VGGT-ρ: ACCELERATING VISUAL GEOMETRY INFERENCE VIA ADAPTIVE VIEW SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Feed-forward visual geometry models (e.g., VGGT-Ω) have recently demonstrated impressive performance in scene reconstruction and camera pose estimation, but their inference cost grows rapidly with the number of views. We therefore investigate whether a sparse subset of views can preserve reconstruction quality. Our study finds that sparse subsets preserve much of the available geometric coverage, while prioritizing feature-space coverage under a fixed distance threshold and restricting selection to local neighborhoods do not consistently improve reconstruction. These observations motivate VGGT-ρ, a training-free framework that selects views for high-resolution reconstruction using farthest-point sampling (FPS) over all candidate views in the feature space of low-resolution images. Image features are extracted from low-resolution images using the frozen VGGT-Ω image encoder; selection starts from a fixed reference view and repeatedly adds the candidate farthest from its nearest selected representative. The selected RGB images are passed in their relative input order to a frozen high-resolution reconstruction model. Across 7 Scenes, NRGBD, ScanNet-50, and Tanks and Temples, using only 15% of candidate views with VGGT-Ω increases mean Chamfer distance over the full-input scene domain by at most approximately 4 mm, while changing pose AUC@30 on the same selected views by -0.1 to $+1.5$ percentage points. On ScanNet-50, core-pipeline speedups reach $13.0\times$–$24.0\times$, with 63.3%–73.8% lower peak allocated GPU memory, including low-resolution encoding and selection overhead. Transfer experiments with FastVGGT and Sparse-VGGT further demonstrate the applicability of view selection across backbones, with quality trade-offs depending on the dataset and input budget. Project page: `https://vggt-rho.github.io/`.

1 INTRODUCTION

Feed-forward visual geometry models infer camera poses and dense scene geometry directly from images (Wang et al., 2024; Yang et al., 2025; Wang et al., 2025; 2026e; Lin et al., 2026). Models such as VGGT use global attention to integrate information across views in a single forward pass. At a fixed image resolution, the computation of dense global attention grows quadratically with the number of views. The resulting latency and memory requirements limit reconstruction from long videos and large image collections (Wang et al., 2025; Shen et al., 2026; Elflein et al., 2026).

Existing acceleration methods reduce token computation through merging or pruning (Shen et al., 2026; Shu et al., 2026; Li et al., 2026). Other approaches reduce cross-view interaction costs through sparse attention (Wang et al., 2026a), process long sequences in overlapping chunks (Deng et al., 2025), or partition input views into diversity-aware groups for chunk-wise inference (Park et al., 2026). We study the number of RGB views entering high-resolution joint inference as a complementary source of efficiency. We ask: *how many candidate views are needed to reconstruct the scene represented by the full input collection?* Our objective is to recover this scene from a smaller input subset with a frozen reconstruction backbone, while keeping the full-scene evaluation domain fixed.

Our experiments show that, on the evaluated video sequences, sparse uniform subsets retain much of the geometric coverage of full-input reconstructions. This observation suggests that reducing the number of input views can lower computational cost while preserving broad scene coverage. Under

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.

a limited view budget, the key is to complement scene content that is insufficiently covered by the selected subset, so each candidate view must be assessed in relation to those already selected.

Low-resolution DINO descriptors provide an inexpensive representation for comparing candidate views. We extract them from the frozen image encoder of VGGT-Ω (Wang et al., 2026b), before multi-view aggregation. For a fixed sparse subset, our exploratory study finds that most omitted views have a nearby selected representative in DINO space. However, increasing coverage under a descriptor-distance threshold does not consistently improve geometric coverage or reconstruction quality. At a matched final budget, completing the same initial subset with farthest-point sampling (FPS) yields modest reductions in mean Chamfer distance relative to random completion. Restricting candidates to local neighborhoods provides no consistent benefit. These findings motivate using continuous descriptor distances to select representatives from the full candidate set.

We introduce VGGT-ρ, a training-free framework that selects high-resolution inputs through global FPS in low-resolution DINO descriptor space. For N candidate images, a user-specified retained-view fraction ρ determines the budget $K = \lceil \rho N \rceil$. Starting from a fixed reference view, FPS repeatedly selects the candidate farthest from its nearest selected representative until the budget is filled. The selected RGB images retain their relative input order and constitute the entire input to a frozen reconstruction model, which predicts depth maps and camera parameters. The selector computes the low-resolution descriptors and their pairwise distance matrix once and reuses them throughout selection, without requiring geometric ground truth.

We evaluate VGGT-ρ across multiple reconstruction models and datasets, comparing inference with 15% of the candidate views against full-input inference in terms of reconstruction quality and core-pipeline latency. The measurements include low-resolution encoding, view selection, subset assembly, and high-resolution inference.

Our contributions are:

- We show that sparse subsets can retain much of the geometric coverage of full-input reconstructions, while higher thresholded descriptor coverage does not consistently translate into better geometric coverage or reconstruction quality.
- We introduce VGGT-ρ, which combines low-resolution DINO descriptors with global FPS to select a prescribed number of high-resolution inputs for a frozen reconstruction model without additional training.
- We evaluate the trade-off between reconstruction quality and core-pipeline latency when reducing input views across multiple models and datasets, using the same full-scene evaluation domain for each comparison.

2 Related Work

Feed-Forward 3D Foundation Models. Feed-forward geometry models have progressed from pairwise prediction and matching (Wang et al., 2024; Leroy et al., 2024) to joint multi-view reconstruction (Yang et al., 2025; Wang et al., 2025). Other models address reference-view dependence, geometric consistency across view configurations, and metric reconstruction with optional geometric inputs (Wang et al., 2026e; Lin et al., 2026; Keetha et al., 2025).

Reconstruction Acceleration. Existing methods reduce token and cross-view computation through token merging (Shen et al., 2026; Shu et al., 2026; Chen et al., 2026; Wang et al., 2026d), sparse attention, K/V subsampling, or compressed descriptor attention (Wang et al., 2026a; Sun et al., 2026; Wang & Xu, 2026), and combinations of merging and K/V reduction (You et al., 2026; Tang et al., 2026). Learned approaches prune image tokens or use a compressed context branch to guide sparse attention (Li et al., 2026; Ren et al., 2026). Good Token Hunting selects diverse views and intra-frame tokens for the K/V context while retaining all query views (Zheng et al., 2026). Our selected RGB subset constitutes the entire high-resolution input. At the sequence level, VGGT-Long processes overlapping chunks, while SwiftVGGT reduces alignment and loop-closure costs (Deng et al., 2025; Lee et al., 2025). DVP partitions the input collection into balanced, diverse groups for reconstruction (Park et al., 2026), whereas our method selects a single budgeted subset.

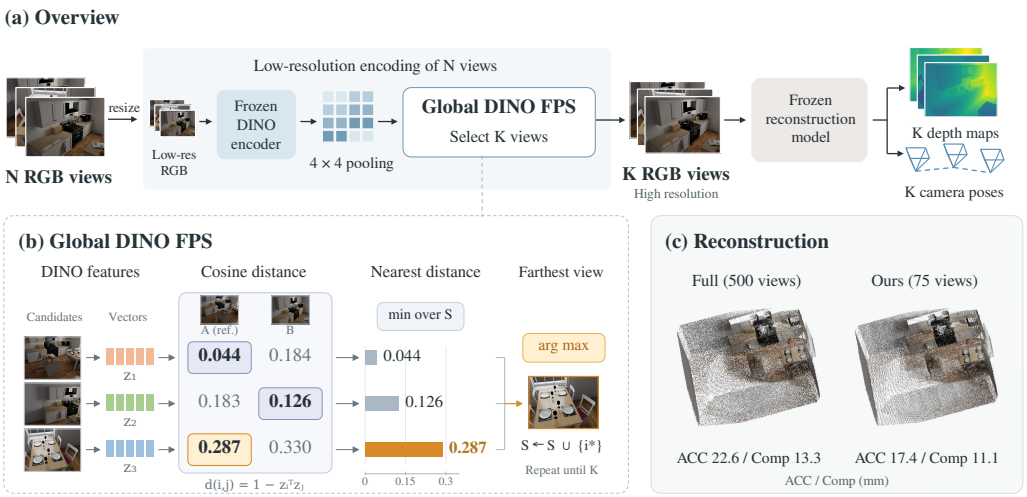

Figure 1: **Overview of VGGT-ρ.** (a) Low-resolution DINO descriptors guide global FPS to select K views for frozen high-resolution reconstruction. (b) In this illustrative FPS example, orange highlights the candidate with the largest nearest-representative cosine distance. (c) VGGT-Ω reconstructions of NRGBD *kitchen* using all 500 views or 75 selected views. ACC and Comp are evaluated against the same full-input GT domain (mm; lower is better).

Streaming Reconstruction and Cache Management. Streaming methods manage historical context through K/V pruning and quantization (Yuan et al., 2026; Su et al., 2026), token compression (Wang et al., 2026c; Liu et al., 2026), or frame retrieval and coherent frame-wise memory (Zou et al., 2026; Xu & Oishi, 2026).

3 METHOD

Given an input sequence $(I_1, \ldots, I_N)$ and a retained-view fraction $\rho \in (0, 1]$, we select $K = \lceil \rho N \rceil$ images for a frozen high-resolution reconstruction model. Let $\mathcal{J} = \{1, \ldots, N\}$ denote the input indices and $r = 1$ the index of the fixed reference view, which counts toward the budget K. Figure 1(a) summarizes the pipeline. We compute low-resolution DINO descriptors for all candidates and apply global farthest-point sampling (FPS) to select K indices. These indices identify the RGB views passed to the frozen high-resolution model in their relative input order.

3.1 PRELIMINARIES

Backbone Architecture and Features. VGGT (Wang et al., 2025) encodes input images into patch tokens, augments them with camera and register tokens, and alternates frame-wise and global attention to predict cameras and dense geometry. VGGT-Ω (Wang et al., 2026b) adopts a DINOv3-initialized encoder, replaces some global-attention layers with register attention, and simplifies the camera and depth heads. Its remaining dense global-attention layers incur an attention cost of $\mathcal{O}(N^2P^2d)$ per layer for N images, P tokens per image, and hidden dimension d, motivating view selection before high-resolution reconstruction. The encoder's patch features are available before multi-view aggregation; we extract them at low resolution to construct the selection descriptors below.

