Assessing the Reconstruction Potential of 3D Vision Foundation Models for Oblique Photogrammetry

Junfan Wang[1], Feng Liu[2,*], Zhihao Jia[1], Han Hu[1,3], Min Chen[1,3], Xuming Ge[1,3], Ping Wen[3,4], Chong Wang[3,4], Qing Zhu[1,3]

[1]Faculty of Geosciences and Engineering, Southwest Jiaotong University, 611756 Chengdu, China
[2]CRSC Communication & Information Group Co., Ltd.
[3]Yunnan Engineering Research Center of 3D Real Scene, Kunming 650500, China
[4]Kunming Engineering Corporation Limited, Kunming 650500, China

Keywords: 3D Vision Foundation Model, Oblique Imagery, Photogrammetric Reconstruction, Evaluation

Abstract

3D vision foundation models, which directly regress 3D geometry from 2D images in an end-to-end manner, have recently attracted growing attention in the computer vision community. However, their potential for oblique 3D reconstruction has not been systematically evaluated. To this end, we establish an automated evaluation pipeline to benchmark these models on oblique imagery. Our experiments reveal that: benefiting from the powerful zero-shot generalization, 3D vision foundation models can robustly estimate camera parameters and generate dense point clouds under sparse-view and low-overlap conditions, with some rivaling traditional photogrammetry configured with redundant observations. Counterintuitively, two-view reasoning foundation models employing explicit PnP-RANSAC for global alignment consistently outperform multi-view reasoning foundation models inferring multi-view relationships via implicit attention mechanism when processing more than 2 views. Notably, incorporating known camera parameters as conditioning inputs, which act as weak supervision rather than rigid geometric constraints, yields only marginal accuracy improvements. Based on ViT architecture, these foundation models face scalability bottlenecks to large-scale and high-resolution oblique imagery, and their prevalent ideal pinhole camera assumption still makes explicit distortion correction an unavoidable preprocessing step.

1. INTRODUCTION

Oblique photogrammetry is a fundamental technology that captures scenes from multiple tilted viewing angles, enabling the simultaneous acquisition of both the top and side textures of ground objects (Remondino and Gerke, 2015). This capability greatly improves the accuracy and realism of 3D reconstruction, and therefore, oblique photogrammetry has been applied in various domains such as urban modeling, disaster assessment, agricultural management, and other industries (Zhang et al., 2020).

Traditional photogrammetric pipelines, typically composed of Structure-from-Motion (SfM) (Schonberger and Frahm, 2016, Pan et al., 2024) for camera parameters estimation and Multi-View Stereo (MVS) (Furukawa et al., 2015) for dense point cloud reconstruction, have formed the backbone of oblique photogrammetry. They recover 3D scene structure from multi-view correspondences under epipolar and projective geometry constraints, thereby ensuring high geometric consistency in reconstruction. Nonetheless, their performance often depends on reliable feature extraction, sufficient image overlap, and stable lighting conditions.

In recent years, the rapid development of deep learning technology has brought many changes to the field of 3D vision. Unlike traditional methods, the feed-forward inference model achieves a direct mapping from images to dense 3D geometry through an end-to-end learning paradigm (Zhang et al., 2025). This novel architecture does not require complex multi-stage process such as feature description and matching, triangulation, bundle adjustment, depth map estimation and fusion, but integrates the entire reconstruction process into a unified neural network. These feed-forward models learn scene geometry priors by training on large-scale and diverse 3D datasets. As a result, they can directly infer the geometric structure of previously unseen scenes using pretrained weights. Owing to this capability, such models are often referred to as 3D vision foundation models. Since these models do not rely on scene-specific optimization or explicit feature extraction, they can overcome the reconstruction limitations of traditional methods in textureless or repetitive regions, producing more complete and smoother geometric results.

By leveraging end-to-end architectures and universal 3D priors, 3D foundation models may overcome the limitations of manual feature engineering and the sequential error propagation found in classical pipelines. This makes their application to photogrammetry a highly promising direction. Such a transition first requires a rigorous assessment of their performance in oblique scenarios. Although the capability of these foundation models has been widely demonstrated on general-purpose multi-view reconstruction benchmarks, their effectiveness in oblique photogrammetric scenarios remains unexplored. In oblique photogrammetry, aerial multi-camera systems capture tilted images along planned flight paths, resulting in viewpoint distributions that are constrained by flight geometry and exhibit considerably sparser overlaps than conventional datasets. Such characteristics pose unique challenges for accurate and consistent 3D reconstruction, motivating this study to systematically evaluate existing 3D vision foundation models under these conditions.
To this end, we design an evaluation pipeline that enables fair and reproducible comparisons of 3D vision foundation models under varying imaging conditions, such as different numbers of input images and overlap ratios. Using our self-collected oblique

* Corresponding author

dataset, we conduct extensive experiments to investigate how different model designs influence reconstruction accuracy. Based on these analyses, we provide new insights into the applicability of 3D vision foundation models to oblique photogrammetry and highlight promising directions for future research.

2. RELATED WORK

3D vision foundation models signify a transition from conventional geometric pipelines to end-to-end, data-driven 3D reconstruction. In this section, We review monocular depth estimation foundation models which aim to infer depth from a single view, and pointmap regression foundation models which predict 3D point clouds directly from multi-view images.

Monoculer Depth Estimation Foundation Models. Depth serves as an indispensable source of information in 3D reconstruction (Arampatzakis et al., 2024). To achieve the goal of estimating depth from a monocular image in arbitrary scenarios without retraining or fine-tuning, DepthAnything (Yang et al., 2024a) takes a data-driven approach to enhance model generalization. It first trains a teacher model on real-world images with depth labels obtained from LiDAR or SfM. The teacher model is then used to generate pseudo-depth labels for a large number of unlabeled real images. A student model is subsequently trained using both real and pseudo-labeled images, allowing it to effectively learn the intrinsic data distribution and maintain robust performance across diverse scenes. Following this, DepthAnythingV2 (Yang et al., 2024b) observes that training with real-world images containing noisy or incomplete depth labels can lead to unavoidable artifacts and loss of fine details. To address this issue, it replaces real images with synthetic data that provide continuous and accurate depth supervision when training the teacher model. However, these depth foundation models predict only relative-scale depth from monocular inputs. When applied to large-scale oblique photogrammetry scenarios containing numerous images, the estimated depths fail to maintain multi-view consistency.

