GEORANK: GEOMETRY-AWARE LOW-RANK COMPRESSION FOR EFFICIENT 3D RECONSTRUCTION

Anonymous authors
Paper under double-blind review

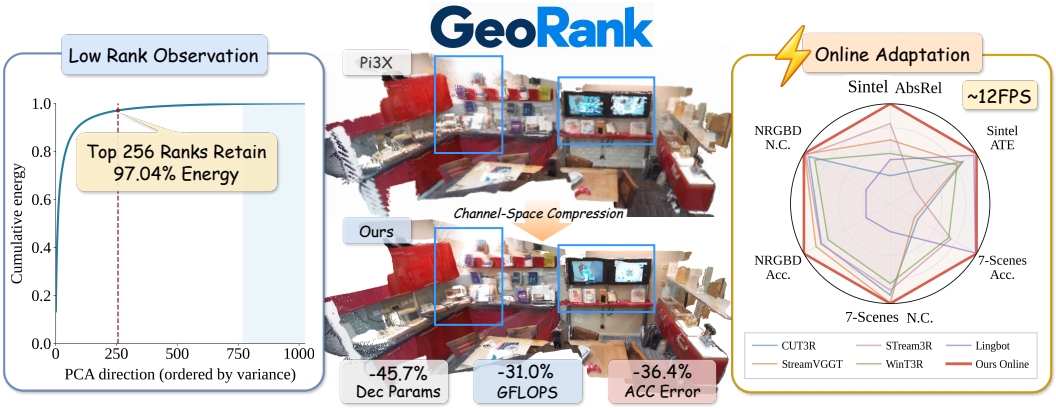

Figure 1: **Overview of GeoRank.** Geometry representations exhibit strong *channel-space* redundancy. GeoRank compresses channel computation while preserving reconstruction quality, maintaining strong online adaptation performance across 3D reconstruction benchmarks.

ABSTRACT

Feed-forward visual geometry transformers have advanced 3D reconstruction, but their large backbones incur substantial computational and memory costs. Existing efficiency methods primarily reduce redundancy in spatial tokens, we instead investigate a complementary and largely overlooked source: *channel-space redundancy*. We find substantial low-rank structure in both encoder representations and decoder operators, but its geometric importance is non-uniform across tasks and not fully captured by activation energy alone. Motivated by this, we introduce **GeoRank**, a *geometry-aware low-rank compression* framework for feed-forward 3D reconstruction. In calibration stage, GeoRank first conducts activation-aware factorization to construct low-rank operators that better preserve the model's effective feature space. Since activation energy does not fully capture geometric importance, we further redistribute a rank budget according to geometric sensitivity. During training, to handle variation beyond static compression, we introduce lightweight adaptive refinement that adjusts the compressed representation to each input and selectively restores information lost during truncation. A geometry-sensitive regularization further stabilizes geometry-critical directions. Furthermore, our factorized representation also extends naturally to streaming inference by caching historical features in compact rank space. Benchmarking results demonstrate that our method substantially reduces model cost while retaining strong quality. Compared with the full model, it reduces the decoder parameter size by 45.7% and GFLOPs by 31.0% while improving 7-Scenes point-map accuracy from 0.055 to 0.035, completion from 0.090 to 0.050. The same compact representation also extends effectively to causal reconstruction with strong online performance. Overall, GeoRank establishes channel-space low-rank compression as a complementary direction for building efficient feed-forward and streaming 3D reconstruction models.

1 INTRODUCTION

3D reconstruction is a fundamental problem in computer vision, supporting applications such as simulation, navigation, and embodied robotics (Guo et al., 2026; 2025; Jia et al., 2024). Recent

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.

feed-forward visual geometry transformers (Wang et al., 2024; 2025a; 2026; 2025b) have substantially simplified multi-view reconstruction by replacing optimization-heavy pipelines with learned feed-forward inference. However, their strong performance increasingly relies on large transformer backbones based high-dimensional features, leading to significant parameter, memory, and computational costs. To address this, existing approaches primarily identify redundancy issues in token or sequence space, reducing the number of tokens, views, or global attention interactions (Shen et al., 2025; Feng et al., 2026). We instead shift our focus from the well-studied token space to *channel space*, and ask: *do high-dimensional geometric representations truly require full-rank computation?*

Taking the state-of-the-art geometric foundation model π^3 (Wang et al., 2026) as a representative example, we perform PCA on its intermediate features and observe a strongly concentrated spectrum: a small fraction of principal directions captures most of the activation energy (Fig. 1). Projecting features onto the leading subspace largely preserves downstream geometry, whereas retaining low-energy directions causes severe degradation (Fig. 2), exhibiting substantial channel-space redundancy. More importantly, our rank-sensitivity and geometry-importance analyses show that this low-rank structure is *geometry-anisotropic*. Different geometric tasks require different number of retained rank: depth remains robust under stronger truncation, while camera pose, particularly rotation, is substantially more sensitive. Meanwhile, activation energy, although strongly correlated with geometric sensitivity, does not uniquely determine importance (Fig. 3). Thus, geometry exhibits a non-uniform low-rank structure, with rank requirements varying across tasks.

Motivated by these findings, we introduce **GeoRank**, a geometry-aware low-rank compression framework for efficient 3D reconstruction. GeoRank operates in two stages: *calibration* constructs a geometry-preserving compressed model and allocates rank where it matters most, while *training* further adapts the compressed subspace to input variation and enhances its geometric fidelity. During calibration, activation-aware factorization first constructs compact operators that preserve the model's effective feature space. Since activation energy alone cannot determine where geometric capacity is most valuable, GeoRank further performs *Geometry-Aware Rank Allocation*, redistributing a fixed rank budget according to downstream geometric sensitivity and preserving more capacity where geometry is most sensitive to compression. During *training*, *Adaptive Geometry Refinement* adapts the compressed representation to input-dependent variations and selectively recovers information lost through truncation, while *Geometry-Sensitive Regularization* stabilizes geometry-critical directions during optimization. The same factorized representation naturally extends to *online streaming reconstruction* by storing historical K/V states in compact rank space, enabling causal inference with reduced cache memory.

