OBJALIGNER: FEED-FORWARD OBJECT POSE ESTIMATION AND RECONSTRUCTION VIA SHARED GEOMETRY

Anonymous authors
Paper under double-blind review

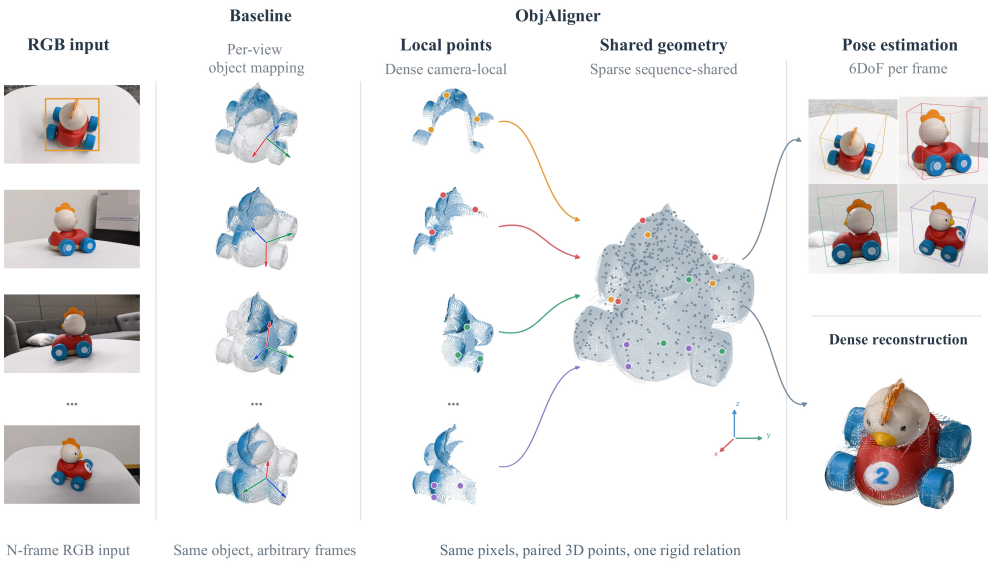

Figure 1: **Overview of ObjAligner.** Pixel-aligned dense camera-local and sparse sequence-shared points yield pose by rigid alignment, with the same transforms fusing geometry for reconstruction.

ABSTRACT

Feed-forward geometry models recover dense 3D structure in camera-centric coordinates, but camera motion alone cannot align an independently moving object across views. We address joint 6D pose estimation and reconstruction with a simple insight: represent the same rigid object through two geometries that play complementary yet asymmetric roles. Dense per-view geometry in camera coordinates preserves the surface detail required for reconstruction, while sparse sequence-shared geometry provides the rigid scaffold needed to determine motion. Their pixel-aligned correspondences turn pose estimation into differentiable rigid alignment, and the same transform maps dense local geometry into the shared frame for reconstruction. The shared frame need not be canonical: it is learned only up to a sequence-common rigid transform, avoiding supervision of absolute object coordinates. We present ObjAligner, a feed-forward framework that takes an RGB sequence and a first-frame target box, without CAD/NOCS models, later-frame annotations, or test-time optimization. Experiments across camera-only, object-only, and joint camera-object motion validate the shared geometric relation for both pose and reconstruction; ObjAligner outperforms matched direct pose regression and transfers to unseen objects and external domains without target-domain adaptation.

1 INTRODUCTION

Jointly recovering the 3D geometry and 6D motion of an unseen rigid object from RGB observations is important for object-level perception in robotic manipulation and embodied interaction (Wen

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2024; 2023). Yet existing methods typically approach pose and reconstruction through various geometric representations. Object pose methods like NOCS (Wang et al., 2019), FoundationPose (Wen et al., 2024), Query6DoF (Wang et al., 2023), and KeyPose (Yu et al., 2025) establish an object reference through canonical coordinates, CAD models, shape priors, or sparse object representations, and then estimate its rigid relation to the camera. Feed-forward reconstruction models like π^3 (Wang et al., 2026), by contrast, predict dense geometry in camera-centric coordinates. For an unseen object without prescribed geometry, neither task-specific representation alone provides the sequence-shared reference needed to both measure object motion and fuse its observations (Huang et al., 2025).

Recent methods begin to close this gap, but they do not remove the need for an object reference. Dynamic methods like D4RT (Zhang et al., 2026a) recover point-level motion without defining a single rigid reference for object pose and fusion. FMOV (Shi et al., 2025) coordinates pose and shape through per-scene optimization, while KV-Tracker (Taher et al., 2026) caches an object-centric keyframe representation for online tracking and reconstruction. These highlight the need for a reference that remains consistent across views without prescribing an unseen object's origin or pose.

Our key insight is to represent the same rigid object with two geometries that play complementary yet asymmetric roles. *Dense local geometry* is defined in the camera coordinate system of each view and preserves the surface detail required for reconstruction, whereas *sparse shared geometry* is expressed in an object coordinate frame shared across the sequence and provides the cross-view geometric reference needed to determine motion. This shared frame is not intended to recover a predefined canonical object frame. We show that the shared object frame is defined only up to a sequence-common rigid transform, and that this freedom does not affect the cross-frame object motion determined by the relation between local and shared geometry, thereby separating the choice of object coordinates from per-view motion. Since the two representations describe the same object, pose is obtained by solving the rigid relation between them, while reconstruction is obtained by applying the same relation to the dense local geometry. Moreover, determining this rigid relation requires only a sufficient set of non-degenerate correspondences. The sparse shared geometry therefore only needs to provide a rigid scaffold, without recovering the complete object surface. This formulation avoids reliance on either a predefined object coordinate system or a complete canonical reconstruction, while unifying pose estimation and reconstruction through the same geometric relation.

Based on this formulation, we introduce ObjAligner, a feed-forward framework for jointly recovering rigid-object pose and dense reconstruction from multi-view RGB observations. Given an RGB sequence and a first-frame target box, ObjAligner uses a target-conditioned multi-view representation to maintain target identity within the full scene and exploits cross-view context to predict sparse shared geometry. The shared points and dense camera-local points are attached to the same image locations, so their correspondences are available by construction. Differentiable weighted rigid alignment then recovers object pose, and the same transform further aggregates dense local geometry from all views into the shared frame to form the final reconstruction. The entire process requires no CAD or NOCS model, later-frame annotation, or test-time optimization. We systematically evaluate ObjAligner under camera-only, object-only, and joint camera-object motion. Under this protocol, ObjAligner obtains the lowest rotation and translation-direction P90 errors and the highest joint AUC@15 on HO3D, COPE119, and UOPE56; on COPE119 it also achieves the lowest reconstruction CD and the highest F-score. Without target-domain adaptation, it further achieves the lowest rotation P90 on OnePose (Appendix 7) while remaining competitive on real-object reconstruction benchmarks.

We make three contributions.

- We formulate joint pose and reconstruction with a sequence-shared object frame whose coordinate freedom preserves camera-relative object motion, avoiding canonical coordinates.
- We introduce ObjAligner, which derives pose by aligning sparse shared and dense local geometry and uses the same transform for reconstruction in one feed-forward pass.
- We evaluate ObjAligner across diverse motion regimes and domains, and assess the shared geometric relation through matched direct-regression and representation ablations.

2 Related Work

Feed-forward visual geometry. DUSt3R (Wang et al., 2024) and MASt3R (Leroy et al., 2024) learn pointmaps and correspondences from image pairs. Follow-up methods (Yang et al., 2025; Wang

et al., 2025a; Keetha et al., 2026; Cabon et al., 2025; Lin et al., 2026a) extend reconstruction to multiple views; π^3 (Wang et al., 2026) predicts camera-local geometry and poses with permutation equivariance. SegVGGT (Qu et al., 2026) also predicts instance segmentation. For objects, FreeSplatter (Xu et al., 2025) predicts shared-frame Gaussians and recovers camera poses through PnP, while NOVA3R (Chen et al., 2026a) predicts amodal point clouds. We focus on the rigid relation between dense local and sparse shared target geometry under independent object motion.

Object pose and reference representations. GigaPose (Nguyen et al., 2024) and SAM-6D (Lin et al., 2024) require object models; FoundationPose (Wen et al., 2024) supports CAD or reference captures. NOCS (Wang et al., 2019), GPV-Pose (Di et al., 2022), Query6DoF (Wang et al., 2023), and KeyPose (Yu et al., 2025) exploit canonical coordinates, geometric voting, or sparse shape priors. CAD-free references include posed RGB images in Gen6D (Liu et al., 2022), SfM models in OnePose/OnePose++ (Sun et al., 2022; He et al., 2022), and an RGB-D anchor in Any6D (Lee et al., 2025). PoseCNN (Xiang et al., 2018) directly regresses pose; GDR-Net (Wang et al., 2021) guides regression with dense correspondences. OPT-Pose (Li et al., 2026) predicts pointmaps and NOCS for relative and canonical absolute pose. ObjAligner instead learns a sequence-specific reference without prescribed canonical coordinates.