Pointmap Regression Foundation Models. (1) *Two-view inference models*: The Pioneering work of this field is DUSt3R (Wang et al., 2024), which proposes a Transformer-based end-to-end framework that directly regresses pointmaps from uncalibrated image pairs, reformulating two-view stereo as a regression problem. MASt3R (Leroy et al., 2024) extends this by adding dense feature prediction and matching loss to improve matching accuracy while maintaining 3D robustness. Both DUSt3R and MASt3R output pairwise point clouds in local coordinate systems, thus require an additional global alignment using PnP-RANSAC optimization. Follow-up MASt3R-SfM (Duisterhof et al., 2025) embed the outputs from MASt3R into conventional SfM pipelines through more systematic and principled approaches. (2) *Multi-view inference models*: Spann3R (Wang and Agapito, 2025) extends DUSt3R by interpreting images as sequential inputs and reconstructing scenes incrementally via a sliding-window network and spatial memory. However, its pairwise processing prevents the correction of early-frame errors, leading to drift over time. Fast3R (Yang et al., 2025) addresses this by employing a transformer architecture with all-to-all attention, enabling global reasoning across all frames, but still rely on post optimization to estimate camera parameters using output pointmaps in global coordinate systems of input views. Subsequent VGGT (Wang et al., 2025) directly use a neural branch to estimate camera pose without post-process. Models above all set the first input view as the origin of global coordinate system, introducing sensitivity to initial view selection that affects reconstruction quality. To address this limitation, π^3 (Wang et al., 2026) employs a permutation-equivariant architecture that produces outputs independent of input image order, predicting affine-invariant camera poses and scale-invariant local pointmaps. Inspired by the widespread use of conditioning inputs in novel-view synthesis and diffusion-based image generation, MapAnything (Keetha et al., 2026) not only supports images as input, but can also selectively utilise a variety of geometric inputs such as camera intrinsic parameters, extrinsic parameters, depth maps, and even local reconstructions. Unlike monoculer depth estimation foundation models that produce independent depth maps with scale ambiguity, pointmap regression foundation models inherently integrate spatial alignment and cross-view constraints within their architecture, which enables a direct migration to photogrammetry while ensuring multi-view consistency. Therefore, we focus on evaluating the performance of the latter in photogrammetry tasks.

3. EVALUATION FRAMEWORK

Overview. To assess the potential of 3D vision foundation models for oblique photogrammetric reconstruction, we design a standardized evaluation workflow comprising three key modules: generation of reference baselines for evaluation (Section 3.1), view selection strategy for diversified test cases construction (Section 3.2), and registration from arbitrary-scale model predictions to metric-scale reference baselines (Section 3.3). As illustrated in Figure 1, our framework first generates reference camera paramters and dense point cloud of each view for evaluation. To account for diverse photogrammetric configurations, a view selection strategy is then applied to construct test cases with varying image counts and overlap ratios. These selected test cases are subsequently processed by 3D vision foundation models to predict their camera poses and dense 3D reconstruction results. The predicted results at arbitrary-scale are then registered to the coordinate system of reference baselines at metric-scale, after which quantitative accuracy metrics are computed for comprehensive performance evaluation.

3.1 Generation of Reference Baselines for Evaluation

To evaluate the practical applicability of 3D vision foundation models in real-world oblique photogrammetry, we utilize reconstruction results generated by established photogrammetric workflows as the benchmark for accuracy assessment. In contrast to synthetic environments created in modeling software like Blender where camera parameters and point clouds are perfectly known, real-world survey data inevitably contain measurement and reconstruction residuals. Although these photogrammetric pipelines do not yield an absolute ground truth, their reconstruction residuals are rigorously maintained within a reliable tolerance for practical surveying and mapping.

Oblique Images ware collected via using an Unmanned Aerial Vehicle (UAV) equipped with an RTK module. During flight, image observations were jointly constrained by GNSS positioning and IMU-recorded platform poses. Based on these constraints, aerial triangulation was performed to estimate reference camera parameters at a real-world metric scale. Subsequently, the dense point cloud was reconstructed via Multi-View Stereo (MVS), followed by Poisson surface reconstruction to yield the mesh model $\mathcal{M}$. All images from the collected dataset are utilized to perform aforementioned workflow, then we generate per-pixel reference dense point clouds for each image using the method described below.

As shown in Figure 2, we render the mesh $\mathcal{M}$ from each camera

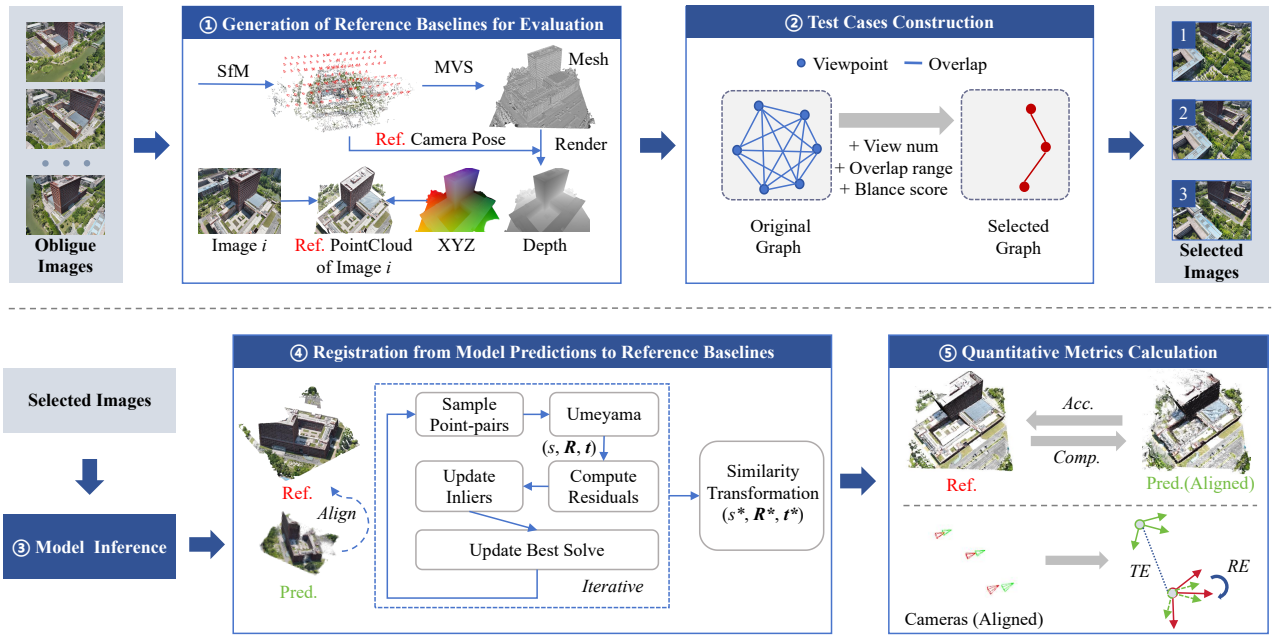

Figure 1. Overall workflow of proposed evaluation framework.

viewpoint V to generate the corresponding reference depth map. Given image $\mathbf{I}_i$ with camera intrinsics $\mathbf{K}_i$ and extrinsics $\mathbf{T}_i$ of view V_i, its depth map $\mathbf{D}_i$ can be obtained by:

$$\mathbf{D}_i = \mathrm{zbuf}(\mathcal{R}_{\mathrm{rast}}(\mathcal{M}, (\mathbf{K}_i, \mathbf{T}_i))) \tag{1}$$