Experiments demonstrate that GeoRank substantially reduces model cost while preserving strong geometric prediction across reconstruction tasks. Compared with full π^3X, GeoRank reduces decoder parameters by 45.7%, GFLOPs by 31.0%. On 7-Scenes, it improves sparse-view point-map mean accuracy from 0.055 to 0.035 and completion from 0.090 to 0.050. Moreover, the online extension achieves strong online streaming reconstruction in almost 12FPS, demonstrating that the same compact channel-space representation naturally transfers to streaming inference. Overall, GeoRank establishes geometry-aware channel-space compression as a complementary efficiency dimension for both feed-forward and streaming 3D reconstruction.

2 Related Work

Feed-forward 3D reconstruction. DUSt3R (Wang et al., 2024) reformulated multi-view reconstruction as point-map regression, while MASt3R (Leroy et al., 2024) strengthened matching with learned descriptors. Fast3R (Yang et al., 2025) scaled feed-forward reconstruction to large image sets, FLARE (Zhang et al., 2025) coupled camera, geometry, and appearance, VGGT (Wang et al., 2025a) unified multiple geometry tasks in a large visual geometry transformer, and CUT3R (Wang et al., 2025b) introduced persistent state for continuous online 3D perception. π^3 and π^3X (Wang et al., 2026) removes the fixed-reference-view assumption through permutation-equivariant processing. Recent efficiency-oriented methods (Shen et al., 2025; Shu et al., 2026) reduce computational redundancy, while QuantVGGT (Feng et al., 2026) and QVGGT (Pan et al., 2026) compress VGGT through quantization. In contrast, we are the first to investigate in the *channel space* of learned geometry representations and decoder operators.

Low-rank Model Compression. Low-rank factorization compresses neural networks by replacing dense weights with compact factors (Denton et al., 2014). Since truncated SVD ignores non-

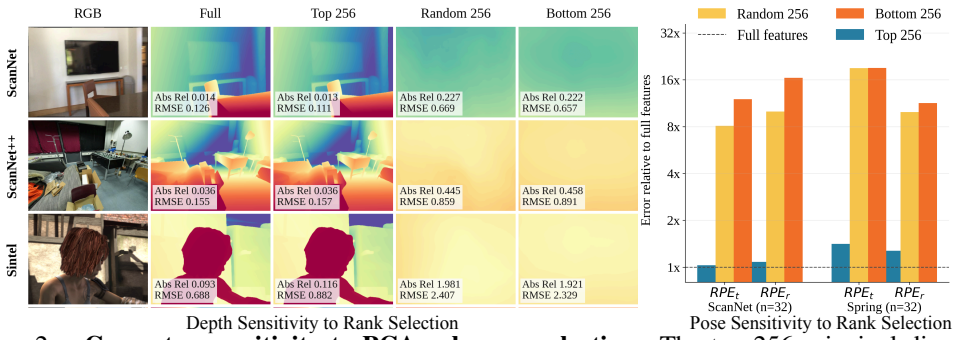

Figure 2: **Geometry sensitivity to PCA subspace selection.** The top-256 principal directions largely preserve geometry prediction, while random or bottom directions cause substantial degradation. These results reveal a compact, geometry-relevant feature subspace.

uniform activation distributions, recent studies further show stronger low-rank feature structure than weights (Yu & Wu, 2023). ASVD (Yuan et al., 2025) uses activation-aware rescaling, while SVD-LLM (Wang et al., 2025d;c) reduces activation-induced truncation error. LoSparse (Li et al., 2023) preserves residual information with sparse components, while Palu (Chang et al., 2025) applies low-rank projection to KV-cache compression. Existing methods mainly determine low-rank structure from weight reconstruction, activation statistics, or task-level sensitivity. We instead introduce geometry-aware sensitivity to identify channel directions important for preserving multi-view geometric relationships and guide rank allocation in feed-forward geometry models.

3 Observation

We first analyze channel-space redundancy in the geometry model, from compact encoder representations to low-rank decoder computation. Next, we examine whether this redundancy is uniform, and show that the required rank varies across geometric tasks and directions.

3.1 Encoder features admit a compact representation

We begin by examining the intrinsic dimensionality of encoder features. Let $x_n \in \mathbb{R}^d$ denote an encoder token and collect N tokens into $X = [x_1, \ldots, x_N]^\top \in \mathbb{R}^{N\times d}$. We center each token as $\bar{x}_n = x_n - \mu$, where $\mu = \frac{1}{N}\sum_n x_n$, and denote the resulting centered feature matrix by $\bar{X} = [\bar{x}_1, \ldots, \bar{x}_N]^\top$. We compute the empirical covariance $C_x = \frac{1}{N}\bar{X}^\top\bar{X} = U\Lambda U^\top$, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. The cumulative variance retained by the leading r principal components is $E_X(r) = \sum_{i=1}^r \lambda_i / \sum_{i=1}^d \lambda_i$. As shown in Fig. 1 (left), a small number of principal directions captures most of the variance, showing strongly concentrated feature spectrum.

We next test whether this compact structure is also relevant to geometry. We project each feature onto the leading r components and reconstruct it in the original feature space as $\hat{x}_n = \mu + U_r U_r^\top (x_n - \mu)$ with $U_r \in \mathbb{R}^{d\times r}$ as the leading r eigenvectors. We then pass the reconstructed features through the unchanged decoder and evaluate pose and depth prediction. As shown in Fig. 2, the leading subspace largely preserves geometric accuracy, whereas random or low-energy directions cause substantial degradation. These results indicate that encoder features lie in a compact, geometry-relevant subspace.

3.2 From feature redundancy to operator compression

The preceding analysis suggests that geometry features occupy a compact effective subspace. We next investigate whether this redundancy can be translated into lower-cost decoder computation. For a linear map $y = Wx$, with $W \in \mathbb{R}^{m\times d}$, applying W to the PCA-reconstructed feature yields

$$W\hat{x} = W\mu + (WU_r)\big(U_r^\top (x - \mu)\big). \tag{1}$$

Thus, if the input lies near an r-dimensional subspace, the transformation can be computed through an r-dimensional intermediate representation, directly linking feature compactness to low-rank operator computation. However, decoder layers experience different activation distributions and need

not be equally compressible. Standard truncated SVD minimizes $\min_{\mathrm{rank}(\widehat{W})\le r} \|W-\widehat{W}\|_F^2$, which weights all input directions equally. In contrast, for layer l input x_l and $\Delta W_l = W_l - \widehat{W}_l$, the expected output error is $\mathbb{E}\big[\|\Delta W_l x_l\|_2^2\big] = \mathrm{tr}\big(\Delta W_l M_l \Delta W_l^\top\big) = \|\Delta W_l M_l^{1/2}\|_F^2$, where $M_l = \mathbb{E}[x_l x_l^\top]$ is the layer-specific second-moment matrix. This shows that low-rank approximation should account for the observed activation distribution rather than weight space alone. We therefore adopt activation-aware low-rank decomposition as the bridge from compact features to compressed decoder operators, as detailed in Sec. 4.1.

