Emergent Extreme-View Geometry in 3D Foundation Models

Yiwen Zhang[1] Joseph Tung[2] Ruojin Cai[3] David Fouhey[2] Hadar Averbuch-Elor[1]
[1]Cornell University [2]New York University [3]Kempner Institute, Harvard University

https://ext-3dfms.github.io/

Abstract

3D foundation models (3DFMs) have recently transformed 3D vision, enabling joint prediction of depths, poses, and point maps directly from images. Yet their ability to reason under extreme, non-overlapping views remains largely unexplored. In this work, we study their internal representations and find that 3DFMs exhibit an emergent understanding of extreme-view geometry, despite never being trained for such conditions. To further enhance these capabilities, we introduce a lightweight alignment scheme that refines their internal 3D representation by tuning only a small subset of backbone bias terms, leaving all decoder heads frozen. This targeted adaptation substantially improves relative pose estimation under extreme viewpoints without degrading per-image depth or point quality. Additionally, we contribute MegaUnScene, a new benchmark of Internet scenes unseen by existing 3DFMs, with dedicated test splits for both relative pose estimation and dense 3D reconstruction. All code and data will be released.

Figure 1. **Do 3D foundation models have an emergent understanding of extreme-views?** The **pretrained** VGGT model was trained primarily on overlapping images. Surprisingly, when tested on non-overlapping image pairs*, the model still produces plausible estimates of relative pose, with nearly half of the pairs yielding a rotation error below 30°. Careful fine-tuning of a small number of parameters **substantially improves results**. Here, for instance, the pretrained model produces incorrect pose, **yielding the red ghost structure**); the fine-tuned model corrects the error.
*All from *unseen* scenes assembled in our MegaUnScene benchmark.

1. Introduction

Much of 3D vision research has been built on a comfortable assumption: that we have access to densely captured, overlapping images of a scene. With such coverage, 3D structure can be inferred by matching pixels, tracking features, and triangulating points. Yet in many real-world scenarios—such as casual captures on mobile devices, historical archives or tourist photo collections—this assumption simply does not hold. Image observations may be sparse or captured from viewpoints with minimal to no visual overlap, making correspondence extraction unreliable. In these under-constrained settings, classical 3D vision pipelines begin to falter, motivating the search for models that can reason about geometry beyond the limits of visual overlap.

Recent 3D foundation models (3DFMs) [33, 53, 57] jointly estimate camera poses, depths, and point maps in a single feedforward pass and represent a radical departure from classical pipelines. These models consist of a shared backbone that attends within and across images, followed by heads that produce per-image 3D geometry and relative poses that place the 3D geometry in a common coordinate frame. 3DFMs eliminate the need for explicit correspondence estimation, but are trained primarily on data that still largely conforms to the comfortable assumption of visual overlap between views. Nonetheless, when tested on views with limited overlap, these methods are surprisingly capable of reasonable predictions (see the error distribution curves in Figure 1), although far from their original performance.

This paper's key premise is that an internal *3D language*—a learned geometric representation—crystallizes in the shared backbone of 3DFMs, with the decoder heads simply translating this language to explicit 3D outputs. Based on this insight, we propose a lightweight alignment scheme that targets this internal language to enhance per-

formance on extreme views. Specifically, we *freeze* the decoder heads and tune *only* the biases of selected backbone layers to minimize rotation loss on camera poses. By merely updating around 80k parameters and training on roughly 65k image pairs for 2 epochs, these billion-parameter 3DFMs improve substantially in handling extreme view cases, predicting accurate relative rotations and, at the same time, per-image depth and point map estimations remain intact. As we show experimentally, many alternative approaches (*e.g.*, unfreezing the decoder heads) lead to performance degradation across other outputs.

To evaluate our system, we contribute a new dataset named *MegaUnScene*. This is necessary because existing in-the-wild datasets [29, 49, 50, 58] lack dedicated test splits and have been used for 3DFM training. MegaUnScene comprises 476 Internet scenes *unseen* by existing 3DFMs, enabling systematic evaluation of 3DFMs under realistic, unconstrained conditions for both relative pose estimation and dense 3D reconstruction.

Our contributions include: (a) a lightweight alignment scheme that targets the internal 3D representations to enable multiple 3DFMs to work well on extreme views: reducing relative rotation error while preserving their pretrained multi-task capabilities; (b) a new benchmark composed of 476 Internet scenes for evaluating 3DFM methods; and (c) a new state-of-the-art in extreme view predictions that reduces median rotation error on sELP [4] (a single camera setting) from $13.2^\circ \rightarrow 9.7^\circ$ and on in-the-wild data (with and without translations) from $42.4^\circ \rightarrow 13.1^\circ$ and $28.4^\circ \rightarrow 11.7^\circ$.

2. Related Work

Extreme Pose Estimation. Traditional relative pose estimation relies on local feature matching and geometric verification, typically assuming sufficient visual overlap between input views. Classical pipelines employ handcrafted descriptors [34, 39] and RANSAC-based matches [17]. Learning-based matchers [13, 15, 32, 40, 48] leverage deep neural networks to learn both feature descriptors and matching, improving robustness to illumination and appearance variations. However, these overlap-dependent methods degrade under wide baselines or non-overlapping settings.

To handle sparse or overlap-free inputs, learning-based pose predictors directly infer 3D relationships from images without explicit correspondences. Diffusion-based estimators [51, 68] and other sparse-view or object-centric approaches [16, 30, 46, 54, 64, 67] demonstrate strong 3D priors but are restricted to constrained domains. Recent works have also explored leveraging generative video priors for pose estimation [8, 35], synthesizing plausible transitions between sparse views to provide useful cues for relative pose reasoning.

At the scene level, earlier extreme-pose methods relied on geometric assumptions [9, 38] or additional input modalities [62, 63]. Several works design learning-based frameworks specifically tailored for the relative rotation task [7, 12]. Bezalel *et al*. [4] recently extend these to handle uncontrolled Internet imagery. However, performance remains limited over extreme real-world image pairs. Our work instead adapts 3D foundation models directly, fine-tuning their geometric representations to become inherently robust to extreme, non-overlapping viewpoints.

3D Foundation Models. Classical 3D reconstruction methods such as Structure-from-Motion (SfM) [1, 21, 41, 47] and Multi-View Stereo (MVS) [18, 42] recover camera poses and 3D structure by matching local features and jointly optimizing via bundle adjustment. Recent feed-forward 3D models have transformed this pipeline by directly predicting camera pose, depth, point map in a single forward pass. DUSt3R [56] pioneered pairwise reconstruction by regressing dense point maps from two input images, followed by [14, 27, 55, 61, 69], which improve scalability and architectural efficiency. Building on these pairwise models, VGGT [53] introduced a unified transformer backbone that jointly infers camera poses, depth, and point maps from multiple views. Subsequent models such as [26, 33, 57] further generalized this paradigm with permutation-equivariant reasoning and metric-scale reconstruction for large-scale scenes. Several works also extend these models for reasoning over dynamic scenes [10, 20, 31], while others explore more efficient paradigms [44, 61].

Despite their impressive performance over diverse datasets, these foundation models are typically trained and evaluated on overlapping or smoothly varying viewpoints. In this work, we demonstrate that these models demonstrate an *emergent* understanding of extreme-view settings, which can be further enhanced with our proposed lightweight alignment scheme that only modifies the shared alternating-attention backbone. This is unlike prior approaches [11, 59] that typically employ a head-only fine-tuning strategy, while the shared backbone remains frozen. However, we demonstrate that this allows for significantly improving performance on extreme-view scenarios, *without* degrading the model's performance over other 3D tasks.

3D Pipeline Evaluation. Existing benchmarks for evaluating scene-scale 3D pipelines usually rely on laser scans [43], controlled camera setups [24], or synthetic rendering [3]. Ideally, 3D pipelines should also handle *unconstrained* inputs captured *in-the-wild*. However, while datasets like MegaDepth [29], WikiScenes [58], MegaScenes [49] and AerialMegaDepth [50] use Internet photos to capture in-the-wild conditions, these datasets are typically used for training and lack dedicated evaluation subsets for dense prediction tasks. This work contributes a new evaluation benchmark assembled from unconstrained collections, enabling evaluation on unseen Internet scenes.

Figure 2. **VGGT Cross-View Attention Maps.** We visualize cross-view attention maps for three image pairs of varying overlap, from high overlap (left) to none (right). For each image pair, we highlight three query regions in I_1 with colored boxes (green, cyan, magenta). Solid boxes indicate region overlap, while dashed boxes indicate no overlap. I_2's corresponding attention maps are shown at the bottom row with like colors. Reconstructed pointmaps are shown towards the right.

3. Method

In order to fine-tune 3D foundation models (3DFMs) for enhanced understanding of extreme geometric configurations, we propose a lightweight alignment framework that preserves the model's pretrained knowledge while improving its robustness to large viewpoint changes. In what follows, we first analyze how 3DFMs internally represent 3D structure (Section 3.1). Building upon our findings, we then introduce a compact fine-tuning scheme that aligns the shared 3D representation for viewpoint robustness (Section 3.2).

3.1. The Internal Language of 3DFMs

3D foundation models (3DFMs) have recently shown remarkable progress in reconstructing scene geometry directly from unstructured images. However, despite their growing adoption, their internal structure has remained largely unexplored. In this section, we first analyze their internal *3D language* via cross-view attention maps, revealing that these models already encode a surprisingly rich understanding of scene geometry within their shared alternating attention backbone. We then perform a fine-grained, layer-level analysis of the backbone to examine which layers contribute the most to this learned representation.