DINO Descriptors. DINO-family encoders provide self-supervised patch representations (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025). We use the DINOv3-initialized image encoder trained within VGGT-Ω and keep its weights frozen during selection (Wang et al., 2026b). All candidate images are encoded at 224×288, producing a 14×18 patch grid before multi-view aggregation. We partition this grid into a 4×4 array of non-overlapping spatial bins. For patch

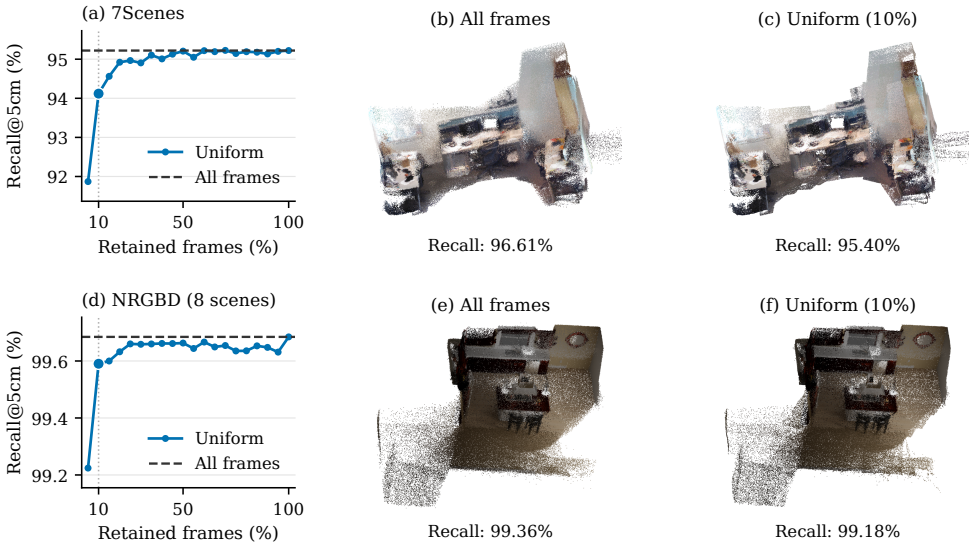

Figure 2: **Uniform subsampling preserves geometric coverage.** (a,d) Mean Recall@5 cm on 7 Scenes (18 sequences) and NRGBD (eight scenes, excluding kitchen). (b,c,e,f) All frames and Uniform (10%) reconstructions of office/seq-02 and complete_kitchen, with per-scene Recall and matched display views. Experimental details are in Appendix A.1.

feature $\mathbf{f}_{i,p}$ at position p in image i, the mean feature of bin B_b is

$$\mathbf{g}_{i,b} = \frac{1}{|B_b|} \sum_{p \in B_b} \mathbf{f}_{i,p}, \qquad b = 1, \ldots, 16. \tag{1}$$

We normalize each bin vector, concatenate the 16 vectors in spatial order, and normalize the result:

$$\mathbf{z}_i = \mathrm{norm}([\mathrm{norm}(\mathbf{g}_{i,1}); \ldots; \mathrm{norm}(\mathbf{g}_{i,16})]), \tag{2}$$

where $\mathrm{norm}(\mathbf{x}) = \mathbf{x}/\|\mathbf{x}\|_2$. This representation retains separate features for different image regions. The cosine distance between candidate views is

$$d(i,j) = 1 - \mathbf{z}_i^\top \mathbf{z}_j. \tag{3}$$

The descriptors and pairwise distance matrix are computed once and reused throughout selection.

3.2 Exploratory Study

Sparse inputs and scene coverage. We first consider three sampling strategies based on input sequence positions: Uniform selects views at approximately equal intervals; Random samples views at random; stratified random partitions the sequence into intervals and randomly samples a view within each interval. We use Uniform to examine how reducing the input views affects geometric coverage, then compare the three strategies in DINO descriptor space.

Figure 2 reports Recall@5 cm against the fixed geometry defined by all candidate inputs, averaged over 18 test sequences of 7 Scenes and eight NRGBD scenes excluding kitchen. Recall rises rapidly at small budgets and then approaches the full-input reference; the paired point clouds illustrate the scene structure retained with 10% of the frames. These observations indicate room to reduce the number of high-resolution inputs. We therefore ask whether inexpensive image descriptors can guide the choice of representative views under a fixed budget.

Descriptor representation and readout cost. We compare camera, register, and patch descriptors read after frame-wise attention blocks with the encoder DINO descriptors defined above. For each readout and scene, a neighborhood threshold is calibrated from the fifth percentile of cosine

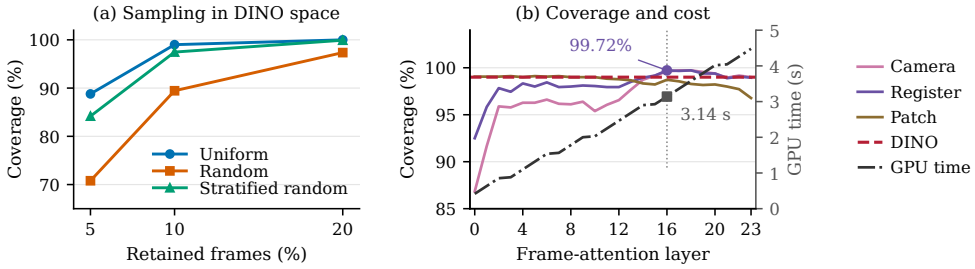

Figure 3: **Descriptor coverage and readout cost.** Four NRGBD scenes, $N = 500$, neighborhood quantile $q = 0.05$. (a) Sampling strategies in DINO space. (b) Uniform $K = 50$ coverage across readouts; the gray curve shows cumulative GPU forward time (right axis). The marker highlights 99.72% Register coverage at 3.14 s. See Appendix A.1 for protocols.

distances between non-reference views. We measure the fraction of omitted views with a selected non-reference representative within the threshold distance; Appendix A.1 defines the diagnostic.

On four NRGBD development scenes with $N = 500$ and Uniform $K = 50$, the selected non-reference views cover 99.00% of omitted views in DINO space. As shown in Figure 3, DINO coverage is close to the highest values achieved by deeper readouts, which require additional aggregation. This diagnostic characterizes how a fixed subset represents the remaining inputs within each descriptor space. Together with DINO's direct availability from the encoder, the result motivates using it as a low-cost representation for subsequent selection experiments.

Selection rules and reconstruction quality. We next test how selection rules in DINO space affect reconstruction quality on all nine NRGBD scenes and all 18 test sequences of 7 Scenes, using up to 500 candidate frames and a final retained fraction of 15%. These datasets have been used for method development. At the same final budget, initializing global FPS with a thresholded-coverage set increases mean CD by 0.012 mm on NRGBD and 0.158 mm on 7 Scenes relative to starting from the reference view alone. With a fixed FPS prefix, continuing FPS to the final budget lowers mean CD by 0.280 mm and 0.063 mm, respectively, compared with random completion from that same prefix. These comparisons show no average CD benefit from coverage initialization and a modest average gain from FPS completion. The full comparisons and protocols appear in Appendix A.2, Table 6.

Candidate restrictions. We isolate candidate-pool effects by fixing coverage initialization and scoring against all selected views. Restricting candidates to large coverage neighborhoods increases mean CD on NRGBD. A separate comparison within that restricted pool finds no consistent mean CD improvement across budgets from using only local selected representatives (Appendix A.2). Together with the initialization and completion results, these observations motivate reference-initialized FPS over all candidates, using all selected views as representatives.

3.3 Global DINO Farthest-Point Sampling

Budgeted selection. Using the distances in Eq. 3, initialize $S = \{r\}$. The distance of an unselected image to its nearest selected representative is

$$\delta_i(S) = \min_{j \in S} d(i, j). \tag{4}$$

We use $\delta_i(S)$ to measure how poorly image i is represented by the current selection in descriptor space. At each step, FPS selects the view with the largest nearest-representative distance:

$$i^\star \in \arg\max_{i \in \mathcal{J} \setminus S} \delta_i(S), \qquad S \leftarrow S \cup \{i^\star\}. \tag{5}$$

Figure 1(b) illustrates this rule. For remaining candidates, update $\delta_i \leftarrow \min\{\delta_i, d(i, i^\star)\}$. Ties favor the smaller input index, and selection continues until $|S| = K$, including when the largest remaining distance is zero.

Table 1: 3D reconstruction on 7 Scenes and NRGBD. ACC and Comp are in meters; Time is in seconds.

Frame budget	Method	7 Scenes				NRGBD			
		ACC↓	Comp↓	NC↑	Time↓	ACC↓	Comp↓	NC↑	Time↓
500	VGGT-Ω	**0.019**	**0.016**	0.629	34.0	0.015	**0.008**	0.875	32.7
	+ Ours	**0.019**	0.017	**0.638**	**2.1**	**0.014**	**0.008**	**0.891**	**2.3**
	FastVGGT	0.023	**0.020**	0.637	26.2	0.024	**0.015**	0.772	26.6
	+ Ours	**0.022**	0.022	**0.643**	**3.5**	**0.022**	0.016	**0.802**	**3.4**
	SparseVGGT	**0.024**	**0.022**	0.661	84.1	0.025	0.015	0.838	81.0
	+ Ours	0.025	0.024	**0.668**	**4.2**	**0.016**	**0.014**	**0.890**	**4.2**
1000	VGGT-Ω	**0.019**	**0.016**	0.628	117.1	0.016	0.009	0.865	114.1
	+ Ours	**0.019**	**0.016**	**0.631**	**4.7**	**0.014**	**0.008**	**0.888**	**4.6**
	FastVGGT	0.024	**0.019**	0.631	61.9	0.034	0.016	0.758	60.5
	+ Ours	**0.023**	0.021	**0.640**	**5.6**	**0.021**	**0.015**	**0.797**	**5.5**
	SparseVGGT	**0.025**	**0.022**	0.650	299.0	0.033	0.016	0.799	290.6
	+ Ours	**0.025**	0.024	**0.666**	**10.2**	**0.017**	**0.014**	**0.882**	**10.0**

High-resolution reconstruction and cost. We assemble the selected RGB images in their relative input order using the input resolution and preprocessing of the chosen reconstruction backbone, then run the frozen model to predict depth and cameras for these K views. Selection requires no additional training or GT inputs, and the reconstruction evaluation domain remains the same as for the full input.