By utilizing mesh rasterization $\mathcal{R}_{\mathrm{rast}}$ to generate the depth buffer zbuf, we explicitly account for surface occlusion. This ensures that the back-projected points from the depth map are guaranteed to be visible from the current viewpoint, effectively filtering out occluded or back-facing elements that would otherwise persist in a naive point-based projection. Then pixels in $\mathbf{I}_i$ with valid depth values are back-projected into 3D space to get XYZ map $\mathbf{X}_i$:

$$\begin{bmatrix} \mathbf{X}_i(u,v) \\ 1 \end{bmatrix} = \mathbf{D}_i(u,v)\mathbf{T}_i\mathbf{K}_i^{-1}[u,v,1]^\top, (u,v) \in \mathbf{I}_i. \tag{2}$$

Since each pixel in the RGB image $\mathbf{I}$ has a one-to-one correspondence with the pixel in the XYZ map $\mathbf{X}$, we can directly obtain the color and the 3D coordinate of each point from the same pixel position in the RGB image and XYZ map, respectively. This allows us to generate a view-specific reference point cloud. For N-views scenarios, the complete reference point cloud $\mathcal{P}_{\mathrm{gt}}$ is obtained by merging all per-view point clouds as below:

$$\mathcal{P}_{\mathrm{gt}} = \bigcup_{i=1}^{N} \{(\mathbf{X}_i(u,v), \mathbf{I}_i(u,v)) \mid (u,v) \in \mathbf{I}_i\}. \tag{3}$$

3.2 View Selection Strategy for Diversified Test Cases Construction

Evaluating the foundation models solely on the entire dataset would lack sufficient diversity in scene configurations. To this end, we adopt a strategy of constructing multiple sub-scenes by selecting view subsets based on varying view counts and overlap intensities. Given that the reference camera parameters and per-pixel 3D point maps for each image were established in the preceding sections, we can flexibly generate the corresponding reference baselines for any sub-scene configuration. The following details our specific strategy for selecting views to construct these diverse test cases.

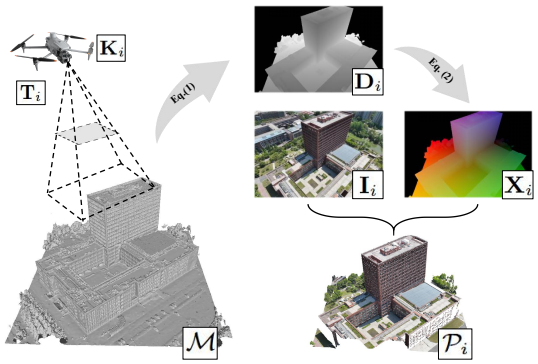

Figure 2. The pipeline of reference baseline generation of target view V_i.

Pairwise Overlap Ratio. To quantitatively evaluate the geometric overlap between two images, we adopt a visibility-based method that leverages both the camera geometry and depth information, rather than relying on simplified frustum intersection or feature-level similarity. Given two images ($\mathbf{I}_i$,$\mathbf{I}_j$) with camera intrinsics ($\mathbf{K}_i, \mathbf{K}_j$) , extrinsics ($\mathbf{T}_i, \mathbf{T}_j$), depth map ($\mathbf{D}_i$, $\mathbf{D}_j$), and XYZ map ($\mathbf{X}_i$, $\mathbf{X}_j$), the 3D points in $\mathbf{X}_i$ are transformed into the camera coordinate frame of $\mathbf{I}_j$:

$$\mathbf{P}_{i\to j}^{camera} = \mathbf{T}_j^{-1} \begin{bmatrix} \mathbf{X}_i(u,v) \\ 1 \end{bmatrix}. \tag{4}$$

The z-coordinate of $\mathbf{P}_{i\to j}^{camera}$ represents the depth value of the spatial point in the camera coordinate system of viewpoint V_j, denoted as $Z_{i\to j}$. The projected pixel coordinate in $\mathbf{I}_j$ is ob-

tained by

$$\mathbf{p}_{i\to j} = \pi(\mathbf{K}_j \mathbf{P}_{i\to j}^{camera}), \quad \pi([x,y,z]^\top) = \left[\frac{x}{z}, \frac{y}{z}\right]^\top. \quad (5)$$

A pixel from view V_i is considered *visible* in view V_j if its reprojected coordinate $\mathbf{p}_{i\to j}$ lies inside the image bounds, and the depth of the reprojected 3D point in view V_j does not exceed $(1+\epsilon)$ times the depth value sampled from $\mathbf{D}_j$ at $\mathbf{p}_{i\to j}$, which provides robustness to reprojection and depth discretization errors.

$$\mathcal{V}_{i\to j} = \left\{ (u,v) \in I_i \;\middle|\; \begin{array}{l} \mathbf{p}_{i\to j} \in I_j, \\ \mathbf{D}_j(\mathbf{p}_{i\to j}) > 0, \\ 0 < Z_{i\to j}(u,v) < \mathbf{D}_j(\mathbf{p}_{i\to j})(1+\epsilon) \end{array} \right\}. \quad (6)$$

ϵ is the tolerance for depth comparison, which is set to 10^{-2} in our experiments. The directional overlap ratio from V_i to V_j is thus defined as

$$O_{i\to j} = \frac{|\mathcal{V}_{i\to j}|}{|N_i|}. \quad (7)$$

where $|N_i|$ denotes the number of valid pixels in $\mathbf{I}_i$. The final overlap metric between two views is given by

$$O_{ij} = \frac{1}{2}\left(O_{i\to j} + O_{j\to i}\right). \quad (8)$$

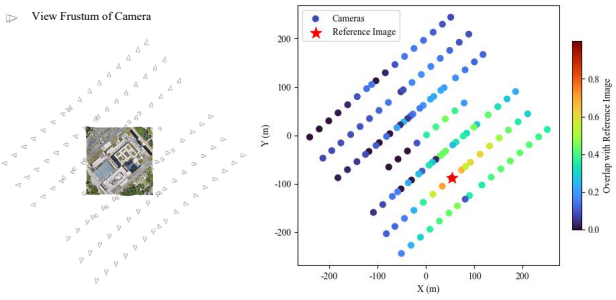

Figure 3. Visualization of the overlap with respect to the reference image.

Figure 3 shows the overlap of images in the scene with respect to a certain reference frame. This visibility-based overlap estimation captures the effective overlapping region between image pairs, as it is inherently occlusion-aware and geometry-consistent. After computing the pairwise overlap between all images, a complete graph can be constructed, where each node represents an image and each edge denotes the overlap ratio between image pairs.