Background: 3DFM Architectural Design. Modern 3DFMs [33, 53, 57] share a common architectural structure, as illustrated in Figure 3. Each model first encodes the input images, extracting patch-level embeddings (optionally concatenated with additional input modality tokens) via an encoder ε. For simplicity, we describe a scenario with two input images, I_1 and I_2, each divided into $N_p = H_p \times W_p$ patches. These N_p patch tokens and then fed into a shared transformer-based backbone, which alternates between *frame* attention blocks and *global* attention blocks. In a frame attention block, patch tokens are flattened and processed independently as:

$$\mathbf{T}^{(i)}_{\text{frame}} \in \mathbb{R}^{N_p \times D}, \quad i \in \{1, 2\}. \tag{1}$$

By contrast, the global attention block concatenates tokens from both images:

$$\mathbf{T}_{\text{global}} = [\mathbf{T}^{(1)}_{\text{frame}}, \mathbf{T}^{(2)}_{\text{frame}}] \in \mathbb{R}^{(2N_p) \times D}. \tag{2}$$

Both frame and global attention use self-attention, which projects these concatenated tokens into queries $\mathbf{Q} = f_Q(\mathbf{T})$, keys $\mathbf{K} = f_K(\mathbf{T})$, and values $\mathbf{V} = f_V(\mathbf{T})$ through learned linear projections f_Q, f_K, f_V. For each layer l and head h, the attention maps are computed as: $\mathbf{A}^{(l)}_h = \text{softmax}(\mathbf{Q}\mathbf{K}^T/\sqrt{d_h})$ where d_h is the dimension of the attention head h.

The output patch tokens are then decoded by task-specific heads, such as the camera head $\mathcal{D}_c$ and the dense prediction head $\mathcal{D}_d$ visualized in Figure 3.

Cross-View Attention. To understand the cross-view dependencies in the alternating attention backbone, we analyze the global attention maps, considering the last layer where the cross-view information is thoroughly fused. Specifically, we select query vectors $\mathbf{q} \in \mathbf{Q}$ located at various spatial locations in I_1, comparing it with all key vectors $\mathbf{k} \in \mathbf{K}$ in I_2 and summing over all the attention heads.

We illustrate the cross-view attention maps in Figure 2, as exemplified by VGGT [53] and three image pairs with varying levels of overlap. For each image pair, we highlight attention maps from three query locations, marked by green, cyan and magenta boxes. As shown, consistent patterns are observed across both *overlapping* and *non-overlapping* image regions. For regions with direct visual overlap (visualized with solid boxes), the high attention areas in I_2 accumulate precisely at the corresponding locations. This demonstrates the model's ability to identify exact visual correspondences within the shared attention backbone. For regions without direct visual overlap (visualized with dashed boxes), the cross-view maps reveal non-trivial structural patterns. For instance, in the middle example, regions green and magenta are not visible in I_2, yet the attention weights accumulate at the image corners that are spatially nearest to the selected regions in the 3D world. Such "near-correspondence" understanding can also be observed for the image pair on the right, which contain no overlapping regions. Furthermore, for query regions that do not have a near-corresponding area (*e.g.*, the rightmost green query), we can see that the model has learned to attend to other symmetry-based cues, such as the gate's overall cur-

Figure 3. **Rotation-based Alignment Framework**. Above we illustrate our lightweight alignment scheme, which supervises the camera head of 3DFMs via a rotation loss $\mathcal{L}$ over predicted and ground truth relative rotation matrices. To preserve pretrained knowledge, we only update the bias terms of a sparse set of layers in the shared alternating attention (AA) backbone.

vature and its fine-grained ornate details, which can further guide the model's 3D understanding in such extreme cases.

These observations suggest that a rich internal representation of the scene's geometry is already constructed within the shared backbone. This representation—which extends far beyond direct visual correspondences—provides an implicit *3D language* that encodes spatial relations, depth, and pose. The task-specific heads, such as the DPT head and camera head, simply act as format converters that read out this shared state into explicit geometric quantities like depth maps, point maps, or camera poses.

Are All Backbone Layers Equally Important? Prior studies have shown that not all layers in Transformer-based architectures contribute equally to the learned representation [5, 22, 25]. Motivated by these findings, we conduct a similar analysis over 3DFMs, quantifying the degree of representational change between neighboring backbone layers. Our results reveal that only a small subset of layers exhibit substantial changes in their feature representations. Interestingly, the layers identified by this analysis largely coincide with those connected to the dense prediction head, implying that such skip connections play a key role in shaping the model's internal 3D language; additional experimental details are provided in the supplementary material.

3.2. Rotation-Based 3DFM Alignment

The analysis we conducted in the previous section suggests that 3DFMs already encode a rich 3D language within their shared attention backbone. Building on this insight, we demonstrate that the model's representation can be significantly strengthened by supervising only on camera pose. Crucially, rather than retraining the task-specific heads, we apply minimal fine-tuning to the alternating attention layers, aligning the model's internal 3D language while avoiding overfitting to individual tasks. An overview of our approach is provided in Figure 3.

Rotation-Only Supervision. Inspired by prior work that design frameworks for estimating relative rotations between non-overlapping image pairs [4, 7, 12], we propose a rotation-based alignment objective over image pairs. Formally, given an image pair (I_1, I_2) with predicted absolute rotation matrices $\mathbf{R}_1^{\text{pred}}$ and $\mathbf{R}_2^{\text{pred}}$, and ground-truth rotations $\mathbf{R}_1^{\text{gt}}$ and $\mathbf{R}_2^{\text{gt}}$, we propose a simple unified objective shared across all architectures:

$$\mathcal{L} = \mathcal{L}_{\text{geo}}\left(\mathbf{R}_{\text{rel}}^{\text{pred}}, \mathbf{R}_{\text{rel}}^{\text{gt}}\right) + \mathbb{1}_{\text{a}}\, \mathcal{L}_{\text{geo}}\left(\mathbf{R}_1^{\text{pred}}, \mathbf{I}\right), \tag{3}$$

where the relative rotations are $\mathbf{R}_{\text{rel}}^{\text{pred}} = \mathbf{R}_2^{\text{pred}}(\mathbf{R}_1^{\text{pred}})^\top$ and $\mathbf{R}_{\text{rel}}^{\text{gt}} = \mathbf{R}_2^{\text{gt}}(\mathbf{R}_1^{\text{gt}})^\top$ and $\mathcal{L}_{\text{geo}}$ denotes the geodesic loss which measures the minimal angular distance between rotation matrices on the $\text{SO}(3)$ manifold. The indicator $\mathbb{1}_{\text{a}}$ activates the optional anchoring term for models assuming a fixed reference frame (e.g., VGGT), enforcing the first image to align with the world coordinate frame. For permutation-invariant architectures (e.g., π^3), $\mathbb{1}_{\text{a}} = 0$, and the loss reduces to the symmetric relative-rotation term.

Note that, unlike the dense point-wise annotations required for training these models, our rotation-only supervision relies on much sparser signals (*i.e.*, relative rotations between image pairs).

Preserving Pretrained Knowledge via Minimal Backbone Fine-Tuning. The proposed rotation-based objective supervises only the camera head, without directly affecting the dense prediction head. This partial supervision can disrupt the geometric alignment between heads, degrading performance on downstream tasks that rely on dense predictions (which we show experimentally). To prevent this, we adopt a minimal backbone fine-tuning strategy that targets both the minimal set of *layers* and *parameters* within the backbone: we fine-tune only the bias terms within the sparse set of backbone layers identified in Section 3.1. Prior work has shown that bias-only fine-tuning in Transformer-based models is often competitive with full fine-tuning [65]. As our experiments show, combining selective layer updates with bias-only tuning allows for effectively aligning the model's internal 3D language for extreme-view reasoning, all while preserving its pretrained multi-task knowledge.

Training Data. We follow the methodology introduced in Bezalel *et al.* [4] to construct a training set of image pairs from scene-level COLMAP reconstructions from MegaScenes [49]. We include pairs with larger translations than [4] in order to train a model that is able to generalize well to pairs with camera translations. We report precise details in the supplementary material.

4. The MegaUnScene Benchmark

Existing benchmarks evaluating 3DFMs typically have scenes with constrained 3D environments—*e.g.*, assuming constant illumination, transient objects, and camera intrinsics. To evaluate 3DFMs on unconstrained inputs captured *in-the-wild*, we create *MegaUnScene*: a new collection of 476 Internet scenes *unseen* by existing models. From these scenes, we assemble two test sets for relative pose estimation and one for dense reconstruction.

Benchmark Construction. We follow the protocol from MegaScenes [49] to curate a 3D reconstruction benchmark of Internet photos. We only include scenes that are not included in MegaScenes [49], verified by cross-referencing unique image and scene names. To improve reconstruction fidelity, we slightly modify the pipeline by using Doppelgangers++ [59] integrated with MASt3R-SfM [14]. Doppelgangers++ disambiguates challenging internet photos depicting ambiguous views, while dense MASt3R [28] matches yield more robust pairwise poses for incremental SfM. Lastly, we run multi-view stereo to obtain dense depth. Additional details are in the supplementary material.

Benchmarking Relative Pose in the Wild. We construct subsets with and without camera translations for evaluating relative pose estimation: *UnScenePairs*, contains image pairs predominant rotational motion, and *UnScenePairs-t* contains image pairs with larger camera baselines. Unlike the ELP test sets [4], these subsets capture unconstrained, in-the-wild image pairs unseen by 3DFMs. We follow the image–pair selection procedure proposed in prior work [4], constructing mutual K-nearest neighbor graphs using camera translation distances. We select a small K for constructing UnScenePairs (setting $K = 5$), and a much larger value to admit pairs with greater translations for constructing UnScenePairs-t (setting $K = 50$). After obtaining the candidate pairs, we assign each pair to an overlap category (Large, Small, None) according to the relative rotation between the two camera poses and their FoV.

For UnScenePairs-t, the increased translation baseline introduces parallax challenges where the classification algorithm may classify image pairs with some overlap into the *None* category. Therefore, to ensure reliable overlap labels under larger baselines, we verify correspondences by querying the filtered image matches from MASt3R-SfM [14] and Doppelgangers++ [59], confirming no matches for *None* pairs and distinguishing *Large* from *Small* overlap based on overall spatial match coverage.

Finally, we manually review all selected pairs and removed those affected by motion blur, significant occlusions, or insufficient geometric structure (e.g., flat painted surfaces). The final UnScenePairs test set has 3,883 (778 non-overlapping) image pairs across 458 scenes and the UnScenePairs-t test set has 2,432 (763 non-overlapping) image pairs across 387 scenes.