Joint object tracking and reconstruction. BundleSDF (Wen et al., 2023) jointly optimizes poses and a neural object field from RGB-D; in-hand scanning (Hampali et al., 2023) and FMOV (Shi et al., 2025) optimize shape and pose from RGB sequences. Point2Pose (Lin et al., 2026b) combines RGB-D tracking, registration, pose-graph optimization, and TSDF fusion. Streaming alternatives include CUT3R's persistent scene state (Wang et al., 2025b), OnlineSplatter's object-centric Gaussians (Huang et al., 2025), and KV-Tracker's cached keyframe features for RGB pose tracking and reconstruction (Taher et al., 2026). HORT (Chen et al., 2025) predicts hand-held shape and hand-relative translation from one image; ForeHOI (Chen et al., 2026b) completes object geometry from video, then estimates pose through rendering, matching, and refinement. ObjAligner jointly processes a view set and couples pose with visible-surface fusion through one rigid alignment.

Dynamic geometry and pose recovery. MonST3R (Zhang et al., 2025a), St4RTrack (Feng et al., 2025), POMATO (Zhang et al., 2025b), Track4World (Lu et al., 2026), and D4RT (Zhang et al., 2026a) recover dynamic geometry or correspondences. Their point-level outputs can support rigid fitting, but do not specify our shared object-frame representation. ProxyPose (Zhang et al., 2026b) generates a proxy video and solves pose using its known geometry. Rigid alignment, differentiable SVD registration (Wang & Solomon, 2019), and global coordinate freedom are established tools. Our contribution is learning sparse shared geometry under sequence-common coordinate freedom and relating it to dense local geometry for both motion and reconstruction.

3 METHOD

3.1 OVERVIEW

Given an RGB sequence $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ and a first-frame box $\mathbf{b}_1$ identifying a rigid target, ObjAligner recovers its 6D motion and a consistent object-level reconstruction in a single feed-forward pass, without a CAD model or later-frame annotation. We build on π^3 (Wang et al., 2026), which predicts per-view camera poses up to a common similarity gauge and pixel-aligned local point maps in each view's camera coordinates. These local point maps share one unknown sequence scale, but cannot by themselves align an independently moving object because camera and object motion jointly determine its cross-frame change. The pipeline of the method is shown in Figure 2.

Our key idea is to derive pose from two geometric descriptions of the same rigid points. At query locations inside the predicted target region of frame i, ObjAligner pairs camera-local points $\mathbf{x}_{ik}^{C_i}$ from π^3 with shared points $\mathbf{y}_{ik}^{\mathcal{S}}$ expressed in a sequence-common frame $\mathcal{S}$. Because both points are attached to the same image location, their correspondence is given by construction, and their rigid relation is

$$\mathbf{Q}_i = \mathbf{T}_{\mathcal{S}\leftarrow C_i}, \qquad \mathbf{y}_{ik}^{\mathcal{S}} = \mathbf{R}_i \mathbf{x}_{ik}^{C_i} + \mathbf{t}_i. \tag{1}$$

Rather than regressing $\mathbf{Q}_i$, ObjAligner solves it by weighted rigid alignment. The same transform then maps visible object geometry from every frame into $\mathcal{S}$ for reconstruction. The frame $\mathcal{S}$ need

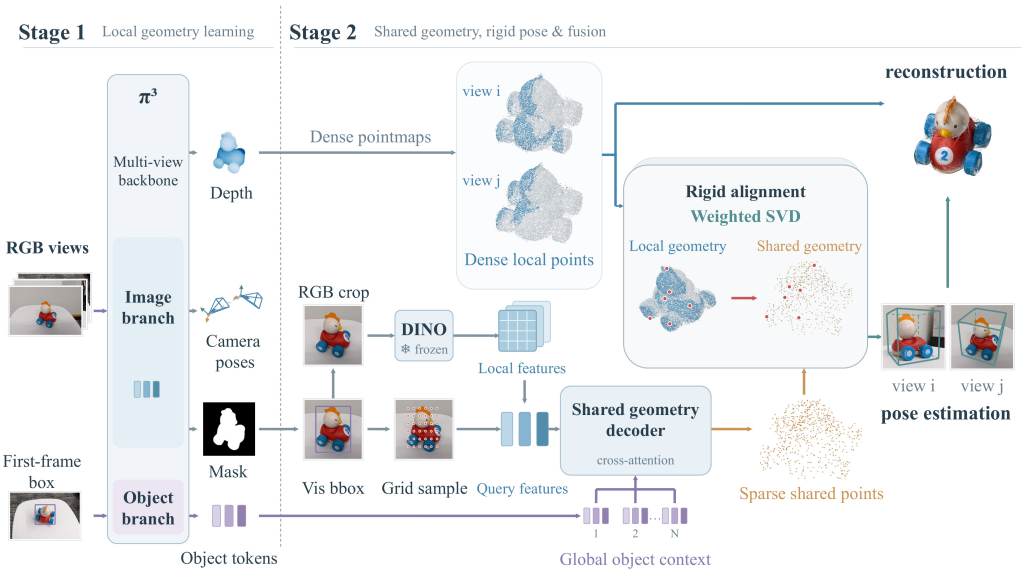

Figure 2: **Pipeline of ObjAligner.** Stage I learns masks and dense local geometry; Stage II learns sparse shared geometry. Weighted SVD recovers rigid transforms that also fuse the dense reconstruction.

only be consistent within the sequence and is not tied to CAD, NOCS, or dataset-defined object axes. Figure 3 illustrates the resulting fusion and coordinate freedom.

3.2 Target-Conditioned Multi-View Encoding

π^3 maps an unordered set of views to corresponding per-view outputs with a fully permutation-equivariant architecture (Wang et al., 2026). It embeds each image into patch tokens and alternates view-wise and global self-attention without a reference-view token or frame-index embedding. Separate decoders then predict the camera poses $\{\mathbf{C}_i\}$ and local point maps $\{\mathbf{X}_i^{C_i}\}$ used by ObjAligner. We preserve this full-image stream and its permutation equivariance conditional on the prompt assignment: permuting the views together with the designation of which view carries $\mathbf{b}_1$ permutes the outputs in the same way. In particular, the object tokens carry no frame-index embedding.

We encode the center and size of $\mathbf{b}_1$ and add the result to an 8×8 grid of learned object tokens. The tokens cross-attend to the evolving image features, carrying target-specific evidence across views rather than reading only the final π^3 outputs. Their multi-view context conditions a mask decoder that predicts a target probability map $\mathbf{M}_i$ in every frame. We threshold $\mathbf{M}_i$ at 0.5 and take the tight enclosing box as the visible region $\mathbf{b}_i$; predictions with fewer than 16 foreground pixels fall back to the prompt region. The resulting regions define only the spatial support of the shared-geometry queries. Thus, $\mathbf{b}_1$ specifies the target once, whereas all later regions are inferred from the sequence; no later-frame ground-truth box or mask enters inference.

3.3 Sequence-Shared Object Geometry

An unknown object has no uniquely defined origin or axes. Instead of predicting dataset-defined canonical coordinates, ObjAligner learns a coordinate frame $\mathcal{S}$ that is chosen freely but shared by all views of the sequence. Within each predicted region $\mathbf{b}_i$, we sample the cell centers of a bbox-relative 12×12 grid, $\mathcal{U}_i = \{\mathbf{u}_{ik}\}_{k=1}^{K}$ with $K = 144$. At each location, we read the camera-local point $\mathbf{x}_{ik}^{C_i}$ from the local point map of π^3 and the corresponding image feature from the multi-view representation. A shared-geometry decoder uses these query features together with the sequence-wide object context to predict $\mathbf{y}_{ik}^{\mathcal{S}} \in \mathbb{R}^3$. Because $\mathbf{x}_{ik}^{C_i}$ and $\mathbf{y}_{ik}^{\mathcal{S}}$ are attached to the same $\mathbf{u}_{ik}$, their within-frame correspondence requires no matching.