Multi-View Scenario Generation. Based on the constructed complete graph, we perform view selection to generate subscenes for evaluation. Given a target number of images k and an overlap ratio range $(O_{\min}, O_{\max})$, we iterate over each node (reference image) $\mathbf{I}_i$ and identify its neighboring nodes whose edge weights satisfy the overlap constraint $O_{ij} \in (O_{\min}, O_{\max})$. Among these, we select the top $(k-1)$ neighbors with the highest overlap values to form a candidate subscene $\mathcal{S}_i = \{I_i\} \cup \mathcal{N}_i$. To evaluate the geometric consistency of each candidate, we compute the proportion of all pairwise edges within $\mathcal{S}_i$ whose overlap values also fall within the desired range:

$$\text{Score}(\mathcal{S}_i) = \frac{\{(p,q) \mid O_{pq} \in (O_{\min}, O_{\max}),\ p,q \in \mathcal{S}_i\}}{\binom{k}{2}}. \quad (9)$$

Finally, the subscene with the highest score is selected as the optimal configuration: $\mathcal{S}^* = \arg\max_{\mathcal{S}_i} \text{Score}(\mathcal{S}_i)$. This strategy ensures that the selected subscene maintains both strong local connectivity to the reference image and uniform inter-view overlaps among all selected images, resulting in a geometrically balanced and well-conditioned setup for evaluating reconstruction performance. We designed the following experimental scenarios: (1) The number of images varies. We constructed five sub-scenes containing 2, 4, 8, 16 and 32 images, respectively. The overlap range of images was fixed to (0.4, 0.6). (2) The viewpoin overlap-ratio varies. We fixed the number of UAV images to 4, and varied the overlap-ratio ranges to (0.1, 0.2), (0.3, 0.4), (0.5, 0.6) and (0.7, 0.8).

3.3 Registration from Arbitrary-Scale Model Predictions to Metric-Scale Reference Baselines

The camera parameters and dense point clouds inferred by the foundation models are expressed in local coordinate systems with arbitrary scale, which differ from the georeferenced coordinate system at the metric scale adopted by the reference baselines. So the outputs of the foundation models must be transformed into the coordinate system of the reference baselines before quantitative evaluation.

For each input image, the foundation model outputs a per-view predicted XYZ map, which is pixel-wise corresponded with the reference XYZ map of the same view. This means that the correspondences between the predicted and reference 3D points are inherently known, so there is no need to perform an additional iterative closest point (ICP) matching step. We estimate a similarity transformation consisting of rotation $\mathbf{R} \in SO(3)$, translation $\mathbf{t} \in \mathbb{R}^3$, and scale $s \in \mathbb{R}^+$ that minimizes the following least-squares objective:

$$\min_{\mathbf{R},\mathbf{t},s} \sum_i \|\mathbf{y}_i - (s\mathbf{R}\mathbf{x}_i + \mathbf{t})\|_2^2, \quad \text{s.t. } \mathbf{R} \in SO(3). \quad (10)$$

where $\{\mathbf{x}_i\}$ and $\{\mathbf{y}_i\}$ denote corresponding points from the predicted and reference 3D point sets, respectively. We adopt the Umeyama (Umeyama, 2002) method to solve this problem efficiently. In practice, a fraction of unreliable predictions may produce outlier correspondences, which may bias the least-squares solution. To enhance robustness, we wrap the Umeyama estimator with a RANSAC (Derpanis, 2010) scheme. Specifically, in each iteration, a minimal subset of corresponded points is randomly sampled to estimate a candidate similarity transform $(s, \mathbf{R}, \mathbf{t})$ using the Umeyama method. The transformation is then applied to all points to compute alignment residuals, and points whose residuals fall below a predefined threshold are treated as inliers. The model yielding the highest inlier count is selected, and a final refinement is performed by re-estimating $(s, \mathbf{R}, \mathbf{t})$ using all inliers.

4. EXPERIMENTS

We conduct experiments using the configurations described in Section 4.1 on the scenarios selected in Section 3.2 to evaluate and analyze the camera parameters estimation and dense reconstruction performance of different models.

4.1 Dataset, Implement Details and Metrics

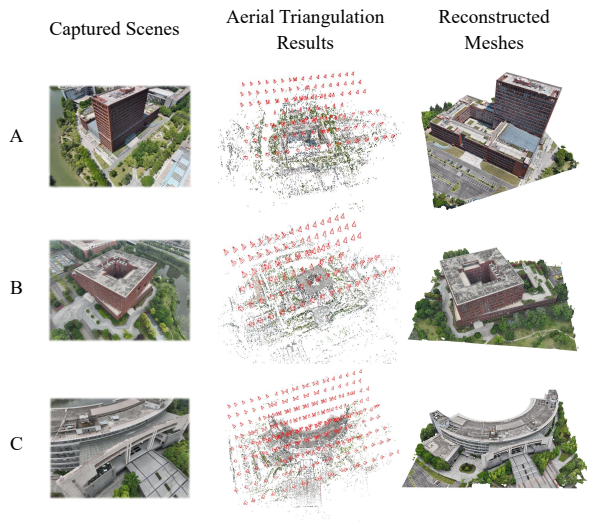

Figure 4. Self-Collected Oblique Imagery Dataset Visualization.

Data Acquisition. As shown in Figure 4, we conducted experiments on a self-collected oblique imagery dataset. The dataset was captured using a DJI Matrice 4E drone, covering three scenes, referred to as A, B, and C. Throughout the flight, the UAV continuously received stable RTK signals, with the RTK system operating in fixed solution mode and achieving a horizontal positioning accuracy of 1 cm + 1 ppm and a vertical accuracy of 1.5 cm + 1 ppm. Images were captured using a wide-angle lens mounted on this UAV, featuring a 4/3-inch CMOS sensor with approximately 20 million effective pixels, and providing an 82° field of view (FOV) and a 24 mm equivalent focal length. To balance acquisition efficiency and multi-view coverage, a five-directional oblique acquisition mode was adopted, consisting of one nadir view and four oblique views with a tilt angle of 45°. The ground sampling distances (GSDs) of the oblique views are 4.56 cm/px, 3.42 cm/px, and 2.28 cm/px, corresponding to the different flight altitudes of the three scenes. Each image has a size of 5280 × 3956 px, and the numbers of collected images for scene A, B, and C are 127, 100, and 278, respectively.

Data Processing. Following the procedure in Section 3.1, reference camera poses and dense point clouds are generated using DJI Terra for aerial triangulation, ContextCapture for mesh reconstruction and Pytorch3D for depth rendering. The reference system follows the WGS84 UTM projection, and is centered at the survey area’s centroid to create a local Cartesian frame with metric-scale for computational efficiency. Aerial triangulation yielded reprojection RMS errors of 0.69 px, 0.65 px, and 0.72 px, respectively. As most of existing foundation models are under the assumption of an ideal pinhole camera without distortion, the captured images were first rectified to remove lens distortion. To satisfy the batch size requirements of these foundation models and improve computational efficiency, the rectified images were subsequently downsampled to 512 × 368 px or 518 × 364 px, ensuring that the size is divisible by 14 or 16, respectively. The intrinsic camera parameters were adjusted accordingly to maintain geometric consistency after resampling.

Compared Models and Configurations. We evaluate DUSt3R, MASt3R, Fast3R, VGGT, π^3 and MapAnything. For DUSt3R and MASt3R, We employ the fully connected image-pair sampling strategy. For Fast3R, We utilize the outputs of the global head for focal length estimation. For VGGT, according to the official recommendation, we obtain the dense point clouds by back-projecting the predicted depth maps and camera poses rather than directly using predicted point maps. For MapAnything, we further investigated the contribution of conditioning input such as known camera parameters to the overall reconstruction accuracy. To this end, we experiment with three input configurations: images-only (denoted as MA-I), images + intrinsics (denoted as MA-IK), images + intrinsics + extrinsics (denoted as MA-IKO). All experiments are performed on a computing platform with NVIDIA GeForce RTX 4090 24G.