Figure 4. Metric scale visualization of *UnSceneRecon* scenes (L→R): Aghavnavank Monastery, Alamgiri Gate, Predjama Castle, and the Ritz Tower. For reference, the person is 2 meters tall.

Benchmarking Dense Predictions in the Wild. We construct *UnSceneRecon*, a dense reconstruction subset comprising 100 in-the-wild reconstructions with metric scale annotations (Figure 4). To curate this subset, we select reconstructions with at least 50 images and fuse depth maps to create a representative point cloud. Human annotators inspect a subset of these reconstructions, focusing on ones with the most registered images, and filter out incorrect or low-quality reconstructions. To enable comparable evaluation metrics across scenes, human annotators also identify a metric scale factor for each reconstruction by cross-referencing distances on Google Maps. Please refer to the supplementary material for additional details.

5. Experiments

In this section, we present both qualitative and quantitative experiments to evaluate the effectiveness of our alignment scheme. We first conduct experiments on extreme relative rotation estimation (Section 5.1). We then examine whether pretrained knowledge is preserved via multiview pose estimation (Section 5.2) and dense reconstruction (Section 5.3). Finally, we present an ablation study comparing alternative fine-tuning strategies (Section 5.4). Implementation details, extended ablations, additional tasks, and interactive visualizations are provided in the supplementary material.

Models Considered. We consider three recent 3DFMs: VGGT [53], WorldMirror (WM) [33] and π^3 [57]. Fine-tuned variants are denoted by the subscript FT.

5.1. Extreme Relative Rotation Estimation

Experimental Details. We evaluate performance on sELP [4] and the in-the-wild sets introduced in the previous sections (UnScenePairs, UnScenePairs-t). sELP is curated from the Cambridge Landmarks dataset [2]. Beyond 3DFMs, we also compare against ExRot [4], a recent frame-

Figure 5. **Qualitative results over UnScenePairs-t**. From left to right, we show the input image pair, spherical projection of relative rotations (black: reference view, blue: ground truth, red: pretrained VGGT, yellow: fine-tuned VGGT), and the corresponding reconstructions (sparse ground truth, dense pretrained and fine-tuned). Please refer to the supplementary material for many additional visualizations.

Table 1. **Extreme Relative Rotation Estimation**. We evaluate performance over non-overlapping image pairs in *s*ELP [4], UnScenePairs, and UnScenePairs-t. As illustrated, our fine-tuned models consistently improve the pretrained 3DFMs.

Method	*s*ELP			UnScenePairs			UnScenePairs-t		
	MRE↓	RA_{15}↑	RA_{30}↑	MRE↓	RA_{15}↑	RA_{30}↑	MRE↓	RA_{15}↑	RA_{30}↑
ExRot	13.23	53.1	59.6	28.48	35.7	50.8	42.45	31.3	43.8
VGGT	92.92	24.2	29.1	31.64	33.8	48.8	46.65	29.1	42.1
VGGT$_{FT}$	**14.21**	**50.9**	**56.6**	**12.71**	**53.6**	**67.9**	**14.48**	**50.6**	**62.1**
WM	68.96	36.3	42.5	19.25	44.1	58.9	21.52	42.6	57.4
WM$_{FT}$	**9.74**	**56.9**	**63.5**	**11.75**	**56.2**	**68.1**	**13.13**	**53.3**	**64.5**
π^3	45.24	43.8	48.3	17.66	46.5	59.4	21.62	43.5	56.8
π^3_{FT}	**11.96**	**53.7**	**60.0**	**12.92**	**54.0**	**69.2**	**13.31**	**53.1**	**65.5**

work tailored for the extreme rotation task. ExRot was previously state-of-the-art over in-the-wild image pairs.

Metrics. Let $\mathbf{R}_{\text{pred}}$ and $\mathbf{R}_{\text{gt}}$ denote the predicted and ground-truth relative rotation matrix. We compute the geodesic error, defined as $\arccos(\frac{1}{2}(\text{tr}(\mathbf{R}_{\text{pred}}^\top \mathbf{R}_{\text{gt}})-1))$. Following prior work [4], we report the median rotation error (MRE) and relative rotation accuracy (RA) at thresholds of 15° and 30°, denoted RA_{15} and RA_{30}.

Evaluation. Results over non-overlapping pairs are reported in Table 1. As shown, our fine-tuned models achieve consistent and substantial improvements across all test sets, establishing a new state of the art in extreme rotation estimation. Interestingly, pretrained models show pronounced variability across test sets, performing far better on UnScenePairs and UnScenePairs-t than on sELP (e.g., VGGT: MRE 31.64/46.65 vs. 92.92). This disparity likely stems from their pretraining on landmark-centric data (specifically MegaDepth [29]), which does not resemble the forward-walking, low-parallax trajectories characteristic of the non-overlapping sELP set. In contrast, our fine-tuned models effectively close this gap, despite using only landmark-centric data, demonstrating improved generalization. This generalization also extends to image pairs with translation (UnScenePairs-t), demonstrating robustness beyond purely rotational motion. Figure 5 illustrates this further, showing camera pose and dense reconstruction for two pairs from UnScenePairs-t. After alignment, the fine-tuned VGGT predicts both accurate relative pose and translation, correcting the significant errors of the pretrained model.

Results on pairs with Large and Small overlap are reported in the supplementary material. Overall, we observe that pretrained and fine-tuned models perform similarly for overlapping pairs, with the fine-tuned variants showing consistent, moderate gains. These findings further show that our alignment procedure improves robustness in extreme-view geometry settings while preserving strong performance in overlapping cases.

Beyond Relative Rotations. In the supplementary material, we also report translation accuracy on UnScenePairs-t. We find that while fine-tuning substantially improves rotation estimation, translation accuracy improvements are more modest. For example, the median translation error (reported in prior work [52]) slightly decreases from 37.28° to 35.79° after fine-tuning VGGT. We also observe the same behavior when supervising full poses rather than rotations exclusively. These results suggest that large translational displacements remain intrinsically challenging and represent an important direction for future work.

5.2. Multiview Pose Estimation

Experimental Details. We evaluate on two scene-scale datasets: RE10K [71], which contains camera trajectories from various video clips, and ETH3D [43], which contains multi-view indoor and outdoor scenes with high-precision ground-truth geometry and camera poses. Following prior work [51, 53, 57], we randomly sample and infer on 10 im-

Table 2. **Multiview Pose Estimation**. We evaluate camera pose angular accuracy over image collections from RealEstate10K [71] and ETH3D [43]. Non-negligible differences ($>$ 1 absolute) between the base and fine-tuned models are in **bold**.

Method	RealEstate10K			ETH3D		
	$RA_{30}\uparrow$	$TA_{30}\uparrow$	$AUC_{30}\uparrow$	$RA_{30}\uparrow$	$TA_{30}\uparrow$	$AUC_{30}\uparrow$
VGGT	99.91	92.75	77.28	96.41	87.01	72.52
$VGGT_{FT}$	99.99	**93.77**	**79.11**	96.41	**94.70**	**79.30**
WM	99.99	95.42	85.93	92.48	**91.45**	**78.24**
WM_{FT}	99.99	95.72	85.52	92.48	90.26	74.95
π^3	99.99	95.52	**87.14**	100.00	95.38	**82.81**
π^3_{FT}	99.99	95.21	85.25	100.00	95.21	79.81

Table 3. **Dense Reconstruction** on UnSceneRecon and ETH3D [43] datasets. The accuracy (ACC) and completion (CMP) are reported in meters. Non-negligible differences ($>5\%$ relative) between the base and fine-tuned models are in **bold**.

Method	UnSceneRecon				ETH3D			
	ACC $\downarrow$		CMP $\downarrow$		ACC $\downarrow$		CMP $\downarrow$	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
VGGT	1.441	1.049	1.403	0.729	0.284	0.194	0.342	0.204
$VGGT_{FT}$	**1.291**	**0.908**	**1.155**	**0.650**	**0.233**	**0.144**	**0.259**	**0.144**
WM	0.933	**0.612**	0.702	0.387	0.285	0.201	0.301	0.169
WM_{FT}	0.972	0.660	0.735	0.368	**0.235**	**0.167**	**0.217**	**0.126**
π^3	**0.716**	**0.466**	**0.635**	**0.377**	**0.187**	**0.125**	0.210	**0.129**
π^3_{FT}	0.791	0.517	0.689	0.403	0.202	0.137	0.219	0.137

ages per scene and compute metrics on all pairs.

Metrics. Following prior work [51, 53, 57], we report the 30° threshold for relative rotation accuracy (RA_{30}), relative translation accuracy (TA_{30}), and AUC (AUC_{30}).

Evaluation. Results are reported in Table 2. As shown, despite fine-tuning on image pairs and a rotation loss alone, rotation and translation accuracies are mostly preserved across all models, indicating that the aligned variants continue to perform well in multi-view settings. While WM and π^3 show slight decreases in AUC after fine-tuning, they remain highly competitive—likely nearing their performance ceiling in multi-view pose estimation, where non-overlapping images are rare. In contrast, VGGT exhibits substantial gains across all metrics. We hypothesize that, as the weakest initial model, VGGT has the greatest capacity for improvement, and hence our lightweight alignment scheme most effectively enriches its internal geometric representation. The dense reconstruction results presented next further support this observation.

5.3. Dense Reconstruction

Experimental Details. We evaluate on UnSceneRecon and ETH3D. For ETH3D, we follow π^3 and sample every 5th image. For UnSceneRecon, we sample 10 images per scene using a graph-based greedy algorithm to ensure sufficient visual overlap. Following prior work [55, 57], we aggregate point head predictions to assemble a point cloud per scene, and align the predicted points to the ground truth using the Umeyama algorithm, followed by iterative closest point.

Metrics. Standard accuracy (ACC) and completion (CMP) metrics are reported following prior work [53, 56, 57].

