LESS CONTEXT, BETTER GEOMETRY: MASKED GEOMETRIC ENCODER FOR ROBUST 3D FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent progress in 3D foundation models has enabled rapid 3D reconstruction and camera calibration by leveraging learned 3D priors from vast amount of spatial data. However, the all-to-all global attention design leads to quadratic complexity and limits long-sequence inference; unconstrained cross-view interactions also can propagate unreliable evidence from occluded or visually similar but geometrically distant views. In this paper, We introduce a *Masked Geometric Encoder* (MGE), which promotes the learning of robust geometric representations under incomplete cross-view context. During training, MGE strategically drops frame tokens from global attention and distills from a pretrained full-context teacher model. This allows the model to learn an intrinsically richer per-frame representation while providing sufficient intermediate supervision to avoid performance degradation. Through extensive experiments, we show that MGE leads to much stronger performance under occlusion and doppelganger views while retaining high performance on standard benchmarks. Such a richer frame representation also leads to more effective token reduction during inference. To this end, we develop a novel *Anchor-Guided Adaptive* token merging technique that preserves representative anchor frames while jointly merging redundant tokens from the remaining views. Compared to other efficient inference approaches, we can achieve inference speedup while consistently maintaining higher reconstruction quality, particularly in limited-view settings.

1 INTRODUCTION

Extracting 3D concepts from images is a fundamental research problem in computer vision, with various applications in autonomous driving, geo-localization, augmented and virtual reality, and robotics (Geiger et al., 2013; Arandjelovic et al., 2016; Newcombe et al., 2011; Mur-Artal et al., 2015). Recent feedforward 3D foundation models (Wang et al., 2024; 2025a; 2026b) have substantially advanced this capability. These models jointly estimate camera parameters and dense scene geometry from multiple images in one forward pass. By encoding geometric priors learned from large-scale 3D data, these models produce scene-level predictions without requiring the iterative, scene-specific optimization conventionally used in Structure-from-Motion (SfM) pipelines.

Despite impressive progress, scalability and robustness remain important challenges for current 3D foundation models. Conventional encoder design aggregates multi-view information through dense global attention. However, additional observations do not always provide reliable geometric evidence. For example, visually similar but geometrically distinct structures, often referred to as Doppelgangers (Cai et al., 2023), can induce incorrect associations across views. Furthermore, the computation grows quadratically with the total number of frames, making long sequences increasingly expensive to process.

Conventional SfM pipelines provide explicit mechanisms for dealing with these challenges. COLMAP (Schönberger & Frahm, 2016) constructs a view graph from selected image pairs and geometrically verified correspondences. As such, we can remove redundant or ambiguous matching pairs in the view graph through additional prior (Schönberger & Frahm, 2016; Arandjelovic et al., 2016; Berton & Masone, 2025; Xiangli et al., 2025), leading to speedup or improved ac-

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.

curacy in calibration. Current 3D foundation models compute all-to-all attention across all frames. Naively removing or sparsifying these global interactions can lead to substantial performance degradation (Wang et al., 2026a; Sun et al., 2026), suggesting that the pretrained representations are highly dependent on dense cross-view connectivity. By drawing an analogy between a calibration view graph and global attention, we know that not every frame-to-frame relationship is required or can even hurt in bundle adjustment. This prompts our question: **Can 3D foundation models learn a more robust representation to missing frames?**

To this end, we introduce Masked Geometric Encoder (MGE), a flexible framework for learning robust geometric representations in feed-forward 3D foundation models. MGE draws inspiration from masked representation learning (He et al., 2022) and teacher–student learning (Caron et al., 2021; Oquab et al., 2023). Instead of masking image content, we randomly remove frame-to-frame connections during global attention, restricting the available cross-view information to the geometric encoder to simulate sparse calibration view graph. Importantly, the underlying scene geometry remains unchanged: the network must still recover consistent 3D structure from the remaining cross-view interactions. Training under varying connectivity patterns therefore discourages excessive reliance on particular view pairs and encourages each frame representation to retain more geometrically informative context.

To supervise learning under incomplete context, we employ a self-distillation strategy that maintains geometric knowledge from the pretrained base model to MGE. Instead of relying on direct ground-truth supervision, which can be sparse (as shown in the Appendix A.1), we align the masked student with the teacher at two complementary levels: intermediate geometric features and predicted geometry. The feature-level objective encourages the student to preserve the teacher's geometric representation despite reduced cross-view access, while the output-level objective constrains its resulting 3D predictions including point maps and camera poses. Consequently, MGE learns to recover the same underlying geometry from different subsets of cross-view evidence, promoting representations that remain stable when individual view relations are missing or unreliable.

Finally, we connect this robust representation learning objective to efficient long-sequence inference through an Anchor-Guided Adaptive (AGA) token-merging strategy. Token merging (Bolya et al., 2022) reduces inference cost by combining similar or redundant tokens before global attention. Since MGE learns richer and more independent per-frame representations, it enables more effective merging than standard token-merging strategies (Shen et al., 2025). During inference, AGA uniformly selects a sparse set of full-resolution anchor frames and use their features to estimate the redundancy of the remaining views. Non-anchor frames are then assigned adaptive merging ratios according to their similarity to the anchor context, such that highly redundant views are compressed more aggressively while complementary views retain more tokens. Compared to standard token merging, our approach leads to significant performance improvement. The advantage is especially pronounced when neighboring views are farther apart, where uniform token reduction (Shen et al., 2025) can discard geometric information and cause severe reconstruction degradation, whereas our method better preserves full-token performance.

To summarize, our contributions are as follows:

1. We introduce MGE, which uses random frame-feature dropping within global attention to train feedforward 3D multi-view encoders under incomplete cross-view context.

2. During MGE training, we apply a self-distillation design from a pretrained base model through intermediate-feature and geometry alignment, providing supervision beyond sparse geometric ground truth that degrades performance.

3. Building on the robust representations learned by MGE, we propose AGA token merging, achieving better reconstruction quality than existing token-merging methods at a comparable computational cost, particularly in limited-view regimes.

4. We perform large scale experiments and demonstrate that MGE is significantly more robust and accurate in challenging evaluation settings such as occlusions, and doppelgangers compared to previous SoTA, while retaining top performance in conventional evaluations.

2 Related Works

Camera Calibration with View Graphs. Traditional structure-from-motion pipelines such as COLMAP (Schönberger & Frahm, 2016) estimate camera poses through pairwise feature matching (Sun et al., 2021; Edstedt et al., 2024; 2025; Shao et al., 2026), geometric verification, and bundle adjustment. These pipelines naturally form a *view graph*, where edges encode reliable geometric relations between image pairs. The quality of this graph is critical: incorrect connections caused by visually similar but geometrically unrelated views can lead to erroneous camera estimates and corrupted reconstructions. A line of work therefore focuses on identifying and removing unreliable pairwise relations before geometric optimization. Doppelgangers (Cai et al., 2023) and Doppelgangers++ (Xiangli et al., 2025), for example, detect visually ambiguous image pairs and prune spurious view-graph edges. These studies show that dense connectivity is not necessarily desirable: selectively restricting cross-view interactions can improve geometric consistency.

Feed-forward 3D Reconstruction. Recent methods replace explicit correspondence construction and geometric optimization with learned geometric prediction. DUSt3R (Wang et al., 2024) regresses dense pointmaps from uncalibrated image pairs, while MASt3R (Leroy et al., 2024) augments this formulation with learned descriptors and matching supervision. Moving beyond pairwise prediction, VGGT (Wang et al., 2025a) jointly estimates cameras, depth, pointmaps, and tracks from multiple images, whereas CUT3R (Wang et al., 2025b) maintains a recurrent scene representation for continuous geometric perception. In these models, pairwise geometric relations are no longer explicitly represented as view-graph edges, but are instead implicitly encoded through learned global attention. This shifts the robustness problem from selecting geometric constraints before optimization to learning how information should propagate across views.

Subsequent feed-forward approaches have also explored different organizations of multi-view inference. GLUEMAP (Pan et al., 2026) combines local feed-forward reconstruction with global SfM, VGGT-Long (Deng et al., 2025) processes views in overlapping subsets, and Diversity-aware VGGT (Park et al., 2026) partitions views according to visual and spatial diversity. These approaches modify which views interact or how their predictions are combined at inference time. In contrast, MGE targets the learned interaction itself, training the geometric encoder under restricted cross-view connectivity so that its representation remains robust to missing or unreliable view relations.

Long-Sequence and Efficient 3D Foundation Models. Existing approaches scale reconstruction through joint prediction, incremental memory, and compressed attention. Fast3R predicts globally aligned geometry from large image collections, while Spann3R and MUSt3R accumulate cross-view information through spatial or multi-layer memory (Yang et al., 2025; Wang & Agapito, 2025; Cabon et al., 2025). FlashVGGT and VGGT-Ω reduce cross-view computation through compact descriptor or scene tokens (Wang & Xu, 2026; Wang et al., 2026b). Other methods modify inference without retraining: STAC compresses the key–value cache for streaming reconstruction, while HTTM and FastVGGT reduce computation through token merging (Wang et al., 2026c;d; Shen et al., 2025). These methods improve scalability through architectural compression, memory management, or inference-time token reduction, with some also improving reconstruction quality. MGE complements these directions by explicitly learning geometric representations that remain reliable under partial cross-view connectivity, and AGA subsequently exploits this property to compress redundant frames adaptively during inference.