Focal Estimation Metric. Given the predicted focal lengths $\{(f_{x,p,i}, f_{y,p,i})\}_{i=1}^{N}$ and the reference values $\{(f_{x,r,i}, f_{y,r,i})\}_{i=1}^{N}$, the error is computed as:

$$e_{\text{focal}} = \sqrt{\frac{1}{2N}\sum_{i=1}^{N}[(f_{x,p,i} - f_{x,r,i})^2 + (f_{y,p,i} - f_{y,r,i})^2]}. \tag{11}$$

This metric measures the RMSE between the predicted and reference focal lengths across all test images. The unit of e_{focal} is pixels.

Pose Estimation Metrics. The accuracy of estimated camera poses is evaluated using the rotational and translational errors between the predicted and reference camera-to-world transformations. Given the predicted pose $(\mathbf{R}_p, \mathbf{t}_p)$ and the reference pose $(\mathbf{R}_r, \mathbf{t}_r)$, the rotation error RE and translation error TE are defined as:

$$\text{RE} = \arccos\left(\frac{\text{trace}(\mathbf{R}_p^\top \mathbf{R}_r) - 1}{2}\right). \tag{12}$$

$$\text{TE} = \|\mathbf{t}_p - \mathbf{t}_r\|_2. \tag{13}$$

where RE is reported in degrees (°) and TE in meters (m).

Dense Reconstruction Metrics. For the dense 3D point cloud, we evaluate reconstruction fidelity using accuracy and completeness. Given the predicted point set $\mathcal{P}$ and the reference point set $\mathcal{G}$, accuracy measures the mean distance from predicted points to the nearest reference points, and completeness is defined vice versa.

$$\text{Accuracy} = \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\min_{g\in\mathcal{G}}\|p - g\|_2, \tag{14}$$

$$\text{Completeness} = \frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}}\min_{p\in\mathcal{P}}\|g - p\|_2. \tag{15}$$

Both metrics are reported in meters (m).

4.2 Intrinsic Evaluation

Method	2 images	4 images	8 images	16 images
DUSt3R	59.389	49.281	44.093	49.078
MASt3R	61.528	29.338	18.236	19.947
Fast3R	24.478	26.847	68.909	15.607
VGGT	105.460	64.886	120.821	82.147
MA-I	28.915	24.903	34.831	22.934
MA-IK	**11.646**	11.654	15.695	10.985
MA-IKO	12.499	**11.041**	**13.316**	**10.854**

Table 1. The RMSE of estimated focals on Scene A. π^3 is not included in this evaluation since its released code does not support camera intrinsic estimation. **Bold** and underline denote the best and second-best results per scene.

As shown in Table 1, even the minimum estimation error persists above 10 pixels, underscoring the performance limitations of current foundation models in precise focal length estimation.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-821-2026 | © Author(s) 2026. CC BY 4.0 License.

Images num.		**2 images**		**4 images**		**8 images**		**16 images**		**32 images**	
Pose Estimation		RE (°)	TE (m)	RE (°)	TE (m)	RE (°)	TE (m)	RE (°)	TE (m)	RE (°)	TE (m)
Scene A	DUSt3R	2.002	21.900	2.249	13.412	1.359	14.267	3.515	20.212	OOM	OOM
	MASt3R	1.456	22.930	**1.047**	8.900	**0.651**	**5.244**	**1.184**	**6.019**	**0.143**	**2.935**
	Fast3R	1.646	8.542	2.463	8.758	5.669	30.693	3.534	10.883	2.429	7.474
	VGGT	2.479	39.874	2.041	18.830	3.397	32.279	2.465	22.572	1.882	18.745
	π^3	**0.244**	**0.838**	2.837	7.224	3.836	10.810	4.028	11.616	3.515	9.196
	MA-I	2.236	9.560	5.633	18.334	12.220	31.482	9.432	27.011	9.241	23.724
	MA-IK	0.660	3.349	5.315	15.158	11.972	31.191	9.155	24.432	9.443	23.071
	MA-IKO	0.638	3.109	2.200	**6.579**	4.670	14.612	3.703	9.610	3.214	8.050
Scene B	DUSt3R	2.104	10.171	3.205	7.756	4.678	11.014	2.830	10.971	OOM	OOM
	MASt3R	2.214	12.842	**1.825**	5.632	**1.115**	**1.351**	1.800	**2.051**	**1.789**	**0.994**
	Fast3R	2.489	9.930	5.417	28.529	11.457	23.422	2.919	15.838	12.122	29.833
	VGGT	3.766	35.536	3.710	27.392	3.524	23.944	2.877	32.747	4.261	32.036
	π^3	1.018	3.498	1.953	**4.417**	1.691	5.999	**1.421**	4.101	2.743	6.554
	MA-I	1.982	3.427	5.199	10.193	12.705	17.030	7.765	15.788	9.610	19.023
	MA-IK	0.988	2.028	4.838	9.912	12.859	17.394	7.861	14.369	9.769	18.215
	MA-IKO	**0.657**	**1.445**	2.289	4.751	7.514	7.965	4.080	7.378	5.796	10.519
Dense Reconstruction		Acc. (m)	Comp. (m)	Acc. (m)	Comp. (m)	Acc. (m)	Comp. (m)	Acc. (m)	Comp. (m)	Acc. (m)	Comp. (m)
Scene A	DUSt3R	1.428	1.161	0.840	0.623	1.468	0.938	0.974	0.362	OOM	OOM
	MASt3R	1.442	1.081	**0.717**	**0.495**	**0.844**	**0.476**	**0.478**	**0.262**	**0.375**	**0.193**
	Fast3R	1.465	1.095	0.978	0.611	1.937	0.994	1.211	0.393	1.211	0.314
	VGGT	1.037	0.771	1.388	0.747	1.542	0.738	1.483	0.439	1.386	0.286
	π^3	**0.953**	**0.721**	1.018	0.532	2.030	0.934	1.259	0.378	1.096	0.241
	MA-I	9.202	1.069	4.842	1.047	5.929	1.219	5.201	0.520	4.830	0.344
	MA-IK	9.435	1.210	4.941	0.748	5.948	1.048	5.293	0.519	4.983	0.360
	MA-IKO	9.593	1.179	4.590	0.777	6.015	1.216	4.837	0.511	4.478	0.291
Scene B	DUSt3R	1.065	0.853	1.023	0.623	1.327	0.789	1.060	0.603	OOM	OOM
	MASt3R	0.967	0.716	0.622	**0.434**	**0.589**	**0.305**	**0.439**	**0.230**	**0.533**	**0.193**
	Fast3R	1.799	1.355	1.458	0.891	2.713	1.212	1.584	0.903	2.644	1.000
	VGGT	1.019	0.883	1.280	0.640	1.397	0.844	1.151	0.468	1.389	0.344
	π^3	**0.736**	**0.633**	**0.619**	0.455	0.755	0.465	0.664	0.287	0.728	0.198
	MA-I	7.162	0.861	5.154	0.770	5.092	1.180	5.710	0.499	5.536	0.431
	MA-IK	7.146	0.821	5.161	0.704	5.193	1.217	5.697	0.480	5.403	0.392
	MA-IKO	7.079	0.845	5.068	0.651	5.124	1.104	5.651	0.499	5.588	0.362

Table 2. Evaluation of pose estimation and dense reconstruction performance of various 3D visual foundation models with different numbers of input views on scene A and scene B . **Bold** and underline denote the best and second-best results per scene, respectively. OOM means that we encountered out-of-memory issues when testing the current model.