3.3 Low-Rank Structure is Geometry-Anisotropic

We find that the decoder is geometry-anisotropically low rank: different geometric tasks tolerate different degrees of rank reduction, and directions with similar activation strength can have different geometric importance. To quantify this effect, we divide the ordered low-rank directions of layer l into rank bands b, with contribution ΔW_{lb}, and define their geometric sensitivity as

$$S_{lb}^{\mathrm{geo}} = \mathbb{E}_x\left[\|J_{\phi,l}\Delta W_{lb}x_l\|_2^2\right], \qquad (2)$$

where $J_{\phi,l}$ measures the response of a downstream geometry objective. As shown in Fig. 3, pose is substantially more sensitive to rank truncation than depth, indicating that different geometric tasks require different amounts of retained capacity. Moreover, activation energy is strongly correlated with geometric sensitivity (Spearman $\rho = 0.82$ for depth and 0.85 for pose), yet considerable variation remains among directions with similar activation energy. Together, these results show that activation statistics reveal the compressible subspace, but do not fully determine where rank is most valuable for geometry. This motivates using geometric sensitivity to guide rank allocation in Sec. 4.1.

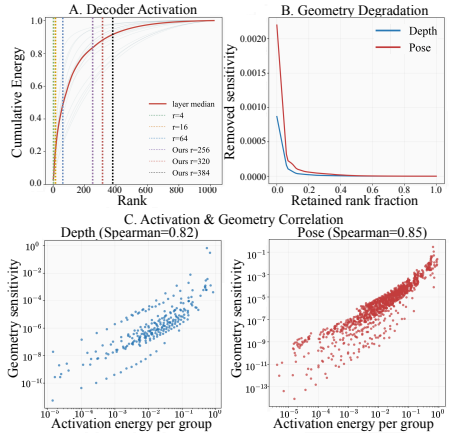

Figure 3: **Geometry-anisotropic low-rank structure.** Different geometric tasks exhibit different rank sensitivity, while activation energy correlates but does not fully determine the geometric importance.

4 GeoRank

We formulate low-rank compression as preserving geometric prediction under a fixed capacity budget. Consider a feed-forward geometry model $\mathcal{F}(\cdot;\mathcal{W})$ with decoder operators $\mathcal{W} = \{W_l\}_{l=1}^L$. We replace each W_l with a rank-constrained approximation $\widehat{W}_l$ and optimize

$$\min_{\widehat{\mathcal{W}},\,\mathbf{r}} \mathbb{E}_{(X,Y)\sim\mathcal{D}}\Big[\mathcal{L}_{\mathrm{geo}}\big(\mathcal{F}(X;\widehat{\mathcal{W}}),Y\big)\Big] \quad \text{s.t.} \quad \mathrm{rank}(\widehat{W}_l)\le r_l, \quad \sum_{l=1}^{L} C_l(r_l) \le B. \qquad (3)$$

Here, $\mathcal{D}$ denotes the reconstruction data distribution, with multi-view images X and geometric supervision Y, $\mathbf{r} = \{r_l\}_{l=1}^L$ denotes the layer-wise rank allocation, $C_l(r_l)$ the cost of retaining rank r_l, and B the total capacity budget. As illustrated in Fig. 4, GeoRank consists of two stages. During calibration, activation-aware factorization constructs compact low-rank bases, followed by geometry-aware rank allocation under a fixed compression budget (Sec. 4.1). During training, adaptive subspace refinement compensates for input-dependent information missed by static truncation (Sec. 4.2), while geometry-sensitive regularization preserves geometry-critical directions during optimization (Sec. 4.3). Finally, for streaming reconstruction, the factorized representation is extended to causal decoding with compressed rank-space KV caching and online distillation (Sec. 4.4).

4.1 Low Rank Calibration

Activation-aware Low-rank Basis. Since SVD ignores the non-uniform activation of input channels, we therefore construct the low-rank basis using activation statistics to better preserve the

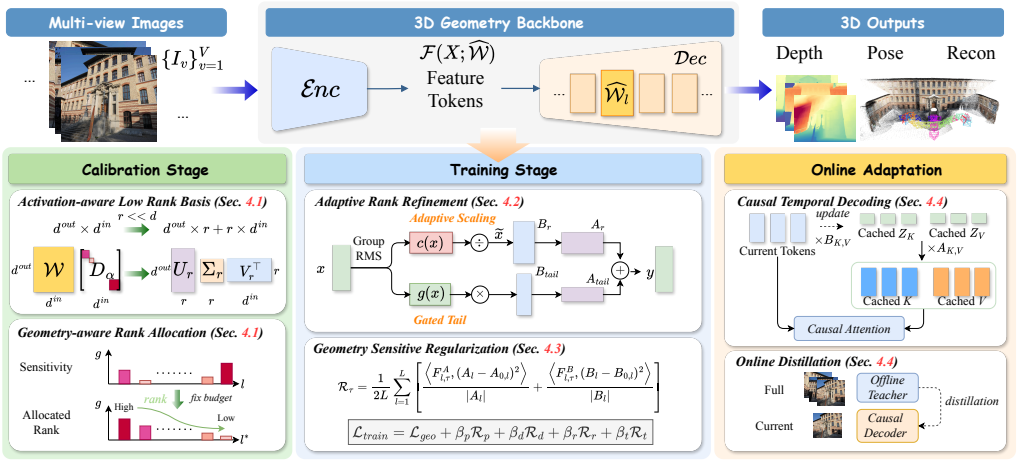

Figure 4: **Pipeline of GeoRank.** We first preprocess the pretrained model once for calibration. During training, we apply adaptive scaling and gated tail for adjusting and recovering discarded information. We further achieve streaming reconstruction with KV cache and online losses.