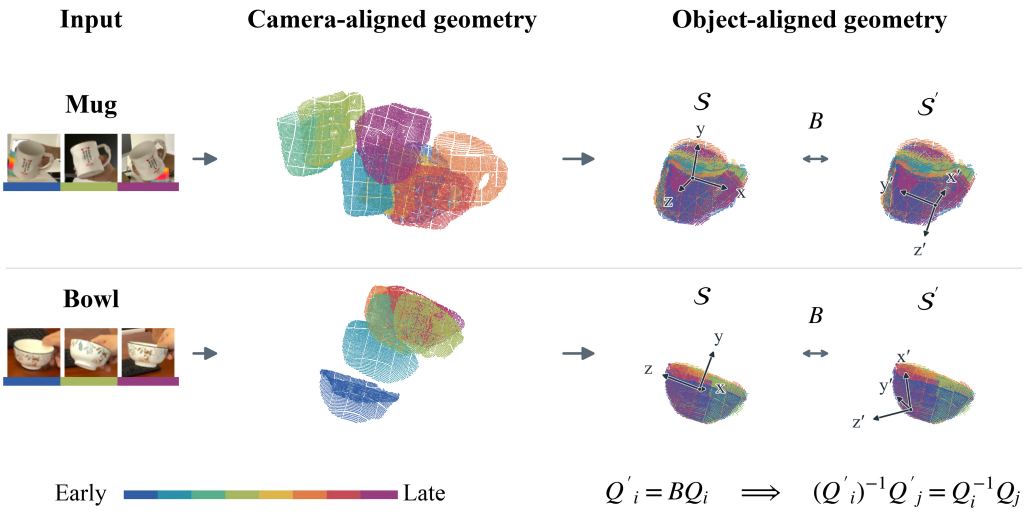

Figure 3: **Shared-frame fusion and coordinate freedom.** Input crops accompany camera- and object-aligned fusion of the same predicted points; colors indicate source views, and the alternative axes illustrate equivalent sequence-wide coordinate choices that preserve relative pose.

The shared designation additionally imposes a cross-frame condition: repeated observations of the same physical surface point must receive the same coordinates in $\mathcal{S}$. The sequence-wide object context couples the per-view predictions, and the consistency objective in Sec. 3.5 enforces this condition geometrically. This cross-frame agreement, rather than the within-frame pairing alone, makes $\mathcal{S}$ one sequence-common object frame.

Local evidence for small targets. Sparse target pixels can be overwhelmed by full-image features. We therefore resample each predicted region as a 168×168 high-resolution crop and add its features residually to the original queries. A second residual connection injects the query-local feature into the shared-geometry decoder without recentering the query, thereby retaining its offset within the region as a position cue. These additions refine only the predicted shared points: they do not alter the query locations, camera-local points, or rigid solver.

Gauge invariance. The shared frame is determined only up to one sequence-common rigid transform. We implicitly lift 3D points to homogeneous coordinates when applying rigid transforms. For any $\mathbf{B} \in SE(3)$, applying

$$\mathbf{y}_{ik}^{\mathcal{S}'} = \mathbf{B}\mathbf{y}_{ik}^{\mathcal{S}}, \qquad \mathbf{Q}'_i = \mathbf{B}\mathbf{Q}_i \tag{2}$$

to homogeneous points leaves both native training quantities unchanged:

$$(\mathbf{Q}'_i)^{-1}\mathbf{Q}'_j = \mathbf{Q}_i^{-1}\mathbf{Q}_j, \qquad (\mathbf{Q}'_0)^{-1}\mathbf{y}_{ik}^{\mathcal{S}'} = \mathbf{Q}_0^{-1}\mathbf{y}_{ik}^{\mathcal{S}}. \tag{3}$$

Thus, the model need not select dataset-defined axes or align its shared frame during training. The local point maps of π^3 retain one global scale ambiguity shared by the sequence. Relative to that scale, the rigid pairing between $\mathbf{x}_{ik}^{C_i}$ and $\mathbf{y}_{ik}^{\mathcal{S}}$ introduces only an $SE(3)$ object-frame gauge, not an additional object-specific or per-frame $Sim(3)$ ambiguity.

3.4 Pose and Reconstruction from Shared Geometry

Let w_{ik} be the object probability at query $\mathbf{u}_{ik}$. ObjAligner recovers the transform from the camera frame to the shared frame by weighted rigid alignment,

$$(\mathbf{R}_i, \mathbf{t}_i) = \arg\min_{\mathbf{R}\in SO(3),\, \mathbf{t}} \sum_{k=1}^{K} w_{ik} \left\| \mathbf{R}\mathbf{x}_{ik}^{C_i} + \mathbf{t} - \mathbf{y}_{ik}^{\mathcal{S}} \right\|_2^2. \tag{4}$$

We compute the weighted centroids and cross-covariance and obtain a proper rotation with differentiable SVD (Wang & Solomon, 2019). A sparse but non-degenerate set of correspondences

is sufficient; a complete object surface is unnecessary. Very narrow, nearly collinear, or poorly distributed visible support can nevertheless make the fit noise-sensitive, while object symmetries can admit equivalent rotations (Hodan et al., 2018). During training, the camera-local points and mask-derived weights are detached in this solver, while gradients through the recovered pose update the shared points. Pose supervision therefore organizes the learned shared geometry instead of altering the geometry predicted by π^3 and used as its reference.

The same transform directly aggregates visible geometry. If $\mathbf{X}_i^{\mathrm{obj}}$ denotes the camera-local points selected by the predicted object mask, then

$$\widetilde{\mathbf{X}}_i^{\mathcal{S}} = \mathbf{Q}_i \mathbf{X}_i^{\mathrm{obj}}, \qquad \mathbf{X}_{\mathrm{fused}}^{\mathcal{S}} = \bigcup_{i=1}^{N} \widetilde{\mathbf{X}}_i^{\mathcal{S}}. \tag{5}$$

Thus, pose and reconstruction use one geometric relation rather than separate prediction heads.

The native $\mathbf{Q}_i$ expresses pose in the arbitrary shared frame. Dataset-frame relative motion is not invariant to this choice, so reporting physical 6DoF pose requires one evaluation-time calibration. Let $\mathbf{A}_i^{\mathrm{gt}} = \mathbf{T}_{C_i \leftarrow O}$ be the ground-truth object-to-camera pose. We estimate a single $\mathbf{B} = \mathbf{T}_{\mathcal{S} \leftarrow O}$ per test sequence by least-squares alignment of $\{\mathbf{Q}_i\}$ to $\{(\mathbf{A}_i^{\mathrm{gt}})^{-1}\}$, such that $\mathbf{Q}_i \simeq \mathbf{B}(\mathbf{A}_i^{\mathrm{gt}})^{-1}$. This is analogous to $SE(3)$ trajectory alignment before reporting ATE (Sturm et al., 2012): it accounts for six degrees of freedom for the entire sequence and cannot absorb per-frame error. We then obtain

$$\mathbf{Q}_i^O = \mathbf{B}^{-1}\mathbf{Q}_i, \qquad \mathbf{O}_i = \mathbf{C}_i(\mathbf{Q}_i^O)^{-1}, \qquad \Delta\mathbf{O}_{ij} = \mathbf{O}_i^{-1}\mathbf{O}_j, \tag{6}$$

where $\mathbf{C}_i = \mathbf{T}_{W \leftarrow C_i}$. Reconstruction and training do not use $\mathbf{B}$; dataset-frame pose evaluation does. For objects with annotated rotational symmetry, we additionally report symmetry-aware rotation metrics without replacing the raw headline evaluation.

3.5 Training Objective

We train ObjAligner with six loss terms,

$$\mathcal{L} = \lambda_{\mathrm{scene}}\mathcal{L}_{\mathrm{scene}} + \lambda_{\mathrm{cam}}\mathcal{L}_{\mathrm{cam}} + \lambda_{\mathrm{mask}}\mathcal{L}_{\mathrm{mask}} + \lambda_{\mathrm{local}}\mathcal{L}_{\mathrm{local}} + \lambda_{\mathrm{rel}}\mathcal{L}_{\mathrm{rel}} + \lambda_{\mathrm{cons}}\mathcal{L}_{\mathrm{cons}}. \tag{7}$$

$\mathcal{L}_{\mathrm{scene}}$ and $\mathcal{L}_{\mathrm{cam}}$ retain the scene-geometry and relative-camera-pose supervision of π^3. $\mathcal{L}_{\mathrm{mask}}$ supervises target visibility, while $\mathcal{L}_{\mathrm{local}}$ supervises object-region camera-local geometry with point, normal, and log-depth constraints. These four terms establish the observations on which shared geometry relies.

The remaining terms organize these observations into one rigid object. Using the object-to-camera poses $\mathbf{A}_i^{\mathrm{gt}}$ defined above, the relative-pose term compares

$$\mathbf{Q}_i^{-1}\mathbf{Q}_j \quad \text{with} \quad \mathbf{A}_i^{\mathrm{gt}}(\mathbf{A}_j^{\mathrm{gt}})^{-1} \tag{8}$$

over frame pairs. To constrain the shared points themselves, we additionally map them into the first camera frame and enforce

$$\mathbf{Q}_0^{-1}\mathbf{y}_{ik}^{\mathcal{S}} \approx \mathbf{A}_0^{\mathrm{gt}}(\mathbf{A}_i^{\mathrm{gt}})^{-1}\,\mathrm{sg}\left(\mathbf{x}_{ik}^{C_i}\right), \tag{9}$$

where $\mathrm{sg}(\cdot)$ stops gradients through the camera-local reference, and $\mathbf{A}_0^{\mathrm{gt}}(\mathbf{A}_i^{\mathrm{gt}})^{-1}$ transports the observed object point to the first-frame camera coordinates. Eq. 8 defines $\mathcal{L}_{\mathrm{rel}}$, and a robust point loss on Eq. 9 defines $\mathcal{L}_{\mathrm{cons}}$. The latter constrains every frame through the same gauge-invariant map to C_0, preventing a frame-dependent drift of the shared geometry without fixing the coordinates of $\mathcal{S}$. Both quantities are gauge invariant by Eq. 3; no training-time gauge alignment is required.