Evaluation. We report dense reconstruction results in Table 3. As shown in the table, our fine-tuned models preserve, and in many cases improve, the reconstruction performance of 3DFMs, despite using only a rotation loss and receiving no supervision on dense outputs. Fine-tuning VGGT yields the largest gains across both datasets, while WM, which starts from a stronger baseline, shows substantial improvements on ETH3D and only minor variations on UnSceneRecon. In contrast, π^3 exhibits some degradation, which we attribute to architectural differences: π^3 lacks a dedicated per-frame camera token like VGGT and WM, forcing camera information to be distributed throughout the image tokens. We hypothesize that a dedicated camera token helps preserve the model's internal langauge when fine-tuning on extreme rotations. Results over DTU [24] and 7Scenes [45] are provided in the supplementary material. These further validate that our alignment scheme enables preserving reconstruction performance.

5.4. Ablation Study

We perform a series of ablations to assess our alignment scheme over two key questions: (1) *Which components should be fine-tuned?* and (2) *How should the backbone be fine-tuned?* For all ablations, we report changes in extreme rotation error (Δ**ROT**) on *UnScenePairs* (using MRE) and changes in reconstruction accuracy (Δ**REC**) on *UnSceneRecon* (by averaging median ACC and CMP) to evaluate both extreme-view performance and dense reconstruction quality. We report results for VGGT and World-Mirror (WM); additional metrics and π^3 results (over a subset of ablations, as this model lacks depth predictions) are provided in the supplementary material. Overall, we observe that π^3 follows the same trends observed in WM.

Part 1: Which Components to Finetune?

Backbone unfreezing is essential. Our alignment strategy updates only the alternating attention backbone, refining the encoded 3D representation. To verify that this component must remain trainable, we freeze the backbone and fine-tune only the camera head ($\mathcal{D}_c$). As shown in the first row of each model block in Table 4, this setting yields degrading prediction for VGGT and very limited improvement for WM on extreme-view rotation estimation comparing to their pre-trained weights. Meanwhile, the point-head outputs also remain unchanged because the frozen backbone feeds iden-

Table 4. Ablation study evaluating rotation (△**ROT**) and reconstruction changes relative to pretrained models. We report reconstruction changes for both point-head predictions (△**REC**$_{\text{PH}}$) and depth-fused predictions (△**REC**$_{\text{Fused}}$). Part 1 studies which components to fine-tune. The camera head is denoted as $\mathcal{D}_c$ and the backbone as AA. Part 2 ablates update strategies for the AA backbone using layer-only (LO), bias-only (BO), or both (LO+BO) updates. Significant improvements ($> 10\%$ error drop) are shown in green, and degradations in red.

Model	Comp.	LO	BO	△ROT	△REC$_{\text{PH}}$	△REC$_{\text{Fused}}$	#Params
Part 1: Component Ablation							
VGGT	$\mathcal{D}_c$	×	×	6.8%	0.0%	0.0%	216.2M
	AA+$\mathcal{D}_c$	×	×	-74.3%	(N/A)	+90.3%	820.9M
	AA	×	×	-69.8%	-33.7%	-9.2%	604.7M
WM	$\mathcal{D}_c$	×	×	-1.4%	0.0%	0.0%	216.2M
	AA+$\mathcal{D}_c$	×	×	-47.5%	(N/A)	+81.5%	820.9M
	AA	×	×	-41.6%	+14.0%	+13.6%	604.7M
Part 2: AA-Only Ablation							
VGGT	AA	×	×	-69.8%	-33.7%	-9.2%	604.7M
	AA (LO)	✓	×	-66.7%	-33.3%	(N/A)	100.8M
	AA (BO)	×	✓	-69.7%	-16.7%	(N/A)	0.4M
	AA (LO+BO)	✓	✓	-59.8%	-12.4%	(N/A)	0.07M
WM	AA	×	×	-41.6%	+14.0%	+13.6%	604.7M
	AA (LO)	✓	×	-40.4%	+18.3%	(N/A)	100.8M
	AA (BO)	×	✓	-42.3%	+12.9%	(N/A)	0.4M
	AA (LO+BO)	✓	✓	-39.0%	+2.9%	(N/A)	0.07M

Figure 6. Fused VGGT predicted pointmaps from unprojecting depth and applying camera extrinsics on UnSceneRecon's *Wat Yai Chai Mongkhon* scene. The **left zoom-in** shows better alignment with a frozen camera head (AA), while the **right zoom-in** shows misalignment with an unfrozen camera head (AA+$\mathcal{D}_c$).

tical features into the dense prediction head. These results prove that the 3D representation is encoded entirely in the backbone rather than in the heads, and therefore the backbone must be unfrozen in order to refine it effectively.

Freezing the camera head preserves multi-task behavior. The camera head should remain frozen to maintain consistency in the model's shared internal 3D representation between geometry and pose reasoning. To verify this, we compare two alignment schemes: (1) unfreezing both the backbone and the camera head, and (2) unfreezing only the backbone. As shown in the second and third rows of Table 4, unfreezing the camera head consistently harms dense reconstruction quality for depth-fused predictions. For instance, freezing the camera head improves VGGT reconstructions by 9.2% on UnSceneRecon, while unfreezing it causes 90.3% degradation. As shown in Figure 6, modifying the camera head breaks the pretrained depth-pose consistency, highlighting the need to keep it frozen during alignment.

Part 2: How to Finetune the Backbone?

Selective bias-only tuning achieves the best trade-off. The backbone makes up the majority of a 3DFM's parameters, yet only a small subset needs to be adapted to effectively minimize rotation error. We find that the best options are to (1) tune only biases, in (2) select layers. To demonstrate that both are necessary for preserving the pretrained multi-task knowledge, we compare several configurations varying along these two dimensions: (1) *which layers* are updated, full-layer tuning vs. layer-only (LO) tuning, and (2) *which parameters* are updated, weight-and-bias tuning vs. bias-only (BO) tuning. The results of these four configurations are summarized in the lower half of Table 4.

Results reveal model-dependent behavior across update strategies. For WM, which already exhibit strong pretrained performance in earlier evaluations, tuning all weights across layers offers only marginal rotation gains but noticeably degrades reconstruction quality, indicating overfitting to the rotation objective. In contrast, updating only the bias terms of selected layers (LO+BO) achieves a comparable 39.0% reduction in rotation error while keeping reconstruction quality nearly unchanged (+2.9%), using just 0.07M parameters. Well-trained models like WM therefore benefit most from minimal adaptation, whereas models like VGGT with greater capacity for improvement can exploit larger updates to refine its internal 3D representation. Overall, these results indicate that fine-tuning only the bias terms of selected backbone layers provides the best balance between geometric refinement and preservation of pretrained multi-task behavior.

6. Conclusion

In this work, we study the emergent capabilities of 3D foundation models to understand extreme-view geometry, revealing that they encode a latent 3D language within their alternating attention backbone that can operate beyond visual overlap. We introduce a lightweight alignment scheme that enhances this internal 3D language through selective bias tuning, which modifies only around 80k parameters, four orders of magnitude smaller than the full model. These findings highlight the untapped potential of 3DFMs: even minimal adaptation unlocks strong 3D reasoning across extreme viewpoints, revealing low-parameter alignment as a promising direction for scalable, real-world 3D perception.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011. 2

[2] Roberto Cipolla Alex Kendall, Matthew Grimes. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV -International Conference on Computer Vision*, pages 2938–2946, 2015. 5

[3] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, 2022. 2

[4] Hana Bezalel, Dotan Ankri, Ruojin Cai, and Hadar Averbuch-Elor. Extreme rotation estimation in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1061–1070, 2025. 2, 4, 5, 6, 12, 15

[5] Srinadh Bhojanapalli, Ayan Chakrabarti, Andreas Veit, Michal Lukasik, Himanshu Jain, Frederick Liu, Yin-Wen Chang, and Sanjiv Kumar. Leveraging redundancy in attention with reuse transformers, 2021. 4

[6] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part VI*, page 611–625, Berlin, Heidelberg, 2012. Springer-Verlag. 18, 20

[7] Ruojin Cai, Bharath Hariharan, Noah Snavely, and Hadar Averbuch-Elor. Extreme rotation estimation using dense correlation volumes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14566–14575, 2021. 2, 4

[8] Ruojin Cai, Jason Y Zhang, Philipp Henzler, Zhengqi Li, Noah Snavely, and Ricardo Martin-Brualla. Can generative video models help pose estimation? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16764–16773, 2025. 2

[9] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3258–3268, 2021. 2

[10] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Easi3r: Estimating disentangled motion from dust3r without training. *arXiv preprint arXiv:2503.24391*, 2025. 2

[11] Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once, 2025. 2

[12] Shay Dekel, Yosi Keller, and Martin Cadik. Estimating extreme 3d image rotations using cascaded attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2588–2598, 2024. 2, 4

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[14] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *International Conference on 3D Vision 2025*, 2025. 2, 5, 12, 13, 18, 19

[15] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2

[16] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, Hanwen Jiang, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object, in any scene, with one reference. *arXiv preprint arXiv:2305.15727*, 2023. 2

[17] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM (CACM)*, 24(6):381–395, 1981. 2

[18] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[19] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, 2013. 18, 20

[20] Jisang Han, Honggyu An, Jaewoo Jung, Takuya Narihira, Junyoung Seo, Kazumi Fukuda, Chaehyun Kim, Sunghwan Hong, Yuki Mitsufuji, and Seungryong Kim. D^2ust3r: Enhancing 3d reconstruction for dynamic scenes, 2025. 2

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[22] Shwai He, Guoheng Sun, Zheyu Shen, and Ang Li. What matters in transformers? not all attention is needed, 2024. 4, 13

[23] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 15

[24] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413. IEEE, 2014. 2, 7, 17, 18, 20

[25] Jiachen Jiang, Jinxin Zhou, and Zhihui Zhu. Tracing representation progression: Analyzing and enhancing layer-wise similarity, 2025. 4

[26] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. Mapanything: Universal feed-forward metric 3d reconstruction, 2025. 2