model's effective feature space. For a linear layer $y = Wx + b$, we collect calibration activations and compute the i-th channel RMS statistics $a_i = \sqrt{\frac{1}{N}\sum_n x_{n,i}^2}$, from which we form the diagonal activation-scaling matrix $D_\alpha = \mathrm{diag}((a_i + \epsilon)^\alpha)$. Following ASVD (Yuan et al., 2025), we decompose the activation-scaled operator and retain its leading rank-r components:

$$WD_\alpha = U\Sigma V^\top, \qquad W \approx W_r := \underbrace{U_r\Sigma_r^{1/2}}_{A_r}\underbrace{\Sigma_r^{1/2}V_r^\top D_\alpha^{-1}}_{B_r}. \tag{4}$$

The original input and output dimensions remain unchanged. Specifically, a weight matrix $W \in \mathbb{R}^{d^{out}\times d^{in}}$, with storage and computational cost proportional to $d^{out}d^{in}$, is replaced by two thin factors $A_r \in \mathbb{R}^{d^{out}\times r}$ and $B_r \in \mathbb{R}^{r\times d^{in}}$, such that $W \approx A_rB_r$. This reduces the corresponding cost from $d^{out}d^{in}$ to $r(d^{out} + d^{in})$. ASVD therefore provides an activation-supported basis and ordering of candidate directions.

Geometry-aware Rank Allocation. Activation-aware factorization identifies directions supported by the input distribution, but does not determine how many ranks each operator requires for downstream geometry. Therefore, we redistribute a fixed ASVD rank budget according to geometric sensitivity. For $W_l \in \mathbb{R}^{d_l^{\text{out}}\times d_l^{\text{in}}}$, the factor cost per retained rank is $c_l = d_l^{\text{in}} + d_l^{\text{out}}$. We preserve the reference factor-parameter budget $B_0 = \sum_{l\in\mathcal{T}} r^0_{f(l)}c_l$. Starting from the calibrated ASVD basis, we partition the ordered directions of each operator into rank bands. Let $I \subseteq \{1,\dots,r_l^{\max}\}$ denote a contiguous group of ordered ASVD directions, for task $t \in \{\text{depth}, \text{point}, \text{pose}\}$, we measure the sensitivity of rank group I in layer l by $G_{l,I}^{(t)} = \frac{\langle \nabla_{h_l}\mathcal{L}_t, \delta h_{l,I}\rangle^2}{\|\nabla_{h_l}\mathcal{L}_t\|_2^2+\epsilon}$, where $\delta h_{l,I}$ is the output perturbation induced by the corresponding ASVD directions. We estimate the importance of retaining each rank band by combining its activation energy and geometric sensitivity as $u_{\ell,I} = \eta E_{\ell,I} + (1-\eta)G_{\ell,I}$, where $\eta \in [0,1]$ controls their relative contributions and $E_{\ell,I}$ denotes the output energy of rank band I in layer ℓ. We then choose layer-wise ranks under the same total parameter budget. Dynamic programming assigns more rank to operators whose additional directions are more important for geometry. Thus, ASVD orders the candidate directions, while geometric sensitivity determines how many are retained.

4.2 Adaptive Geometry Refinement

The allocation above captures dataset-level importance using calibration statistics, but the relevant activation distribution may vary across scenes, view configurations, and image conditions. We therefore introduce two lightweight input-conditioned modules: Adaptive Scaling reweights the input channels before low-rank projection to accommodate input-dependent variations, while Gated Tail selectively restores information discarded by rank truncation.

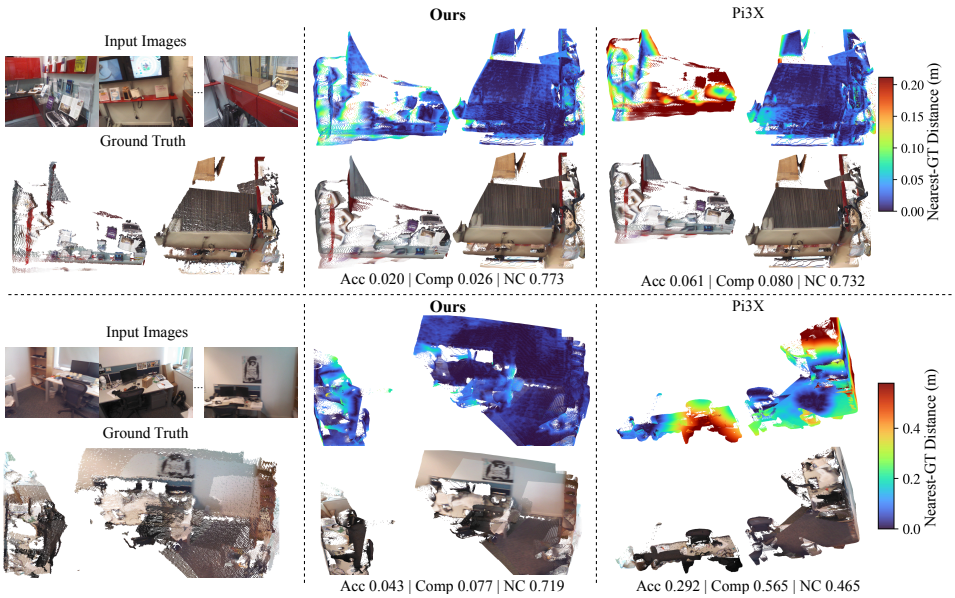

Figure 5: **Qualitative comparison with π^3X on 7-Scenes.** Given the same sparse input views, GeoRank reconstructs more complete and geometrically coherent scene structure than full π^3X.

Adaptive Scaling. The calibrated low-rank basis is shared across inputs, although their channel activation patterns may differ. To adapt the retained subspace with minimal overhead, we partition the channels of each token x into G groups and summarize group j by its RMS activation, $h_j(x) = \sqrt{\frac{1}{|\mathcal{G}_j|}\sum_{i\in\mathcal{G}_j} x_i^2}$, $j = 1, \ldots, G$. A lightweight MLP predicts bounded group-wise scaling factors $c(x) = 1 + \tanh f_\theta(h(x))$, yielding $\widetilde{x}_i = x_i / c_{g(i)}(x)$ and $\widetilde{y} = A_r(B_r\widetilde{x}) + b$. The MLP is zero-initialized so that $c(x) = 1$ initially, preserving the calibrated low-rank solution before adaptation.