For descriptor dimension d_z, constructing the distance matrix requires $\mathcal{O}(N^2 d_z)$ operations and $\mathcal{O}(N^2)$ storage. Given the matrix, incremental nearest-distance updates require $\mathcal{O}(NK)$ operations for selection. The dense global-attention term in high-resolution inference decreases from $\mathcal{O}(N^2P^2d)$ to $\mathcal{O}(K^2P^2d)$. Core-pipeline timing includes low-resolution encoding, descriptor and distance computation, selection, high-resolution subset assembly, and high-resolution inference.

4 Experiments

We evaluate whether selecting 15% of candidate views preserves reconstruction and pose quality while reducing computation, comparing frozen reconstruction backbones with and without our selector.

4.1 Experimental Setup

Datasets and metrics. We evaluate 3D reconstruction on 7 Scenes (Shotton et al., 2013), Neural RGB-D (NRGBD) (Azinović et al., 2022), and ScanNet-50 (Dai et al., 2017; Shen et al., 2026), and camera pose estimation on ScanNet-50 and Tanks and Temples (T&T) (Knapitsch et al., 2017). Reconstruction metrics include accuracy (ACC), completeness (Comp), normal consistency (NC), and Chamfer distance (CD). Pose evaluation reports AUC@5/15/30 on ScanNet-50 and AUC@30 on T&T. Results are averaged equally over scenes or test sequences within each dataset.

Comparisons. We compare the quality and cost of VGGT-Ω, FastVGGT, and SparseVGGT with and without our selector. The main experiments retain 15% of candidate views. For each backbone, Full and + Ours use the same weights, input resolution, and preprocessing. Reconstruction is evaluated over the same full-scene domain, and pose comparisons use the same selected views. Candidate budgets are specified in the result tables; sampling and evaluation protocols are detailed in Appendix B.

Implementation and timing. The selector uses the frozen VGGT-Ω image encoder (Wang et al., 2026b) at 224×288, without additional training. Experiments run on NVIDIA H200 GPUs. Re-

Table 2: 3D reconstruction on ScanNet-50. CD is in meters; Time is in seconds.

Method	$N = 1000$		$N = 500$		$N = 300$	
	CD ↓	Time ↓	CD ↓	Time ↓	CD ↓	Time ↓
VGGT-Ω	0.206	115.7	**0.216**	33.8	**0.198**	13.2
+ Ours	**0.204**	**4.8**	**0.216**	**1.9**	0.202	**1.0**
FastVGGT	0.192	67.0	**0.186**	26.3	**0.173**	11.8
+ Ours	**0.189**	**5.6**	0.197	**3.4**	0.187	**2.2**
SparseVGGT	0.250	295.3	0.198	83.8	**0.176**	32.7
+ Ours	**0.192**	**10.1**	**0.192**	**4.1**	0.186	**2.4**

ported times for all models include low-resolution encoding, view selection, and high-resolution inference. Additional implementation details are given in Appendix B.4.

4.2 3D Reconstruction

Tables 1 and 2 report reconstruction quality and latency on the RGB-D datasets and ScanNet-50, respectively. With 15% retention, VGGT-Ω largely preserves full-input reconstruction quality. On 7 Scenes, ACC remains approximately 0.019 m at both candidate budgets and NC improves, while Comp increases from 0.016 to 0.017 m at the 500-frame budget and remains approximately 0.016 m at 1000. On NRGBD, ACC and NC improve at both budgets, with Comp maintained or improved. ScanNet-50 shows a similar quality range: CD changes from 0.198/0.216/0.206 m to 0.202/0.216/0.204 m at candidate limits of 300/500/1000.

The selector also transfers to other reconstruction backbones, with dataset-dependent quality trade-offs. At the 1000-frame NRGBD budget, ACC decreases from 0.034 to 0.021 m for FastVGGT and from 0.033 to 0.017 m for SparseVGGT, with improvements in Comp and NC. On 7 Scenes, both backbones improve NC but incur higher completeness error. For FastVGGT on ScanNet-50, CD decreases from 0.192 to 0.189 m at $N = 1000$, but increases by 0.015 and 0.011 m at $N = 300$ and $N = 500$. Thus, reducing the input preserves or improves several quality measures, while completeness and overall reconstruction error can deteriorate in some settings.

Figure 1(c) compares NRGBD *kitchen* reconstructions shown from the same viewpoint and evaluated against the same full-input GT domain. Selecting 75 of 500 views reduces ACC from 22.6 to 17.4 mm and Comp from 13.3 to 11.1 mm. Appendix C shows additional scenes and backbones.

4.3 Camera Pose Estimation

Tables 3 and 4 compare pose quality on the same selected views. On ScanNet-50, VGGT-Ω improves all three AUC metrics at $N = 500$ and $N = 1000$. At $N = 1000$, AUC@5/15/30 increases from 23.7/59.1/76.1 to 24.8/60.4/76.9; at $N = 300$, the metrics decrease by 0.4/0.2/0.1 percentage points. FastVGGT likewise improves at the largest budget and declines at the two smaller budgets. On complete T&T sequences, mean AUC@30 increases from 76.2 to 77.7 for VGGT-Ω and from 95.9 to 96.5 for SparseVGGT, while FastVGGT changes from 93.0 to 92.9. These results show that inference with 15% of candidate views can retain pose quality on the selected frames, with the direction of the change depending on the backbone and input budget.

4.4 Efficiency

We examine runtime and memory scaling on ScanNet-50 at candidate limits of 300, 500, and 1000, using VGGT-Ω with and without selection. Core-pipeline latency decreases from 13.23/33.80/115.74 s to 1.01/1.93/4.82 s, corresponding to $13.0\times$/$17.5\times$/$24.0\times$ speedups computed from the unrounded dataset-mean times. These measurements include selection overhead, which grows from 0.335 to 1.126 s across the tested budgets, while selected-input high-resolution inference grows from 0.680 to 3.689 s. The reduction in high-resolution computation therefore outweighs the additional selection cost throughout this range.

Table 3: Camera pose estimation on ScanNet-50, evaluated on the same selected views. AUC values are percentages.

Method	$N = 1000$			$N = 500$			$N = 300$		
	AUC_5	AUC_{15}	AUC_{30}	AUC_5	AUC_{15}	AUC_{30}	AUC_5	AUC_{15}	AUC_{30}
VGGT-Ω	23.7	59.1	76.1	24.8	59.8	76.5	**25.8**	**61.4**	**77.5**
+ Ours	**24.8**	**60.4**	**76.9**	**25.5**	**60.9**	**77.2**	25.4	61.1	77.4
FastVGGT	23.4	58.4	75.4	**25.4**	**60.6**	**77.0**	**25.8**	**60.9**	**77.1**
+ Ours	**24.5**	**59.5**	**76.2**	24.6	59.8	76.5	23.4	58.5	75.6
SparseVGGT	19.7	52.8	70.6	24.0	58.4	75.2	25.6	60.3	76.7
+ Ours	**24.8**	**59.9**	**76.6**	**25.8**	**61.0**	**77.3**	**26.3**	**61.3**	**77.5**

Table 4: Camera pose estimation on complete T&T sequences, evaluated on the same selected views. AUC@30 is in percent; Time is in seconds.

Method	Barn	Caterpillar	Courthouse	Ignatius	Meetingroom	Truck	Avg. ↑	Time ↓
VGGT-Ω	92.0	84.1	36.0	77.1	85.6	**82.3**	76.2	39.96
+ Ours	**92.4**	**84.7**	**43.4**	**77.6**	**85.7**	**82.3**	**77.7**	**2.54**
FastVGGT	**95.3**	93.1	83.2	**95.8**	**94.3**	96.5	**93.0**	25.80
+ Ours	94.4	**93.2**	**83.8**	94.9	94.1	**97.0**	92.9	**3.02**
SparseVGGT	97.9	96.9	86.5	97.7	97.7	**98.9**	95.9	58.16
+ Ours	**98.2**	**97.1**	**88.3**	**98.1**	**98.2**	98.8	**96.5**	**3.60**

Selection also reduces peak GPU memory. The maximum allocated memory across the 50 scenes and measured rounds decreases from 21.29/32.61/60.92 GiB to 7.81/10.14/15.97 GiB at the three candidate limits, a reduction of 63.3%–73.8%. These peaks cover the complete algorithm, including the resident model and low-resolution selection. At the 1000-frame limit, peak reserved memory likewise decreases from 97.21 to 21.16 GiB. Appendix Table 8 reports the stage timings and both memory peaks.

4.5 Ablation Studies

We study selection rules and retained-view budgets with frozen VGGT-Ω, using up to $N = 500$ uniformly sampled candidates and 384×512 reconstruction. All selectors retain the reference view and preserve input order. Reconstruction uses the same full-input evaluation domain as the main RGB-D experiments; detailed protocols are given in Appendix B.

Selection at a matched budget. Table 5 compares selectors at 15% retention. Coverage + Local FPS uses the same DINO descriptors and first greedily selects an initial set that covers all candidate views under a descriptor-distance threshold. It then restricts further selection to fixed large neighborhoods and applies FPS using the selected views within each neighborhood as distance references until the same view budget is reached (Appendix A.2). Global FPS lowers mean CD relative to this variant by 1.08 mm on 7 Scenes and 1.13 mm on NRGBD, while also reducing mean completeness error and latency. Among the matched-budget selectors, it has the lowest mean completeness error on both datasets. Its mean CD is lower than Random on both datasets and lower than Uniform on NRGBD; Uniform performs better on 7 Scenes. The comparison favors global FPS over the restricted local variant, with dataset-dependent quality differences relative to simple sampling.