3 Methods

3.1 Preliminary: Multi-View Encoder in Feed-Forward 3D Reconstruction

Recent feed-forward 3D foundation models (Wang et al., 2025a; Lin et al., 2025; Wang et al., 2026b; Keetha et al., 2026; Wang et al., 2026e) follow an encoder–decoder formulation. Given a set of N input images $\mathcal{I} = \{I_i\}_{i=1}^{N}$, an image encoder, such as DINOv2, first maps each image into a sequence of visual tokens $\mathbf{X}_i \in \mathbb{R}^{M \times d}$, where M is the number of tokens per image and d is the feature dimension. A multi-view encoder then alternates between frame attention and global attention for L layers. Frame attention updates tokens independently within each image, while

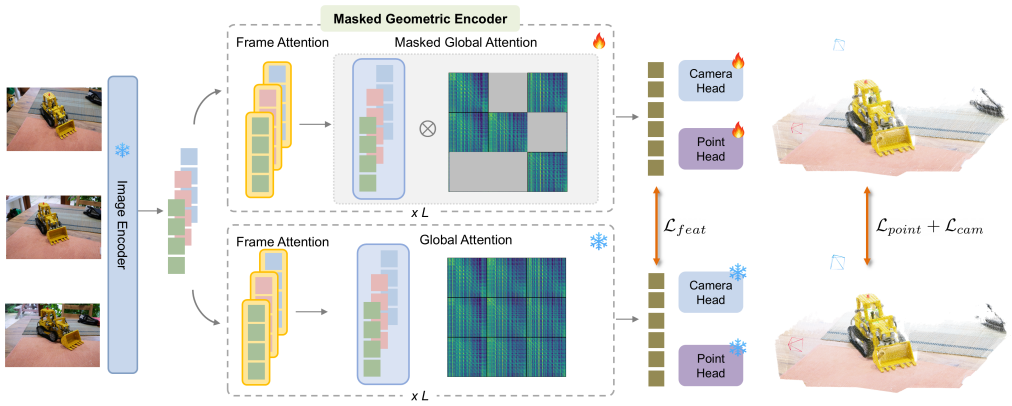

Figure 1: **Method overview.** During training, MGE masks the global attention randomly to limit the all-to-all cross-view information flow. To promote richer intermediate feature representation, we further introduce constrain MGE to produce similar 3D features compared to the full-attention teacher encoder output.

global attention jointly processes tokens from all views:

$$\mathbf{F}_i^l = \mathrm{SelfAttn}_{\mathrm{frame}}(\mathbf{G}_i^l), \quad \mathbf{G}^l = \mathrm{SelfAttn}_{\mathrm{global}}\left([\mathbf{F}_1^l; \ldots; \mathbf{F}_N^l]\right), \quad l = 1, \cdots, L, \tag{1}$$

where $\mathbf{G}_i^1 = \mathbf{X}_i$. The output $\mathbf{G}^l$ is partitioned into per-frame representations $\{\mathbf{G}_i^l\}_{i=1}^N$,and used as the input to the next frame attention layer.

By alternating frame and global attention, the multi-view encoder progressively integrates local appearance information with multi-view geometric context. The resulting per-frame representations $\mathbf{G}_i^L$ are then decoded by task-specific prediction heads to recover scene geometry and camera information, including point maps, depth maps, confidence maps, and camera poses.

3.2 Masked Geometric Encoder

Global attention provides unrestricted cross-view information flow in conventional 3D foundation models; however, prior works (Arandjelovic et al., 2016; Cai et al., 2023; Xiangli et al., 2025) in camera calibration have demonstrated that eliminating certain cross-view pairs from bundle adjustment can be beneficial. In dense acquisition, many cross-view pairs are redundant, while some pairs can be harmful due to ambiguous features. By removing these pairs from a view graph, we can obtain better efficiency and higher performance.

As shown in Fig 1, to improve robustness to unreliable cross-view information and and overall efficiency, we introduce a *Masked Geometric Encoder* that explicitly regularizes frame-to-frame attention during training. Let $\mathbf{A} \in \{0,1\}^{N \times N}$ denote a frame-level connectivity mask, where $\mathbf{A}_{ij} = 1$ indicates that view i is allowed to attend to view j, and $\mathbf{A}_{ij} = 0$ indicates that the corresponding cross-view interaction is removed. During training, we randomly generate $\mathbf{A}$ according to a predefined mask ratio ρ. Specifically, for each off-diagonal entry, we sample

$$\mathbf{A}_{ij} \sim \mathrm{Bernoulli}(1-\rho), \qquad i \neq j. \tag{2}$$

The diagonal entries are always preserved, i.e., $\mathbf{A}_{ii} = 1$, so that each frame can attend to itself. We note that $\mathbf{A}$ can be realized in forms other than a Bernoulli distribution, and provide additional experiments on other realizations in the Appendix A.8.

The frame-level mask is expanded to the token level and applied to the global-attention operation. Given query, key, and value matrices $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, the masked global attention is defined as

$$\mathrm{MaskAttn}_{\mathbf{A}}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + \mathbf{M}_{\mathbf{A}}\right)\mathbf{V}, \tag{3}$$

where $\mathbf{M}_{\mathbf{A}}$ is the token-level attention mask induced by the frame-level connectivity matrix $\mathbf{A}$. For a query token belonging to frame i and a key token belonging to frame j, the corresponding mask

value is

$$(\mathbf{M}_{\mathbf{A}})_{ij} = \begin{cases} 0, & \text{if } \mathbf{A}_{ij} = 1, \\ -\infty, & \text{if } \mathbf{A}_{ij} = 0. \end{cases} \tag{4}$$

The masking operation is applied exclusively to global attention, while frame attention remains unchanged. Specifically, we replace the standard self-attention operation for global attention in Eq. 1 with the masked attention operator defined above:

$$\mathbf{G}^{(l)} = \mathrm{MaskAttn}_{\mathbf{A}}\left([\mathbf{F}_1^{(l)}; \ldots; \mathbf{F}_N^{(l)}]\right). \tag{5}$$

Therefore, each image can still fully model its intra-frame spatial structure, while cross-frame information exchange is restricted according to the sampled connectivity pattern. During training, different frame-to-frame connections are removed across iterations, forcing the encoder to reconstruct consistent 3D geometry from varying subsets of cross-view information. MGE is a model-agnostic framework that can be integrated into feed-forward 3D architectures with cross-view attention. We use π^3 as the primary instantiation throughout the main paper and provide additional experiments on VGGT in the Appendix A.4.

3.3 Geometry and Feature Distillation

We train the masked student model by distilling geometric predictions from a frozen pretrained teacher. Given a multi-view sequence, the teacher is evaluated with dense global attention, while the student uses the masked global attention introduced above. We use the teacher predictions as supervision instead of sparse ground-truth annotations, and further validate this design choice through the ablations in Sec. 4.7 and analysis in the Appendix A.1.

Geometry distillation. Following previous works (Wang et al., 2025a; 2026e), the geometric loss consists of a scale-invariant point-map loss, a surface-normal loss, a confidence loss, and a relative camera-pose loss:

$$\mathcal{L}_{\mathrm{geo}} = \mathcal{L}_{\mathrm{points}} + \lambda_{\mathrm{normal}}\mathcal{L}_{\mathrm{normal}} + \lambda_{\mathrm{conf}}\mathcal{L}_{\mathrm{conf}} + \lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cam}}. \tag{6}$$

We describe the point-map loss below and provide the definitions of the normal, confidence, and camera-pose losses in the Appendix A.2. Since the point maps are defined up to a scene-level scale, following Wang et al. (2026e), we align the student prediction to the teacher pseudo target with a single scale factor s^* . Let $\mathbf{p}_{i,j}^S$ and $\mathbf{p}_{i,j}^T$ denote the student and teacher 3D points at pixel j of view i, respectively, and let $d_{i,j}^T$ denote the depth of the teacher point. The point-map distillation loss is then defined as

$$\mathcal{L}_{\mathrm{points}} = \frac{1}{3NHW}\sum_{i=1}^{N}\sum_{j=1}^{HW}\frac{1}{d_{i,j}^T}\left\|s^*\mathbf{p}_{i,j}^S - \mathrm{sg}\left(\mathbf{p}_{i,j}^T\right)\right\|_1, \tag{7}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient.