Gated Tail. Adaptive scaling can recondition the retained subspace, but cannot recover information removed by rank truncation. We therefore introduce a compact residual branch $r_{\text{tail}}(x) = A_{\text{tail}}B_{\text{tail}}x$ and compute the final output as $y = A_r(B_r\widetilde{x}) + b + g_\theta(h(x))\, r_{\text{tail}}(x)$, where $A_{\text{tail}} \in \mathbb{R}^{d^{\text{out}}\times s}$ and $B_{\text{tail}} \in \mathbb{R}^{s\times d^{\text{in}}}$ form a rank-s residual projection, and $g_\theta(h(x))$ controls its input-dependent contribution. The tail output and gate are initialized near zero, so the initial model recovers the calibrated low-rank operator. During training, the tail is activated only when additional residual capacity is useful. Together, Adaptive Scaling adjusts information *within* the retained subspace, while the Gated Tail recovers complementary information *beyond* it.

4.3 Geometry-Sensitive Regularization

Fisher-Weighted Anchoring. After calibration, we jointly optimize the retained factors, adaptive scaling, and gated tail while freezing the remaining backbone. However, fine-tuning may shift the compressed subspace away from its calibrated solution, particularly along geometry-critical directions. We therefore constrain each factor according to its task sensitivity. For each geometry task $\tau \in \{\text{depth}, \text{point}, \text{rot}, \text{trans}\}$, we estimate diagonal empirical Fisher weights over K calibration batches and regularize the calibrated factors as

$$F_{\Theta_l,\tau} = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{\partial \mathcal{L}_\tau^{(k)}}{\partial \Theta_l}\right)^2, \qquad \mathcal{R}_\tau = \frac{1}{2L}\sum_{l=1}^{L}\sum_{\Theta\in\{A,B\}} \frac{\left\langle F_{\Theta_l,\tau}, (\Theta_l - \Theta_{0,l})^2\right\rangle}{|\Theta_l|}, \tag{5}$$

where $\Theta_{0,l}$ denotes the corresponding calibrated factor. Geometry-sensitive parameters are anchored more strongly to their calibrated values, while less-sensitive parameters remain freer to adapt.

Training Objective. The complete objective is

$$\mathcal{L}_{\text{train}} = \mathcal{L}_{\text{geo}} + \sum_\tau \beta_\tau \mathcal{R}_\tau, \quad \mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{point}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}}, \tag{6}$$

where $\mathcal{L}_{\text{geo}}$ is adopted from π_3 (Wang et al., 2026). Thus, rank allocation determines where capacity is retained, while geometry-sensitive regularization preserves its geometric utility during adaptation.

Table 1: **Point map estimation on 7-Scenes and NRGBD.**

Method	Params	7-Scenes						NRGBD					
		Acc. ↓		Comp. ↓		N.C. ↑		Acc. ↓		Comp. ↓		N.C. ↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
FLARE (Zhang et al., 2025)	1400M	0.085	0.057	0.145	0.107	0.696	0.780	0.053	0.024	0.051	0.025	0.877	0.988
VGGT (Wang et al., 2025a)	1191M	0.044	0.025	0.056	0.033	0.733	0.845	0.051	0.029	0.066	0.038	0.890	0.981
FastVGGT (Shen et al., 2025)	1158M	0.054	0.033	0.069	0.040	0.719	0.824	0.055	0.035	0.080	0.053	0.855	0.970
QuantVGGT (Feng et al., 2026)	1191M	0.048	0.026	0.063	0.036	0.758	0.865	0.053	0.027	0.070	0.039	0.885	0.985
DepthAnythingV3 (Lin et al., 2025)	1690M	0.048	0.031	0.069	0.051	0.757	0.865	0.031	0.019	0.032	0.016	0.923	0.996
LiteVGGT (Shu et al., 2026)	1158M	0.050	0.030	0.072	0.045	0.732	0.839	0.041	0.023	0.047	0.021	0.872	0.982
π^3X (Wang et al., 2026)	992M	0.055	0.036	0.090	0.067	0.744	0.844	0.030	0.018	0.032	0.016	0.908	0.994
Ours	786M	0.035	0.021	0.050	0.033	0.758	0.867	0.026	0.015	0.028	0.014	0.904	0.992

Table 2: **Monocular and video depth estimation comparison.** Monocular depth is evaluated on Sintel and NYU-v2, while video depth is evaluated on Bonn and KITTI.

Method	**Monocular Depth**				**Video Depth**			
	Sintel		NYU-v2		Bonn		KITTI	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
FLARE (Zhang et al., 2025)	0.606	0.402	0.089	0.898	0.152	0.790	0.356	0.570
VGGT (Wang et al., 2025a)	0.335	0.599	0.056	0.951	0.057	0.966	0.062	0.969
FastVGGT (Shen et al., 2025)	0.331	0.602	0.055	0.953	0.048	0.975	0.060	0.962
QuantVGGT (Feng et al., 2026)	0.333	0.608	0.055	0.953	0.087	0.906	0.157	0.802
DepthAnythingV3 (Lin et al., 2025)	0.363	0.610	0.086	0.904	0.046	0.974	0.053	0.977
LiteVGGT (Shu et al., 2026)	0.312	0.618	0.052	0.959	0.061	0.973	0.055	0.972
π^3X (Wang et al., 2026)	0.284	0.622	0.052	0.958	0.038	0.975	0.035	0.986
Ours	0.276	0.624	0.055	0.955	0.039	0.978	0.039	0.981

4.4 Online Adaptation to Compressed KV Caches

Rank-Space KV Caching. Following (Zhuo et al., 2026), we adapt the temporal decoder and camera head to causal inference, processing frames sequentially while reusing cached history. For each factorized temporal QKV operator at layer l, we reuse the adaptive low-rank representation from Sec. 4.2, where $z_t^l = B_l \widetilde{x}_t^l$ and $\bar{u}_t^l = g_l(h(x_t^l))B_l^{\text{tail}} x_t^l$ denote the core and gated-tail coordinates. Instead of caching full-width K/V features, we store these compact coordinates and reconstruct the required attention features as $P_{\mathcal{S}}^l = A_{l,P} z_{\mathcal{S}}^l + A_{l,P}^{\text{tail}} \bar{u}_{\mathcal{S}}^l$, with $P \in \{Q, K, V\}$. Here, $\mathcal{S} = \{t\}$ for the current query and $\mathcal{S} = \mathcal{C}_t$ for cached keys and values. This keeps historical states in compact rank space and reconstructs them only when needed by attention.