Retained-view budget. Figure 4 varies retention from 10% to 25% on NRGBD while holding the candidate inputs and GT domain fixed. Global FPS attains its lowest mean CD among these budgets at 15% (21.59 mm, 1.903 s). Increasing retention to 25% raises core-pipeline time to 3.497 s and mean CD to 22.09 mm. Thus, retaining more views increases computation without a monotonic

Table 5: **Selection strategies at 15% retention on 7 Scenes and NRGBD.** CD and Comp are in millimeters; Time is in seconds. Random results are averaged over 10 and 20 seeds, respectively.

	7 Scenes			NRGBD		
Method	CD↓	Comp↓	Time↓	CD↓	Comp↓	Time↓
Full (100%)	34.54	15.99	33.518	23.62	8.48	32.346
Uniform	**35.35**	17.02	**1.381**	22.02	8.12	**1.336**
Random	36.10	18.03	1.393	21.88	8.39	1.371
Coverage + Local FPS	36.71	18.52	2.015	22.72	8.67	1.950
Global DINO FPS (ours)	35.63	**16.75**	1.956	**21.59**	**8.01**	1.903

Bold: best among 15% selectors.

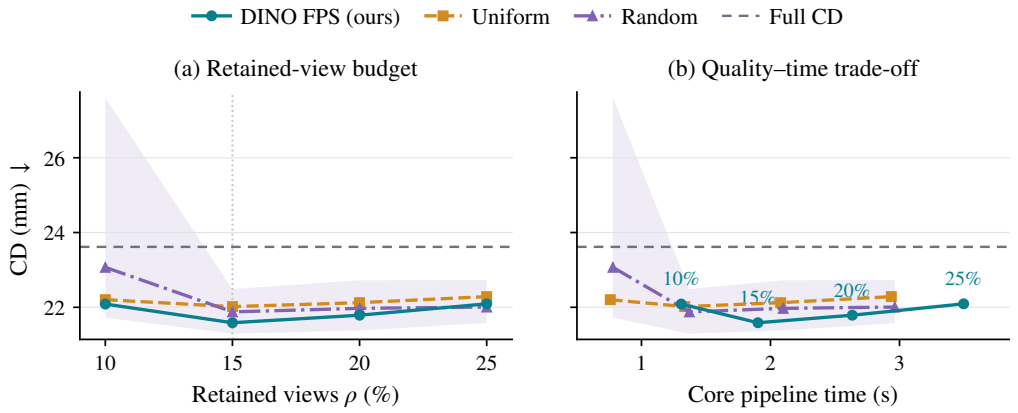

Figure 4: **Retained-view budget and computational cost.** Means over nine NRGBD scenes, $N \leq 500$, with a fixed full-input GT domain. Markers span 10%, 15%, 20%, and 25% retention from left to right on each curve. Random results are averaged over 20 seeds per scene; shading shows the range of seed-wise dataset means, not a confidence interval. Dashed horizontal lines show Full CD; lower CD and time are better.

improvement in reconstruction quality on this panel. Appendix A provides component comparisons, coverage diagnostics, stage timings, and per-unit quality differences.

5 Conclusion

We presented VGGT-ρ, a training-free framework for reducing the number of views participating in high-resolution joint reconstruction. Our experiments show that sparse subsets retain much of the scene's geometric coverage, while additional views or higher thresholded descriptor coverage do not consistently improve reconstruction. Guided by these findings, VGGT-ρ uses low-resolution DINO descriptors and reference-initialized global FPS to select a prescribed number of inputs for a frozen reconstruction backbone.

Across four benchmarks, retaining 15% of candidate views largely preserves VGGT-Ω reconstruction quality over the full-input scene domain and pose quality on the selected views. For this backbone, core-pipeline latency and peak GPU memory decrease substantially even after accounting for selection overhead. Applying the same selector to FastVGGT and SparseVGGT demonstrates compatibility with token merging and sparse attention, with quality trade-offs that depend on the dataset and budget. Together, these results support input-view selection as a practical way to control reconstruction cost that complements acceleration within the backbone.

REPRODUCIBILITY STATEMENT

The view selection procedure is described in Section 3.3. The experimental setup is specified in Section 4.1, and Appendix B details input sampling, reconstruction and pose evaluation, and implementation settings. Protocols for the exploratory studies and additional ablations are provided in Appendix A.

AI USE STATEMENT

The authors conceived the main research idea and designed the exploratory experiments. Generative AI tools assisted with code implementation, monitoring experimental runs, analysis of experimental results, numerical verification, creation of scientific figures, literature retrieval and discovery, drafting of manuscript sections, translation, language editing, and building the project website. Generative AI was not used for any research or writing tasks other than those listed above. The authors take responsibility for the final content of this paper, including its claims, results, references, and material produced with AI assistance.

REFERENCES

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Azinovic_Neural_RGB-D_Surface_Reconstruction_CVPR_2022_paper.html.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. URL https://arxiv.org/abs/2104.14294.

Yutian Chen, Yuheng Qiu, Ruogu Li, Jay Patrikar, and Sebastian Scherer. Co-Me: Confidence guided token merging for visual geometric transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14590–14599, 2026.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Dai_ScanNet_Richly-Annotated_3D_CVPR_2017_paper.html.

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. URL https://arxiv.org/abs/2507.16443.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3D reconstruction at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36464–36474, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Elflein_VGG-T3_Offline_Feed-Forward_3D_Reconstruction_at_Scale_CVPR_2026_paper.html.

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. doi: 10.48550/arXiv.2509.13414. URL https://arxiv.org/abs/2509.13414.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. URL https://tanksandtemples.org/.

Jungho Lee, Minhyeok Lee, Sunghun Yang, Minseok Kang, and Sangyoun Lee. SwiftVGGT: A scalable visual geometry grounded transformer for large-scale scenes. *arXiv preprint arXiv:2511.18290*, 2025. URL https://arxiv.org/abs/2511.18290.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024.

Haotang Li, Zhenyu Qi, Shaohan Henry Wang, Kebin Peng, Zi Wang, Qing Guo, Sen He, and Huanrui Yang. PaceVGGT: Pre-alternating-attention token pruning for visual geometry transformers. *arXiv preprint arXiv:2605.08371*, 2026. URL https://arxiv.org/abs/2605.08371.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Xuanyi Liu, Chunan Yu, Deyi Ji, Qi Zhu, Lingyun Sun, Xuanfu Li, Jin Ma, Tianrun Chen, and Lanyun Zhu. StreamCacheVGGT: Streaming visual geometry transformers with robust scoring and hybrid cache compression. *arXiv preprint arXiv:2604.15237*, 2026. URL https://arxiv.org/abs/2604.15237.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Jinsoo Park, Donggyu Choi, Ahyun Seo, Minsu Cho, and Jeany Son. Diversity-aware view partitioning for scalable VGGT. *arXiv preprint arXiv:2607.01885*, 2026. URL https://arxiv.org/abs/2607.01885.

Weining Ren, Xiao Tan, and Kai Han. Speed3R: Sparse feed-forward 3D reconstruction models. *arXiv preprint arXiv:2603.08055*, 2026. URL https://arxiv.org/abs/2603.08055.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Fast visual geometry transformer. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://openreview.net/forum?id=asl8NJlIMe.

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.

Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, and Xiao-Xiao Long. LiteVGGT: Boosting vanilla VGGT via geometry-aware cached token merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36422–36432, 2026.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Zunhai Su, Weihao Ye, Hansen Feng, Keyu Fan, Jing Zhang, Dahai Yu, Zhengwu Liu, and Ngai Wong. Xstreamvggt: Extremely memory-efficient streaming vision geometry grounded transformer with kv cache compression. *Journal of the Society for Information Display*, 34(5):535–545, 2026.

Xianbing Sun, Zhikai Zhu, Zhengyu Lou, Bo Yang, Jinyang Tang, Liqing Zhang, He Wang, and Jianfu Zhang. AVGGT: Rethinking global attention for accelerating VGGT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 251–260, 2026.

Zecheng Tang, Jiaye Fu, Qiankun Gao, Haijie Li, Yanmin Wu, Jiaqi Zhang, Siwei Ma, and Jian Zhang. Spark3R: Asymmetric token reduction makes fast feed-forward 3D reconstruction. *arXiv preprint arXiv:2605.06270*, 2026. URL https://arxiv.org/abs/2605.06270.

Chung-Shien Brian Wang, Christian Schmidt, Jens Piekenbrinck, and Bastian Leibe. Block-sparse global attention for efficient multi-view geometry transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14546–14555, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21486–21499, 2026b. URL https://arxiv.org/abs/2605.15195.

Runze Wang, Yuxuan Song, Youcheng Cai, and Ligang Liu. STAC: Plug-and-play spatio-temporal aware cache compression for streaming 3D reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7567–7576, 2026c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024.

Weitian Wang, Lukas Meiner, Rai Shubham, Cecilia De La Parra, and Akash Kumar. HTTM: Head-wise temporal token merging for faster VGGT. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26379–26388, 2026d. URL https://openaccess.thecvf.com/content/CVPR2026/html/Wang_HTTM_Head-wise_Temporal_Token_Merging_for_Faster_VGGT_CVPR_2026_paper.html.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026e.

Zipeng Wang and Dan Xu. FlashVGGT: Efficient and scalable visual geometry transformers with compressed descriptor attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21826–21835, 2026.

Zhisong Xu and Takeshi Oishi. FrameVGGT: Coherence-preserving memory for bounded streaming geometry. *arXiv preprint arXiv:2603.07690*, 2026. doi: 10.48550/arXiv.2603.07690. URL https://arxiv.org/abs/2603.07690v3.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025.

Jinhao You, Shuo Lyu, Zhuohang Lyu, Tanxuan Li, Zibo Zhao, Jiaxiang Hu, Kai Tang, and Yichen Guo. RegimeVGGT: Layer-wise spatially preserving redundancy removal for visual geometry grounded transformer. *arXiv preprint arXiv:2606.18439*, 2026. doi: 10.48550/arXiv.2606.18439. URL https://arxiv.org/abs/2606.18439.

Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. InfiniteVGGT: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026. doi: 10.48550/arXiv.2601.02281. URL https://arxiv.org/abs/2601.02281.