[27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r, 2024. 2

[28] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024. 5, 12

[29] 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*, pages 2041–2050, 2018. 2, 6, 12

[30] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 2

[31] Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second, 2025. 2

[32] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching atokens are frozen. *arXiv preprint arXiv:2306.13643*, 2023. 2

[33] Yifan Liu, Zhiyuan Min, Zhenwei Wang, Junta Wu, Tengfei Wang, Yixuan Yuan, Yawei Luo, and Chunchao Guo. Worldmirror: Universal 3d world reconstruction with any-prior prompting, 2025. 1, 2, 3, 5, 18, 19, 20

[34] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 2

[35] Qing Mao, Tianxin Huang, Yu Zhu, Jinqiu Sun, Yanning Zhang, and Gim Hee Lee. Posecrafter: Extreme pose estimation with hybrid video synthesis. *arXiv preprint arXiv:2510.19527*, 2025. 2

[36] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 18, 20

[37] E. Palazzolo, J. Behley, P. Lottes, P. Giguère, and C. Stachniss. ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. *arXiv*, 2019. 18, 20

[38] Chris Rockwell, Justin Johnson, and David F Fouhey. The 8-point algorithm as an inductive bias for relative pose prediction by vits. In *2022 International Conference on 3D Vision (3DV)*, pages 1–11. IEEE, 2022. 2

[39] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[40] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[41] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2, 12

[42] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 12

[43] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2538–2547, 2017. 2, 6, 7

[44] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 2

[45] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images, 2013. 7, 17, 18, 20

[46] Samarth Sinha, Jason Y Zhang, Andrea Tagliasacchi, Igor Gilitschenski, and David B Lindell. Sparsepose: Sparse-view camera pose regression and refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21349–21359, 2023. 2

[47] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 2

[48] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[49] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. *arXiv preprint arXiv:2406.11819*, 2024. 2, 4, 5, 12, 15

[50] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21674–21684, 2025. 2

[51] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 2, 6, 7

[52] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment, 2024. 6, 16

[53] 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*, pages 5294–5306, 2025. 1, 2, 3, 5, 6, 7, 18, 19, 20

[54] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 2

[55] Qianqian Wang*, Yifei Zhang*, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025. 2, 7, 18

[56] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. *arXiv preprint arXiv:2312.14132*, 2023. 2, 7

[57] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua

Shen, and Tong He. π^3: Scalable permutation-equivariant visual geometry learning, 2025. 1, 2, 3, 5, 6, 7, 15, 17, 18, 19, 20

[58] Xiaoshi Wu, Hadar Averbuch-Elor, Jin Sun, and Noah Snavely. Towers of babel: Combining images, language, and 3d geometry for learning multimodal vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 428–437, 2021. 2

[59] Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features, 2025. 2, 5, 12, 13, 18, 19

[60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021. 12

[61] Jianing Yang, Alexander Sax, Kevin J Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 2

[62] Zhenpei Yang, Jeffrey Z Pan, Linjie Luo, Xiaowei Zhou, Kristen Grauman, and Qixing Huang. Extreme relative pose estimation for rgb-d scans via scene completion. In *CVPR*, 2019. 2

[63] Zhenpei Yang, Siming Yan, and Qixing Huang. Extreme relative pose network under hybrid representations. In *CVPR*, 2020. 2

[64] Zhenpei Yang, Zhile Ren, Miguel Angel Bautista, Zaiwei Zhang, Qi Shan, and Qixing Huang. Fvor: Robust joint shape and pose optimization for few-view object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2497–2507, 2022. 2

[65] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022. 4

[66] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 18

[67] Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In *European Conference on Computer Vision*, pages 592–611. Springer, 2022. 2

[68] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 2

[69] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21936–21947, 2025. 2

[70] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 12

[71] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 6, 7

Supplementary Material

We refer readers to the interactive visualizations at the accompanying `viewer.html` that show randomly-selected results for all three 3D foundation models (pretrained and fine-tuned) on the *relative rotation estimation* and the *dense reconstruction* test sets. In this document, we provide details regarding our proposed benchmark (Section A), additional implementation details (Section B) and provide additional experiments and results (Section C).

A. The MegaUnScene Benchmark

We first provide additional details relating to the curation of MegaUnScene (Section A.1). We then provide information on how we construct our three test sets: UnScenePairs (Section A.2), UnScenePairs-t (Section A.3), and UnSceneRecon (Section A.4).

A.1. Initial Curation

We provide additional details on the benchmark construction pipeline for MegaUnScene. First, we describe our construction pipeline, organized in four subsections: identifying scenes, sparse reconstruction, obtaining depth maps, and ensuring unseen scenes. Finally, we summarize MegaUnScene's overall scene statistics.

Scene identification. We first identify candidate scenes, each corresponding to an image collection, that we want to reconstruct. As mentioned in the main paper, we follow the MegaScenes [49] dataset curation pipeline to find scenes and their corresponding image collection from Wikimedia Commons and its sister site, Wikidata. In order to avoid scene overlap with the MegaScenes dataset, we query Wikidata with different high-level classes than in those used in MegaScenes. We filter out all Wikimedia Commons categories whose names intersect with MegaScenes, as category names are unique. This results in approximately 340,000 candidate scenes. For each scene, we follow MegaScenes and download images from all Wikimedia Commons subcategories with a max depth of four.

Sparse Reconstruction. We then reconstruct each candidate scene with at least 50 images using Doppelgangers++ (DGPP) [59] integrated with MASt3R-SfM [14] as mentioned in the main paper. In this pipeline, MASt3R [28] is used for image retrieval and matching, followed by match pruning using the DGPP classifier with a threshold of 0.8. We use COLMAP [41] for incremental SfM, manhattan-world alignment, and image undistortion. As Internet photos are noisy, not all images in the scene's image collection are registered to a reconstruction; thus we filter again for reconstructions that contain at least 50 images.

Obtaining Depth Maps. We obtain semi-dense depth maps with COLMAP's stereo fusion [42]. As described in MegaDepth [29], multi-view stereo depth maps typically contain artifacts from reconstruction ambiguities, especially on unconstrained internet photos. We clean these depth maps by following the same depth refinement protocol in MegaDepth. The protocol is slightly modified by replacing the PSPNet [70] segmentation model with the more recent SegFormer [60] in the semantic filtering step. For more details, please refer to Algorithm 1 of MegaDepth's supplementary material.

Ensuring Unseen Scenes. As a postprocessing step after reconstruction, we check for image conflicts between MegaUnScene and MegaScenes's base 9M image set (a superset of MegaScenes' 2M image subset that is reconstructed). This is possible as Wikimedia Commons ensures that each image has a unique filename. We only use MegaUnScene reconstructions with no image conflicts for UnSceneRecon, and reconstructions with less than 10% image conflict for both UnScenePairs and UnScenePairs-t. For UnScenePairs, we only select image pairs where neither image intersects with MegaScenes. For release, we note all conflicting images registered to reconstructions.

Benchmark statistics. From the dataset curation pipeline, we identify 758 reconstructions across 658 scenes with ≥ 50 images and $< 10\%$ image overlap with MegaScenes. Of these, we release 485 reconstructions across 476 scenes to be used for evaluation in our three new test sets: UnScenePairs, UnscenePairs-t, and UnSceneRecon. A breakdown of dataset statistics is provided in Table 5.

A.2. UnScenePairs Test Set

Prior work [4] introduced the wELP ("in-the-wild" Extreme Landmark Pairs) test set curated from MegaDepth [29]. However, all three 3D foundation models that we fine-tuned were pretrained on MegaDepth. To provide evaluation on a distinct data source but in the same camera-centric distribution, we follow the same pipeline to filter image pairs on MegaUnScene.

As introduced in our paper, the pipeline identifies image pairs with negligible translation and predominant rotation using mutual K-nearest neighbor graphs ($K = 5$) constructed from the distances between camera translation. Mutual neighbors ensure that only pairs that are consistently close in translation space are preserved. Each surviving image pair is then assigned an overlap level (Large, Small, and None) with the following algorithm:

Given the relative rotation matrix $\mathbf{R}$ between two cam-

Table 5. **MegaUnScene Statistics.** Statistics for our benchmark and three MegaUnScene test sets: UnScenePairs, UnScenePairs-t, and UnSceneRecon. For UnScenePairs, and UnScenePairs-t, we report the number of image pairs extracted for each overlap level. For UnSceneRecon, we report the number of reconstructions. At the bottom, we summarize the total number of unique MegaUnScene scenes and reconstructions across the three test sets.

			#Pairs			
Subset	# Scenes	K	Large	Small	None	Total
UnScenePairs	458	5	1,878	1,227	778	3,883
UnScenePairs-t	387	50	1,146	523	763	2,432

Subset	# Scenes	# Recons	Notes
UnSceneRecon	96	100	Human-annotated metric scale

Overall	# Scenes	# Recons	Notes
MegaUnScene	476	485	Counts after de-duplication

eras and their respective FoVs, the overlap category o is determined by:

$$o = \begin{cases} \textit{Large} & |\gamma| < \frac{\text{fov}_x^1 + \text{fov}_x^2}{4} \wedge |\beta| < \frac{\text{fov}_y^1 + \text{fov}_y^2}{4} \\ \textit{None} & |\gamma| > \frac{\text{fov}_x^1 + \text{fov}_x^2}{2} \wedge |\beta| > \frac{\text{fov}_y^1 + \text{fov}_y^2}{2} \\ \textit{Small} & \text{otherwise} \end{cases} \tag{4}$$

where γ and β denote the relative yaw and pitch angles extracted from $\mathbf{R}$ using Euler decomposition.

After the pipeline, we manually review all selected pairs and removed those affected by motion blur, occlusions, or insufficient geometric structure. The resulting UnScenePairs statistics is shown in Table 5.

A.3. UnScenePairs-t Test Set

As discussed in our paper, we also construct *UnScenePairs-t* from MegaUnScene with the same pipeline but use $K = 50$ mutual nearest neighbors to evaluate performance on pairs with larger camera translations. We then implement correspondence-based verification using Doppelgangers++ [59] + MASt3R-SfM [14]'s reconstructed geometry and checking the verified inlier matches.