Bounded Historical Cache. For bounded-memory inference, the cache retains at most C frames. We use $C = R + M$, where R denotes recent frames and M historical anchors; setting $M = 0$ gives a recent-only cache. The same frame-retention policy is used by the temporal decoder and camera head, while rank-space compression is applied to the temporal K/V cache.

Online Distillation. We train the causal model with sequential chunks and recurrent cache updates, together with distillation from a frozen dense π^3X teacher:

$$\mathcal{L}_{\text{dist}} = \lambda_{\text{feat}} \mathcal{L}_{\text{feat}}^{\text{dist}} + \lambda_{\text{point}} \mathcal{L}_{\text{point}}^{\text{dist}} + \lambda_{\text{pose}} \mathcal{L}_{\text{pose}}^{\text{dist}}, \quad \mathcal{L}_{\text{online}} = \mathcal{L}_{\text{train}} + \lambda_{\text{D}} \mathcal{L}_{\text{dist}}. \tag{7}$$

The resulting model therefore retains the same geometry-aware low-rank representation while extending it to causal, bounded-memory reconstruction.

5 Experiments

5.1 Experimental setup

Implementation Details. All experiments are implemented using Pytorch (Paszke et al., 2019) on NVIDIA 6000 PRO GPU. We apply GeoRank to a recent geometry model π^3X (Wang et al., 2026). During calibration, the fixed reference ranks are $(384, 320, 256, 256)$ for $(\text{qkv}, \text{proj}, \text{fc1}, \text{fc2})$. We sample 512 batches of calibration set from the training dataset we use, with a sequence length from 10 to 30 per batch. During training, the default configuration uses 2–24 input views. We use AdamW (Kingma & Ba, 2014) with a peak learning rate of 1e-6 for 50 epochs. Input images follow the preprocessing and resolution of the corresponding baseline models.

Datasets. We evaluate GeoRank on feed-forward visual geometry reconstruction with depth, point, and camera-pose outputs. The current training mixture includes ScanNet (Dai et al., 2017), ScanNet++ (Yeshwanth et al., 2023), ARKitScenes (Baruch et al., 2021), BlendedMVS (Yao et al., 2020),

Table 3: **Camera pose estimation on Co3Dv2, Sintel, TUM-dynamics, and ScanNet.** RRA, RTA, and AUC are evaluated with a threshold of 30°.

Method	Co3Dv2			Sintel			TUM-dynamics			ScanNet		
	RRA ↑	RTA ↑	AUC ↑	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓
FLARE (Zhang et al., 2025)	96.38	93.76	73.99	0.207	0.090	3.015	0.026	0.013	0.475	0.064	0.023	0.971
VGGT (Wang et al., 2025a)	98.96	97.13	88.59	0.167	0.062	0.491	0.012	0.010	0.311	0.035	0.015	0.382
FastVGGT (Shen et al., 2025)	99.44	97.35	87.66	0.166	0.070	0.521	0.013	0.011	0.322	0.031	0.015	0.415
QuantVGGT (Feng et al., 2026)	99.52	97.17	88.10	0.202	0.096	0.710	0.025	0.021	0.416	0.040	0.023	0.589
DepthAnythingV3 (Lin et al., 2025)	99.42	98.38	90.51	0.138	0.057	0.369	0.012	0.009	0.306	0.033	0.014	0.406
LiteVGGT (Shu et al., 2026)	99.24	97.19	87.35	0.186	0.074	0.537	0.014	0.011	0.324	0.044	0.019	0.455
π^3X (Wang et al., 2026)	99.45	97.70	87.82	0.125	0.041	0.293	0.011	0.008	0.309	0.031	0.013	0.362
Ours	99.48	97.94	87.50	0.115	0.040	0.344	0.011	0.008	0.307	0.033	0.013	0.369

Table 4: **Cross-backbone low-rank compression and refinement.** FT refers to fine-tuning.

Backbone	Variant	FT	Model Size		Runtime Efficiency			Pose (TUM)			Vid. Depth (Sintel)	
			Dec. (M) ↓	Total (M) ↓	Images/s ↑	Peak GPU (MiB) ↓	GFLOPs ↓	ATE ↓	RPE$_t$ ↓	RPE$_r$ ↓	AbsRel ↓	RMSE ↓
VGGT	Full	×	604.73	1190.60	11.86	15386.7	32657.8	0.012	0.010	0.311	0.227	5.547
	SVD	×	233.53	819.40	13.98	14026.0	21838.2	0.014	0.011	0.329	0.288	5.867
	+ ASVD	×	233.53	819.40	13.85	14026.0	21838.2	0.014	0.012	0.327	0.265	5.795
	+ Geo. Alloc.	×	233.53	819.40	13.72	14027.8	21838.2	0.014	0.011	0.324	0.247	5.642
π^3X	Full	×	453.55	992.47	32.73	8743.1	31984.8	0.011	0.008	0.309	0.202	4.162
	SVD	×	213.82	752.73	46.93	6418.7	21043.7	0.016	0.010	0.320	0.209	4.956
	+ ASVD	×	213.82	752.73	47.08	6418.7	21043.7	0.015	0.010	0.315	0.204	4.675
	+ Geo. Alloc.	×	213.82	752.73	48.33	6418.7	21043.7	0.014	0.010	0.310	0.198	4.531
π^3X	**Ours**	✓	246.33	786.25	47.07	6509.9	22058.2	0.011	0.008	0.307	0.193	4.073

MegaDepth (Li & Snavely, 2018), MVS-Synth (Huang et al., 2018), Spring (Mehl et al., 2023), HyperSim (Roberts et al., 2021), Waymo (Sun et al., 2020), VirtualKITTI2 (Cabon et al., 2020), PointOdyssey (Zheng et al., 2023), OmniObject3D (Wu et al., 2023), and WildRGBD (Xia et al., 2024). We further highlight the table cell to show best, second best, and third best.