Shuhong Zheng, Michael Oechsle, Erik Sandström, Marie-Julie Rakotosaona, Federico Tombari, and Igor Gilitschenski. Good token hunting: A hitchhiker's guide to token selection for visual geometry transformers. *arXiv preprint arXiv:2605.23892*, 2026. URL https://arxiv.org/abs/2605.23892.

Zichen Zou, Xiaosong Jia, Zuxuan Wu, and Yu-Gang Jiang. Attention itself could retrieve: RetrieveVGGT for training-free long-context streaming 3D reconstruction via query-key similarity retrieval. *arXiv preprint arXiv:2605.09644*, 2026. URL https://arxiv.org/abs/2605.09644.

A Exploratory Studies and Additional Ablations

A.1 Coverage Diagnostics

Geometric coverage. For Figure 2, we uniformly sample up to $N = 500$ candidate frames from each complete sequence; thin_geometry has $N = 395$. Uniform retains the reference frame and samples the remaining $K - 1$ frames at approximately equal intervals, with $K = \max(1, \lfloor \rho N \rfloor)$ in this diagnostic. All configurations use 384×512 reconstruction and evaluation resolution and the same full-input GT domain, with paired Sim(3) followed by ICP alignment. Recall@5 cm is the fraction of GT evaluation points within 5 cm of a predicted point. Curves average evaluation units equally; highlighted points mark 10% retention. Both qualitative examples use 500 frames for Full and 50 for Uniform, with the same view, scale, and point size within each scene.

Descriptor construction. For the readout comparison in Figure 3, Camera, Register, and Patch use the states immediately after each frame-attention block and before the corresponding inter-frame block, with zero-indexed layers $\ell \in \{0, \ldots, 23\}$. These states include the encoder computation and any preceding aggregation blocks. At 224×288 input resolution, each view has one camera token, 16 aggregator register tokens, and a 14×18 patch grid, all with feature dimension 1024. Let $\mathbf{c}_i^{(\ell)}$ and $\mathbf{r}_{i,s}^{(\ell)}$ denote the camera token and register token s of view i at layer ℓ. Using the ℓ_2 normalization operator defined in Eq. 2, we construct

$$\begin{aligned} \mathbf{z}_{i,\text{Camera}}^{(\ell)} &= \operatorname{norm}\left(\mathbf{c}_i^{(\ell)}\right), \\ \mathbf{z}_{i,\text{Register}}^{(\ell)} &= \operatorname{norm}\left(\left[\operatorname{norm}(\mathbf{r}_{i,1}^{(\ell)}); \ldots; \operatorname{norm}(\mathbf{r}_{i,16}^{(\ell)})\right]\right). \end{aligned} \tag{6}$$

Register tokens are concatenated in their fixed token order. For Patch, we apply the spatial mean pooling in Eq. 1 to the patch states at layer ℓ, then normalize each pooled vector, concatenate the vectors in spatial order, and normalize the concatenation as in Eq. 2. DINO uses the same construction on the encoder's output patch features before multi-view aggregation and remains fixed across the layer comparison. Both spatial constructions use row boundaries $(0, 3, 7, 10, 14)$ and column boundaries $(0, 4, 9, 13, 18)$, defining 16 non-overlapping bins ordered by row and then column. Camera descriptors have dimension 1024; Register, Patch, and DINO descriptors each have dimension $16 \times 1024 = 16384$. Readout tensors, spatial pooling, and normalization of individual register tokens or pooled bins use FP32; normalization of each complete descriptor and cosine-distance computation use FP64 in this diagnostic. All four descriptor types use the cosine distance in Eq. 3.

Descriptor coverage and timing. For Figure 3, we use $N = 500$ frames at 224×288 resolution from four NRGBD scenes: complete_kitchen, green_room, grey_white_room, and staircase. Uniform uses K equally spaced input positions, rounded down to integer indices. For descriptor readout ϕ, let d_ϕ denote its cosine distance as in Eq. 3, and let $V = \mathcal{J} \setminus \{r\}$. We calibrate the threshold separately for each scene and readout:

$$\tau_\phi = \operatorname{Quantile}_{0.05}\left(\{d_\phi(i, j) : i, j \in V,\ i < j\}\right), \tag{7}$$

using linear interpolation over all unordered pairs of non-reference views. The closed descriptor neighborhood is

$$\mathcal{B}_\phi(s) = \{i \in \mathcal{J} : d_\phi(i, s) \le \tau_\phi\}. \tag{8}$$

Table 6: Initialization and completion comparisons. Mean paired ΔCD (left minus right), in mm; negative values favor the left strategy. Both comparisons match $K = \lceil 0.15N \rceil$. W/T/L gives the number of units in which the left strategy has lower/equal/higher CD, after averaging random-completion results over seeds within each unit.

		NRGBD		7 Scenes	
Comparison	Budget	ΔCD	W/T/L	ΔCD	W/T/L
H + FPS $-$ Global FPS	K	$+0.012$	5/0/4	$+0.158$	7/0/11
Global FPS $-$ P + Random	K	-0.280	6/0/3	-0.063	12/0/6

For a selected set S and omitted set $U = \mathcal{J} \setminus S$, the fixed-subset diagnostic reports

$$\mathrm{Cov}_\phi(S) = \frac{1}{|U|} \sum_{i \in U} \mathbf{1}\left[\min_{s \in S \setminus \{r\}} d_\phi(i, s) \le \tau_\phi \right]. \tag{9}$$

The reference is retained in S but excluded as a diagnostic anchor; each omitted view is counted once. For Random and stratified random, we average results over three seeds within each scene, then average scenes equally. The marker identifies the earliest maximum of Register coverage. We measure timing on an H200 using one warmup pass and three measured forward passes per scene, then report the mean of the per-scene medians. The cumulative GPU time includes image encoding and preceding aggregation, excluding transfers, selection, and high-resolution reconstruction.

A.2 Selection Ablations

Initialization and completion comparisons. Table 6 uses all nine NRGBD scenes and all 18 official test sequences of 7 Scenes. Candidates are sampled uniformly over each complete valid sequence up to $N = 500$; thin_geometry has $N = 395$, and all other units have $N = 500$. For NRGBD kitchen, we exclude frames 1150–1155 before sampling. At 15% retention, the final budget is $K = 75$, or $K = 60$ for thin_geometry. Starting from the reference view, the coverage rule repeatedly adds the view that covers the most currently uncovered inputs under $\mathcal{B}_{\mathrm{DINO}}$. It stops as soon as all inputs are covered, yielding the set H. Coverage-gain ties favor the candidate that minimizes the mean nearest-selected descriptor distance after addition, followed by the smaller input index. This stopping rule covers all of $\mathcal{J}$ and includes the reference as an anchor, whereas Eq. 9 scores only omitted views using non-reference anchors. For each unit, P is the first $h = |H|$ views from the same reference-initialized global FPS trajectory; h ranges from 28 to 35 in NRGBD and from 25 to 38 in 7 Scenes. Continuing P along that trajectory to K is identical to reference-initialized global FPS to K. All FPS completions use global candidates and score distances against all currently selected views. Random completion retains P and samples the remaining $K - h$ views uniformly without replacement from the full remaining input, using 20 seeds per unit. We average seeds within each unit and then average over nine scenes or 18 sequences, respectively. All selected sets are passed to the high-resolution model in input order. Reconstruction uses the fixed-N GT domain and the median-scale plus ICP protocol in Appendix B.2. CD is the sum of the two directed mean distances and is reported in millimeters in Table 6.

Continuing FPS from P improves mean CD over random completion from the same prefix. For 7 Scenes, the improvement in completeness is partly offset by worse accuracy.

GT-sample visibility coverage. The geometric diagnostic uses original GT depth and cameras for the same N candidate frames. GT points are sampled at an image stride of 4, averaged within 1 cm voxels, and subsampled without replacement to at most 50,000 points using seed 0. A sample is visible from a selected camera when it projects inside the image and its projected depth agrees with the observed depth within 5 cm. GT-sample coverage is the fraction of these fixed samples visible from at least one selected camera. It measures the selected cameras' observations using GT geometry, while Recall@5 cm measures the reconstructed point cloud against the evaluation geometry. Although H attains full descriptor coverage, its GT-sample visibility coverage is lower than that of the matched-size FPS prefix P on every 7 Scenes sequence and on average in NRGBD. Figure 5 pairs these measurements with reconstruction CD at the same prefix size. Relative to P, H

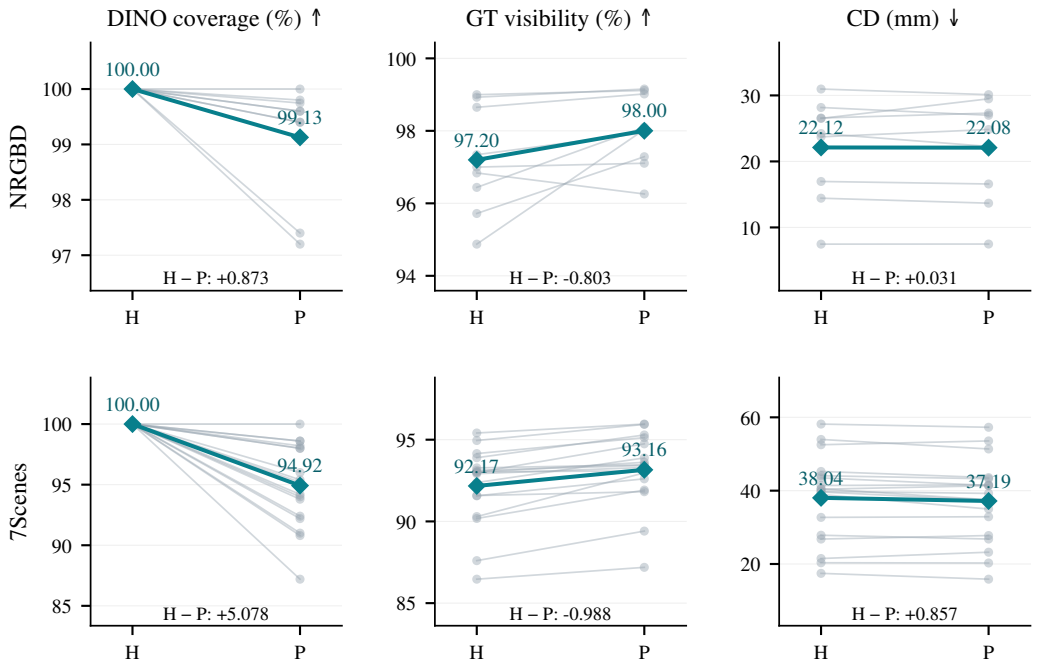

Figure 5: **Descriptor coverage, GT visibility, and reconstruction at matched prefix sizes.** H is the hard-coverage set and P is the first $|H|$ global-FPS views, including the reference in both cases. Gray lines pair individual scenes or sequences; teal diamonds show equal-unit means. Descriptor coverage scores all candidate inputs with all selected anchors, including the reference. GT visibility uses fixed GT samples and is distinct from reconstruction recall. Annotations report mean $H - P$ differences: percentage points for coverage and mm for CD. H has lower/equal/higher CD than P in 4/0/5 NRGBD scenes and 6/0/12 sequences of 7 Scenes. Vertical scales vary by panel.

increases mean descriptor coverage by 0.873 and 5.078 percentage points on NRGBD and 7 Scenes, respectively, while reducing GT visibility by 0.803 and 0.988 percentage points. Its mean CD is higher by 0.031 mm and 0.857 mm, respectively, so full coverage under the descriptor threshold does not ensure better geometric coverage or reconstruction.