Additionally, with larger camera translations included as a consequence of setting $K = 50$, the same scene structure may appear at vastly different scales depending on camera-to-scene distance, where a telephoto lens from afar and a wide-angle lens nearby could capture the same 3D structure at incompatible image resolutions. To ensure scale consistency, we extend the basic FoV threshold with three criteria: (1) *both* horizontal and vertical FoV differences below 15° independently, (2) focal length ratio $\max(f_x^1, f_x^2)/\min(f_x^1, f_x^2) < 2.5$ to catch zoom differences, and (3) image resolution ratio $\max(w_1h_1, w_2h_2)/\min(w_1h_1, w_2h_2) < 3.0$ to prevent sensor size discrepancies from obscuring focal length mismatches.

Finally, similar to UnScenePairs filtering, we also manually inspected UnScenePairs-t and removed noisy image pairs. Statistics are shown in Table 5.

A.4. UnSceneRecon Test Set

We construct a user interface for human annotators to annotate each reconstruction, as depicted in Figure 7. Annotators are first instructed to visually assess reconstruction quality to see determine if the reconstruction is realistic based on the images; they label the reconstruction "good" or "bad" accordingly. If a reconstruction is good, they are instructed to draw a line on the reconstruction, find the corresponding points on Google Maps in satellite view, and annotate the metric scale (as shown on Figure 7). In practice, we instruct annotators to only label one line to estimate the metric scale. From this process, we label 100 reconstructions across 96 scenes with metric scale annotations, as shown in Table 5.

B. Implementation Details

We first describe how we select backbone layers for finetuning (Section B.1), then outline the construction of the training set (Section B.2), followed by our training configuration (Section B.3), and finally provide the full evaluation protocols used across all tasks (Section B.4).

B.1. Backbone Layer Selection

To quantify the degree of representational change between neighboring backbone layers, we follow the layer similarity pipeline conducted by prior work [22] and run forward passes on ten image pairs, measuring the similarity between the input and output token representations of each layer. The input and output tokens $\mathbf{T}_l^{\text{in}}$ and $\mathbf{T}_l^{\text{out}}$ correspond to either the frame tokens $\mathbf{T}_{\text{frame}}^{(i)}$ or the global tokens $\mathbf{T}_{\text{global}}$, as defined in the method section. For a layer l, we compute the cosine similarity:

$$\text{sim}_l = \frac{\mathbf{T}_l^{\text{in}} \cdot \mathbf{T}_l^{\text{out}}}{\|\mathbf{T}_l^{\text{in}}\| \cdot \|\mathbf{T}_l^{\text{out}}\|}. \tag{5}$$

We run the pipeline on the frame and global attention blocks of VGGT, WorldMirror (WM), and π^3 separately, and the resulting similarity curves are shown in Figure 8. As can be observed in the figure, the curves of WM and VGGT exhibit similar similarity drops, which is expected given that WM inherits both the architecture and weight initialization of VGGT. We also find that the similarity drop regions—where the curve shows a pronounced decline—coincide with the intermediate layers commonly used for dense predictions, namely layers 4, 11, 17, and 23. We therefore adopt this fixed set for these models.

Figure 7. **UnSceneRecon Reconstruction Annotator**. We show the annotator webpage on MegaUnScene's Krzyżtopór Castle scene (left) and its corresponding Google Maps satellite view (bottom right). At the top-left of the viewer, we depict a randomly sampled set of 10 images. On the bottom-left, we show the corresponding 3D model from unprojecting the depths of these 10 images in the global coordinate frame. On the top right of the page, annotators label whether the reconstruction is good or not, and also make a metric estimate of the reconstruction. The metric estimate is done by drawing a line on the reconstruction (shown at the bottom-left, at the top of the building's 3D model in red), measuring the corresponding distance in Google Maps (shown on the bottom right), and pasting the measurement in the corresponding field on the top-right of the viewer. In this example, the annotator labeled "Yes" that the reconstruction is good, and that the annotated red line is 27.14 meters. We provide links to the Wikimedia Commons page, as well as a Google Maps page that searches the scene name, to help annotators identify the correct location on Google Maps.

Figure 8. **Layer Analysis.** Cosine similarity curves of the input–output representations for pretrained VGGT, WorldMirror (WM), and π^3. For VGGT and WM, the curves span 24 layers, where each layer corresponds to a frame-global block pair. For π^3, the curve spans 18 layers. Our fine-tuning focuses on layers with pronounced similarity drops; see Section B.1 for additional details.

For π^3 which doesn't include such skip connections, we select layers by detecting local minima in the similarity curve (using peak detection on the inverted signal) and expanding around each minimum to include adjacent layers with low similarity scores. This ensures that both the most transformative layers and their contextually relevant neighbors are captured. The selection criterion is defined as:

$$\mathcal{L} = \bigcup_{i \in \mathcal{M}} \left\{ i \pm k : s_{i \pm k} \leq \bar{s} - \frac{\sigma_s}{2},\ k \in [1, \delta] \right\}, \qquad (6)$$

where $\mathcal{M}$ is the set of detected local minima, $\bar{s}$ and σ_s are the mean and standard deviation of similarity scores, and δ controls the neighborhood expansion radius. In practice, we use $\delta = 2$. This yields a selected subset which includes frame layers 4, 12–16 and global layers 13–15.

B.2. MegaScenes Train Set

For the train set, we use the same pipeline described in Section A.2 with $K = 50$ to filter image pairs from scene-level COLMAP reconstructions in MegaScenes [49] and employ a balanced subsampling strategy to ensure uniform pair selection across overlap categories. We first cap each scene at a maximum of 40 pairs to prevent scene-level bias, then subsample to achieve exact balance across the three overlap categories. The final train set contains 64,584 image pairs (21,528 for each overlap category) across 3,284 scenes.

B.3. Training Details

We use the same training configuration for all three models. Training is performed with the AdamW optimizer using a learning rate of 5×10^{-5} and a weight decay of 1×10^{-4}. We train on four NVIDIA RTX A6000 GPUs in a distributed setting with a per-GPU batch size of 1. For each selected layer, we update only the bias parameters in the attention and MLP modules—specifically, the biases of the query–key–value projection (attn.qkv.bias), the attention output projection (attn.proj.bias), and the two MLP fully connected layers (mlp.fc1.bias and mlp.fc2.bias). To stabilize training, we apply gradient clipping with a maximum norm of 1.0 to mitigate gradient explosion.

Additionally, for the ablation checkpoints that unfreezes both weights and biases of a backbone layer or the camera head, we apply LoRA [23] with rank = 24, alpha = 48, and dropout = 0.1 for parameter efficient fine-tuning.

B.4. Evaluation Protocols

In this subsection, we provide additional details on our evaluation protocols and settings.

General Evaluation Settings. For all non-ablation evaluations, the base models we use for VGGT, WorldMirror, and π^3 are the publicly available checkpoints from HuggingFaceHub with the following names:

- `facebook/VGGT-1B`
- `tencent/HunyuanWorld-Mirror`
- `yyfz233/Pi3`

The fine-tuned models are the Layer-Only (LO) and Bias-Only (BO) checkpoints.

Relative Rotation Evaluation Settings. We preprocess input images using each architecture's provided functions: WorldMirror and VGGT crop images to width 518; π^3 proportionally scales with a pixel limit, preserving aspect ratio. All three ensure image width and height are multiples of 14 after preprocessing. The predicted quaternions are converted into rotational matrices for evaluation.

For ExRot [4], we evaluate on their publicly available model on their GitHub page. For all datasets, the images are downsized such that the longer dimension is 256 pixels, then center zero-padded to be size 256x256.

Multiview Pose Estimation, Monocular Depth, and Dense Reconstruction Settings. We also preprocess input images with each architecture's provided function. For multview pose estimation and monocular depth, we follow π^3's [57] protocol and downsize images to a target size of 512 pixels, with dimensions adjusted to be divisible by 14 through rounding. For dense reconstruction, the target size is 518 pixels. As UnSceneRecon is the only dataset with variable aspect ratio, we downsize the longest edge to 518

and center zero-pad to be 518x518 pixels for the dense reconstruction evaluation.

UnSceneRecon Graph-Based Image Sampling for Dense Reconstruction Evaluation. In the main paper, we mention that we subsample images from UnSceneRecon using a graph-based greedy algorithm for dense reconstruction evaluations. This is because UnSceneRecon scenes typically have widely distributed camera poses that capture different portions of the scene. Random image sampling often leads to poor overlap—*i.e.*, selecting images from disjoint locations on opposite sides of the scene—that are implausible for reconstruction pipelines to reasonably reconstruct. Our graph-based approach ensures connectivity across images while maintaining diversity.

We construct an image connectivity graph for each sparse reconstruction, where nodes represent images and edges connect image pairs with at least 30 shared 3D points and a translation of at least 5 meters. We initialize with a random node in the largest connected component, then greedily sample images based by selecting neighbors that maximize a score combining 80% connectivity and 20% diversity. Here, "connectivity" is the normalized node degree (degree divided by the maximum degree in the graph); "diversity" is the average distance from a candidate to the nodes of all selected images, normalized by the maximum translation across all edges in the graph.

C. Experiments and Results

We begin by evaluating relative camera pose across both overlapping and non-overlapping settings (Section C.1), then present additional dense reconstruction experiments on multiple benchmarks (Section C.2), followed by monocular depth evaluations (Section C.3), and conclude with expanded ablation studies analyzing alternative fine-tuning strategies (Section C.4).

C.1. Relative Camera Pose

Evaluations on Large/Small Overlapping Pairs. As shown in Table 6, we also evaluate the three fine-tuned models on large and small overlapping image pairs from *s*ELP, UnScenePairs, and UnScenePairs-t. All models achieve comparable, and sometimes slightly improved, rotation accuracy. This demonstrates that our fine-tuning procedure does not compromise performance on overlapping image pairs. Since the pretrained models already produce strong rotation estimates when overlap is present, fine-tuning preserves this capability.

