-0.02)	0.0154 (-0.0001)
DA3	raw	99.50	0.0217	93.48	0.0610	99.15	0.0090
DA3	shared	99.49 (-0.01)	0.0218 (+0.0001)	93.48 (+0.01)	0.0610 (-0.0000)	99.15 (-0.00)	0.0089 (-0.0000)
DA3	camera	99.69 (+0.19)	0.0207 (-0.0010)	93.49 (+0.01)	0.0612 (+0.0002)	99.14 (-0.00)	0.0090 (+0.0000)
DA3	recon.	98.96 (-0.54)	0.0217 (-0.0000)	93.49 (+0.02)	0.0609 (-0.0000)	99.14 (-0.00)	0.0089 (-0.0000)
DA3	depth	99.70 (+0.20)	0.0209 (-0.0008)	93.47 (-0.01)	0.0614 (+0.0004)	99.14 (-0.01)	0.0092 (+0.0003)
VGGT-Ω	raw	99.91	0.0138	94.65	0.0539	99.67	0.0098
VGGT-Ω	shared	99.91 (+0.00)	0.0138 (-0.0000)	94.63 (-0.02)	0.0538 (-0.0000)	99.67 (+0.00)	0.0098 (-0.0000)
VGGT-Ω	camera	99.91 (-0.00)	0.0139 (+0.0002)	94.61 (-0.04)	0.0539 (+0.0000)	99.67 (+0.00)	0.0099 (+0.0000)
VGGT-Ω	depth	99.91 (+0.00)	0.0138 (+0.0000)	94.62 (-0.02)	0.0538 (-0.0001)	99.67 (+0.00)	0.0097 (-0.0001)

Table 14: Single-pass specialists: dynamic-scene depth. Configurations are defined in Table 9; raw references and protocols match Table 6. Each row uses complete outputs from one model inference. $\delta_{1.25}$ is in percent; AbsRel is unscaled. Depth uses one scale-and-shift alignment per sequence.

Model	Method	DyCheck		Sintel		TUM-Dyn.	
		$\delta_{1.25}\uparrow$	AbsRel↓	$\delta_{1.25}\uparrow$	AbsRel↓	$\delta_{1.25}\uparrow$	AbsRel↓
VGGT	raw	93.41	0.0640	62.86	0.2808	92.84	0.0803
VGGT	shared/recon.	94.65 (+1.24)	0.0601 (-0.0039)	62.84 (-0.02)	0.2715 (-0.0093)	92.87 (+0.03)	0.0813 (+0.0011)
VGGT	camera	92.73 (-0.68)	0.0691 (+0.0051)	62.63 (-0.23)	0.2785 (-0.0022)	92.88 (+0.03)	0.0806 (+0.0003)
VGGT	depth	93.95 (+0.53)	0.0612 (-0.0028)	64.60 (+1.74)	0.2620 (-0.0187)	92.99 (+0.15)	0.0794 (-0.0009)
DA3	raw	95.71	0.0536	67.01	0.2591	94.49	0.0830
DA3	shared	95.73 (+0.02)	0.0536 (-0.0000)	67.34 (+0.33)	0.2585 (-0.0006)	94.50 (+0.01)	0.0828 (-0.0002)
DA3	camera	95.54 (-0.17)	0.0548 (+0.0012)	65.22 (-1.80)	0.2822 (+0.0231)	93.96 (-0.54)	0.0817 (-0.0012)
DA3	recon.	95.58 (-0.13)	0.0525 (-0.0012)	63.60 (-3.41)	0.2725 (+0.0134)	94.47 (-0.03)	0.0834 (+0.0004)
DA3	depth	96.41 (+0.70)	0.0518 (-0.0019)	66.93 (-0.08)	0.2643 (+0.0052)	94.22 (-0.28)	0.0809 (-0.0021)
VGGT-Ω	raw	98.60	0.0279	92.36	0.0831	97.39	0.0350
VGGT-Ω	shared	98.60 (+0.01)	0.0278 (-0.0001)	92.91 (+0.55)	0.0815 (-0.0016)	97.41 (+0.02)	0.0350 (-0.0001)
VGGT-Ω	camera	98.61 (+0.01)	0.0279 (-0.0000)	92.77 (+0.41)	0.0820 (-0.0011)	97.41 (+0.01)	0.0350 (+0.0000)
VGGT-Ω	depth	98.61 (+0.01)	0.0278 (-0.0001)	92.95 (+0.59)	0.0814 (-0.0017)	97.42 (+0.02)	0.0350 (-0.0000)