Feature distillation. In addition to output-level geometry supervision, we distill features before the prediction heads. Let $\mathbf{G}^{T,L}$ and $\mathbf{G}^{S,L}$ denote the final-layer features of the teacher and student. We split them into register tokens $\mathbf{R}$ and image tokens $\mathbf{Z}$,

$$\mathbf{G}^{T,L} = [\mathbf{R}^T; \mathbf{Z}^T], \qquad \mathbf{G}^{S,L} = [\mathbf{R}^S; \mathbf{Z}^S], \tag{8}$$

We apply cosine-distance losses to the two token groups:

$$\mathcal{L}_{\mathrm{feat}} = \left(1 - \cos\left(\mathbf{R}^S, \mathrm{sg}\left(\mathbf{R}^T\right)\right)\right) + \lambda_{\mathrm{img}}\left(1 - \cos\left(\mathbf{Z}^S, \mathrm{sg}\left(\mathbf{Z}^T\right)\right)\right), \tag{9}$$

where $\cos(\cdot,\cdot)$ denotes the mean cosine similarity over tokens. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{geo}} + \lambda_{\mathrm{feat}}\mathcal{L}_{\mathrm{feat}}. \tag{10}$$

3.4 Anchor-Guided Adaptive Token Merging

A key challenge for 3D foundation models is the quadratic cost of global attention, which limits scalability to long image sequences. Existing methods typically reduce this cost through token-level compression (Shen et al., 2025; Wang & Xu, 2026; Shu et al., 2026). However, token-level similarity does not necessarily imply geometric redundancy. Repetitive structures, such as windows

on a building facade, may produce similar features and thus be merged into a single token despite corresponding to different 3D locations. Aggressive token merging can therefore remove spatially distinct geometric evidence. To better preserve scene structure, we introduce Anchor-Guided Adaptive token merging (AGA) to compress at the frame level instead of token level.

Given N input frames, we uniformly sample a subset as anchor frames $\mathcal{A}$, which are kept at full token resolution and are distributed across the input sequence to provide a coarse but complete coverage. To measure redundancy, we mean-pool the DINOv2 tokens $\mathbf{X}_i$ into a frame descriptor and compute the maximum cosine similarity between each non-anchor frame and the anchor frames:

$$\mathbf{z}_i = \frac{1}{P}\sum_{p=1}^{P}\mathbf{X}_{i,p}, \qquad s_i = \max_{a\in\mathcal{A}} \frac{\mathbf{z}_i^\top \mathbf{z}_a}{\|\mathbf{z}_i\|, \|\mathbf{z}_a\|}, \quad i \notin \mathcal{A}. \tag{11}$$

A high value of s_i indicates that frame i is well represented by the anchor context and is likely to contain redundant information. Conversely, a low value of s_i suggests that the frame provides complementary content and should preserve more tokens.

Similarity-adaptive merging ratio. For each non-anchor frame $i \notin \mathcal{A}$, we map the redundancy score to a merging ratio as $r_i = ks_i + b$, where k and b are calibrated such that $\sum_{i\notin\mathcal{A}} \lfloor r_i P \rfloor = R_{\text{target}}$ under the constraint $0 \le r_i \le r_{\max}$. Anchor frames use $r_a = 0$ and remain uncompressed.

Token merging and recovery. For each non-anchor frame $i \notin \mathcal{A}$, we apply ToMe-style bipartite matching (Bolya et al., 2022) to its unprotected patch tokens and merge $\lfloor r_i P \rfloor$ tokens, obtaining a compressed representation $\widetilde{\mathbf{F}}_i = \text{Merge}(\mathbf{F}_i, r_i)$. Global attention is then computed over full-resolution anchor tokens and compressed non-anchor tokens during inference:

$$\widetilde{\mathbf{G}} = \text{SelfAttn}_{\text{global}}\left([\{\mathbf{F}_a\}_{a\in\mathcal{A}}; \{\widetilde{\mathbf{F}}_i\}_{i\notin\mathcal{A}}]\right). \tag{12}$$

The merged non-anchor outputs are finally restored to the original token layout using the recorded merge assignments, i.e., $\mathbf{G}_i = \text{Unmerge}(\widetilde{\mathbf{G}}_i)$, and are passed to the subsequent layers and prediction heads without architectural modification.

4 Experiments

4.1 Training Datasets

We distill our model on a large-scale multi-domain mixture following the data protocol of prior feed-forward 3D foundation models (Zhuo et al., 2026). The training set consists of 13 datasets, including ARKitScenes (Baruch et al., 2021), BlendedMVS (Yao et al., 2020), Co3Dv2 (Reizenstein et al., 2021), HyperSim (Roberts et al., 2021), MegaDepth (Li & Snavely, 2018), MVS-Synth (Huang et al., 2018), OmniObject3D (Wu et al., 2023), PointOdyssey (Zheng et al., 2023), ScanNet (Dai et al., 2017), Spring (Mehl et al., 2023), Virtual KITTI (Cabon et al., 2020), Waymo (Sun et al., 2020), and WildRGB (Xia et al., 2024). This mixture covers indoor and outdoor scenes, object-centric and scene-level captures, synthetic and real-world data, as well as both static and dynamic sequences. Such diversity is important for distilling robust geometric priors under masked cross-view connectivity. Details of model training can be found in Appendix A.3.

4.2 Evaluation datasets and metrics

We evaluate MGE on four aspects: robustness to visual ambiguity (Sec. 4.3), robustness to occlusion (Sec. 4.4), token merging performance (Sec. 4.5), and standard reconstruction performance (Sec. 4.6). Qualitative point-map visualizations are provided in Figure 2, Figure 6 and Figure 9. We further conduct ablation studies to analyze the contribution of each component (Sec. 4.7).

For 3D reconstruction task, predicted point maps are aligned to the ground truth using a similarity transformation followed by ICP. We report accuracy (Acc.), completeness (Comp.), chamfer distance (CD) and normal consistency (NC), where lower Acc./Comp./CD and higher NC are better. For camera trajectories, we report Sim(3)-aligned absolute trajectory error (ATE), relative translation error (RPE-t), and relative rotation error (RPE-r). For relative camera poses estimation, we report pose AUC@30, RRA@15 and RTA@15. Best and second-best results are shown in **bold** and underline.

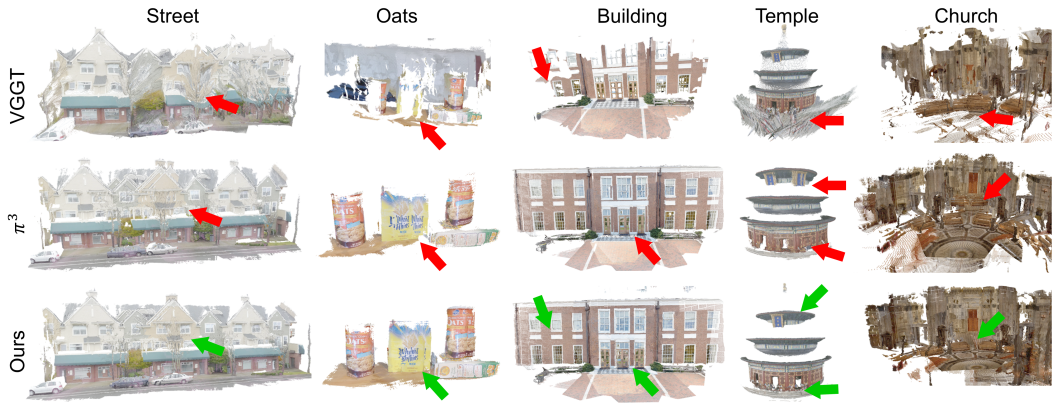

Figure 2: **Qualitative comparison of multi-view 3D reconstruction.** Compared with VGGT and π^3, our method produces more consistent reconstructions in standard scenes (Street, Oats) and in regions with repeated or visually similar structures, including building doors, temple signboards, and church benches, where the baselines often produce shifted or duplicated geometry.

4.3 Robustness to Visual Ambiguity

Visual ambiguities, i.e., regions with similar appearance but different geometric locations, often disrupt camera calibration. Since no official multi-view Doppelganger benchmark exists, we build a camera-pose benchmark under visual ambiguities from Heinly et al. (2014), using COLMAP reconstructions with Doppelgangers++ (Xiangli et al., 2025) as pseudo ground truth. The benchmark contains 43 subsets, each consisting of 15 geometrically consistent reference views and 5 visually similar but geometrically inconsistent "Doppelganger" views. Each feed-forward model processes all 20 images jointly in a single forward pass. We also include LoFTR (Sun et al., 2021) as a traditional matching-based baseline, which estimates correspondences independently for each image pair and then performs geometric verification, triangulation, and global pose optimization.

As shown in Table 1, LoFTR achieves the highest AUC@30 on the reference pairs. Since we use sequential pairwise matching, unrelated Doppelganger views have limited impact on the reconstruction of the reference images. However, some images fail to register, and the performance on reference–doppelganger cross pairs is much lower. Among the feed-forward methods, MGE achieves the best overall performance, with the strongest results on all-pair, reference-pair, and cross-pair evaluations. Qualitative camera-pose visualizations are provided in Appendix Figure 7.

Table 1: **Visual ambiguity evaluation.** We report relative-pose accuracy over all image pairs, reference-only pairs, and reference–doppelganger cross pairs. Higher is better.