Candidate-pool comparisons. Table 7 separates initialization, candidate eligibility, and distance references on both datasets. Variant A is reference-initialized global FPS; B initializes with H and completes with global FPS; C further restricts the completion candidates; D also uses local distance references and corresponds to Coverage + Local FPS in Table 5. The comparisons use the same hard-coverage prefix H and threshold as above. For every non-reference center in H, its region consists of its non-reference DINO neighbors, including the center itself. We rank centers in descending order of the number of other non-reference neighbors, breaking ties by input index, and retain the first $\lceil 0.25(h-1) \rceil$ regions. Their union, excluding already selected views, forms the fixed candidate pool. The candidate-pool comparison retains all of S as distance references and changes only the eligible candidates. The local-reference variant uses S intersected with each region, takes the largest candidate-to-nearest-representative distance across the regions, and updates all regions containing the added view. Regions remain fixed; ties favor the smaller region-center index and then the smaller candidate index. The comparisons retain 10%, 15%, 20%, or 25% of actual inputs on NRGBD and 15% on 7 Scenes, using the reconstruction protocol of Table 6. Changing initialization alone (B−A) has mixed effects across the evaluated settings. Restricting candidates (C−B) increases mean CD in every setting, with the smallest difference at 10% retention on NRGBD. Changing to local references within that pool (D−C) gives smaller, mixed differences. These comparisons support the use of unrestricted candidates; they do not establish a consistent benefit from hard-coverage initialization or local references.

Table 7: **Component-wise paired quality differences.** All entries are equal-unit mean differences in mm; negative values favor the added change. B−A changes initialization, C−B restricts candidates, and D−C changes distance references. Each pair shares the same input, retained-view budget, and full-input GT domain.

		B−A		C−B		D−C	
Dataset	Retention	ΔCD	ΔComp	ΔCD	ΔComp	ΔCD	ΔComp
NRGBD	10%	−0.418	−0.145	+0.022	−0.049	+0.143	+0.050
	15%	+0.012	−0.006	+0.844	+0.578	+0.275	+0.093
	20%	−0.147	−0.042	+1.542	+0.927	−0.013	−0.018
	25%	−0.178	−0.034	+1.440	+1.105	+0.052	+0.018
7 Scenes	15%	+0.158	+0.166	+0.974	+1.626	−0.048	−0.023

Table 8: **ScanNet-50 runtime and peak memory.** VGGT-Ω on H200 with Full input or 15% selection. Time is in seconds and memory in GiB (2^{30} bytes); lower is better. The candidate limit is an upper bound, and K gives the actual high-resolution input count or range across scenes.

			Time (s)			Peak memory (GiB)	
Limit	Method	K	Total	Selection	High-res.	Allocated	Reserved
300	Full	300	13.233	0.000	13.233	21.29	32.32
	+ Ours	45	1.015	0.335	0.680	7.81	9.39
500	Full	430–500	33.797	0.000	33.797	32.61	50.87
	+ Ours	65–75	1.927	0.565	1.362	10.14	12.75
1000	Full	430–1000	115.736	0.000	115.736	60.92	97.21
	+ Ours	65–150	4.815	1.126	3.689	15.97	21.16

Times are averaged equally across scenes; memory values are maxima across scenes. Stage statistics are summarized independently and are not summed to obtain Total.

A.3 Timing and Per-Unit Results

Runtime and memory scaling on ScanNet-50. Table 8 reports runtime and memory for VGGT-Ω with and without Global DINO FPS at 15% retention, using the same ordered candidate inputs for each scene and budget. Reported times are averaged equally across the 50 scenes. Total time is measured over the complete core pipeline; selection includes low-resolution encoding, descriptor and distance computation, FPS, and high-resolution subset assembly. We report the maximum PyTorch allocated and reserved memory peaks across all scenes, including the resident model. Peak statistics are reset before the algorithm starts and read after the predictions return to the CPU, without resetting between selection and high-resolution inference. A single VGGT-Ω model is reused for selection and reconstruction. Allocated memory measures tensor storage, whereas reserved memory also includes the allocator cache.

Stage timings at a matched budget. At 15% retention, low-resolution encoding, descriptor computation, selection, and subset assembly take 0.649 s on 7 Scenes and 0.626 s on NRGBD for Global FPS, compared with 0.709 s and 0.677 s for Coverage + Local FPS. High-resolution inference remains similar across selectors at the matched budget (1.306–1.315 s on 7 Scenes and 1.261–1.286 s on NRGBD). Uniform and Random have lower selection costs because they require no descriptor extraction. Stage statistics use the same units and seed averaging as Table 5, which reports directly measured core-pipeline times; image preparation is excluded.

Per-scene and per-sequence comparisons. Figure 6 shows all paired CD differences underlying the 15% comparisons. Global FPS improves over Coverage + Local FPS in seven of nine NRGBD scenes and 11 of 18 sequences of 7 Scenes. Against Uniform, these counts are five of nine and six of 18; against Random, they are six of nine and 12 of 18. On NRGBD, Global FPS lowers mean CD

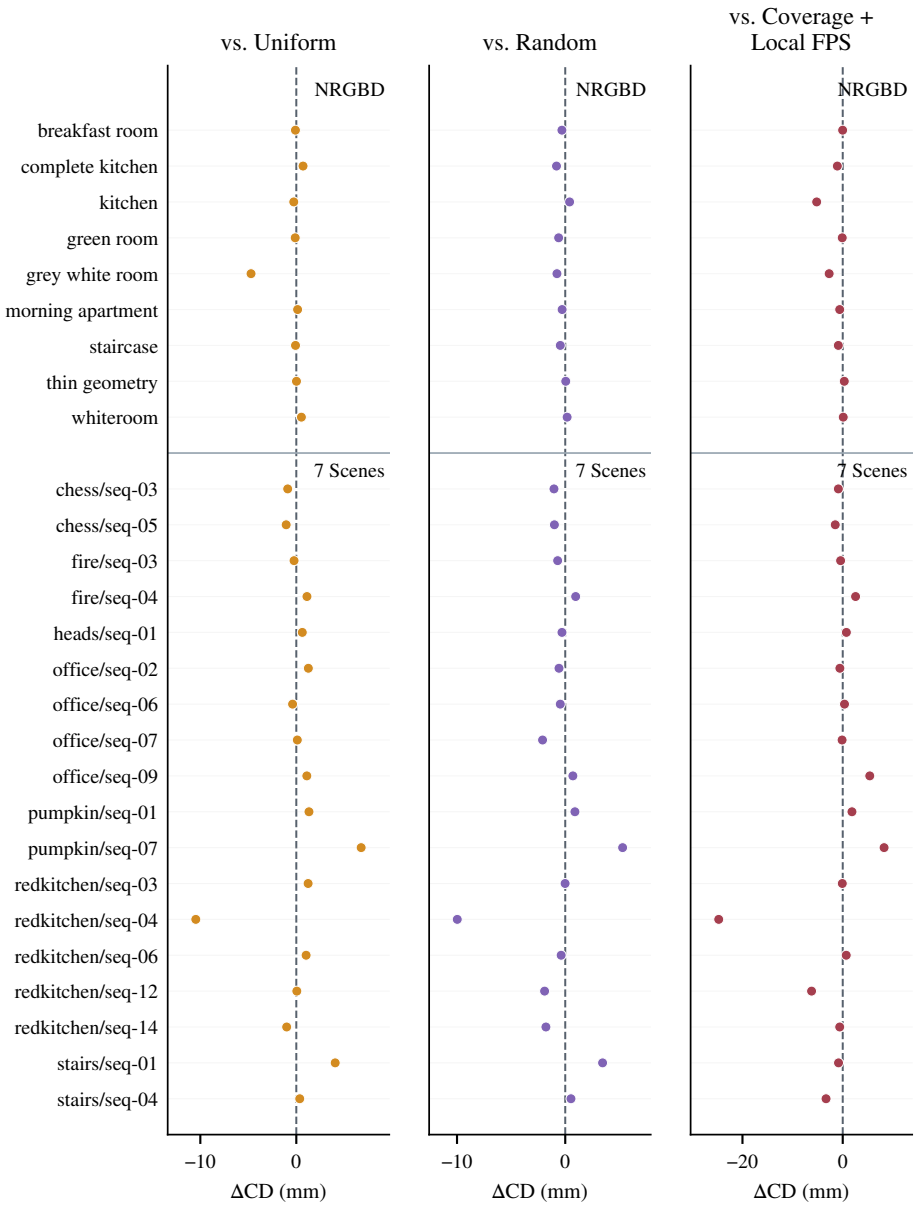

Figure 6: **Paired reconstruction differences for every evaluation unit.** Each point shows the CD of Global DINO FPS minus that of the indicated comparator at 15% retention; negative values favor Global FPS. Random results are averaged over 20 seeds per NRGBD scene and 10 seeds per 7 Scenes sequence before subtraction. All comparisons use matched candidates, retained counts, and the same full-input GT domain. The dashed line marks zero; horizontal scales vary by comparator.

by 0.291 mm relative to the 20-seed Random mean. On 7 Scenes, the difference from Coverage + Local FPS ranges from -24.738 to $+8.247$ mm, indicating substantial variation across sequences.