Translation Evaluation on UnScenePairs-t. For UnScenePairs-t, ground-truth relative translations are also available. Following prior work [52], we evaluate translation accuracy using the angular error. Let $\mathbf{t}_{21}$ and $\mathbf{t}^{\star}_{21}$ denote the predicted and ground-truth translation vectors from camera 2 to camera 1. The error is defined as

$$\mathrm{err}_T = \arccos\left(\frac{\left|\mathbf{t}_{21}^{\top}\mathbf{t}^{\star}_{21}\right|}{\|\mathbf{t}_{21}\|\;\|\mathbf{t}^{\star}_{21}\|}\right). \tag{7}$$

As shown in Figure 6, we report the Median Translation Error (MTE) and the Translation Accuracy (TA) at thresholds of 15° and 30°, denoted TA_{15} and TA_{30}. Since our loss only supervises over the predicted rotation, the translation error and accuracy stay the roughly the same after fine-tuning, with small improvement on the non-overlapping pairs for all three models. As discussed in the paper, this suggests that all fine-tuning framework performs minimal updates to the pretrained weights without overfitting to predicting extreme relative rotation.

Supervising over Full Pose. We also experiment with adding a loss term to supervise relative translation in addition to rotation. Each of the three models addresses scale ambiguity differently: VGGT normalizes all camera translations in the training data, WorldMirror provides a normalized camera pose through prior prompting, and π^3 computes a scale factor by aligning its predicted point map with the ground truth. We follow VGGT by normalizing each translation vector using the mean distance from the ground truth sparse 3D points to the point-cloud center, noting that VGGT measures distances to the origin while our dataset does not anchor the first image. Additionally for π^3, we adopt a similar scaling strategy. However, with only two input images, aligning a partial predicted point map to the full ground-truth scene is unstable. We instead apply a simple scaling heuristic. For each image pair i, let $\mathbf{t}^i_{\mathrm{pred}}$ and $\mathbf{t}^i_{\mathrm{gt}}$ denote the predicted and ground-truth relative translations in world coordinate. We compute a scale factor

$$s_i = \frac{\left\|\mathbf{t}^i_{\mathrm{gt}}\right\|_2}{\left\|\mathbf{t}^i_{\mathrm{pred}}\right\|_2}, \tag{8}$$

and rescale the prediction as

$$\tilde{\mathbf{t}}^i_{\mathrm{pred}} = s_i\,\mathbf{t}^i_{\mathrm{pred}}, \tag{9}$$

We add an L1 loss on the translation vectors and keep the same geodesic loss for rotations. For VGGT and WorldMirror, the translation loss anchors the first image as the reference frame and compute the loss on the two absolute predicted translations where the first image's translation should be [0,0,0] in world coordinate. For π^3, we instead compute the loss on the relative translation vectors. Results in Table 7 show that this additional translation supervision provides essentially no improvement in translational or rotational accuracy, and occasionally will lead to worse performances. This illustrates that our proposed rotation-based objective can better align models to extreme-view geometries, in comparison to the full pose objective used by prior

Table 6. Expanded comparison of sELP, UnScenePairs, and UnScenePairs-t benchmarks across VGGT, WorldMirror (WM), π^3, and their fine-tuned variants. MRE and MTE report the median rotation and translation errors in degrees. RA_{15}/RA_{30} and TA_{15}/TA_{30} indicate the percentage of predictions whose rotation or translation errors are below 15° or 30°, respectively.

		sELP			UnScenePairs			UnScenePairs-t					
Overlap	Method	MRE	RA_{15}	RA_{30}	MRE	RA_{15}	RA_{30}	MRE	RA_{15}	RA_{30}	MTE	TA_{15}	TA_{30}
Large	VGGT	0.75	99.7	99.7	0.99	99.7	99.8	1.07	99.9	100.0	2.44	92.7	97.0
	$VGGT_{FT}$	0.95	99.9	100.0	1.04	99.6	99.7	1.08	99.9	99.9	2.24	94.0	97.6
	WM	0.58	99.7	99.7	1.23	96.8	99.7	1.20	99.0	99.7	3.08	87.4	94.3
	WM_{FT}	0.82	99.7	99.7	1.40	98.2	99.5	1.15	99.4	99.7	3.34	88.7	94.6
	π^3	0.68	99.8	99.8	1.24	96.8	99.2	1.09	99.2	99.8	3.39	86.7	94.1
	π^3_{FT}	0.93	99.9	99.9	1.23	99.4	99.8	0.99	99.9	100.0	3.05	88.7	95.9
Small	VGGT	1.86	96.9	97.8	1.93	96.7	98.0	2.01	98.8	99.8	9.69	62.7	74.0
	$VGGT_{FT}$	2.12	98.8	99.4	1.95	97.5	98.8	2.04	100.0	100.0	9.18	61.4	75.1
	WM	1.20	97.7	98.3	2.37	95.4	98.6	1.98	97.7	99.8	11.15	55.8	72.8
	WM_{FT}	1.92	98.1	98.2	2.65	96.5	98.9	2.39	97.9	99.6	11.75	56.4	72.1
	π^3	1.50	97.3	97.7	2.31	94.4	98.5	1.77	98.1	99.8	9.68	60.6	75.3
	π^3_{FT}	2.29	98.4	98.7	2.43	96.7	99.1	1.99	99.0	100.0	10.91	57.5	75.5
None	VGGT	92.92	24.2	29.1	31.64	33.8	48.8	46.65	29.1	42.1	37.28	25.3	42.2
	$VGGT_{FT}$	14.21	50.9	56.5	12.71	53.6	67.9	14.48	50.6	62.1	35.79	26.5	44.0
	WM	68.96	36.3	42.5	19.25	44.1	58.9	21.52	42.6	57.4	33.83	27.6	45.2
	WM_{FT}	9.74	56.9	63.5	11.75	56.2	68.1	13.13	53.3	64.5	33.42	27.8	46.7
	π^3	45.24	43.8	48.3	17.66	46.5	59.4	21.62	43.5	56.8	33.16	29.4	45.7
	π^3_{FT}	11.96	53.7	60.0	12.92	54.0	69.2	13.31	53.1	65.5	32.05	31.4	47.7

work. As mentioned in the paper, our finding also shows that predicting large translational displacements between two images is intrinsically hard and remains to be an important future line of work.

C.2. Dense Reconstruction

Additional Results: DTU and 7Scenes. We provide additional dense reconstruction experiments on the object-centric DTU [24] and indoor 7Scenes [45] datasets. We follow π^3 [57] and sample every 5th image from DTU (10 images per scene). For 7Scenes, we sample every 200th image, corresponding to π^3's dense-view evaluation setting (16 scenes with 25 images, 2 scenes with 13 images). We report accuracy (ACC) and completion (COMP) as in the main paper.

We depict results in Table 9. As shown on DTU, despite fine-tuning on Internet photos, all our finetuned models are able to generalize to an object-centric dataset: VGGT shows minimal change from fine-tuning, while WM and π^3 show minimal performance loss. Furthermore, on 7Scenes, the reconstruction performance for all models is extremely similar before and after fine-tuning. This is validated by the qualitative results in Figure 9. Thus, our finetuned models still generalize to indoor scenes with minimal impact on dense image inputs.

Additional Results: UnSceneRecon. We show additional qualitative results on UnSceneRecon in Figure 10; input images and 3D model visualizations for π^3 are shown in the accompanying `viewer.html`. These scenes are selected from the 100 reconstructions in UnSceneRecon using a random sampler. As shown, there is negligible difference in reconstruction quality between the base and finetuned models. Note that when the base model reconstructs poorly, the finetuned model does too, as exemplified by VGGT and WorldMirror on Kamerlengo-0 and Naubat Khana (Red Fort)-0. The scale alignment of $VGGT_{FT}$ appears much worse than base VGGT on Kamerlengo-0, since automatic alignment has greater variance when aligning incorrect point clouds to the ground truth. These reconstructions demonstrate the difficulty that current 3DFMs have in accurately reconstructing Internet photos, emphasizing the importance of a test set of real-world, unconstrained settings.

Figure 9. **DTU and 7Scenes Examples**. We show reconstruction results from the base and finetuned VGGT [53], WorldMirror (WM) [33], and π^3 [57] models on DTU's [24] scan1 and scan4 and 7Scenes's [45] chess-seq03 and fire-seq03. The ground-truth reconstruction, obtained using Doppelgangers++ [59] and MASt3R-SfM [14] as further detailed in the text, are shown in the first column. The predicted scenes are automatically aligned to the ground truth per the evaluation protocol discussed in the main paper.

C.3. Monocular Depth Estimation

We additionally evaluate monocular depth to determine if fine-tuning alters the performance of the 3DFM's dense prediction heads, as it directly alters the internal representations they decode.

Experimental Details. We test four datasets: Sintel [6], Bonn [37], KITTI [19], and NYU-v2 [36]. Sintel and Bonn contain synthetic scenes; KITTI contains real outdoor driving scenes; NYU-v2 contains real indoor scenes. Following prior work [55, 57, 66], we align each predicted depth to the ground truth with per-frame median scaling. For VGGT and WorldMirror, we directly use the depth head outputs for evaluation. For π^3, since the model does not have a depth head, we obtain the depth by taking the z-values of the model's point map prediction.

Metrics. We report absolute relative error (AbsRel) and threshold accuracy below 1.25 (δ_1) like in [55, 57, 66].

Results. We show monocular depth results in Table 8. Remarkably, all fine-tuned models perform similarly to their base counterparts across all datasets, with VGGT demonstrating minor improvements. This indicates that the frozen dense prediction heads remain effective at decoding the altered internal representations, despite fine-tuning with only rotation loss and no depth supervision.