Method	All			Reference			Cross		
	AUC@30	RRA@15	RTA@15	AUC@30	RRA@15	RTA@15	AUC@30	RRA@15	RTA@15
LoFTR	**59.60**	60.34	64.68	**88.22**	89.13	**91.30**	16.86	16.90	24.78
VGGT	55.56	<u>76.40</u>	<u>65.95</u>	65.75	<u>93.47</u>	75.90	<u>42.21</u>	<u>50.95</u>	<u>53.95</u>
π^3	51.23	70.73	62.48	59.42	85.40	69.50	39.10	47.41	52.34
MGE	<u>58.68</u>	**77.52**	**70.07**	<u>70.51</u>	**94.99**	<u>81.48</u>	**42.72**	**51.60**	**54.60**

4.4 Robustness to Occlusion

We evaluate occlusion robustness on ETH3D (Schops et al., 2017), progressively corrupting 10%/30%/50% views using a centered white patch to formulate small/medium/large occlusion scenarios. We evaluate dense reconstruction and camera pose under increasing levels of occlusion, considering both full-attention and token-merging methods. Table 2 shows that our method performs best under medium and large occlusion. As occlusion increases from clean to large, VGGT and π^3 suffer 119% and 128% increases in accuracy error, respectively. Their fast variants show smaller relative degradation, but start from substantially worse reconstruction quality because of token merging and therefore remain less accurate overall. In contrast, our masked cross-view training improves representation robustness: the point accuracy error increases by only 48%, while pose estimation even improves by 4% under heavy occlusion. AGA further benefits from this robustness by

Table 2: **Occlusion evaluation on ETH3D.**

Condition	Clean			Small			Medium			Large		
Methods	Acc. ↓	Comp. ↓	NC. ↑	Acc. ↓	Comp. ↓	NC. ↑	Acc. ↓	Comp. ↓	NC. ↑	Acc. ↓	Comp. ↓	NC. ↑
VGGT	0.149	0.124	0.849	0.178	0.133	0.841	0.268	0.168	0.818	0.327	0.224	0.799
π^3	0.122	**0.092**	**0.872**	**0.114**	**0.092**	**0.871**	0.181	0.144	0.855	0.278	0.195	0.823
MGE	**0.116**	0.101	0.871	0.121	0.096	0.863	**0.142**	**0.107**	**0.860**	**0.171**	**0.129**	**0.847**
FastVGGT	0.313	0.351	0.800	0.321	0.281	0.791	0.383	0.359	0.783	0.423	0.314	0.767
Fast-π^3	0.163	0.141	0.837	0.184	0.156	0.824	0.188	0.155	**0.829**	0.225	0.172	0.816
MGE+AGA	**0.156**	**0.138**	**0.838**	**0.154**	**0.126**	**0.832**	**0.175**	**0.136**	0.823	**0.182**	**0.143**	**0.819**
Methods	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓
VGGT	0.716	0.887	2.122	0.738	0.941	2.155	0.758	0.961	2.948	0.898	1.227	4.673
π^3	0.248	0.334	2.047	**0.155**	**0.230**	1.679	0.264	0.355	4.502	0.361	0.458	7.969
MGE	**0.214**	**0.319**	**1.750**	0.159	0.250	**1.614**	**0.199**	**0.286**	**2.140**	**0.205**	**0.293**	**2.043**
FastVGGT	1.178	1.529	8.490	1.147	1.486	8.085	1.213	1.552	10.537	1.290	1.659	10.699
Fast-π^3	0.499	0.668	3.140	0.542	0.711	3.170	0.587	0.765	3.894	0.556	0.774	5.034
MGE+AGA	**0.448**	**0.602**	**2.379**	**0.432**	**0.561**	**2.526**	**0.456**	**0.631**	**2.819**	**0.427**	**0.589**	**2.798**

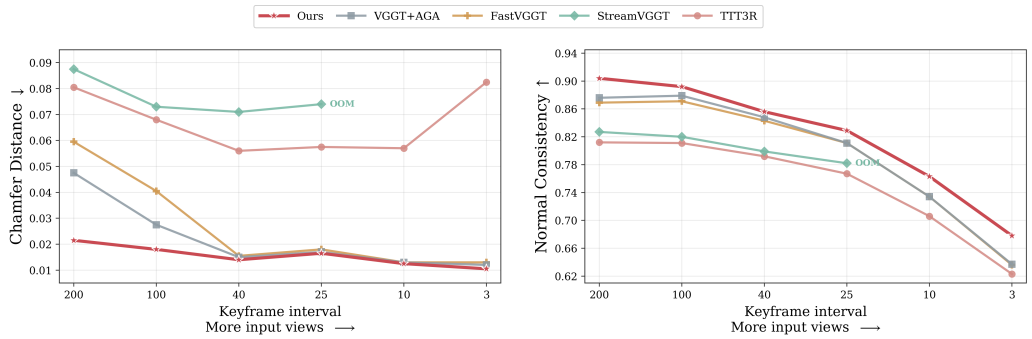

Figure 3: **Comparison of 3D Reconstruction on NRGBD.** Following FastVGGT (Shen et al., 2025), we uniformly sample keyframes with strides from 200 to 3.

retaining more tokens from less-occluded views, leading to a smaller performance drop as occlusion becomes more severe.

4.5 Token Merging Performance

Following FastVGGT (Shen et al., 2025), we evaluate 3D reconstruction on NRGBD (Azinović et al., 2022) by measuring the distance between predicted point maps and the ground-truth point cloud. Figure 3 shows that our method consistently outperforms streaming approaches, including TTT3R (Chen et al., 2026) and StreamVGGT (Zhuo et al., 2026), as well as existing token-merging method FastVGGT Shen et al. (2025) across different sequence lengths. Applying AGA to the original VGGT already yields consistent gains over FastVGGT, while combining AGA with MGE further improves robustness, particularly in limited-view settings where aggressive token merging tends to discard important geometric evidence. Detailed quantitative results are provided in the Appendix A.6.

4.6 Standard Benchmarks

We evaluate camera pose estimation on Sintel (Bozic et al., 2021), TUM-Dynamics (Sturm et al., 2012), and ScanNet (Dai et al., 2017), following the standard pose error metric used in prior work (Zhang et al., 2025a; Wang et al., 2025b). As shown in Table 3, the masked student remains highly competitive with the pretrained teacher across all datasets, matching its performance on TUM-Dynamics and slightly improving results on Sintel and ScanNet. The occasional gains further suggest that the proposed MGE can act as an effective regularizer rather than merely approximating the dense teacher. Further quantitative evaluations on additional benchmarks are provided in the Appendix A.7.

4.7 Ablation Study

Mask Designs. We ablate the masking ratio on NRGBD in Table 4.The results reveal a trade-off between representation independence and reconstruction fidelity. Stronger masking encourages each frame to rely less on dense cross-view communication, but excessive sparsification removes useful

Table 3: **Camera pose estimation over standard benchmarks.**

Method	Sintel			TUM-dynamics			ScanNet (seen)		
	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓
Fast3R (Yang et al., 2025)	0.371	0.298	13.75	0.090	0.101	1.425	0.155	0.123	3.491
CUT3R (Wang et al., 2025b)	0.217	0.070	0.636	0.047	0.015	0.451	0.094	0.022	0.629
FLARE (Zhang et al., 2025b)	0.207	0.090	3.015	0.026	0.013	0.475	0.064	0.023	0.971
VGGT (Wang et al., 2025a)	0.167	0.062	0.491	0.012	0.010	0.311	0.035	0.015	0.382
π^3 (Wang et al., 2026e)	0.074	0.040	0.282	0.014	0.009	0.312	0.031	0.013	**0.347**
MGE	**0.071**	**0.033**	**0.234**	0.014	0.009	**0.306**	**0.029**	**0.012**	0.348

Table 4: **Effect of mask ratio ρ.** We compare different masking ratios under two keyframe sampling strides.

Setting	stride=100		stride=10	
	CD ↓	NC ↑	CD ↓	NC ↑
$\rho = 0.3$	**0.0150**	**0.9184**	0.0134	0.7607
$\rho = 0.5$	0.0159	0.9104	**0.0126**	**0.7660**
$\rho = 0.7$	0.0182	0.8983	0.0156	0.7440

Table 5: **Ablation of merging designs on ETH3D.** All variants retain the same overall token ratio for a fair comparison.

Method	Acc. ↓	Comp. ↓	NC ↑
MGE+FastVGGT-style merging	0.308	0.311	0.815
MGE+Replace with anchor frames	0.230	0.256	0.838
VGGT+AGA	0.305	0.311	0.821
MGE+AGA	**0.216**	**0.243**	**0.842**

geometric context and degrades reconstruction even when all tokens are retained at inference. Conversely, weaker masking preserves reconstruction quality but reduces the effectiveness of subsequent token merging on long sequences. In practice, moderate masking provides the best balance, with $\rho = 0.3$–0.5. We further compare different masking patterns during training in the Appendix A.8.