B Evaluation Protocols

B.1 Inputs and Sampling

RGB-D reconstruction experiments use candidate-frame limits of 500 and 1000, while ScanNet-50 uses limits of 300, 500, and 1000 to evaluate quality and computational cost at different input sizes. For 7 Scenes and NRGBD, we sample uniformly across each complete valid sequence, including both endpoints and rounding sample positions down to integer indices. ScanNet-50 follows the 50-scene split and fixed-stride sampling of FastVGGT (Shen et al., 2026). For a sequence of M valid frames and a limit B, the stride is $s = \max(1, \lfloor (M-1)/(B-1) \rfloor)$; we retain the first frame and sample every s frames from the second frame onward, adding at most $B-1$ frames. Sequences shorter than the limit use all valid frames; a valid ScanNet frame has both an RGB image and a valid GT pose.

The T&T pose experiments use all images from six complete sequences: Barn, Caterpillar, Courthouse, Ignatius, Meetingroom, and Truck. For all datasets, candidate images follow their original frame order, with the first frame serving as the reference. Full and Ours share the same candidate set. Ours retains $K = \lceil 0.15N \rceil$ frames based on the actual candidate count N, including the reference view and preserving relative input order. Reconstruction inputs follow the resolution and preprocessing of each backbone, identically for Full and + Ours. The Omega-based selector uses a separate 224×288 bicubic resize of each candidate image. Selected frame identities are used to assemble the corresponding inputs for the reconstruction backbone.

B.2 3D Reconstruction Evaluation

Reconstruction evaluation measures how well the selected views recover the complete scene. Full and Ours therefore share the same reference geometry: GT depth from all N input frames for RGB-D, and the complete scene mesh for ScanNet-50. Each method's point cloud is obtained by back-projecting the depth predictions of its reconstructed frames using its predicted camera parameters, while the evaluation domain remains fixed.

RGB-D alignment. For 7 Scenes, we use GT depth registered to RGB. For both RGB-D datasets, GT depth and intrinsics are mapped to the evaluation image plane using the corresponding preprocessing transform. We retain valid GT depths in $[0.001, 10]$ m and apply the corresponding validity mask to predicted points. For the NRGBD kitchen scene, frames 1150–1155 are excluded before candidate sampling because of anomalous GT depth.

Following the median normalization used by FastVGGT (Shen et al., 2026), we first remove scale and depth offsets in the reference-camera coordinate system, then perform rigid point-to-point ICP. Normalization uses each method's predicted frames and their corresponding GT: coordinate-wise medians define the centers, median distances to these centers define the radii, and the radius ratio determines the scale, with translation along the depth axis. ICP operates in the common N-frame GT reference system, starting from the identity transform with a 0.1 m correspondence threshold and at most 30 iterations. It stops when the absolute changes in both fitness and RMSE fall below 10^{-6}. This step adjusts rigid pose, while evaluation continues to cover the complete reference geometry.

ScanNet-50 alignment. We estimate a global Sim(3) transform from predicted camera centers and their corresponding GT centers using the Umeyama method, then apply it to the entire predicted point cloud. Full uses all N camera correspondences, whereas Ours uses the selected K correspondences. The aligned point cloud is compared directly with the vertices of the complete scene mesh, without subsequent ICP. We retain predicted points with confidence at least 1.0.

Quality metrics. Let P and G denote the aligned predicted and reference point sets. Surface accuracy and completeness are measured by the two directed mean nearest-neighbor distances:

$$\mathrm{ACC}(P, G) = \frac{1}{|P|} \sum_{p \in P} \min_{g \in G} \|p - g\|_2, \tag{10}$$

$$\mathrm{Comp}(P, G) = \frac{1}{|G|} \sum_{g \in G} \min_{p \in P} \|g - p\|_2. \tag{11}$$

RGB-D evaluation reports ACC, Comp, and normal consistency (NC). NC averages the absolute normal dot products of nearest-neighbor correspondences in each direction, then averages the two directional scores; normals are estimated after ICP using 30 neighbors. ScanNet-50 reports $\mathrm{CD} = \mathrm{ACC} + \mathrm{Comp}$. All distances are in meters.

The predicted and reference point sets are each sampled without replacement to at most 999,999 points using random seed 0, before ICP for RGB-D and after similarity alignment for ScanNet-50. Final metrics are averaged with equal weight over the 18 test sequences of 7 Scenes, the nine NRGBD scenes, and the 50 ScanNet-50 scenes, respectively.

B.3 Camera Pose Evaluation

Pose evaluation compares relative geometry between the same views. For each scene, we restrict Full's N-frame predictions to the K frames selected by Ours and evaluate both methods on the same $K(K-1)/2$ unordered camera pairs. ScanNet-50 uses the dataset's GT poses; T&T uses the official COLMAP reference poses matched by image identifier.

For world-to-camera transforms T_i and T_j, the relative pose is $T_i T_j^{-1}$. The error for each pair is the larger of the rotation angle error and the unsigned translation-direction angle error. Opposite translation directions are treated as equivalent, and a zero translation baseline is assigned a direction error of 90°. Evaluation uses the predicted relative poses directly, without the global alignment used for reconstruction.

We compute the cumulative error distribution in 1° bins and report AUC as the mean cumulative fraction within each angular threshold. The denominator includes all camera pairs, and the final bin includes the threshold endpoint. We report AUC@5, AUC@15, and AUC@30 for ScanNet-50 and AUC@30 for T&T, averaging scenes equally and expressing results as percentages.

B.4 Implementation Details

For each backbone, Full and + Ours share the same pretrained weights, numerical precision, and backbone acceleration parameters. We use the pretrained `VGGT-Omega-1B-512` checkpoint for both VGGT-Ω reconstruction and the selector's image encoder, keeping all weights frozen. VGGT-Ω high-resolution reconstruction uses 384×512 inputs (height $\times$ width), while descriptor extraction uses 224×288 inputs. For VGGT-Ω, model weights and prediction heads use FP32, while the aggregator uses BF16. The selector initializes global DINO FPS with the reference view and selects from all candidates until the budget K is filled. Selection uses the low-resolution descriptors and distance matrix computed in the current round.

FastVGGT uses the `model_tracker_fixed_e20.pt` checkpoint with BF16 autocast. Token merging starts at the first global-attention block (`merging`=0), with `merge_ratio`=0.9. SparseVGGT uses the `vggt_1B_commercial.pt` checkpoint, with `sparse_ratio`=0.5, `cdf_threshold`=0.97, average pooling (`pool_mode=avg`), and Hilbert reordering disabled (`use_hilbert=False`). Sparse attention is applied to all 24 global-attention blocks (indices 0–23). Its weights and prediction heads use FP32, the aggregator uses BF16 autocast, and the SpargeAttn kernel uses INT8 queries and keys with FP16 values and accumulation.

Both transferred backbones use their native `crop` preprocessing with a target width of 518 pixels: bicubic resizing preserves the aspect ratio up to rounding the height to a multiple of 14, followed by a vertical center crop when the height exceeds 518 pixels. We use the depth and camera heads for both models.

C Qualitative Reconstruction Results

Figures 7–14 compare Full and + Ours on eight scenes from NRGBD, 7 Scenes, and ScanNet. Each scene uses $N = 500$ candidate views and the same ordered set of $K = 75$ views selected by global DINO FPS for all three backbones. The shared GT appears once on the left; the remaining columns show VGGT-Ω, FastVGGT, and SparseVGGT, with Full above and + Ours below.

All panels within a scene share the display camera, scale, point size, and sampled point count. We randomly sample one quarter of the image pixels per frame to form the predicted display clouds, using identical pixel samples for shared frames on the same image grid. For each panel, we sample points without replacement to the same count: the minimum of 2,000,000 and the available point counts across the GT and all six predictions. The display window follows the full GT bounds with 3% padding. We reuse the saved evaluation alignments described in Appendix B.2: median normalization and ICP for RGB-D, and camera-based Sim(3) for ScanNet. RGB-D reference geometry covers all candidate frames. For ScanNet, the displayed colored reference uses the dense `vh_clean` mesh, while quality evaluation uses the `vh_clean_2` mesh vertices. Panel annotations report the original scene metrics, with ACC and Comp converted to millimeters and CD in meters; display sampling and cropping do not change these metrics.

The examples show that the quality trade-off varies across scenes and backbones. In NRGBD *white-room*, all three backbones have lower ACC with selected inputs. In 7 Scenes *office/seq-02*, all three have higher completeness error, consistent with the trade-off observed in the dataset-level results.

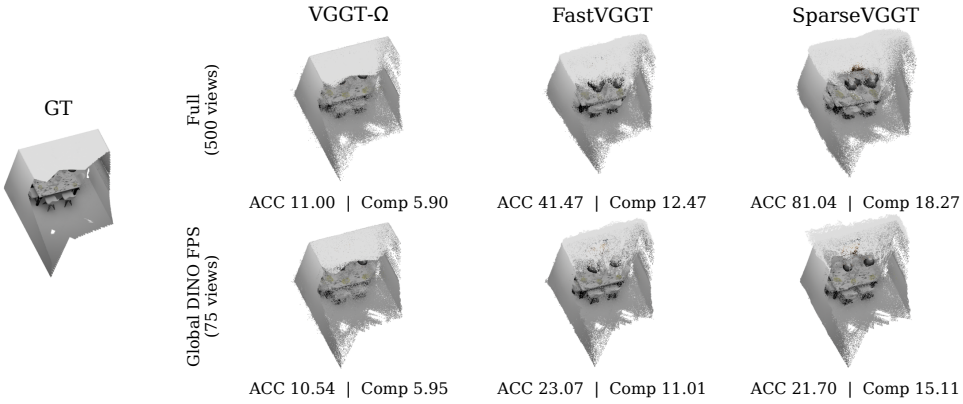

Figure 7: NRGBD *breakfast_room*. ACC and Comp are in millimeters; lower is better.