Stage I: Target-conditioned geometry pretraining. Starting from a pretrained π^3 model, we freeze the visual encoder and optimize the remaining scene, camera, mask, and object-local geometry pathways with $\lambda_{\mathrm{rel}} = \lambda_{\mathrm{cons}} = 0$. Local support is disabled in this stage. The goal is to establish reliable target regions and camera-local geometric references before imposing object motion.

Stage II: Shared-geometry pose learning. We initialize from Stage I, retain its supervision, enable the high-resolution crop and query-local residual, and activate $\mathcal{L}_{\mathrm{rel}}$ and $\mathcal{L}_{\mathrm{cons}}$. The former constrains the recovered rigid transforms, whereas the latter enforces the reference-frame consistency described above. The two-stage schedule therefore learns pose through shared geometry rather than through an independent 6D pose head.

Table 1: **Joint camera-object motion on COPE119 Test35 and UOPE56.** Pose errors are in degrees, CD in cm, and AUC/F-score in %; all methods share the selected views.

Method	COPE119 Test35 (seen)					UOPE56				
	R-P90↓	T_{dir}-P90↓	AUC@15↑	CD↓	F-score↑	R-P90↓	T_{dir}-P90↓	AUC@15↑	CD↓	F-score↑
π^3 + black-bbox	<u>26.30</u>	<u>57.45</u>	4.78	2.015	29.34	29.13	<u>59.55</u>	7.38	1.550	39.99
MapAnything	27.73	61.17	<u>7.63</u>	<u>1.338</u>	<u>43.77</u>	28.08	63.27	7.64	**1.223**	49.64
Fast3R	36.60	67.61	6.76	1.498	43.56	30.68	60.84	<u>9.29</u>	<u>1.247</u>	**52.51**
DA3	–	–	–	4.810	15.85	–	–	–	4.140	18.34
FreeSplatter-O	–	–	–	1.862	33.10	–	–	–	1.598	45.54
MUSt3R	28.17	61.78	5.28	2.567	31.25	<u>27.70</u>	62.80	4.03	1.658	39.57
CUT3R	45.34	69.22	3.86	3.094	22.22	30.62	64.87	7.15	2.240	29.48
ObjAligner	**13.77**	**34.00**	**27.84**	**1.233**	**51.35**	**16.70**	**36.57**	**21.18**	1.257	<u>52.01</u>

Table 2: **Object pose on HO3D-v3 with shared random-start stride-24 windows.** Errors are in degrees; all AUCs are percentages. ADD/ADD-S AUC integrates up to 10 cm.

Method	R-P90↓	T_{dir}-P90↓	AUC@15↑	ADD AUC↑	ADD-S AUC↑
π^3 + black-box	<u>69.53</u>	<u>59.30</u>	<u>6.81</u>	<u>51.38</u>	<u>77.08</u>
MapAnything	103.79	68.29	6.04	46.49	67.96
Fast3R	91.45	73.45	2.31	40.18	58.81
MUSt3R	77.72	68.49	2.14	37.51	59.44
CUT3R	78.87	68.95	4.69	43.64	60.71
ObjAligner	**60.17**	**52.25**	**11.01**	**59.67**	**82.32**

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Protocol. Both training stages use SeqMix-30k, NAVI (Jampani et al., 2023), TACO-ego (Liu et al., 2024), OPT3D (Wu et al., 2017), and COPE119 Train84 (Meng et al., 2025); sampling weights are given in Appendix B.3. Tests on SeqMix, NAVI, and COPE119 use held-out samples from training sources and are marked "(seen)." HO3D-v3, UOPE56, OnePose, and YCBInEOAT are external tests, with no target-domain training, fine-tuning, or checkpoint selection. Unless noted otherwise, ObjAligner jointly processes $N = 8$ selected RGB views with a target box in the first view. Inference uses no depth, CAD model, ground-truth camera or object pose, later-frame target annotation, or test-time optimization.

Metrics. To compare predicted poses with ground truth in the benchmark-defined object frame, we fit a single sequence-wide SE(3) alignment as described in Sec. 3.4. We then compare relative motion over all view pairs, reporting mean rotation error R_{mean}, mean translation error T_{mean}, and AUC@15 when relative translation is non-degenerate. Reconstruction evaluates visible geometry fused by the predicted rigid transforms using Chamfer Distance (CD) and F-score (Tatarchenko et al., 2019). For COPE119 and UOPE56, the reference surface is fused from ground-truth depth, masks, and poses of the selected views; available scans or CAD models are used only as evaluation references. Metric definitions and supplementary measures are given in Appendix C.

Baselines. We compare against scene-level reconstruction models π^3 (Wang et al., 2026) and VGGT (Wang et al., 2025a), and object-level methods FreeSplatter-O (Xu et al., 2025) and NOVA3R (Chen et al., 2026a). Pose comparisons include depth- or reference-assisted FoundationPose (Wen et al., 2024) and Point2Pose (Lin et al., 2026b) (†), sequence-optimized FMOV (Shi et al., 2025), online KV-Tracker (Taher et al., 2026), and D4RT (Zhang et al., 2026a) with downstream rigid fitting (‡). External benchmarks additionally include reference-based OnePose++ (He et al., 2022) and Any6D (Lee et al., 2025); ObjAligner-Direct is a direct-pose control built on the same Stage-I base. To accommodate baseline-specific inputs without later-frame annotations, we use SAM 2 (Ravi et al., 2025) to propagate masks from the first-view box. Each baseline then receives black-bbox, foreground-only, or full RGB as appropriate, whereas ObjAligner predicts masks from full RGB (Appendix B.2).

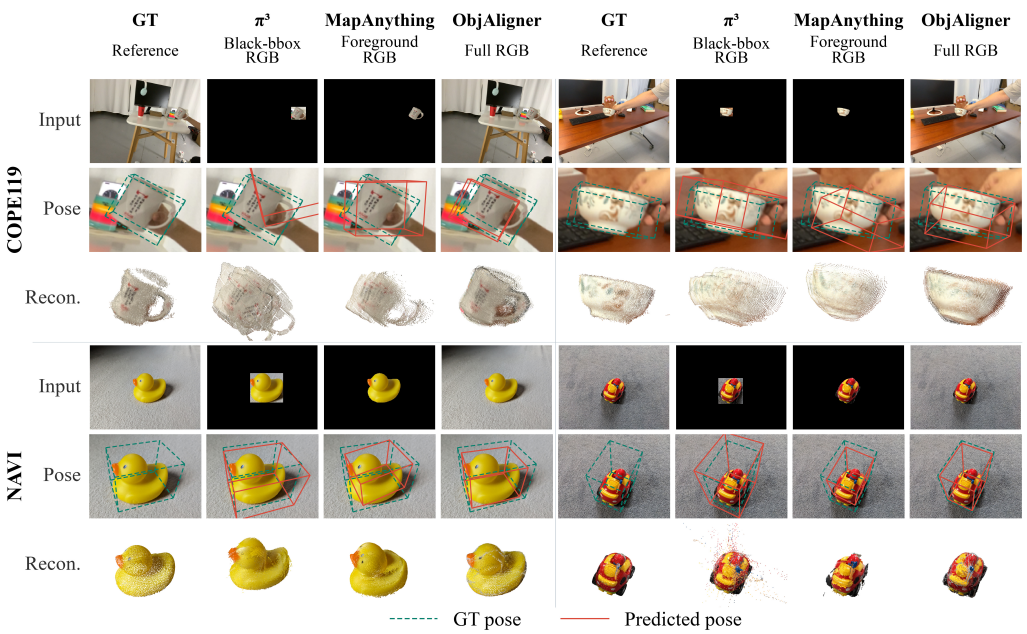

Figure 4: **Qualitative comparison of pose and reconstruction on COPE119 and NAVI.** ObjAligner improves pose under object motion while retaining competitive static-object reconstruction.

4.2 Joint Camera-Object Motion

Joint camera-object motion tests whether the learned shared frame can separate target motion from camera motion while also supporting cross-view surface fusion. We evaluate on COPE119 Test35, a held-out split of a training source, and UOPE56, an external test set without target-domain adaptation (Meng et al., 2025). Both pose and reconstruction are evaluated from the same predictions.