Merging Designs. We evaluate different merging designs on ETH3D (Schops et al., 2017), which contains many low-texture and repetitive regions. Starting from the FastVGGT-style baseline, we replace uniformly protected tokens with uniformly sampled anchor frames, and then further introduce adaptive frame-wise merging. All variants retain the same overall token ratio for a fair comparison. As shown in Table 5, anchor-frame protection substantially improves reconstruction quality over token-level protection, while adaptive merging provides additional gains. Comparing AGA to both VGGT and MGE weights, the marginal gains show that the representations learned by MGE are more robust to aggressive frame-level compression.

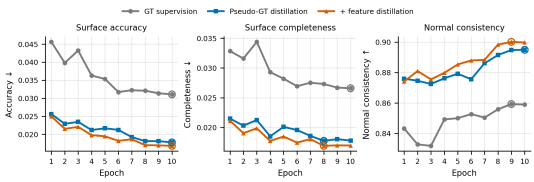

Figure 4: Ablation of various training strategies on NGRBD

Distillation Strategies. To assess the effectiveness of our training strategy, we compare direct ground-truth supervision, output-level pseudo supervision, and additional feature supervision on NRGBD. All variants are trained with the same optimization budget, and intermediate checkpoints are evaluated to compare convergence. Figure 4 shows that direct supervision converges more slowly and yields consistently higher reconstruction error. Pseudo supervision improves both convergence and reconstruction quality, while feature supervision further narrows the gap to the teacher.

5 Conclusion

In this paper, we presented a masked geometric encoder for improving the robustness and efficiency of feed-forward 3D reconstruction. By randomly masking frame-to-frame connections during fine-tuning, the model is encouraged to recover consistent geometry from incomplete cross-view communication and to rely less on individual, potentially misleading interactions. A teacher-guided distillation strategy preserves the geometric capability of the pretrained model while enabling more efficient optimization of the masked student. At inference time, we further exploit the resulting frame-level redundancy through anchor-guided adaptive token merging, preserving complete anchor views while allocating token budgets according to cross-view redundancy.

Extensive experiments show that MGE preserves dense-teacher reconstruction quality while improving robustness to occlusion and visually ambiguous Doppelganger views. AGA further improves long-sequence efficiency, especially in low-texture and repetitive scenes. Together, these results demonstrate that regularizing cross-view connectivity is an effective way to improve the robustness and scalability of feed-forward 3D foundation models.

AI USE STATEMENT

In this work, we used generative AI tools to improve the language and readability of this manuscript. All scientific contributions, including the core ideas, methodology, experimental design, and interpretation of results, are the original work of the authors. All authors have thoroughly reviewed the final version and take full responsibility for its content and claims.

REFERENCES

Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5297–5307, 2016.

Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6280–6291. IEEE, 2022.

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.

Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2852–2858. IEEE, 2025.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Niessner. Transformerfusion: Monocular rgb scene reconstruction using transformers. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1403–1414. Curran Associates, Inc., 2021.

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

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. Must3r: Multi-view network for stereo 3d reconstruction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1050–1060. IEEE, 2025.

Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 34–44, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

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, 2026.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

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.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. Roma v2: Harder better faster denser feature matching. *arXiv preprint arXiv:2511.15706*, 2025.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Jared Heinly, Enrique Dunn, and Jan-Michael Frahm. Correcting for duplicate scene structure in sparse 3d reconstruction. In *European conference on computer vision*, pp. 780–795. Springer, 2014.

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.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction; map-anything. github. io. In *2026 International Conference on 3D Vision (3DV)*, pp. 499–509. IEEE, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pp. 71–91. Springer, 2024.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

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

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.

Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE international symposium on mixed and augmented reality*, pp. 127–136. Ieee, 2011.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7855–7862. IEEE, 2019.

Linfei Pan, Johannes Lutz Schönberger, and Marc Pollefeys. Global structure-from-motion meets feedforward reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jinsoo Park, Donggyu Choi, Ahyun Seo, Minsu Cho, and Jeany Son. Diversity-aware view partitioning for scalable vggt. In *European Conference on Computer Vision*, pp. 547–564. Springer, 2026.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891. IEEE, 2021.

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.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3260–3269, 2017.

Zhimin Shao, Abhay Yadav, Rama Chellappa, and Cheng Peng. Spider: Spatial image correspondence estimator for robust calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 253–263, 2026.

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

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*, pp. 2930–2937, 2013.

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.

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

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

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.

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*, pp. 251–260, 2026.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5283–5293. IEEE, 2025.

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*, pp. 14546–14555, 2026a.

Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. In *2025 International Conference on 3D Vision (3DV)*, pp. 78–89. IEEE, 2025.

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

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

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522. IEEE, 2025b.

Runze Wang, Yuxuan Song, Youcheng Cai, and Ligang Liu. Stac: Plug-and-play spatio-temporal aware cache compression for streaming 3d reconstruction. *arXiv preprint arXiv:2603.20284*, 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, June 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*, pp. 26379–26388, 2026d.

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*, volume 2026, pp. 10481–10497, 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*, pp. 21826–21835, 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.

Yuqi Wu, Wenzhao Zheng, Jie Zhou, and Jiwen Lu. Point3r: Streaming 3d reconstruction with explicit spatial pointer memory. *Advances in Neural Information Processing Systems*, 38:69675–69699, 2026.

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.

Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27166–27175. IEEE, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935. IEEE, 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.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Forrester Cole, Deqing Sun, Ming-Hsuan Yang, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, volume 2025, pp. 82863–82886, 2025a.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21936–21947. IEEE, 2025b.

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.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8535–8546. IEEE, 2025.

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.

A Appendix

A.1 Dense Pseudo-Ground-Truth Supervision

Fig. 5 qualitatively compares the input RGB images (left), dense pseudo-ground truth predicted by the pretrained teacher (middle), and the available ground-truth depth maps (right) from MegaDepth (Li & Snavely, 2018). For visualization, the pseudo-depth is aligned to the ground-truth scale independently for each view, and both depth maps use the same color range. The MegaDepth annotations contain missing and invalid regions, shown in black, whereas the pretrained 3D feed-forward model produces spatially dense depth estimates over the complete image. Despite being dense, the pseudo-ground truth preserves the principal scene structure, object boundaries, and relative depth ordering observed in the available annotations, which encourages us to use it as supervision.

A.2 Training loss

We train the masked student by distilling the geometric predictions of a pretrained π^3 teacher. Given a multi-view sequence $\mathcal{I} = \{I_i\}_{i=1}^N$, the frozen teacher is evaluated with dense global attention, while the student uses the masked global attention. Following the π^3 output formulation, the prediction heads decode the final per-frame representations into camera poses, local point maps, and confidence maps:

$$\left(\mathbf{T}_i^T, \mathbf{P}_i^T, \mathbf{C}_i^T\right) = h_{\theta_T}(\mathbf{G}_i^T), \qquad \left(\mathbf{T}_i^S, \mathbf{P}_i^S, \mathbf{C}_i^S\right) = h_{\theta_S}(\mathbf{G}_i^S), \qquad i = 1, \ldots, N. \tag{13}$$

Here, $\mathbf{T}_i \in SE(3)$ denotes the camera pose, $\mathbf{P}_i \in \mathbb{R}^{H\times W\times 3}$ denotes the pixel-aligned local point map in the coordinate system of view i, and $\mathbf{C}_i \in \mathbb{R}^{H\times W}$ denotes the confidence map.

Normal distillation. Following π^3, we further supervise local surface geometry with a normal loss. Let $\mathbf{n}_{i,j}^S$ and $\mathbf{n}_{i,j}^T$ denote the surface normals computed from the student and teacher point maps, respectively. We define

$$\mathcal{L}_{\text{normal}} = \frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{HW} \arccos\left(\mathbf{n}_{i,j}^S \cdot \text{sg}\left(\mathbf{n}_{i,j}^T\right)\right). \tag{14}$$

Confidence distillation. We supervise the student confidence map using a binary target derived from the pseudo geometric error. Specifically, the target confidence is set to one if the scale-aligned point-map error is below a threshold ϵ, and zero otherwise:

$$y_{i,j} = \mathbb{1}\left[\frac{1}{z_{i,j}^T}\left\|s^* \mathbf{x}_{i,j}^S - \text{sg}\left(\mathbf{x}_{i,j}^T\right)\right\|_1 < \epsilon\right]. \tag{15}$$

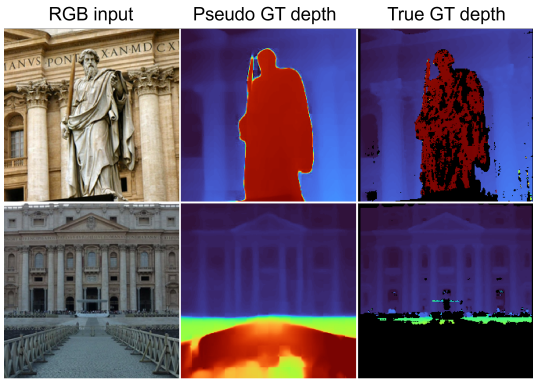

Figure 5: Qualitative comparison of supervision signals on MegaDepth. From left to right: RGB input, scale-aligned π^3 pseudo-GT depth, and dataset GT depth. π^3 supplies dense, spatially coherent geometry where the original annotations are incomplete; black pixels indicate missing GT.