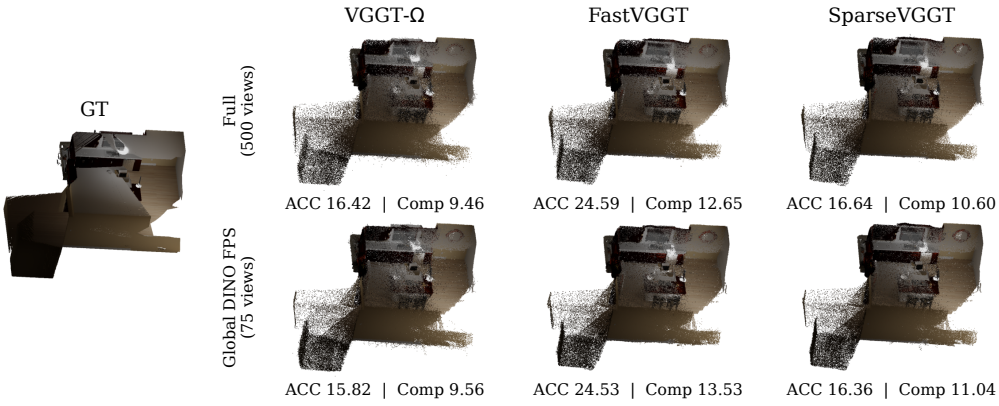

Figure 8: NRGBD *complete_kitchen*. ACC and Comp are in millimeters; lower is better.

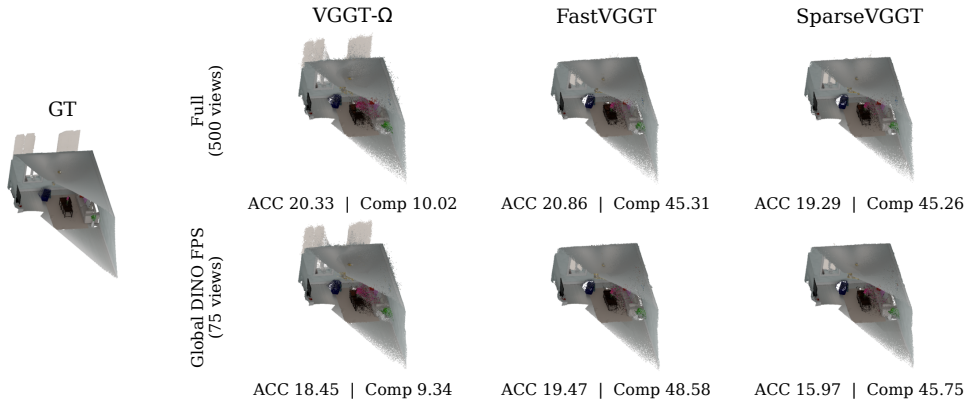

Figure 9: NRGBD *green_room*. ACC and Comp are in millimeters; lower is better.

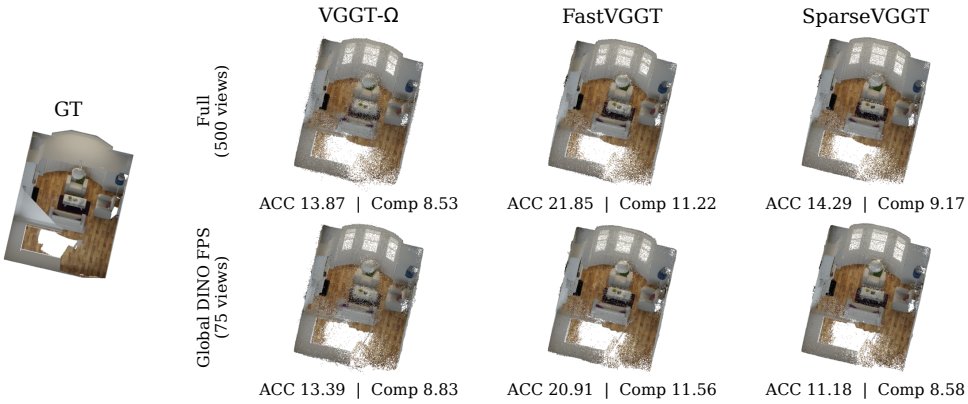

Figure 10: NRGBD *whiteroom*. ACC and Comp are in millimeters; lower is better.

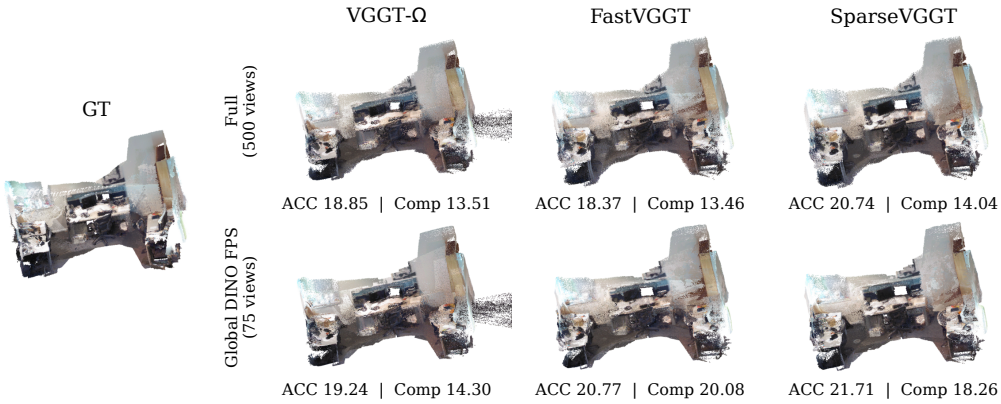

Figure 11: 7 Scenes *office/seq-02*. ACC and Comp are in millimeters; lower is better.

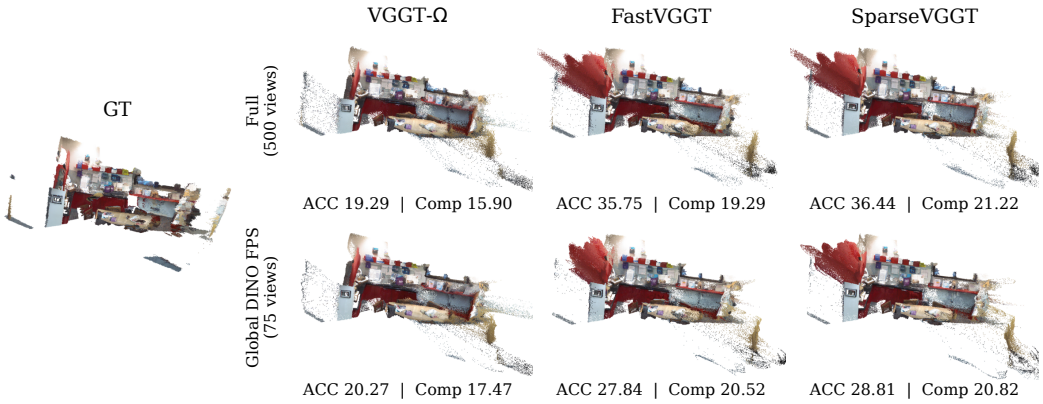

Figure 12: 7 Scenes *redkitchen/seq-03*. ACC and Comp are in millimeters; lower is better.

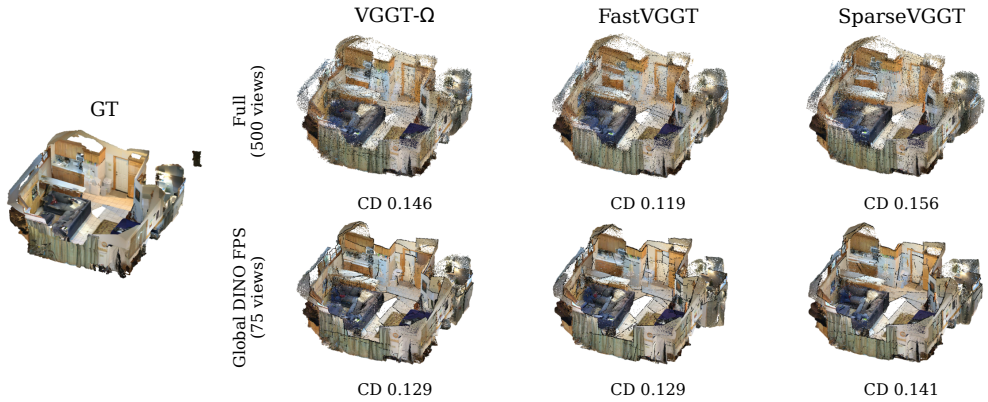

Figure 13: ScanNet *scene0000_00*. CD is in meters; lower is better.

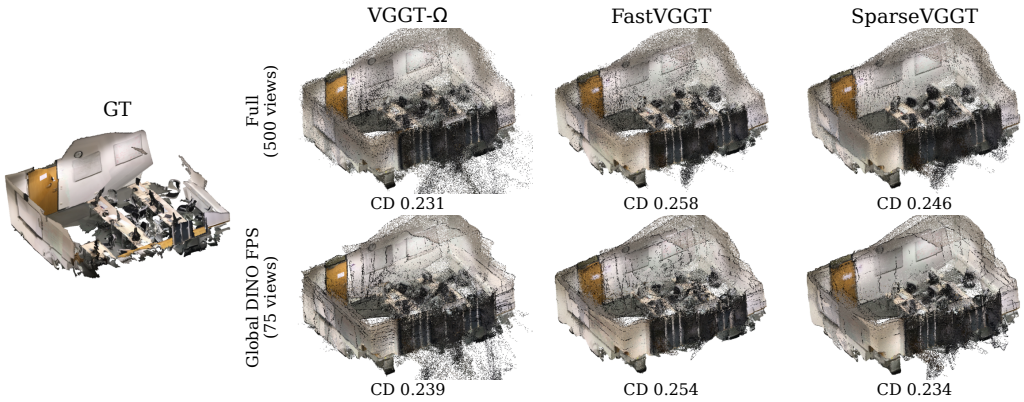

Figure 14: ScanNet *scene0056_00*. CD is in meters; lower is better.