Table 1 shows that the main gain is in pose: ObjAligner more than doubles the best competing AUC@15 on both COPE119 and external UOPE56. The persistence of this gain without target-domain adaptation is consistent with a sequence-shared frame providing a stable reference for target motion beyond scene geometry. The same transforms produce the best visible-surface reconstruction on COPE119, but not on UOPE56, where CD and F-score remain close to the leaders. This divergence clarifies the scope of the shared-local relation: rigid alignment organizes observations across views, while errors in dense camera-local points and foreground support still limit the fused surface.

4.3 Object Pose under Real World Motion

HO3D-v3 evaluates pose under real hand-object interaction, where occlusion changes visible target support. This setting probes shared-geometry alignment with incomplete observations.

Table 2 shows that ObjAligner outperforms all baselines on both relative-motion and model-based pose measures, with AUC@15 reaching 11.01 versus 6.81 for the strongest competitor. The advantage is not confined to angular agreement: ADD and symmetry-tolerant ADD-S also improve, indicating that the estimated transforms place object geometry more accurately despite incomplete observations. This agreement across metrics is consistent with a sequence-shared reference helping to connect partially occluded views. At the same time, the 60.17° rotation P90 reveals a substantial high-error tail, so improved aggregate accuracy does not imply uniformly reliable pose recovery.

4.4 Visible-Surface Reconstruction and Static-Object Pose

We evaluate visible-surface reconstruction on static-object NAVI (Jampani et al., 2023) and hand-interaction HO3D-v3, and static-object pose on the external OnePose benchmark (Sun et al., 2022). These settings test the shared-frame formulation when object motion is absent or visibility is restricted.

Table 3 shows that ObjAligner closely matches π^3 on NAVI and outperforms most baselines on HO3D reconstruction. Together with its leading HO3D pose accuracy (Table 2), these results indicate that

Table 3: **Visible-surface reconstruction on NAVI and HO3D-v3, and static-object pose estimation on OnePose ($n = 80$).** All OnePose methods share 8-frame center stride-24 windows. Bold and underlined items indicate the best and second-best results; – indicates no pose evaluation.

Method	NAVI (seen)			HO3D-v3			OnePose (pose)		
	CD↓	L2↓	F-score↑	CD↓	L2↓	F-score↑	R-P90↓	T_{dir}-P90↓	AUC@15↑
π^3 + black-bbox	**0.148**	<u>0.0832</u>	**91.24**	1.256	4.5998	38.61	<u>14.43</u>	**21.77**	**27.72**
MapAnything	0.305	0.3334	70.97	1.130	5.9380	44.77	24.45	44.13	11.81
Fast3R	0.250	0.3259	81.41	0.972	4.6788	**49.63**	56.66	58.79	7.76
DA3	0.185	0.5275	89.92	2.020	12.9378	29.37	–	–	–
FreeSplatter-O	0.369	0.3900	67.42	1.295	4.7247	34.35	–	–	–
MUSt3R	0.275	0.2109	72.61	**0.817**	**1.7105**	44.38	34.50	41.99	14.28
CUT3R	0.798	4.9125	45.57	3.466	98.0821	32.43	22.52	40.82	12.77
ObjAligner	<u>0.157</u>	**0.0643**	<u>91.18</u>	<u>0.831</u>	<u>2.0990</u>	<u>47.93</u>	**14.23**	<u>25.39</u>	<u>23.44</u>

Table 4: **Structural ablation across controlled joint motion, real reconstruction, and real joint-motion pose.** Black-box and Direct use the five-source Stage-I base; Full uses the final Stage-II checkpoint. † denotes disabling Zoom and LocalSkip only at inference.

Variant	SeqMix diagnostic		NAVI		COPE119 Test35		
	AUC@15↑	F-score↑ (%)	CD↓ (cm)	F-score↑ (%)	$R_{\mathrm{mean}}\downarrow(^\circ)$	$T^{\mathrm{dir}}_{\mathrm{mean}}\downarrow(^\circ)$	AUC@15↑
Adapted π^3	13.18	9.15	<u>0.236</u>	<u>78.08</u>	<u>15.12</u>	<u>29.24</u>	<u>8.96</u>
Direct (frozen base)	<u>16.30</u>	<u>13.24</u>	0.576	49.22	26.43	44.05	1.71
w/o Local Support†	0.43	0.83	0.760	43.03	39.97	47.29	1.65
Full ObjAligner	**62.74**	**44.28**	**0.157**	**91.18**	**8.25**	**17.16**	**27.84**

the same predicted transforms support both motion estimation and visible-surface fusion under hand interaction, despite substantial changes in visible support across views. On OnePose, ObjAligner achieves the lowest rotation P90 but trails π^3 in translation-direction and joint AUC. The shared object frame therefore offers its clearest benefit when independently moving object observations must be related across views, rather than for static-object pose in general.

4.5 Ablation Study

We conduct a detailed ablation analysis of ObjAligner to evaluate its key design choices in Table 4. We compare the full model with foreground-isolated π^3 (adapted π^3) and direct pose regression from the frozen Stage-I base (direct), and assess its dependence on target-local evidence by disabling Zoom and LocalSkip only at inference (w/o local support).

Full ObjAligner improves both pose and reconstruction over the frozen-base Direct variant, consistent with the benefit of deriving pose from shared geometry rather than an independent pose head. Its gain over adapted π^3 further indicates that foreground isolation alone is insufficient. Disabling local support at inference sharply degrades both outputs. Because this variant is not retrained, the result establishes the trained model's reliance on local evidence, not the performance of an architecture trained without that branch. The same rigid transforms recover pose and fuse visible geometry, so evaluating both outputs tests whether estimated motion supports a coherent surface. Pose-only metrics miss errors in local geometry, whereas reconstruction-only metrics may obscure motion errors on geometrically ambiguous objects. Additional sensitivity analyses are included in Appendix E.

5 Conclusion

We presented ObjAligner, a feed-forward framework for joint 6D pose estimation and reconstruction of unseen rigid objects from RGB sequences. It represents the target through dense camera-local geometry and sparse geometry in a sequence-shared object frame. Rigid alignment between corresponding points yields pose, and the same transform fuses local points for reconstruction. Because the shared frame needs only to be consistent within a sequence, this formulation does not require canonical coordinates or test-time optimization. Extensive experiments show that ObjAligner outperforms strong baselines in pose under joint motion and hand interaction while remaining competitive in reconstruction. These results support sequence-shared geometry as a basis for both tasks.

AI USE STATEMENT

We used generative AI tools for language polishing and to assist in generating portions of the research code. All AI-assisted code was manually reviewed, tested, and verified to ensure its correctness. We did not use generative AI tools to generate the core scientific conclusions or experimental results of this work. We reviewed all AI-assisted content and verified the factual statements, mathematical formulations, and experimental results reported in the paper. The authors take full responsibility for the final content of this paper and its correctness.

REPRODUCIBILITY STATEMENT

We have taken several steps to support the reproducibility of this work. The model formulation, training objectives, and experimental setup are described in Secs. 3–4. Appendix B provides detailed implementation and training specifications, including the model configuration, optimization procedure, data mixture, checkpoint selection, baseline input adaptations, and rigid-alignment procedure. Appendix C specifies the evaluation metrics, dataset protocols, alignment procedures, and reconstruction scoring, while Appendix E reports the sensitivity to key inference-time choices. The supplementary code provides the implementation, configuration files, data-preparation utilities, and training and evaluation scripts used in our experiments; required external datasets and pretrained initialization are documented in the accompanying instructions.

REFERENCES

Yohann Cabon, Lucas Stoffl, Leonid Antsfeld, Gabriela Csurka, Boris Chidlovskii, Jerome Revaud, and Vincent Leroy. MUSt3R: Multi-view Network for Stereo 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2503.01661.

Weirong Chen, Chuanxia Zheng, Ganlin Zhang, Andrea Vedaldi, and Daniel Cremers. NOVA3R: Non-pixel-aligned Visual Transformer for Amodal 3D Reconstruction. In *International Conference on Learning Representations*, 2026a. URL https://arxiv.org/abs/2603.04179.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. ForeHOI: Feed-forward 3D Object Reconstruction from Daily Hand-Object Interaction Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8868–8879, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_ForeHOI_Feed-forward_3D_Object_Reconstruction_from_Daily_Hand-Object_Interaction_Videos_CVPR_2026_paper.html.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: Monocular Hand-held Objects Reconstruction with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6046–6057, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Chen_HORT_Monocular_Hand-held_Objects_Reconstruction_with_Transformers_ICCV_2025_paper.html.

Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. GPV-Pose: Category-Level Object Pose Estimation via Geometry-Guided Point-Wise Voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6781–6791, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Di_GPV-Pose_Category-Level_Object_Pose_Estimation_via_Geometry-Guided_Point-Wise_Voting_CVPR_2022_paper.html.

Haiwen Feng, Junyi Zhang, Qianqian Wang, Yufei Ye, Pengcheng Yu, Michael J. Black, Trevor Darrell, and Angjoo Kanazawa. St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8503–8513, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Feng_St4RTrack_Simultaneous_4D_Reconstruction_and_Tracking_in_the_World_ICCV_2025_paper.html.

Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. URL https://arxiv.org/abs/1906.02739.

Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, and Vincent Lepetit. In-Hand 3D Object Scanning From an RGB Sequence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17079–17088, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Hampali_In-Hand_3D_Object_Scanning_From_an_RGB_Sequence_CVPR_2023_paper.html.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35103–35115. Curran Associates, Inc., 2022. doi: 10.52202/068431-2544. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e43f900f571de6c96a70d5724a0fb565-Paper-Conference.pdf.

Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. BOP: Benchmark for 6D Object Pose Estimation. In *European Conference on Computer Vision (ECCV)*, 2018. URL https://arxiv.org/abs/1808.08319.

Mark He Huang, Lin Geng Foo, Christian Theobalt, Ying Sun, and De Wen Soh. OnlineSplatter: Pose-Free Online 3D Reconstruction for Free-Moving Objects. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2510.20605.

Varun Jampani, Kevis-kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, Andre Araujo, Ricardo Martin Brualla, Kaushal Patel, Daniel Vlasic, Vittorio Ferrari, Ameesh Makadia, Ce Liu, Yuanzhen Li, and Howard Zhou. NAVI: Category-Agnostic Image Collections with High-Quality 3D Shape and Pose Annotations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 76061–76084. Curran Associates, Inc., 2023. doi: 10.52202/075280-3323. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/efc90033e6e1b05485312dd09fe302b8-Paper-Datasets_and_Benchmarks.pdf.

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. In *International Conference on 3D Vision (3DV)*, 2026. URL https://arxiv.org/abs/2509.13414.

Taeyeop Lee, Bowen Wen, Minjun Kang, Gyuree Kang, In So Kweon, and Kuk-Jin Yoon. Any6D: Model-free 6D Pose Estimation of Novel Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11633–11643, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_Any6D_Model-free_6D_Pose_Estimation_of_Novel_Objects_CVPR_2025_paper.html.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2406.09756.

Weihang Li, Lorenzo Garattoni, Fabien Despinoy, Nassir Navab, and Benjamin Busam. Object Pose Transformer: Unifying Unseen Object Pose Estimation. *arXiv preprint arXiv:2603.23370*, 2026. URL https://arxiv.org/abs/2603.23370.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the Visual Space from Any Views. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/pdf/9d69fc515ade062e6a76a0e4c7e6835ad3e21402.pdf.

Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27906–27916, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_SAM-6D_Segment_Anything_Model_Meets_Zero-Shot_6D_Object_Pose_Estimation_CVPR_2024_paper.html.

Tzu-Yuan Lin, Ho Jae Lee, Kevin Doherty, Yonghyeon Lee, and Sangbae Kim. Point2Pose: Occlusion-Recovering 6D Pose Tracking and 3D Reconstruction for Multiple Unknown Objects Via 2D Point Trackers. *arXiv preprint arXiv:2604.10415*, 2026b. URL https://arxiv.org/abs/2604.10415.

Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6D: Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://arxiv.org/abs/2204.10776.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking Generalizable Bimanual Tool-ACtion-Object Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_TACO_Benchmarking_Generalizable_Bimanual_Tool-ACtion-Object_Understanding_CVPR_2024_paper.html.

Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4World: Feedforward World-centric Dense 3D Tracking of All Pixels. In *European Conference on Computer Vision (ECCV)*, 2026. URL https://arxiv.org/abs/2603.02573.

Xiangting Meng, Jiaqi Yang, Mingshu Chen, Chenxin Yan, Yujiao Shi, Wenchao Ding, and Laurent Kneip. DynOPETs: A Versatile Benchmark for Dynamic Object Pose Estimation and Tracking in Moving Camera Scenarios. *IEEE Robotics and Automation Letters*, 2025. URL https://arxiv.org/abs/2503.19625.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. URL https://github.com/autonomousvision/occupancy_networks.

Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. GigaPose: Fast and Robust Novel Object Pose Estimation via One Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9903–9913, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Nguyen_GigaPose_Fast_and_Robust_Novel_Object_Pose_Estimation_via_One_CVPR_2024_paper.html.

Jinyuan Qu, Hongyang Li, and Lei Zhang. SegVGGT: Joint 3D Reconstruction and Instance Segmentation from Multi-View Images. *arXiv preprint arXiv:2603.19926*, 2026. URL https://arxiv.org/abs/2603.19926.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment Anything in Images and Videos. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.00714.

Haixin Shi, Yinlin Hu, Daniel Koguciuk, Juan-Ting Lin, Mathieu Salzmann, and David Ferstl. Free-Moving Object Reconstruction and Pose Estimation with Virtual Camera. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2405.05858.

Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IEEE/RSJ International Conference on Intelligent*

Robots and Systems (IROS), pp. 573–580, 2012. doi: 10.1109/IROS.2012.6385773. URL https://doi.org/10.1109/IROS.2012.6385773.

Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-Shot Object Pose Estimation Without CAD Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6825–6834, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Sun_OnePose_One-Shot_Object_Pose_Estimation_Without_CAD_Models_CVPR_2022_paper.html.

Marwan Taher, Ignacio Alzugaray, Kirill Mazur, Xin Kong, and Andrew Davison. KV-Tracker: Real-Time Pose Tracking with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28990–28999, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Taher_KV-Tracker_Real-Time_Pose_Tracking_with_Transformers_CVPR_2026_paper.html.

Maxim Tatarchenko, Stephan R. Richter, Rene Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What Do Single-View 3D Reconstruction Networks Learn? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Tatarchenko_What_Do_Single-View_3D_Reconstruction_Networks_Learn_CVPR_2019_paper.html.

Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-Guided Direct Regression Network for Monocular 6D Object Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16611–16621, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_GDR-Net_Geometry-Guided_Direct_Regression_Network_for_Monocular_6D_Object_Pose_CVPR_2021_paper.html.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized Object Coordinate Space for Category-Level 6D Object Pose and Size Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Normalized_Object_Coordinate_Space_for_Category-Level_6D_Object_Pose_and_CVPR_2019_paper.html.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VGGT_Visual_Geometry_Grounded_Transformer_CVPR_2025_paper.html.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3D Perception Model with Persistent State. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10510–10522, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Continuous_3D_Perception_Model_with_Persistent_State_CVPR_2025_paper.html.

Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, and Wenyu Liu. Query6DoF: Learning Sparse Queries as Implicit Shape Prior for Category-Level 6DoF Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14055–14064, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Query6DoF_Learning_Sparse_Queries_as_Implicit_Shape_Prior_for_Category-Level_ICCV_2023_paper.html.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_DUSt3R_Geometric_3D_Vision_Made_Easy_CVPR_2024_paper.html.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.13347.

Yue Wang and Justin M. Solomon. Deep Closest Point: Learning Representations for Point Cloud Registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Wang_Deep_Closest_Point_Learning_Representations_for_Point_Cloud_Registration_ICCV_2019_paper.html.

Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E. Bekris. se(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. URL https://arxiv.org/abs/2007.13866.

Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 606–617, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Wen_BundleSDF_Neural_6-DoF_Tracking_and_3D_Reconstruction_of_Unknown_Objects_CVPR_2023_paper.html.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wen_FoundationPose_Unified_6D_Pose_Estimation_and_Tracking_of_Novel_Objects_CVPR_2024_paper.html.

Po-Chen Wu, Yueh-Ying Lee, Hung-Yu Tseng, Hsuan-I Ho, Ming-Hsuan Yang, and Shao-Yi Chien. A Benchmark Dataset for 6DoF Object Pose Tracking. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR Adjunct)*, 2017. URL https://media.ee.ntu.edu.tw/research/OPT/.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *Robotics: Science and Systems (RSS)*, 2018. URL https://arxiv.org/abs/1711.00199.

Jiale Xu, Shenghua Gao, and Ying Shan. FreeSplatter: Pose-free Gaussian Splatting for Sparse-view 3D Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25442–25452, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_FreeSplatter_Pose-free_Gaussian_Splatting_for_Sparse-view_3D_Reconstruction_ICCV_2025_paper.html.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3D Reconstruction of 1000+ Images in One Forward Pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21924–21935, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Fast3R_Towards_3D_Reconstruction_of_1000_Images_in_One_Forward_CVPR_2025_paper.html.