The confidence loss is

$$\mathcal{L}_{\text{conf}} = \frac{1}{NHW} \sum_{i=1}^{N} \sum_{j=1}^{HW} \text{BCE}\left(\mathbf{C}_{i,j}^{S}, y_{i,j}\right). \tag{16}$$

Relative camera-pose distillation. Because the predicted camera poses are defined up to a global similarity transformation, we supervise the student using relative poses between views. For each ordered pair $i \neq j$, we compute

$$\widehat{\mathbf{T}}_{i\leftarrow j}^{S} = \left(\mathbf{T}_{i}^{S}\right)^{-1} \mathbf{T}_{j}^{S}, \qquad \widehat{\mathbf{T}}_{i\leftarrow j}^{T} = \left(\mathbf{T}_{i}^{T}\right)^{-1} \mathbf{T}_{j}^{T}. \tag{17}$$

Let $\widehat{\mathbf{R}}_{i\leftarrow j}^{S}$ and $\widehat{\mathbf{t}}_{i\leftarrow j}^{S}$ denote the relative rotation and translation of the student, and let $\widehat{\mathbf{R}}_{i\leftarrow j}^{T}$ and $\widehat{\mathbf{t}}_{i\leftarrow j}^{T}$ denote the corresponding teacher targets. The rotation loss is

$$\mathcal{L}_{\text{rot}}(i,j) = \arccos\left(\frac{\text{Tr}\left(\left(\text{sg}\left(\widehat{\mathbf{R}}_{i\leftarrow j}^{T}\right)\right)^{\top} \widehat{\mathbf{R}}_{i\leftarrow j}^{S}\right) - 1}{2}\right). \tag{18}$$

The translation loss uses the same scale factor s^* as the point-map alignment:

$$\mathcal{L}_{\text{trans}}(i,j) = H_{\delta}\left(s^{*}\widehat{\mathbf{t}}_{i\leftarrow j}^{S} - \text{sg}\left(\widehat{\mathbf{t}}_{i\leftarrow j}^{T}\right)\right), \tag{19}$$

where $H_{\delta}(\cdot)$ denotes the Huber loss. The camera distillation loss is

$$\mathcal{L}_{\text{cam}} = \frac{1}{N(N-1)} \sum_{i\neq j} \left(\mathcal{L}_{\text{rot}}(i,j) + \lambda_{\text{trans}}\mathcal{L}_{\text{trans}}(i,j)\right). \tag{20}$$

$\mathcal{L}_{\text{points}}$ is defined as Eq. 7. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{points}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}} + \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}. \tag{21}$$

A.3 Implementation Details

We initialize the weights for the encoder and the alternating attention module from the pretrained teacher model, and we keep the encoder frozen during training. For our loss function, we set the weights for each component as follows: $\lambda_{\text{normal}} = 1.0, \lambda_{\text{cam}} = 0.1, \lambda_{\text{feat}} = 0.1, \lambda_{\text{img}} = 100$. We employed the AdamW optimizer and a hybrid schedule of linear warm-up (first 0.5 epochs) followed by cosine decay, reaching a peak learning rate of 1e-6. For each training iteration, we randomly sampled a batch of 24 frames from diverse training scenes. Following Point3R (Wu et al., 2026),

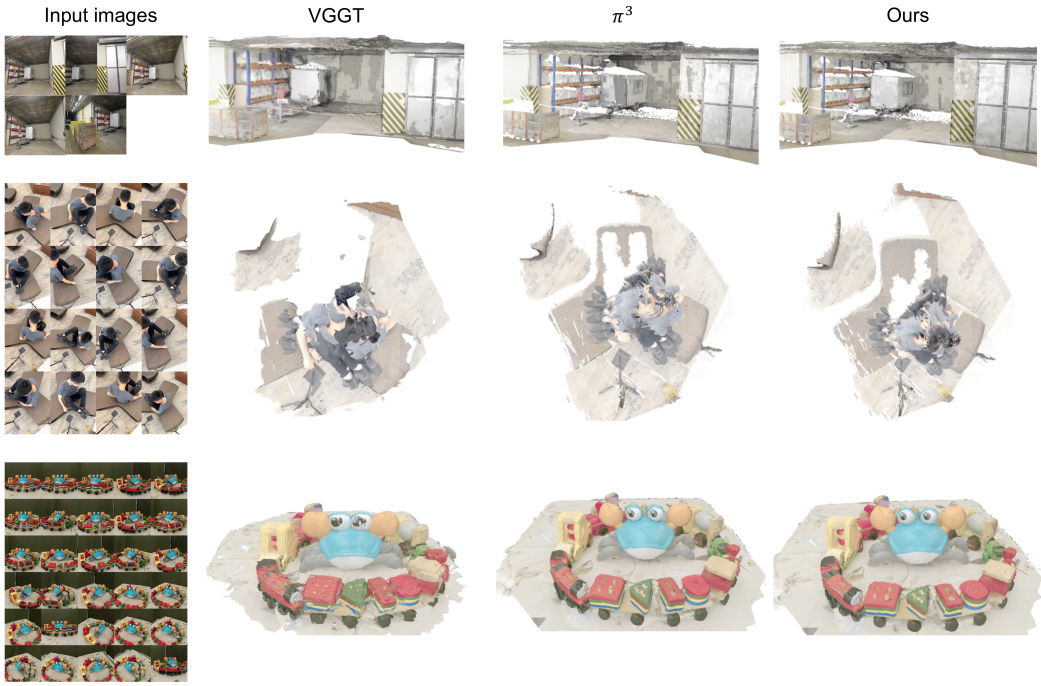

Figure 6: More qualitative comparison of multi-view 3D reconstruction. The second row highlights a representative occlusion case, where a person moves over the bench and causes severe cross-view occlusions. Our method better recovers the occluded part and preserves its geometry.

we processed input images with variable aspect ratios while resizing the maximum edge length to 518 pixels. Training was conducted on 4 NVIDIA H200 GPUs for 6 days. We train the confidence head separately. After completing the main training stages, we freeze the rest of the network and train the confidence head in isolation by setting $\lambda_{\text{conf}} = 0.05$. This stage converges rapidly and has no impact on the model's overall performance.

A.4 Generalization to VGGT.

Our masked training framework is not specific to π^3 and can be applied to other feed-forward 3D reconstruction models. We further validate this by using VGGT (Wang et al., 2025a) as the teacher model. Unlike π^3, VGGT represents the scene in the coordinate system of the first input camera and uses the first frame as the reference view. We therefore preserve the connection from every frame to the first frame when sampling the random attention mask, while masking the remaining cross-frame interactions in the same way as our default setting.

Following the output formulation of VGGT (Wang et al., 2025a), the teacher predicts camera parameters $\mathbf{g}_i^T$, depth maps $\mathbf{D}_i^T$, and point maps $\mathbf{P}_i^T$, which are used as pseudo supervision for the masked student. The geometry distillation objective is

$$\mathcal{L}_{\text{geo}} = \mathcal{L}_{\text{cam}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{pmap}}. \tag{22}$$

For camera prediction, we follow VGGT and use a Huber loss,

$$\mathcal{L}_{\text{cam}} = \sum_{i=1}^{N} \left\| \mathbf{g}_i^S - \text{sg}(\mathbf{g}_i^T) \right\|_\epsilon, \tag{23}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient on the teacher prediction. For depth and point maps, we follow the uncertainty-weighted formulation used by VGGT:

$$\mathcal{L}_{\mathrm{depth}} = \sum_{i=1}^{N} \left\| \mathbf{\Sigma}_i^D \odot (\mathbf{D}_i^S - \mathbf{D}_i^T) \right\|_1 + \left\| \mathbf{\Sigma}_i^D \odot (\nabla \mathbf{D}_i^S - \nabla \mathbf{D}_i^T) \right\|_1 - \alpha \log \mathbf{\Sigma}_i^D, \tag{24}$$

$$\mathcal{L}_{\mathrm{pmap}} = \sum_{i=1}^{N} \left\| \mathbf{\Sigma}_i^P \odot (\mathbf{P}_i^S - \mathbf{P}_i^T) \right\|_1 + \left\| \mathbf{\Sigma}_i^P \odot (\nabla \mathbf{P}_i^S - \nabla \mathbf{P}_i^T) \right\|_1 - \alpha \log \mathbf{\Sigma}_i^P. \tag{25}$$

We use the same feature distillation as Eq. 9. Therefore, the overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{pmap}} + \mathcal{L}_{\mathrm{depth}} + \lambda_{\mathrm{cam}} \mathcal{L}_{\mathrm{cam}} + \lambda_{\mathrm{feat}} \mathcal{L}_{\mathrm{feat}}. \tag{26}$$

We refer to the resulting model as MGE-VGGT and evaluate it on the ETH3D occlusion benchmark. As shown in Table 6, MGE-VGGT remains close to the original VGGT under the clean setting and becomes increasingly better as the occlusion level increases. For dense reconstruction, VGGT's accuracy error increases from 0.149 to 0.327 from clean to large occlusion, whereas MGE-VGGT increases from 0.158 to 0.287. A similar trend is observed for completeness.