F TASK ROUTING VERSUS MIXING GEOMETRIC OUTPUTS

Define VGGT specialists $\mathcal{P} = [12, 12]/3$ for camera, $\mathcal{R} = [5, 8]/4$ for reconstruction, and $\mathcal{D} = [5, 9]/4$ for depth. E, I, Z denote extrinsics, intrinsics, and depth. Task routing uses each pass's complete relevant output; geometric mixing assembles a reconstruction from different passes. All rows retain all five reconstruction benchmarks and repaired ETH3D support. The mixed-output row labels specify only the reconstruction inputs: every such row reports camera metrics from $\mathcal{P}$ and depth metrics from $\mathcal{D}$. Consequently, $E_{\mathcal{P}} + (I, Z)_{\mathcal{R}}$ requires a third pass $\mathcal{D}$ for its independently evaluated depth output, even though its reconstruction uses only $\mathcal{P}$ and $\mathcal{R}$.

Table 15: Multi-pass VGGT aggregate metrics. Positive entries denote improvement from raw. Dist./AbsRel are relative reductions (%); other changes are percentage points. Passes count full model executions, not internal K.

Method	Passes	ΔAUC3	ΔAUC30	ΔF1	Dist. red.	$\Delta\delta$	AbsRel red.
Unified $\mathcal{R}$	1	+.762	+.127	+2.283	+2.56	+.226	+1.85
$E_{\mathcal{P}} + (I, Z)_{\mathcal{D}}$	2	+1.125	+.710	+.756	-.23	+.385	+2.50
$(E, I)_{\mathcal{P}} + Z_{\mathcal{D}}$	2	+1.125	+.710	+.642	-.46	+.385	+2.50
$E_{\mathcal{P}} + (I, Z)_{\mathcal{R}}$	3	+1.125	+.710	-.094	-1.08	+.385	+2.50
Task-routed $\mathcal{P}/\mathcal{R}/\mathcal{D}$	3	+1.125	+.710	+2.283	+2.56	+.385	+2.50

The three-pass task-routed system improves 28 of 39 cells from raw, compared with 26 for unified $\mathcal{R}$. Against unified $\mathcal{R}$ directly, routing has 20 wins, ten losses, and nine ties. Thus task specialization improves aggregate trade-offs without eliminating all regressions or preserving the cost of a single pass. The mixed-output results are consistent with a cross-pass calibration/geometry mismatch, but do not independently identify its causal source.

G ADDITIONAL REPRESENTATION EVIDENCE

Similarity conventions and controls. Raw features are collected on all 30 HiRoom scenes (400 images). Let $z_{i,v,t}$ denote token t in view v after backbone unit i, and let T select a token group. The scene-level similarity is

$$S_T(i, j) = \mathbb{E}_{v,t \in T} \frac{\langle z_{i,v,t}, z_{j,v,t} \rangle}{\|z_{i,v,t}\|_2 \|z_{j,v,t}\|_2}. \tag{4}$$

We average scenes equally, then average distinct layer pairs within a window; single-layer windows have no defined internal similarity. The analysis covers 1,008 configurations: 183 VGGT, 459 Pi3, 318 DA3, and 48 MapAnything configurations. All raw matrices use corresponding-token cosine within a model; token grids across different native model resolutions are not matched to one another. Features are normalized in FP32 and scene reductions use high-precision accumulation. A pooled-patch cosine is a different statistic from mean corresponding-patch cosine. We measure each backbone's native token groups: patches throughout; camera and registers in VGGT and VGGT-Ω; registers in Pi3; the CLS and camera slots in DA3's respective stages; and the per-view CLS and scene-scale tokens in MapAnything. VGGT-Ω inter-frame operations that bypass patches, and MapAnything frame blocks that bypass the scale token, leave those tokens unchanged by construction, producing perfect-similarity entries at the corresponding steps.

For each window we first average its gains over the complete set $K = 2, 3, 4$. We then group windows by their length and update strategy; DA3's single-view and multi-view windows form separate strategy groups, while MapAnything's equivalent wrapper implementations are pooled. Windows without defined internal similarity and groups containing fewer than two windows are excluded. Within a group of size n, similarity and gain are ranked separately, with average ranks for ties, and transformed to $\tilde{r} = (r - (n + 1)/2)/n$. The reported association is the Pearson correlation of these transformed ranks concatenated across groups. For the window-start control, we transform start-index ranks in the same way to form a vector p, subtract the fitted component $p(p^\top x)/(p^\top p)$ from each rank vector x, and correlate the two residual vectors. This removes one linear start-rank effect, not every dependence on depth position. The additional window-start adjustment reduces the VGGT and Pi3 F1 associations from 0.717/0.675 to 0.664/0.320.