C.4. Extra Ablations

We show an expanded version of our ablation results in Table 10, with metrics for VGGT [53], WorldMirror (WM) [33], and π^3 [57]. For clarity, we show columns for whether models are trained on select layers only (**LO**) and bias only (**BO**); we additionally denote whether the reconstruction metrics use the point head or not (using fused point maps from unprojected depth) in the **PH** (point head) column. We use the same $\triangle$**REC**$_{\text{PH}}$ and $\triangle$**REC**$_{\text{Fused}}$ metrics as in the main paper's ablation table. We show the same rotation metrics: **MRE**, **RA**$_{15}$, and **RA**$_{30}$, as well as the median reconstruction metrics: **ACC** and **COMP**, as discussed in the main paper. **REC** is the average of **ACC** and **COMP**.

As discussed in the main paper, π^3 performs similarly to WorldMirror (WM) in the ablations: regarding how to fine-tune the backbone, using select-layers only and bias-only provides a good trade-off between rotation and reconstruction performance (-26.8 for $\triangle$**ROT** and 9.2 for $\triangle$**REC**). Switching the former option to *all layers* or the latter option to *weights and biases* leads to a performance degradation in **REC** (from a 9.2 $\triangle$**ROT** to 14.0 and 15.8, respectively). Interestingly, our fine-tuned π^3 model on all layers and weights and biases exhibits abnormally strong performance (a $\triangle$**ROT** of -37.7 and $\triangle$**REC** of 9.9), and is an outlier in the trends we see across all three models.

We also show π^3's fine-tuning results when unfreez-

Figure 10. **Random UnSceneRecon Examples**. We show reconstruction results from the base and finetuned VGGT [53], WorldMirror (WM) [33], and π^3 [57] models on five randomly selected UnSceneRecon scenes. The selected scenes (top to bottom) are Pampanga Provincial Capitol-0, Kamerlengo-0, Bodiam Castle-0, Marmashen Monastery-1, and Naubat Khana (Red Fort)-0. The ground-truth reconstruction, obtained using Doppelgangers++ [59] and MASt3R-SfM [14] as further detailed in the text, are shown in the first column. The predicted scenes are automatically aligned to the ground truth per the evaluation protocol discussed in the main paper.

ing only the camera head $\mathcal{D}_c$. Unlike VGGT [53] and WM [33], which directly infer 3D points in global space, π^3 uses the extrinsic predictions of the camera head to transform predicted points in local coordinates to global coordinates. Consequently, we see that fine-tuning the camera head leads to worse $\triangle$**REC** metrics (567.4) compared other fine-tuning schemes; this indicates that the performance of $\mathcal{D}_c$ is destroyed. At the same time, we see that $\mathcal{D}_c$ does not have much capacity to align to our target task of extreme rotation estimation, reflected by a negligible $\triangle$**ROT** of 4.1.

Table 7. Evaluation of fine-tuned VGGT, WM, and π^3 models trained with additional translation loss (TL) on the UnScenePairs-t test set. Comparing with our final fine-tuned checkpoints (FT), the results show that translation supervision offers no improvement in relative translation and rotation accuracy.

	Method	UnScenePairs-t					
		MRE	RA_{15}	RA_{30}	MTE	TA_{15}	TA_{30}
Large	$VGGT_{FT}$	1.08	99.9	99.9	2.24	94.0	97.6
	$VGGT_{TL}$	1.15	99.8	99.8	3.08	93.0	96.9
	WM_{FT}	1.15	99.4	99.7	3.34	88.7	94.6
	WM_{TL}	1.22	99.4	99.7	3.85	88.0	94.2
	π^3_{FT}	0.99	99.9	100.0	3.05	88.7	95.9
	π^3_{TL}	1.52	99.8	100.0	6.93	80.6	92.8
Small	$VGGT_{FT}$	2.04	100.0	100.0	9.18	61.4	75.1
	$VGGT_{TL}$	2.36	99.8	100.0	10.34	59.1	73.6
	WM_{FT}	2.39	97.9	99.6	11.75	56.4	72.1
	WM_{TL}	2.43	97.3	99.8	12.47	54.3	72.1
	π^3_{FT}	1.99	99.0	100.0	10.91	57.5	75.5
	π^3_{TL}	2.38	99.0	100.0	14.71	51.8	70.0
None	$VGGT_{FT}$	14.48	50.6	62.1	35.79	26.5	44.0
	$VGGT_{TL}$	14.75	50.2	61.6	37.75	22.2	39.5
	WM_{FT}	13.13	53.3	64.5	33.42	27.8	46.7
	WM_{TL}	13.36	52.0	63.0	38.50	23.2	40.2
	π^3_{FT}	13.31	53.1	65.5	32.05	31.4	47.7
	π^3_{TL}	15.50	49.0	63.8	33.24	27.8	46.0

Table 8. **Monocular Depth Estimation** on Sintel [6], Bonn [37], KITTI [19], and NYU-v2 [36] datasets. Non-negligible differences (> 5% relative) between the base and fine-tuned models are in **bold**.

Method	Sintel		Bonn		KITTI		NYU-v2	
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑
VGGT [53]	0.335	0.597	0.053	0.970	0.082	0.947	0.056	0.951
$VGGT_{FT}$	**0.316**	0.621	**0.050**	0.974	**0.077**	0.952	**0.052**	0.955
WM [33]	0.340	0.625	0.066	0.963	0.093	0.930	0.053	0.957
WM_{FT}	0.336	0.633	0.064	0.964	0.089	0.933	0.053	0.957
π^3 [57]	0.280	0.617	0.048	0.974	0.059	0.971	0.054	0.956
π^3_{FT}	0.287	0.596	0.048	0.974	0.061	0.969	0.054	0.956

Table 9. **Dense Reconstruction** on DTU [24] and 7Scenes [45] datasets. Non-negligible differences (> 5% relative) between the base and fine-tuned models are in **bold**.

Method	DTU				7Scenes			
	ACC ↓		CMP ↓		ACC ↓		CMP ↓	
	Mean	Med.	Mean	Med.	Mean	Med.	Mean	Med.
VGGT [53]	1.185	0.716	2.215	1.302	0.020	0.007	0.030	0.014
$VGGT_{FT}$	1.178	0.711	2.189	1.256	**0.018**	0.007	0.029	0.014
WM [33]	**1.033**	**0.573**	1.759	0.790	**0.016**	0.007	0.028	0.013
WM_{FT}	1.121	0.614	1.761	0.783	0.017	0.007	0.027	**0.012**
π^3 [57]	**1.152**	**0.622**	1.797	0.631	0.016	0.007	0.022	0.011
π^3_{FT}	1.396	0.748	1.768	0.634	0.016	0.007	**0.020**	**0.008**

Table 10. **Full Ablation Table.** The full table for the ablation evaluating rotation (△**ROT**) and reconstruction (△**REC**) changes relative to pretrained models. The camera head is denoted as $\mathcal{D}_c$ and the backbone as AA. We show all fine-tuned results for layer-only (LO), bias-only (BO), or both (LO+BO) updates. We denote whether we use the point head or not (using fused unprojected depths) in the **PH** column for reconstruction metrics. **REC** is the average of **COMP** and **ACC**; we report median values only. The reconstruction delta △**REC** is **REC**'s percent change compared to the base model. Significant improvements (> 10% error drop) for △**ROT** and △**REC** are shown in green, and degradations in red.

Model	Comp.	LO	BO	PH	△ROT	MRE	RA_{15}	RA_{30}	$\triangle REC_{PH}$	$\triangle REC_{Fused}$	REC	ACC	COMP	#Params
VGGT	$\mathcal{D}_c$	×	×	✓	6.8	33.79	28.2	46.4	0.0	(N/A)	0.889	1.049	0.729	216.2M
	AA+$\mathcal{D}_c$	×	×	×	-74.3	8.14	65.5	78.5	(N/A)	90.3	1.692	1.384	2.000	820.9M
	AA	×	×	×	-69.8	9.57	60.4	73.9	(N/A)	-9.2	0.808	0.961	0.654	604.7M
	AA	×	×	✓	-69.8	9.57	60.4	73.9	-33.7	(N/A)	0.590	0.687	0.493	604.7M
	AA	✓	×	✓	-66.7	10.55	57.5	70.8	-33.3	(N/A)	0.593	0.727	0.459	100.8M
	AA	×	✓	✓	-69.7	9.60	61.3	75.6	-16.7	(N/A)	0.741	0.912	0.570	0.4M
	AA	✓	✓	✓	-59.8	12.71	53.6	67.9	-12.4	(N/A)	0.779	0.908	0.650	0.07M
WM	$\mathcal{D}_c$	×	×	✓	-1.4	18.98	43.8	60.7	0.0	(N/A)	0.500	0.612	0.387	216.2M
	AA+$\mathcal{D}_c$	×	×	×	-47.5	10.10	60.7	75.8	(N/A)	81.5	0.907	1.087	0.727	820.9M
	AA	×	×	×	-41.6	11.24	58.0	71.3	(N/A)	13.6	0.567	0.704	0.431	604.7M
	AA	×	×	✓	-41.6	11.24	58.0	71.3	14.0	(N/A)	0.570	0.716	0.423	604.7M
	AA	✓	×	✓	-40.4	11.48	55.9	70.0	18.3	(N/A)	0.591	0.719	0.463	100.8M
	AA	×	✓	✓	-42.3	11.11	56.9	70.6	12.9	(N/A)	0.564	0.702	0.426	0.4M
	AA	✓	✓	✓	-39.0	11.75	56.2	68.1	2.9	(N/A)	0.514	0.660	0.368	0.07M
π^3	$\mathcal{D}_c$	×	×	✓	4.1	18.39	44.0	59.3	567.4	(N/A)	2.812	2.755	2.869	2.1M
	AA	×	×	✓	-37.7	11.00	59.5	73.0	9.9	(N/A)	0.463	0.501	0.425	453.5M
	AA	✓	×	✓	-31.3	12.13	55.0	69.8	14.0	(N/A)	0.480	0.562	0.398	113.4M
	AA	×	✓	✓	-39.6	10.66	60.2	73.8	15.8	(N/A)	0.488	0.555	0.421	0.3M
	AA	✓	✓	✓	-26.8	12.92	54.0	69.2	9.2	(N/A)	0.460	0.517	0.403	0.08M