MapAnything yields focal length estimates closest to the reference when provided with known camera parameters as conditioning input, whereas VGGT shows the largest error throught. As the number of input images increases from 2 to 8, the focal estimation accuracy of MASt3R and DUSt3R improves, suggesting that methods relying on iterative post-optimization to estimate focal length benefit from larger image sets, which may provide richer geometric constraints for effective explicit supervision. However, further increasing the image count yields diminishing returns, with the error stabilizing at a plateau rather than continuing its downward trend. In contrast, for remaining methods employing neural network for focal length prediction, the correlation between error and the number of input images is negligible. Notably, MapAnything exhibited the most stable performance with minimal error fluctuations, demonstrating insensitivity to image count, while Fast3R and VGGT displayed significant stochasticity. For MapAnything, incorporating known camera intrinsics and extrinsics as model's conditioning inputs further improves accuracy, but residual deviations from the reference values still remain.

Unlike traditional SfM that typically initialize intrinsics from EXIF metadata and refine them through rigorous Bundle Adjustment, these foundation models primarily rely on statistical regression learned from large-scale datasets. Such models lack explicit geometric constraints, making them sensitive to domain shifts in image acquisition. Moreover, existing foundation models successfully maintain multi-view consistency via PnP refinement or global attention mechanism, but may achieve a geometrically consistent reconstruction within a scale-ambiguous coordinate system. Due to the projective equivalency between focal length and scene scale, the model often converges to a consistent but physically inaccurate solution. In summary, current foundation models remain inadequate for reliable camera intrinsic estimation and cannot yet replace traditional geometric methods, their inability to achieve sub-pixel precision precludes their direct application in high-accuracy photogrammetry.

4.3 Orientation Evaluation

Image Number Varies. As shown in Table 2, when the number of input images is limited to 2, π^3 and MapAnything achieve the best results. Specifically, π^3 attains the lowest rotation error of 0.244° and translation error of 0.838m in Scene A, while MA-IKO achieves the lowest rotation error of 0.657° and translation error of 1.445m in Scene B. When the number of input images exceeds 4, MASt3R demonstrates superior scalability and robustness, maintaining stable accuracy with increasing image numbers. In particular, when 32 input images are used in Scene A, MASt3R achieves the lowest rotation and translation errors of 0.143° and 2.935m, respectively. Figure 5 visualizes the estimated trajectories of different models on scene B when the number of input images is 16, MASt3R shows the best alignment with the reference. The remarkable performance

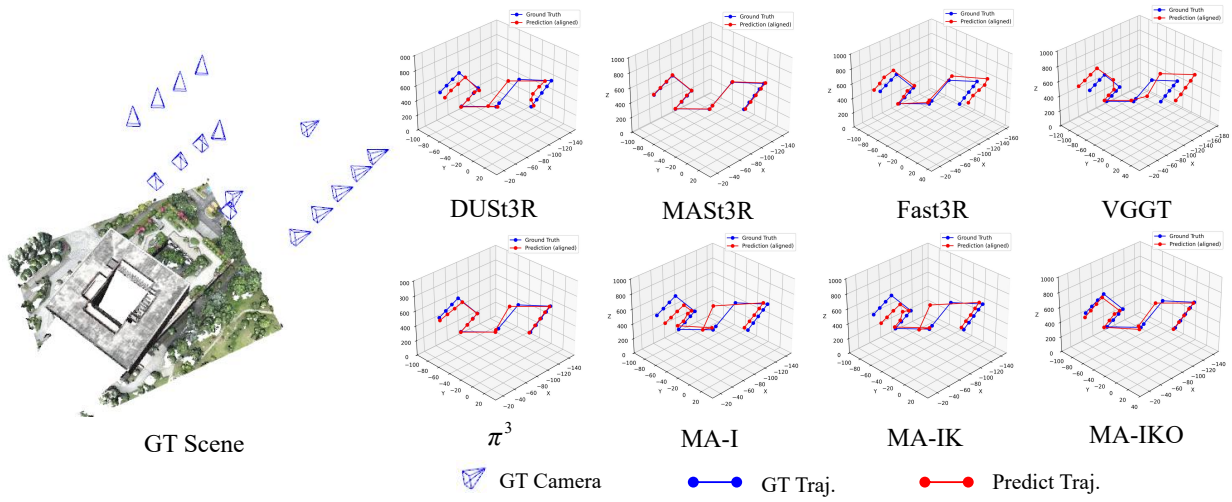

Figure 5. The visualization of estimated trajectories for scene B with 16 input views. MASt3R achieves the lowest translation error.

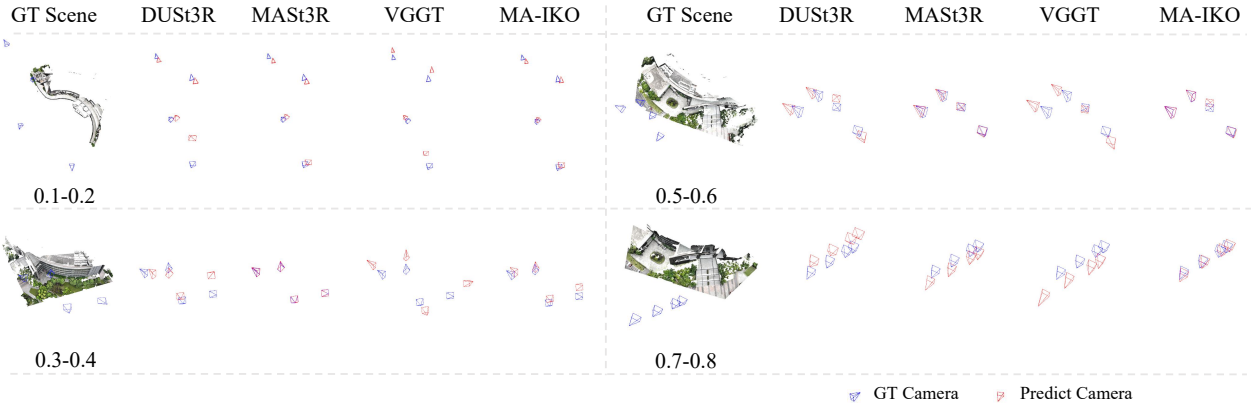

Figure 6. The visualization of estimated camera frustums for scene C with 4 input views under varying overlap ratios. It can be observed that MASt3R and MA-IKO exhibit higher stability to variations in image overlap compared with the other models.

of MASt3R mainly stems from its explicit optimization strategy for multi-view alignment. MASt3R performs two-view inference to generate point maps for all possible image pairs and subsequently applies a global PnP-RANSAC optimization, iteratively refining intrinsic and extrinsic parameters until all pairs are aligned in a unified coordinate frame. In contrast, models like VGGT and its variants rely on implicit neural attention mechanism to predict camera parameters. Without an explicit optimization step, they tend to be less stable when handling complex urban structures and large viewpoint variations typical of oblique photogrammetry.