Sheng Yu, Di-Hua Zhai, and Yuanqing Xia. KeyPose: Category-Level 6D Object Pose Estimation with Self-Adaptive Keypoints. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9653–9661, 2025. doi: 10.1609/aaai.v39i9.33046. URL https://ojs.aaai.org/index.php/AAAI/article/view/33046.

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently Reconstructing Dynamic Scenes One D4RT at a Time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7382–7392, 2026a. URL https://openaccess.thecvf.com/

content/CVPR2026/html/Zhang_Efficiently_Reconstructing_Dynamic_Scenes_One_D4RT_at_a_Time_CVPR_2026_paper.html.

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. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.03825.

Ruihang Zhang, Felix Taubner, Pooja Ravi, Kiriakos N. Kutulakos, and David B. Lindell. ProxyPose: 6-DoF Pose Tracking via Video-to-Video Translation. *arXiv preprint arXiv:2607.06555*, 2026b. URL https://arxiv.org/abs/2607.06555.

Songyan Zhang, Yongtao Ge, Jinyuan Tian, Guangkai Xu, Hao Chen, Chen Lv, and Chunhua Shen. POMATO: Marrying Pointmap Matching with Temporal Motions for Dynamic 3D Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5680–5689, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_POMATO_Marrying_Pointmap_Matching_with_Temporal_Motions_for_Dynamic_3D_ICCV_2025_paper.html.

A THEORETICAL PROPERTIES OF SHARED GEOMETRY

This section expands the coordinate-freedom argument in Secs. 3.3–3.4. The purpose is not to claim that rigid alignment itself is new, but to make explicit why a sequence-common shared frame is sufficient for joint pose estimation and reconstruction, why it need not coincide with dataset-defined object axes, and under what conditions sparse correspondences determine the required rigid relation.

A.1 NOTATION AND ASSUMPTIONS

For frame i, let C_i denote its camera coordinate system and let $\mathcal{S}$ denote the sequence-shared object coordinate system. At a query location k, the camera-local and shared points are respectively $\mathbf{x}_{ik}^{C_i}$ and $\mathbf{y}_{ik}^{\mathcal{S}}$. The rigid transform recovered by Eq. 4 is

$$\mathbf{Q}_i = \mathbf{T}_{\mathcal{S}\leftarrow C_i} \in SE(3), \qquad \mathbf{y}_{ik}^{\mathcal{S}} = \mathbf{Q}_i \mathbf{x}_{ik}^{C_i}, \tag{10}$$

where homogeneous coordinates are used whenever a transform is applied to a 3D point. Let $\mathbf{A}_i^{\mathrm{gt}} = \mathbf{T}_{C_i\leftarrow O_i}$ denote the ground-truth object-to-camera transform under the benchmark convention. The learned frame $\mathcal{S}$ is not required to equal the benchmark object frame.

A.2 SEQUENCE-COMMON COORDINATE FREEDOM LEAVES RELATIVE MOTION UNCHANGED

Proposition 1 (invariance to a sequence-common rigid coordinate change). Let $\mathbf{B} \in SE(3)$ be any rigid transform applied identically to the shared coordinates of every frame, such that

$$\mathbf{y}_{ik}^{\mathcal{S}'} = \mathbf{B}\mathbf{y}_{ik}^{\mathcal{S}}, \qquad \mathbf{Q}_i' = \mathbf{B}\mathbf{Q}_i. \tag{11}$$

Then the relative transforms induced by the shared–local relation are unchanged:

$$(\mathbf{Q}_i')^{-1}\mathbf{Q}_j' = \mathbf{Q}_i^{-1}\mathbf{Q}_j, \tag{12}$$

and the first-frame consistency coordinates used by Eq. 9 are also unchanged:

$$(\mathbf{Q}_0')^{-1}\mathbf{y}_{ik}^{\mathcal{S}'} = \mathbf{Q}_0^{-1}\mathbf{y}_{ik}^{\mathcal{S}}. \tag{13}$$

Proof. By substitution,

$$(\mathbf{Q}_i')^{-1}\mathbf{Q}_j' = (\mathbf{B}\mathbf{Q}_i)^{-1}(\mathbf{B}\mathbf{Q}_j) \tag{14}$$

$$= \mathbf{Q}_i^{-1}\mathbf{B}^{-1}\mathbf{B}\mathbf{Q}_j = \mathbf{Q}_i^{-1}\mathbf{Q}_j. \tag{15}$$

Similarly,

$$(\mathbf{Q}_0')^{-1}\mathbf{y}_{ik}^{\mathcal{S}'} = (\mathbf{B}\mathbf{Q}_0)^{-1}(\mathbf{B}\mathbf{y}_{ik}^{\mathcal{S}}) \tag{16}$$

$$= \mathbf{Q}_0^{-1}\mathbf{B}^{-1}\mathbf{B}\mathbf{y}_{ik}^{\mathcal{S}} = \mathbf{Q}_0^{-1}\mathbf{y}_{ik}^{\mathcal{S}}. \tag{17}$$

Thus one global choice of origin and axes for $\mathcal{S}$ cancels from both supervision paths. Importantly, the same $\mathbf{B}$ must be used for every frame; a frame-specific transform would not cancel from Eq. 12. □

Consequence. The supervision needs to determine only a shared relation across the sequence, not dataset-wide canonical axes. This is precisely the distinction between an admissible sequence-common coordinate freedom and an inadmissible frame-dependent drift.

Scale is a separate ambiguity. The proposition holds at a fixed sequence scale. If both local and shared points are multiplied by the same positive scalar s, then $s\mathbf{y} = \mathbf{R}(s\mathbf{x}) + s\mathbf{t}$; the rotation is unchanged and the translation is multiplied by s. Thus the monocular sequence scale and the rigid choice of object axes are distinct. A sequence-common $Sim(3)$ used for reconstruction scoring accounts for both, whereas the shared branch introduces no independent per-frame scale. Translation-direction scores do not establish metric translation accuracy.

A.3 Reconstruction is equivariant to the shared-frame choice

Proposition 2 (reconstruction equivariance). Let the fused reconstruction in the learned shared frame be

$$\mathbf{X}^{\mathcal{S}}_{\text{fused}} = \bigcup_i \mathbf{Q}_i \mathbf{X}^{\text{obj}}_i. \tag{18}$$

Under the common change of coordinates in Eq. 11, the new fused reconstruction satisfies

$$\mathbf{X}^{\mathcal{S}'}_{\text{fused}} = \mathbf{B}\mathbf{X}^{\mathcal{S}}_{\text{fused}}. \tag{19}$$

Proof. For each frame,

$$\mathbf{Q}'_i \mathbf{X}^{\text{obj}}_i = \mathbf{B}\mathbf{Q}_i \mathbf{X}^{\text{obj}}_i. \tag{20}$$

Taking the union over frames gives Eq. 19. Therefore the choice of $\mathcal{S}$ changes only the global coordinates in which the reconstruction is expressed; it does not alter its intrinsic shape or cross-view consistency. □

Implication for evaluation. A ground-truth fused cloud may be transformed into the same predicted shared frame for visualization or geometric comparison, provided this use of ground truth remains strictly evaluation-only. The predicted fusion itself continues to use only predicted local geometry, the predicted target mask, and the predicted $\mathbf{Q}_i$.

A.4 Why sparse shared geometry is sufficient

Proposition 3 (three non-collinear correspondences determine a rigid transform). Assume noise-free correspondences $(\mathbf{x}_k, \mathbf{y}_k)$ with positive weights, and suppose at least three source points $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are non-collinear. If two proper rigid transforms $(\mathbf{R}, \mathbf{t})$ and $(\widetilde{\mathbf{R}}, \widetilde{\mathbf{t}})$ map all three points to the same targets, then the transforms are identical.

Proof. For $k \in \{1, 2, 3\}$,

$$\mathbf{R}\mathbf{x}_k + \mathbf{t} = \widetilde{\mathbf{R}}\mathbf{x}_k + \widetilde{\mathbf{t}}. \tag{21}$$

Subtract the equation for $k = 1$ from those for $k = 2, 3$:

$$\mathbf{R}(\mathbf{x}_k - \mathbf{x}_1) = \widetilde{\mathbf{R}}(\mathbf{x}_k - \mathbf{x}_1), \qquad k \in \{2, 3\}. \tag{22}$$

Hence $\widetilde{\mathbf{R}}^{-1}\mathbf{R}$ fixes two linearly independent vectors. A proper 3D rotation that fixes two independent vectors must be the identity, so $\mathbf{R} = \widetilde{\mathbf{R}}$. Substituting back gives $\mathbf{t} = \widetilde{\mathbf{t}}$. □

Practical interpretation. The shared branch need not reproduce every visible surface point. It needs enough well-distributed correspondences to make the rigid fit observable. In practice, redundancy beyond the minimal three points is useful because prediction noise, partial visibility, and imperfect masks make the fit statistical rather than exact. Conversely, nearly collinear support or extremely small visible regions can make the solution poorly conditioned even when many query points are present. The uniqueness statement assumes fixed, correct correspondences; it does not resolve the correspondence ambiguity of a symmetric object. Non-collinear coplanar points already determine a proper rigid transform in the noise-free setting, so planarity alone is not a failure of uniqueness.