Table 16: Length/strategy-controlled rank association between raw patch similarity and HiRoom gain, averaging $K = 2, 3, 4$ within each window. Overall is relative distance reduction. These are descriptive associations of overlapping windows, not independent-sample significance tests.

Model	AUC3	AUC30	F1	Overall	Used windows
VGGT	.876	.761	.717	.780	40
Pi3	.568	.600	.675	.646	135
DA3	.172	.322	.296	.190	91
MapAnything	.049	-.066	-.197	.002	15

The division into earlier and later layers shown in Figures 3 and 5 uses the recursive maximum-cut procedure released by Jacobs et al. (2025), with $\alpha = 0.10$. It yields VGGT $[0, 12]/[13, 23]$ and Pi3 $[0, 10]/[11, 17]$ for patch features; these are similarity-based divisions, not architectural feature-cache locations. It forms edge weights $D_{ij} = 1 - S(i, j)$ and recursively splits a current group by its maximum-weight cut only when that cut exceeds the fixed threshold $\epsilon = \alpha \sum_{i<j} D_{ij}$, computed once from the full matrix. Groups with fewer than four layers are not split further; child groups are not constrained to be contiguous. Increasing α stops finer splits without changing the cosine matrix. At a finer Pi3 partition, even the useful main-table window spans two resulting groups. The classification therefore depends on the partitioning threshold. Under the displayed division, Pi3 window $[8, 11]/3$ contains earlier and later layers and improves F1, while the neighboring $[7, 10]/3$ lies entirely in the earlier stage and does not. The counts of improving earlier/later VGGT windows and their median gains in the main text use the mean gain over $K = 2, 3, 4$ for each window.

Raw direction and norms. Here Gℓ denotes the global-attention output of VGGT pair ℓ. Mean patch cosine is 0.96744 between adjacent outputs inside G5–G8, but 0.59961 from G22 to G23. Thus the small raw directional changes in G5–G8 are local to that part of the backbone. In VGGT, adjacent G5–G8 camera/register cosines are .99647/.99636. The same token groups have G12-versus-G23 cosines .10599/.09225 and G22-versus-G23 cosines .64124/.63389. Mean per-scene norm ratios G23/G0 are 66.14, 66.70, and 16.60 for camera, registers, and patches. A norm ratio, an adjacent direction change, and alignment to a distant endpoint are different observables.

Loop traces and uncertainty. Five VGGT windows $[5, 8]$, $[8, 11]$, $[9, 12]$, $[12, 15]$, and $[13, 16]$, all $K = 3$ with all-token updates, have complete 30-scene trajectories. Observers distinguish replay proposals from accepted states and verify unchanged native predictions. For successive updated states we measure $1 - \cos(h_j, h_{j+1})$; at each later pair i we measure $1 - \cos(z^{\text{loop}}_{i,v,t}, z^{\text{raw}}_{i,v,t})$. The plotted measurements use the patch-token components of these states, averaging corresponding tokens within each scene and then weighting scenes equally. For patch tokens, the consecutive-state difference decreases from 0.0154 to 0.0095 in $[5, 8]$, and from 0.2318 to 0.0195 in $[13, 16]$, whose F1 is 28.63 points below raw. The final-pair loop/raw differences are 0.00405, 0.07198, and 0.06475 for $[5, 8]$, $[12, 15]$, and $[13, 16]$, respectively. Later-layer feature norms can approach raw norms while their directions remain different. The plotted intervals use 2,000 paired scene-bootstrap draws, seed 42, and are pointwise 95% percentile intervals rather than simultaneous bands. The cache scan intervals are exploratory and are not adjusted for selecting the best window or for multiple comparisons.

Windows containing DPT feature-cache layers. The scan evaluates all 20 length-four VGGT windows starting at pairs 0 through 19 on HiRoom, with $K = 3$ and the paired feature-cache update in Equation 3. The DPT feature-cache indices are 4, 11, 17, and 23; no scanned window includes the final index 23. Eight windows contain no feature-cache layer, three end at one, six contain one strictly internally, and three start at one. The Spearman correlations between window similarity and AUC@3, AUC@30, F1, and distance improvement are 0.860, 0.895, 0.850, and 0.821. These overlapping windows provide descriptive associations rather than independent samples. The best-F1 window $[3, 6]$ does not outperform $[5, 8]/3$ on every task: the latter has better pose and Overall. The strict-internal versus no-cache F1 difference is $+2.24$ points across the full groups, but the Overall difference is $+1.12$ cm (worse). The local control compares starts $c-2, c-1$ (a feature-cache layer strictly inside the window) with $c-4, c+1$ (no feature-cache layer) around each feature-cache index $c \in \{4, 11, 17\}$. We average the two windows on each side within each scene and weight the three feature-cache locations equally, giving both sides the same mean start $c - 1.5$ at each location. This