Image Overlap Ratio Varies. As shown in Figure 7, at very low overlap levels (e.g., 0.1–0.2), the rotation and translation errors are relatively high. As the image overlap increases, the pose estimation errors of most models exhibit a consistent downward trend to varying degrees. Notably, MASt3R and MA-IKO consistently maintain superior pose estimation accuracy, the visualization results in Figure 6 also corroborate our findings. For two-view reasoning models, MASt3R benefits from the design of its matching head, which produces explicit correspondences between input image pairs. Compared with DUSt3R, these additional correspondences help achieve more reliable camera parameter estimation. For multi-view reasoning models, MapAnything gains advantages by incorporating known camera parameters as inputs. These conditioning parameters provide geometric guidance, allowing the model’s predictions to be closer to ground truth and thus improving pose estimation accuracy.

Method	e_{focal}(px)	RE(°)	TE(m)	Acc.(m)	Comp.(m)
DUSt3R	43.802	1.359	14.267	1.468	0.938
+fixed-K	0	1.835	5.083	1.469	0.979
+fixed-KO	0	0	0	1.578	1.089
MA	34.027	12.220	31.482	5.929	1.219
+input-K	15.396	11.972	31.191	5.948	1.048
+input-KO	12.994	4.670	14.612	6.015	1.216

Table 3. Performance of DUSt3R and MapAnything on scene A with 8 input views. fixed-K: Intrinsics fixed to reference during post-optimization; fixed-KO: Intrinsics and extrinsics both fixed to reference during post-optimization; input-K: Images with intrinsics as input; input-KO: Images with intrinsics and extrinsics as input.

Using Known Camera Parameters. As shown in Table 3, incorporating known camera intrinsics or extrinsics generally improves accuracy of both DUSt3R and MapAnything. However, the way these parameters are utilized fundamentally differs. MapAnything adopts camera parameters as conditioning

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-821-2026 | © Author(s) 2026. CC BY 4.0 License.

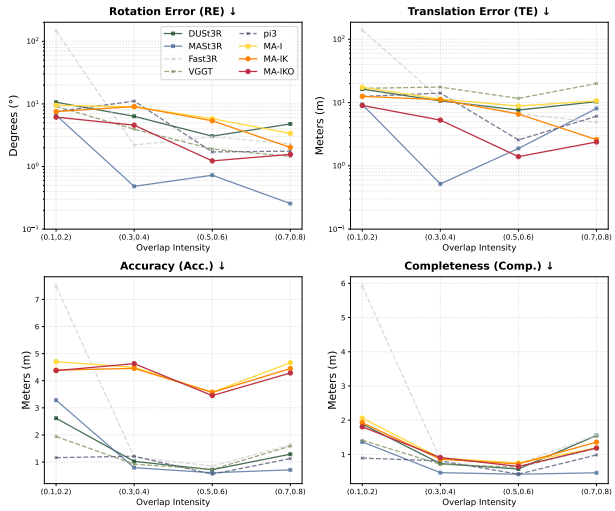

Figure 7. Evaluation of pose estimation and dense reconstruction performance of 3D visual foundation models on scene C under different image overlaps.

inputs that provide informative cues rather than explicit constraints. As a result, the estimated poses tend to approach the ground truth but are not strictly aligned with it, leaving inevitable residual discrepancies. For instance, even with reference intrinsics provided, the estimated focal length still deviates by 15.396 pixels, indicating that the outputs are primarily shaped by the model’s learned priors, with the provided parameters serving only as soft conditioning cues. In contrast, DUSt3R implements a constraint-based mechanism, allowing the camera parameters to be explicitly fixed to their reference values during post-processing.

Collectively, these foundation models demonstrate remarkable efficacy in scenarios characterized by minimal image count and extremely low overlap, conditions under which traditional photogrammetric methods typically fail. While the pose estimation quality of certain models approaches that of traditional frameworks utilizing dense observations, their absolute geometric residuals remain non-negligible for high-precision surveying. Therefore, rather than a direct replacement for professional-grade mapping, these models are better positioned as robust providers of initial priors in extreme scenarios.

4.4 Dense Reconstruction Evaluation

As shown in Table 2 and Figure 7, MASt3R and π^3 achieve the highest point cloud accuracy and completeness across nearly all image input configurations. MASt3R yields the best accuracy when more than two input images are available, while π^3 performs best in the two-view case. Figure 8 illustrates that the reconstructed point clouds contain many outliers, and the linear structures of buildings exhibit noticeable distortions relative to the reference. Among all methods, MapAnything produces the most geometric distortions, whereas MASt3R, Fast3R, VGGT, and π^3 better preserve building contours and straight-line features, demonstrating stronger robustness in oblique photogrammetric settings. Under extremely low overlap conditions (e.g., 0.1–0.2), where conventional methods typically fail, all models are still capable of generating scene-level reconstructions. For MapAnything, incorporating camera intrinsics provides little or even negative benefit for dense reconstruction compared to using image as inputs alone. With full camera parameters, MapAnything still exhibits noticeably higher reconstruction errors than the other models. In essence, all these models regress dense point maps in a similar manner, the difference lies in whether these points are expressed in a local or global coordinate system. Hence, the precision of the reconstructed point cloud primarily depends on two aspects: pairwise regression and multi-view alignment. MASt3R outperforms DUSt3R in reconstruction accuracy, validating the efficacy of its matching head in enhancing multi-view alignment. Notably, π^3’s reference-free design for implicit alignment represents a superior strategy among current multi-view reasoning models.

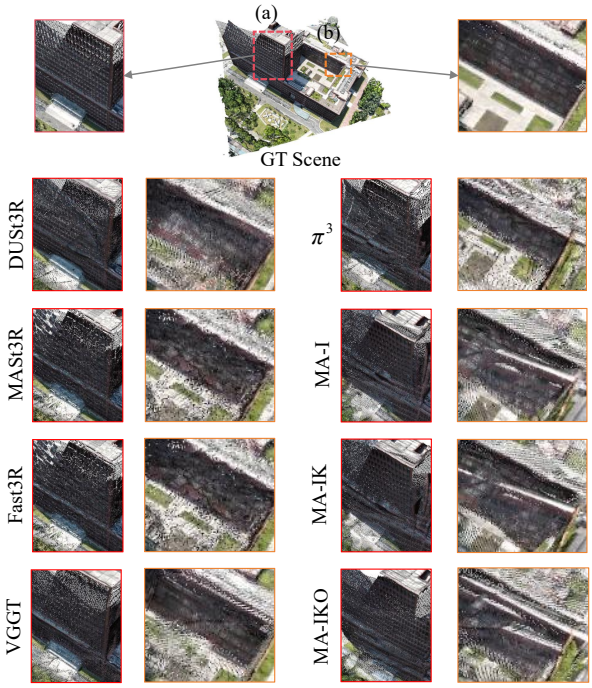

Figure 8. The visualization of estimated dense point clouds for scene A with 8 input views. Observing the highlighted regions (a) and (b), we find that MASt3R, Fast3R, VGGT and π^3 better preserve the regularity of building edges and façades, whereas the other models exhibit noticeable distortions in these areas.