5.2 Benchmark Results

We validate our method on the benchmark following (Wang et al., 2026; Zhuo et al., 2026). We compare with the recent state-of-the-art reconstruction network and the efficient-oriented research.

Point-map estimation. As shown in Table 1, GeoRank achieves state-of-the-art performance across both datasets despite reducing the model size from 992M to 786M, substantially improving the full π^3X baseline. Figure 5 demonstrates our better accuracy and robustness over the full π_3X under sparse view reconstruction. These results indicate that channel compression can remove substantial redundancy without degrading reconstruction, improving geometric accuracy and completeness.

Depth estimation. Table 2 demonstrates that GeoRank preserves strong depth estimation performance on both monocular and video depth across datasets. On Sintel, our GeoRank achieves the best results among all methods. On Bonn, GeoRank remains close to the full π^3X model in Abs Rel while improving $\delta < 1.25$. These results show that GeoRank substantially reduces model capacity while largely preserving or improving depth estimation quality.

Camera pose estimation. As shown in Table 3, GeoRank remains competitive with full π^3X across diverse camera motions, achieving the best ATE and RPE-t on Sintel, tied-best RPE-t and second-best ATE on TUM-dynamics, while maintaining strong ScanNet performance. These results show that GeoRank largely preserves camera-motion geometry under channel compression, particularly for trajectory estimation.

Online Streaming Reconstruction. As shown in Table 5, GeoRank achieves strong streaming performance across depth, pose, and point-map estimation. It obtains the best Sintel and Bonn AbsRel, the lowest Sintel ATE, RPE-r, and RPE-t, and the best 7-Scenes accuracy and normal consistency. On NRGBD, it achieves the best results across all point-map metrics. Meanwhile, GeoRank runs at 11.92 FPS with 786M parameters, showing that the compact low-rank representation transfers effectively to causal reconstruction while retaining competitive efficiency.

5.3 Ablation Studies

Cross-backbone Compression Generalization. Table 4 evaluates our compression pipeline on VGGT and π^3X. Under matched training-free setting, ASVD improves over weight-only SVD, while Geometry-aware Allocation further improves most depth and pose metrics on both backbones

Table 5: **Quantitative comparison on streaming reconstruction.**

Method	Video Depth Sintel AbsRel ↓	δ_1 ↑	Depth Bonn AbsRel ↓	δ_1 ↑	Camera Pose Sintel ATE ↓	RPE-r ↓	RPE-t ↓	Point Map 7-Scenes Acc. ↓	Comp. ↓	N.C. ↑	Point Map NRGBD Acc. ↓	Comp. ↓	N.C. ↑	Efficiency Params (M) ↓	FPS ↑
CUT3R (Wang et al., 2025b)	0.421	47.9	0.063	96.2	0.213	0.621	0.066	0.126	0.154	0.727	0.099	0.076	0.837	793	11.42
StreamVGGT (Zhuo et al., 2026)	0.323	65.7	0.052	97.1	0.239	0.667	0.083	0.129	0.115	0.751	0.084	0.074	0.861	950	3.59
STream3R (Lan et al., 2026)	0.278	66.8	0.058	97.1	0.360	1.268	0.088	0.096	0.113	0.711	0.097	0.084	0.853	1191	1.73
WinT3R (Li et al., 2026)	0.360	57.0	0.061	94.9	0.231	1.024	0.090	0.098	0.192	0.690	0.117	0.150	0.802	750	12.55
TTT3R (Chen et al., 2026b)	0.427	52.3	0.057	96.8	0.207	0.700	0.078	0.101	0.134	0.678	0.098	0.074	0.818	793	7.88
LingBot-Map (Chen et al., 2026a)	0.377	58.0	0.056	95.9	0.181	1.996	0.147	0.074	0.879	0.527	0.221	1.322	0.502	1158	2.60
Ours Online	0.224	63.7	0.043	97.7	0.173	0.615	0.065	0.073	0.117	0.753	0.051	0.048	0.866	786	11.92

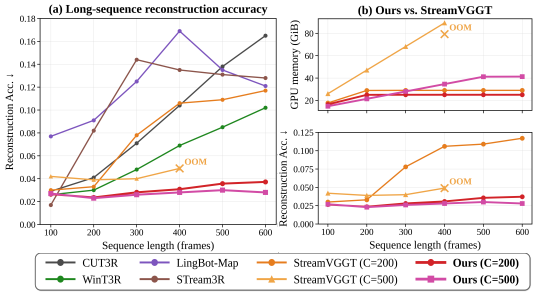

Figure 6: **Long-sequence online inference** compares accuracy from 100 to 600 frames and memory–accuracy scaling against StreamVGGT.

Table 6: **Component ablation of GeoRank.** Each component progressively improves reconstruction and camera-pose accuracy. The full model achieves the best across all metrics.

Variant	Acc ↓	Comp ↓	ATE ↓	RPE-r ↓
Full π^3X	0.0557	0.0910	0.1378	17.016
Full π^3X w/ ft.	0.0387	0.0577	0.1145	12.310
Calib. Init.	0.0421	0.0608	0.1110	11.447
+ Adaptive Scaling	0.0385	0.0554	0.1032	10.154
+ Gated Tail	0.0372	0.0528	0.0988	9.812
+ Geo. Reg. (GeoRank)	**0.0348**	**0.0496**	**0.0936**	**8.724**

without additional model or compute cost. On π^3X, fine-tuning the full GeoRank further improves over the training-free allocation in all depth and pose metrics, while retaining a compact model and high throughput. These results demonstrate that the proposed compression strategy generalizes across geometry backbones and benefits further from geometry-aware refinement.

Component ablation. Table 6 validates the effects of each component on 7-Scenes. Starting from the calibrated low-rank model, adaptive scaling improves all metrics by adapting the retained subspace to each input, while the gated tail further recovers information lost during rank truncation. Geometry-sensitive regularization provides an additional gain, yielding the best results across all metrics. Notably, the complete GeoRank also outperforms the full π^3X models, indicating that geometry-aware compression can preserve, and even improve, reconstruction quality.