matches mean position but not all nonlinear effects of position or of the similarity-based division into earlier and later layers. This local matching control yields $+3.33$ F1 points. Separately, excluding windows that contain both earlier and later layers from the full-group comparison yields $+0.51$ with an interval spanning zero; AUC@30 is lower by 1.37 points and Overall worse by 1.67 cm. The differences depend on metric and control; absence of a uniform penalty is not evidence of equivalence. VGGT's camera head reads the final backbone features, while the auxiliary feature-cache states supply the dense heads and never feed back into attention. Changing only the auxiliary readout policy for a fixed window that excludes the final feature-cache layer therefore leaves camera predictions unchanged.

H Implementation validation and reproducibility

Token-restricted updates. In VGGT-Ω, a binary token mask M restricts which tokens accept the loop update. Starting from the native window input h_0, each iteration forms a candidate u_j and merges it with the raw endpoint $a = g_{s:e}(h_0)$:

$$\begin{aligned} u_j &= h_j + \big(g_{s:e}(h_j) - h_j\big)/K, \\ h_{j+1} &= M \odot u_j + (1 - M) \odot a. \end{aligned} \tag{5}$$

After K masked updates, we apply the anchor blend in Equation 2, then merge its output with a using the same mask. Unselected tokens retain their raw window-endpoint values but participate in every attention operation where the native architecture includes them. Thus patch-only updates do not remove camera or register tokens from attention.

Replay granularity. VGGT-Ω whole-window (block) mode loops the composition $g_{s:e}$. Per-pair (layer) mode completes the K updates independently at each pair before passing its updated output to the next pair. Frame-only and inter-only controls use a single pair index and loop only the selected sublayer; its counterpart executes once in the native order. These controls preserve the original order of attention operations.

Native-output checks. The wrapper leaves learned tensors and vendored model implementations unchanged. $K = 1$ and wrapper removal recover native outputs. DA3/MapAnything adapters are checked against official inference APIs, including exact geometry output equality in bounded real-checkpoint tests. MapAnything's endpoint update is independently checked at native feature captures; outputs for all 24 configurations also supported by the original wrapper remain unchanged.

For VGGT, all 183 original cache-free settings pass old/new wrapper equality on CPU reference blocks in FP32 and BF16. A full 30-scene regression of $[5, 8]/4$ compares three prediction heads and four caches, totaling 210 bitwise-equal tensors. Its pose JSON matches the historical run; reconstructed F1/Overall differ slightly (60.43/8.88 cm versus 60.42/8.89 cm) because the historical Open3D mesh sampling is not fully deterministic. Tiny reconstruction changes are therefore not evidence of changed inference. Final metric records, input lists, and configuration completion records are retained; current manifests additionally bind implementation and checkpoint identities where available. The earlier ETH3D support-ablation records lack complete checkpoint and runtime fingerprints. Incomplete or non-finite outputs are not treated as valid results.

Independent language-model check. Before the geometric transfer, the same basic wrapper reproduces a Qwen3-4B SciQ zero-shot result: normalized accuracy on 1,000 test questions increases from 93.40 to 95.10 with window $[15, 18]/3$. The corresponding published values are 93.30 and 95.10 (Chen et al., 2026). This check uses log-likelihood multiple-choice scoring without a generation-time KV cache; it does not validate free-form generation or all language-model settings.

Rendering inputs. The qualitative reconstructions in Figure 1 use all selected inputs: 10 views per HiRoom scene, 49 per DTU scan, and 100 for the ScanNet++ scene.

Benchmarking pipeline. A single-loading pipeline overlaps preparation, storage, and CPU evaluation while retaining metric definitions. A separate raw-suite comparison reduces the same 13-task workload from 4,049 to 1,528 seconds on a single-GPU-equivalent basis through loading reuse and

CPU/I/O overlap. This 2.65× pipeline improvement is not a speedup of the loop operator or a raw-versus-loop latency comparison. On the measured raw suite, 27 of 31 output JSON files are byte-identical to the previous pipeline; the remaining four reconstruction files show small point-sampling variation. HiRoom CPU pose/reconstruction evaluation falls from 466.36 to 268.00 seconds. Uncontrolled multi-process TSDF fusion did not provide stable gains, so its conservative worker setting is retained. Pipeline throughput and model inference overhead are reported separately.