The difference is more pronounced for camera pose. Under large occlusion, MGE-VGGT reduces ATE from 0.898 to 0.805 and RPE-r from 4.673 to 2.111 compared with VGGT. While the two models perform similarly under clean and small occlusion, VGGT degrades rapidly under medium and large occlusion, whereas MGE-VGGT remains relatively stable. These results show that the robustness brought by masked cross-view training also transfers to VGGT, despite its different architecture and prediction heads, suggesting that MGE is not specific to our default π^3 teacher.

We also evaluated MGE-VGGT on other standard benchmarks, which are reported in Appendix A.7.

Table 6: **Occlusion evaluation on ETH3D for VGGT.**

Condition	Clean			Small			Medium			Large		
Methods	Acc. ↓	Comp. ↓	NC. ↑	Acc. ↓	Comp. ↓	NC. ↑	Acc. ↓	Comp. ↓	NC. ↑	Acc. ↓	Comp. ↓	NC. ↑
VGGT	**0.149**	**0.124**	**0.849**	0.178	0.133	**0.841**	0.268	0.168	**0.818**	0.327	0.224	**0.799**
MGE-VGGT	0.158	0.132	0.840	**0.175**	**0.128**	0.836	**0.252**	**0.160**	0.808	**0.287**	**0.185**	0.792
Methods	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓	ATE ↓	RPE-t ↓	RPE-r ↓
VGGT	**0.716**	0.887	2.122	0.738	0.941	2.155	0.758	0.961	2.948	0.898	1.227	4.673
MGE-VGGT	0.723	**0.885**	**1.790**	**0.733**	**0.899**	**1.774**	**0.757**	**0.926**	**1.936**	**0.805**	**0.995**	**2.111**

A.5 Visual Ambiguity Evaluation.

A.5.1 Benchmark

We first run COLMAP with Doppelgangers++ to obtain pseudo ground-truth camera poses and scene structure. For each scene, we identify an anchor pair with low normal-pair probability and negligible COLMAP covisibility, which defines two visually similar but geometrically separated sides of the scene. Starting from one anchor, we construct a 15-view reference set using farthest-point sampling over COLMAP camera centers. Reference views are required to be consistent with the reference side and geometrically connected to the selected set. We then sample 5 Doppelganger views from the opposite side, choosing views that are visually similar to the reference set but geometrically inconsistent with it. To reduce redundancy, we remove duplicate or highly overlapping subsets and manually verify the retained samples. This procedure produces 43 subsets, each containing 15 reference views and 5 Doppelganger views.

A.5.2 Camera Visualizations

We visualize the predicted camera poses from two scenes by projecting them onto the 2D ground plane, as shown in Figure 7. VGGT and π^3 often place the reference and Doppelganger cameras on the same side of the scene whereas MGE better preserves the separation between the two geometrically distinct groups, consistent with its higher cross-pair pose accuracy.

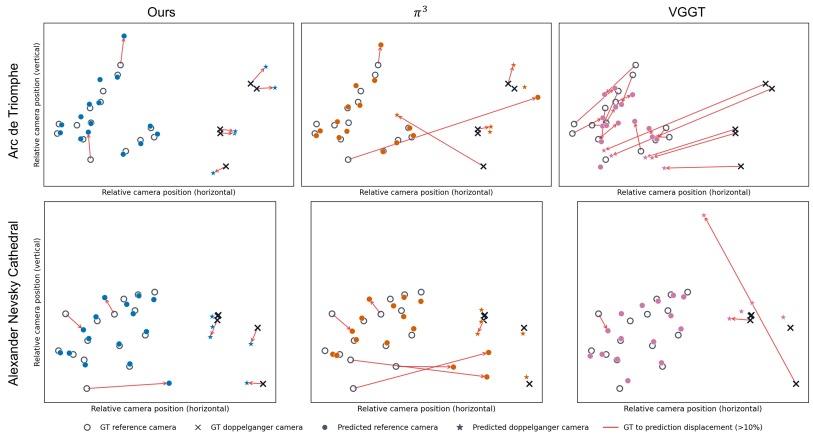

Figure 7: Visualizations of camera poses on visual ambiguity evaluation.

Table 7: Dense reconstruction results under different keyframe strides. Lower is better for Acc, Comp, and CD, while higher is better for NC.

Method	stride=200				stride=100				stride=40			
	Acc ↓	Comp ↓	CD ↓	NC ↑	Acc ↓	Comp ↓	CD ↓	NC ↑	Acc ↓	Comp ↓	CD ↓	NC ↑
TTT3R	0.097	0.064	0.0805	0.812	0.085	0.051	0.0680	0.811	0.079	0.033	0.0560	0.792
StreamVGGT	0.092	0.083	0.0875	0.827	0.086	0.060	0.0730	0.820	0.090	0.052	0.0710	0.799
FastVGGT	0.047	0.072	0.0595	0.869	0.034	0.047	0.0405	0.871	0.016	0.015	0.0155	0.843
VGGT+AGA	0.041	0.054	0.0475	0.876	0.027	0.028	0.0275	0.879	0.016	0.014	0.0150	0.848
Ours	**0.021**	**0.022**	**0.0215**	**0.904**	**0.018**	**0.018**	**0.0180**	**0.892**	**0.014**	**0.014**	**0.0140**	**0.856**

Method	stride=25				stride=10				stride=3			
	Acc ↓	Comp ↓	CD ↓	NC ↑	Acc ↓	Comp ↓	CD ↓	NC ↑	Acc ↓	Comp ↓	CD ↓	NC ↑
TTT3R	0.081	0.034	0.0575	0.767	0.085	0.029	0.0570	0.706	0.123	0.042	0.0825	0.623
StreamVGGT	0.096	0.052	0.0740	0.782	OOM				OOM			
FastVGGT	0.019	0.017	0.0180	0.811	0.015	0.011	0.0130	0.734	0.016	0.010	0.0130	0.636
VGGT+AGA	0.018	0.016	0.0170	0.811	0.014	0.012	0.0130	0.734	0.015	0.009	0.0120	0.637
Ours	**0.017**	**0.016**	**0.0165**	**0.829**	**0.014**	**0.011**	**0.0125**	**0.763**	**0.012**	**0.009**	**0.0105**	**0.678**

A.6 Quantitative metrics for token merging comparison

We provide the detailed metrics in Table 7 which are visualized in Figure 3. Our method consistently outperforms existing long-sequence baselines across different keyframe strides. The advantage is especially pronounced under sparse keyframe sampling, where our model achieves substantially lower reconstruction errors and higher normal consistency than TTT3R, StreamVGGT, and FastVGGT. As the sequence becomes denser, our method remains robust and continues to deliver the best or comparable performance across all metrics. Notably, StreamVGGT runs out of memory for denser settings, whereas our method scales to all evaluated strides. We have a favorable balance between reconstruction quality and long-sequence scalability.

A.7 Evaluation on More Standard benchmarks

A.7.1 Point Map Estimation

We evaluate the quality of reconstructed multi-view point maps on the scene-level 7-Scenes (Shotton et al., 2013) and NRGBD (Azinović et al., 2022) datasets under both sparse and dense view conditions (Wang et al., 2025b; 2026e). Consistent with prior works (Azinović et al., 2022; Wang et al., 2024; 2025b), we report Accuracy (Acc.), Completion (Comp.), and Normal Consistency (NC) in Table 8. Our MGE models achieve significant improvements on 7-scenes with the masked training especially under sparse input, and achieve comparative results on NRGBD to full-attention models.

Table 8: Point map estimation on 7-Scenes and NRGBD.

Method	View	7-Scenes						NRGBD					
		Acc. ↓		Comp. ↓		NC ↑		Acc. ↓		Comp. ↓		NC ↑	
		Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
Fast3R (Yang et al., 2025)	*sparse*	0.095	0.065	0.144	0.089	0.673	0.759	0.135	0.091	0.163	0.104	0.759	0.877
CUT3R (Wang et al., 2025b)		0.093	0.049	0.102	0.051	0.704	0.805	0.104	0.041	0.079	0.031	0.822	0.968
FLARE (Zhang et al., 2025b)		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)		0.044	0.025	0.056	0.033	0.733	0.845	0.051	0.029	0.066	0.038	0.890	0.981
π^3 (Wang et al., 2026e)		0.047	0.029	0.075	0.049	0.742	0.841	**0.026**	**0.015**	**0.028**	**0.014**	**0.916**	**0.992**
MGE-VGGT		**0.037**	**0.023**	**0.054**	**0.032**	0743	0.851	0.050	0.028	0.060	0.033	0.891	0.990
MGE-π^3		0.044	0.027	0.066	0.042	**0.744**	**0.846**	0.027	0.016	0.030	0.016	0.901	0.989
Fast3R (Yang et al., 2025)	*dense*	0.040	0.017	0.056	0.018	0.644	0.725	0.072	0.030	0.050	0.016	0.790	0.934
CUT3R (Wang et al., 2025b)		0.023	0.010	0.027	**0.008**	0.669	0.764	0.086	0.037	0.048	0.017	0.800	0.953
FLARE (Zhang et al., 2025b)		0.019	0.007	0.026	0.013	0.684	0.785	0.023	0.011	0.018	0.008	0.882	0.986
VGGT (Wang et al., 2025a)		0.022	0.008	0.026	0.012	0.666	0.760	0.017	0.010	0.015	0.005	0.893	**0.988**
π^3 (Wang et al., 2026e)		0.016	0.007	**0.022**	0.011	**0.689**	**0.792**	0.015	0.008	**0.013**	**0.005**	**0.898**	0.987
MGE-VGGT		0.019	0.008	0.027	0.009	0.683	0.784	0.014	0.007	0.016	0.006	0.889	0.987
MGE-π^3		**0.016**	**0.007**	0.022	0.010	0.687	0.789	**0.014**	**0.007**	0.014	0.006	0.897	0.988

Table 9: Video depth estimation on Sintel, Bonn and KITTI.