Online Inference Analysis. Figure 6 evaluates online reconstruction on sequences of up to 600 frames on 7-Scenes. For cache-based methods, C denotes the maximum number of historical frames retained during causal inference. Under matched cache budgets, GeoRank consistently achieves lower reconstruction error than StreamVGGT while using less memory. Its temporal cache stores historical K/V states in rank space with cache rank $r = 128$, while the number of retained frames remains bounded by C. As sequence length grows, GeoRank maintains more stable accuracy and memory usage; with $C = 500$, it remains operational through 600 frames, whereas StreamVGGT runs out of memory beyond 400 frames. These results demonstrate that rank-space KV caching enables scalable long-sequence reconstruction with bounded memory.

6 Conclusion

We presented GeoRank, a geometry-aware low-rank compression framework for efficient feed-forward 3D reconstruction. Our analysis reveals substantial channel-space redundancy in geometry transformers, but also shows that this low-rank structure is geometry-anisotropic. Different geometric tasks require different amounts of retained rank, and activation strength alone does not fully reflect downstream geometric importance. Building on this observation, GeoRank combines geometry-aware rank allocation, adaptive subspace refinement, and geometry-sensitive regularization to preserve capacity where it matters most for geometry. The resulting compact representation substantially reduces model cost while retaining strong depth, pose, and point-map performance; on π^3X, it reduces decoder parameters by 45.7% and GFLOPs by 31.0%, while improving sparse-view reconstruction on 7-Scenes. The same low-rank representation further extends naturally and demonstrates its effectiveness for online reconstruction.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing tasks, specifically to polish language and enhance readability. The authors take full responsibility for the scientific content, research, and analysis, and the LLM was not used for generating scientific ideas or designing experiments.

REFERENCES

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Chi-Chih Chang, Wei-Cheng Lin, Chien-Yu Lin, Chong-Yan Chen, Yu-Fang Hu, Pei-Shuo Wang, Ning-Chi Huang, Luis Ceze, Mohamed Abdelfattah, and Kai-Chiang Wu. Palu: Kv-cache compression with low-rank projection. In *International Conference on Learning Representations*, volume 2025, pp. 50222–50249, 2025.

Lin-Zhuo Chen, Jian Gao, Yihang Chen, Ka Leong Cheng, Yipengjing Sun, Liangxiao Hu, Nan Xue, Xing Zhu, Yujun Shen, Yao Yao, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026a.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. In *International Conference on Learning Representations*, volume 2026, pp. 50694–50718, 2026b.

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*, pp. 5828–5839, 2017.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. Exploiting linear structure within convolutional networks for efficient evaluation. *Advances in neural information processing systems*, 27, 2014.

Weilun Feng, Haotong Qin, Mingqiang Wu, Chuanguang Yang, Yuqi Li, Xiangqi Li, Zhulin An, Libo Huang, Yulun Zhang, Michele Magno, et al. Quantized visual geometry grounded transformer. In *International Conference on Learning Representations*, volume 2026, pp. 59817–59838, 2026.

Jiaxin Guo, Wenzhen Dong, Tianyu Huang, Hao Ding, Ziyi Wang, Haomin Kuang, Qi Dou, and Yun-Hui Liu. Endo3r: Unified online reconstruction from dynamic monocular endoscopic video. *arXiv preprint arXiv:2504.03198*, 2025.

Jiaxin Guo, Tongfan Guan, Wenzhen Dong, Wenzhao Zheng, Wenting Wang, Yue Wang, Yeung Yam, and Yun-Hui Liu. Salon3r: Structure-aware long-term feedforward 3d reconstruction from unposed images. *International Journal of Computer Vision*, 134(8):383, 2026.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830. IEEE, 2018.

Mingkai Jia, Qingwen Zhang, Bowen Yang, Jin Wu, Ming Liu, and Patric Jensfelt. Beautymap: Binary-encoded adaptable ground matrix for dynamic points removal in global maps. *IEEE Robotics and Automation Letters*, 9(7):6256–6263, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yushi Lan, Yihang Luo, Fangzhou Hong, Shangchen Zhou, Honghua Chen, Zhaoyang Lyu, Bo Dai, Shuai Yang, Chen Change Loy, and Xingang Pan. Stream3r: Scalable sequential 3d reconstruction with causal transformer. In *International Conference on Learning Representations*, volume 2026, pp. 42746–42768, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024.

Yixiao Li, Yifan Yu, Qingru Zhang, Chen Liang, Pengcheng He, Weizhu Chen, and Tuo Zhao. Losparse: Structured compression of large language models based on low-rank and sparse approximation. In *International Conference on Machine Learning*, pp. 20336–20350. PMLR, 2023.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2041–2050. IEEE, 2018.

Zizun Li, Jianjun Zhou, Yifan Wang, Haoyu Guo, Wenzheng Chang, Yang Zhou, Haoyi Zhu, Junyi Chen, Chunhua Shen, and Tong He. Wint3r: Window-based streaming reconstruction with camera token pool. In *International Conference on Learning Representations*, volume 2026, pp. 59686–59702, 2026.

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

Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991. IEEE, 2023.

Zhizhen Pan, Hesong Wang, and Huan Wang. Qvggt: Post-training quantized visual geometry grounded transformer. *arXiv preprint arXiv:2605.31124*, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.

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

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451. IEEE, 2020.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

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*, pp. 20697–20709, 2024.

Xin Wang, Samiul Alam, Zhongwei Wan, Hui Shen, and Mi Zhang. Svd-llm v2: Optimizing singular value truncation for large language model compression. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4287–4296, 2025c.

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. In *International Conference on Learning Representations*, volume 2025, pp. 19299–19319, 2025d.

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*, 2026.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 803–814. IEEE, 2023.

Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22378–22389. IEEE, 2024.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Y. 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*, 2025.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1787–1796. IEEE, 2020.

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*, pp. 12–22, 2023.

Hao Yu and Jianxin Wu. Compressing transformers: Features are low-rank, but weights are not! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 11007–11015, 2023.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. Asvd: Activation-aware singular value decomposition for compressing large language models. *arXiv preprint arXiv:2312.05821*, 2025. URL https://arxiv.org/abs/2312.05821.

Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views. *arXiv preprint arXiv:2502.12138*, 2025.

Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19798–19808. IEEE, 2023.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming visual geometry transformer. In *International Conference on Learning Representations*, volume 2026, pp. 88055–88072, 2026.