In summary, these foundation models demonstrate robust zero-shot capabilities by generating dense point clouds even under sparse observation conditions that typically cause traditional photogrammetric methods to fail. However, the geometric distortions observed along architectural boundaries and facade details remain non-negligible, precluding their direct application in scenarios requiring high-fidelity metric precision. Consequently, these models are more effectively utilized as a supplementary source to densify under-reconstructed regions where traditional pipelines suffer from insufficient observations. Furthermore, current foundation models primarily employ a per-pixel alignment strategy for 2D-to-3D mapping, where each pixel in every input image corresponds to an individual 3D point. When processing multi-view images, the final output is often a naive superposition of these per-view point maps. This mechanism leads to spatial redundancy and repeated computations in overlapping regions, which may be a bottleneck that should be addressed before such models can be efficiently deployed in large-scale oblique photogrammetry applications.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-821-2026 | © Author(s) 2026. CC BY 4.0 License.

4.5 Memory Consumption under Varying Count of Input Images

We evaluated the peak GPU VRAM consumption of tested foundation models under varying input image counts (N), with resolutions fixed at 512 × 368 px or 518 × 364 px. The results are summarized in Table 4. DUSt3R begins with a lightweight footprint (2.567 GB) but hits a critical memory wall at N =32. This is because it must load and optimize a quadratic number of pairwise point maps and confidence maps during the global alignment (GA) stage. MASt3R extends this limit to 64 images through a more efficient optimization backend, yet eventually succumbs to the same combinatorial bottleneck. π^3 emerges as the most scalable architecture, maintaining the lowest memory footprint among these models and successfully processes 256 images (15.975 GB) without OOM. Traditional aerial blocks often contain thousands of images, far exceeding the current limits of these foundation models. So existing foundation models still require significant memory optimization or divide-and-conquer strategies for large-scale oblique mapping scenarios.

Images Num.	2	4	8	16	32	64	128	256
DUSt3R	2.567	2.581	3.934	9.501	**OOM**	**OOM**	**OOM**	**OOM**
MASt3R	2.640	2.666	2.759	3.265	5.759	19.324	**OOM**	**OOM**
Fast3R	4.190	4.707	5.790	7.928	10.612	12.258	15.708	**OOM**
VGGT	6.819	7.220	8.069	9.765	13.177	20.015	**OOM**	**OOM**
π^3	5.451	5.542	5.684	6.008	6.667	7.998	10.678	15.975
MapAnything	3.743	4.481	5.941	8.872	14.727	**OOM**	**OOM**	**OOM**

Table 4. Comparison of peak GPU VRAM (GB) consumption during inference for foundation models with varying numbers of input images. **OOM** means out of memory.

5. CONCLUSION

This study investigates the potential of existing 3D vision foundation models for oblique photogrammetry. The models are evaluated on diverse sub-scenes derived from our self-collected oblique dataset and compared with traditional photogrammetric reconstructions. We find that (1) under sparse-view and low-overlap conditions where conventional pipelines usually fail, 3D vision foundation models can robustly estimate camera poses and dense point clouds benefiting from their powerful zero-shot generalization. (2) For multi-view (more than 2 views) oblique imagery, two-view reasoning models achieve reconstruction accuracy closer to that of traditional method than multi-view reasoning models. (3) Models that take camera intrinsics or extrinsics as conditioning inputs show improved performance, but these parameters act only as weak supervision rather than fixed constraints as in PnP optimization, so inherent residuals still persist. (4) Current foundation models tend to fail when processing large image sets due to excessive memory consumption. (5) Most models only support undistorted pinhole cameras and fixed input size, limiting their applicability to real-world oblique datasets.

Future research should focus on: (1) Employ divide-and-conquer strategies to integrate 3D vision foundation models into large-scale oblique reconstruction workflows. A hybrid approach can be designed as: partitioning oblique datasets into subsets, reconstructing each subset using foundation models, and globally aligning local point clouds via registration. (2) Develop these foundation models that support arbitrary resolutions and diverse camera models to reduce preprocessing and able to fully exploit high-resolution imagery. (3) Incorporate photogrammetric principles such as multi-view geometric consistency into network architectures or training strategies to enhance robustness and geometric precision on complex oblique datasets.

ACKNOWLEDGEMENTS

This paper is supported by the National Natural Science Foundation of China (Project No. U25A20772), and the Natural Science Foundation of Sichuan Province under Grant 2026NSFSCZY0054.

References

Arampatzakis, V., Pavlidis, G., Mitianoudis, N., Papamarkos, N., 2024. Monocular Depth Estimation: A Thorough Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4), 2396-2414.

Derpanis, K. G., 2010. Overview of the RANSAC Algorithm. *Image Rochester NY*, 4(1), 2–3.

Duisterhof, B. P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J., 2025. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *2025 International Conference on 3D Vision (3DV)*, IEEE, 1–10.

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

Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., Luiten, J., Lopez-Antequera, M., Bulò, S. R., Richardt, C., Ramanan, D., Scherer, S., Kontschieder, P., 2026. MapAnything: Universal feed-forward metric 3D reconstruction. *International Conference on 3D Vision (3DV)*, IEEE.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with mast3r. *European Conference on Computer Vision*, Springer, 71–91.

Pan, L., Baráth, D., Pollefeys, M., Schönberger, J. L., 2024. Global structure-from-motion revisited. *European Conference on Computer Vision*, Springer, 58–77.

Remondino, F., Gerke, M., 2015. Oblique aerial imagery–a review. *Photogrammetric week*, 15number 12, Wichmann/VDE Verlag, 75–81.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Umeyama, S., 2002. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4), 376–380.

Wang, H., Agapito, L., 2025. 3d reconstruction with spatial memory. *2025 International Conference on 3D Vision (3DV)*, IEEE, 78–89.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025. Vggt: Visual geometry grounded transformer. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. Dust3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T., 2026. π^3: Scalable permutation-equivariant visual geometry learning. *The Fourteenth International Conference on Learning Representations*.

Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M., 2025. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21924–21935.

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H., 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10371–10381.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024b. Depth anything v2. *Advances in Neural Information Processing Systems*, 37, 21875–21911.

Zhang, J., Li, Y., Chen, A., Xu, M., Liu, K., Wang, J., Long, X.-X., Liang, H., Xu, Z., Su, H., Theobalt, C., Rupprecht, C., Vedaldi, A., Zhou, K., Liang, P. P., Lu, S., Zhan, F., 2025. Advances in Feed-Forward 3D Reconstruction and View Synthesis: A Survey.

Zhang, X., Zhao, P., Hu, Q., Ai, M., Hu, D., Li, J., 2020. A UAV-based panoramic oblique photogrammetry (POP) approach using spherical projection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 159, 198-219.