Method	Sintel		Bonn		KITTI	
	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑	Abs Rel ↓	$\delta < 1.25$ ↑
DUSt3R (Wang et al., 2024)	0.662	0.434	0.151	0.839	0.143	0.814
MASt3R (Leroy et al., 2024)	0.558	0.487	0.188	0.765	0.115	0.848
MonST3R (Zhang et al., 2025a)	0.399	0.519	0.072	0.957	0.107	0.884
Fast3R (Yang et al., 2025)	0.638	0.422	0.194	0.772	0.138	0.834
MVDUSt3R (Tang et al., 2025)	0.805	0.283	0.426	0.357	0.456	0.342
CUT3R (Wang et al., 2025b)	0.417	0.507	0.078	0.937	0.122	0.876
Aether (Zhu et al., 2025)	0.324	0.502	0.273	0.594	0.056	0.978
FLARE (Zhang et al., 2025b)	0.729	0.336	0.152	0.790	0.356	0.570
VGGT (Wang et al., 2025a)	0.299	0.638	0.057	0.966	0.062	0.969
π^3 (Wang et al., 2026e)	0.233	0.664	0.049	0.975	**0.038**	**0.986**
MGE-VGGT	0.283	0.663	0.054	0.970	0.061	0.972
MGE-π^3	**0.232**	**0.676**	**0.049**	**0.975**	0.038	0.985

A.7.2 Video Depth Evaluation

For the task of video depth estimation, we use the Sintel (**?**), Bonn (Palazzolo et al., 2019)), and KITTI (Geiger et al., 2013) datasets, and report the Absolute Relative Error (Abs Rel) and the prediction accuracy at a threshold of $\delta < 1.25$, following previous works (Wang et al., 2025b; 2026e). We align the video depth to the ground truth with a scale per sequence. As reported in Table 9, compared with the full-attention baseline, our masked-trained model yields consistent improvements on Sintel and Bonn, achieving new SoTA on them while remaining competitive on KITTI.

A.8 Ablation on Masking pattern

We further compare random masking with Sliding-Window Attention (SWA) during masked training. The corresponding mask patterns are visualized in Figure 8, with comparable masking ratios used for a fair comparison. We evaluate both strategies on NRGBD (Azinović et al., 2022) using keyframe strides of 100 and 10, representing limited- and redundant-view settings, respectively. As shown in Table 10, random masking consistently outperforms SWA under both settings, with a larger

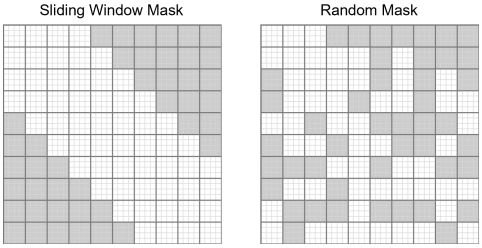

Figure 8: Visualization of two mask patterns.

Table 10: **Ablation of mask pattern.** We compare the sliding-window attention (SWA) mask with the proposed random masking strategy under different keyframe strides.

Shape	stride=100		stride=10	
	CD ↓	NC ↑	CD ↓	NC ↑
SWA	0.0152	0.9125	0.0135	0.7902
Random	**0.0145**	**0.9194**	**0.0117**	**0.8000**

advantage for denser sequences. We attribute this to the limited propagation range of local sliding windows: similar to sequential matching, information must be relayed through intermediate views, whereas random connectivity enables more diverse long-range interactions and better approximates exhaustive multi-view reasoning.

A.9 VISUALIZATIONS

Figure 9 presents qualitative reconstruction results on the 13 ETH3D scenes of our MGE+AGA and FastVGGT. Compared with FastVGGT, our method generally recovers more complete scene geometry while maintaining a coherent global structure. The improvement is particularly visible in challenging scenes such as courtyard, delivery area, and relief 2, where FastVGGT produces missing regions or disconnected fragments, whereas our reconstruction preserves a larger portion of the observed surfaces.

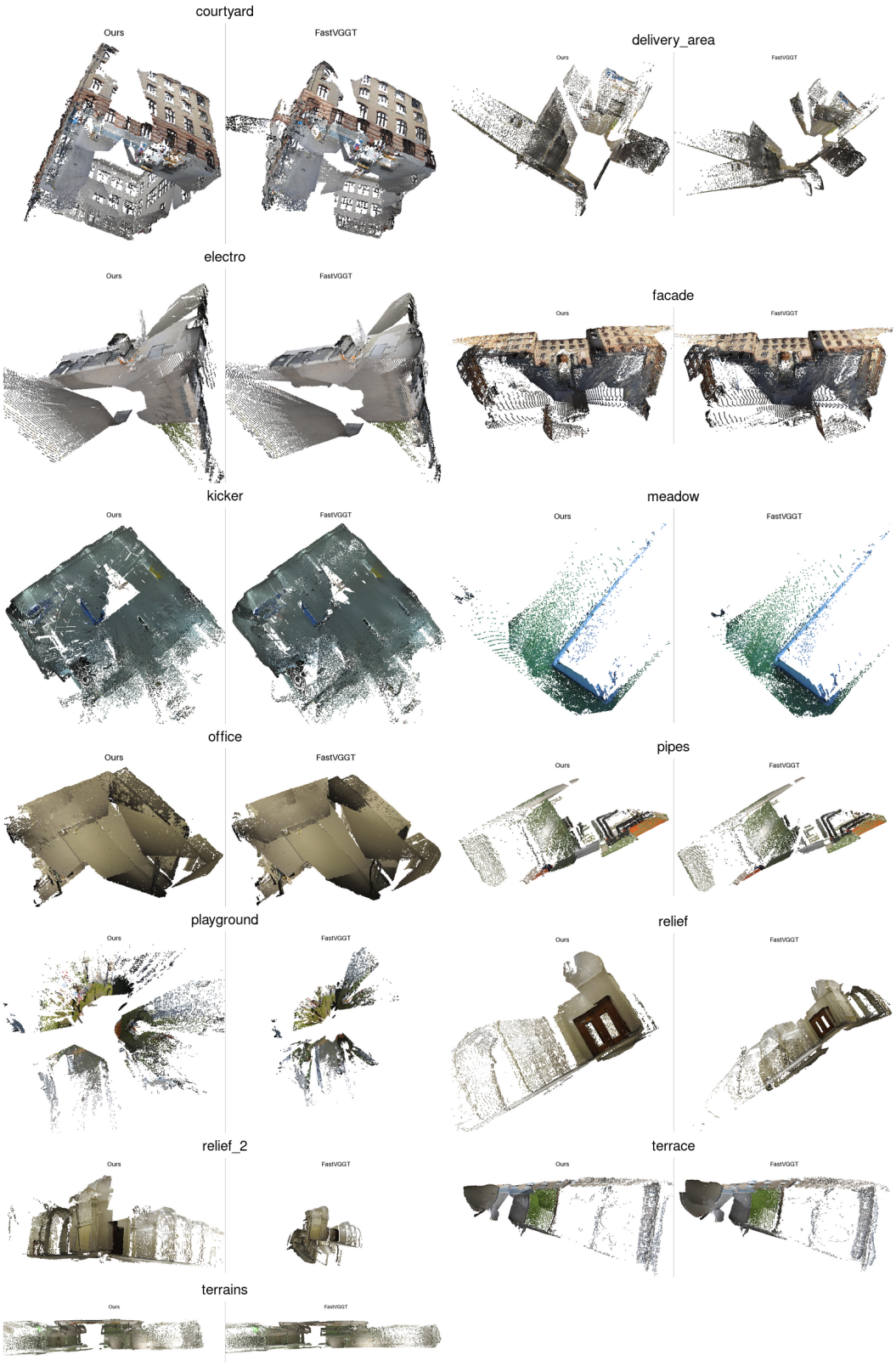

Figure 9: Qualitative comparison on ETH3D. For each scene, our reconstruction is shown on the left and FastVGGT on the right. Our method generally produces more complete and globally coherent geometry, with fewer missing surfaces and disconnected structures.