A.5 A sequence-wide calibration cannot hide frame-dependent errors

Proposition 4 (single-calibration identifiability). Suppose dataset-frame object poses satisfy $\mathbf{Q}^\star_i = \mathbf{B}^\star(\mathbf{A}^{\text{gt}}_i)^{-1}$ for one sequence-common $\mathbf{B}^\star$. For arbitrary predictions $\widehat{\mathbf{Q}}_i$, there exists one rigid calibration $\mathbf{B}$ that makes every calibrated pose exact,

$$\mathbf{B}^{-1}\widehat{\mathbf{Q}}_i = (\mathbf{A}^{\text{gt}}_i)^{-1} \quad \forall i, \tag{23}$$

if and only if the quantity $\widehat{\mathbf{Q}}_i \mathbf{A}^{\text{gt}}_i$ is identical for all frames.

Proof. If Eq. 23 holds, then $\widehat{\mathbf{Q}}_i \mathbf{A}^{\text{gt}}_i = \mathbf{B}$ for every i. Conversely, if $\widehat{\mathbf{Q}}_i \mathbf{A}^{\text{gt}}_i$ is a constant transform $\mathbf{B}$, then Eq. 23 follows immediately. □

Consequence. The evaluation calibration removes only the unobservable sequence-common frame choice. It cannot independently correct different frames. Any frame-dependent rotation or translation error remains in the relative trajectory and pairwise motion metrics.

A.6 Degeneracy and symmetry

The above arguments assume that the visible shared support is geometrically informative. In practice, nearly collinear support, extremely small visible regions, or repeated structures can make the weighted rigid alignment poorly conditioned. Object symmetry introduces a related ambiguity because multiple rotations may be physically equivalent. We keep raw rotation metrics as the primary report. The headline angular evaluation uses benchmark rotations without symmetry correction; Appendix D.3 additionally reports symmetry-aware rotation.

A.7 How pose and local geometry jointly affect fusion

Proposition 5 (pointwise fusion error bound). Fix one evaluation scale and common coordinate frame. Let $\mathbf{x}$ be a true camera-local point, $\widehat{\mathbf{x}}$ its prediction, and $(\mathbf{R}, \mathbf{t})$, $(\widehat{\mathbf{R}}, \widehat{\mathbf{t}})$ the true and predicted local-to-shared transforms. If their rotation discrepancy is $\theta \in [0, \pi]$, then

$$\|\widehat{\mathbf{R}}\widehat{\mathbf{x}} + \widehat{\mathbf{t}} - (\mathbf{R}\mathbf{x} + \mathbf{t})\|_2 \leq \|\widehat{\mathbf{x}} - \mathbf{x}\|_2 + 2\sin(\theta/2)\|\mathbf{x}\|_2 + \|\widehat{\mathbf{t}} - \mathbf{t}\|_2. \tag{24}$$

Proof. Add and subtract $\widehat{\mathbf{R}}\mathbf{x}$, apply the triangle inequality, and use $\|\widehat{\mathbf{R}}\mathbf{v}\|_2 = \|\mathbf{v}\|_2$ and $\|\widehat{\mathbf{R}} - \mathbf{R}\|_2 = 2\sin(\theta/2)$. □

The bound separates local geometric error from pose error. It applies to a corresponding visible point, not to mask false positives or unobserved surfaces. Those additionally affect nearest-neighbor reconstruction scores. Better angular pose therefore need not improve every surface metric, and translation-axis agreement alone does not bound the translation-vector term.

B Implementation and Reproducibility Details

B.1 Final model and inference contract

ObjAligner is initialized from π^3 and takes RGB views with one first-view target box. Table 5 summarizes its configuration. The 12×12 query grid in Sec. 3.3 is the training setting; the benchmark tables use 32×32 queries and fusion threshold 0.75. The structural ablation (Table 4) and qualitative examples retain the original 12×12, 0.5 setting.

Optimization and inputs. Both stages freeze the visual encoder; Stage II initializes from the best Stage-I checkpoint and updates the remaining enabled pathways. AdamW uses $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and no decay for biases or one-dimensional parameters. The cosine OneCycle schedule has no warm-up, division factors 1,000 and 100, and momentum range 0.85–0.95. Training samples N uniformly from 2–12, aspect ratios from 0.5–2, and image areas from 100,000–200,000 pixels, rounded to 14-pixel patches. Augmentation uses color jitter, JPEG degradation, blur, crop margin 16, and focal factor 0.9. Main-benchmark inputs are 448×588 (height $\times$ width), except OnePose (518×518).

Computational resources. Both training stages use four NVIDIA RTX 4090 GPUs with bfloat16 mixed precision and one sequence per GPU. The complete 20-epoch Stage-I and 50-epoch Stage-II runs take approximately 13.8 and 33.9 hours, respectively, including validation and checkpoint saving.

Checkpoint selection. The final checkpoint minimizes validation loss (Stage-II epoch index 45). Validation includes NAVI validation objects and COPE119 Test35; these domains are marked seen. COPE119 training uses the complementary 84 sequences. The other reported evaluation datasets do not enter target-domain training or checkpoint selection.

Query sampling and rigid fitting. The decoder samples bbox-relative cell centers bilinearly with `align_corners=False`: border padding for image features, zero padding for local points and

Table 5: Model and training configuration used by the final formulation.

Component	Setting
Base model	original π^3 pretrained model
Point head	π^3 local point-map head
Object-token grid	8×8 (64 tokens per view)
Shared query grid (training / refined inference)	12×12 / 32×32
High-resolution target crop	168×168
Patch size	14
Target mask loss	foreground/background balanced BCE
Training views	dynamic $N = 2 \dots 12$
Default evaluation views	$N = 8$
Evaluation seed	3407
Stage-I logical length	10,000 samples/epoch
Stage-I epochs / learning rate	20 / 10^{-4}
Stage-II epochs / learning rate	50 / 5×10^{-5}
Shared decoder	5 blocks, width 1024, 16 heads
Shared decoder output	512 channels, linear 3D head
Optimizer	AdamW; weight decay 0.05
Devices / sequences per device	4 / 1
Gradient accumulation / clipping	1 / global norm 1.0
Precision	bfloat16 (SVD computed in float32)
Optimizer steps per epoch	1,250
Stage-I / Stage-II schedule steps	25,000 / 62,500

mask probabilities. It uses a 64×64 RoPE range, no frame-index embedding, zero bbox/zoom padding, and unit zoom residual scale. For weighted centroids $\bar{\mathbf{x}}, \bar{\mathbf{y}}$, the rigid fit is

$$\mathbf{H} = \frac{\sum_k w_k (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{y}_k - \bar{\mathbf{y}})^\top}{\sum_k w_k} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{25}$$

$$\mathbf{R} = \mathbf{V}\,\mathrm{diag}(1, 1, \det(\mathbf{V}\mathbf{U}^\top))\mathbf{U}^\top, \qquad \mathbf{t} = \bar{\mathbf{y}} - \mathbf{R}\bar{\mathbf{x}}. \tag{26}$$

Inference removes nonfinite points, source depths $\leq 10^{-6}$, and weights below 0.05; it returns identity if the remaining weight sum is ≤ 6. Centroid denominators are clamped at 10^{-6}. Training adds 10^{-4} to finite-point weights and checks for at least six finite points. Bbox extraction and calibration support use mask threshold 0.5; refined fusion uses 0.75; fewer than 16 foreground pixels trigger bbox fallback.

B.2 Baseline input adaptations

SAM 2 (Ravi et al., 2025) propagates masks from the first-view box without later-frame GT annotations. π^3 retains RGB inside each mask's tight box and blacks out the rest (*black-bbox*); MapAnything, Fast3R, MUSt3R, and CUT3R retain only foreground RGB. DA3 uses full RGB and extracts masked geometry; FreeSplatter-O resizes the foreground onto white. ObjAligner uses full RGB and predicts its own masks. Thus, views and initial annotation are shared, but input adaptation differs. The adapted black-box control in structural/controlled experiments instead uses frozen Stage-I masks. For OnePose, the SAM2 initialization additionally derives a positive point at the supplied box center; it requires no extra annotation.

B.3 Final five-source training mixture

Table 6 lists sampling allocations per logical epoch. NAVI uses its object-based training split; TACO uses ego-view data with derived SAM2 masks; OPT3D uses its prepared training manifest. COPE119 rotations are projected to $SO(3)$.

B.4 Expanded object-geometry objective

The compact objective in Eq. 7 groups several lower-level geometry terms. The saved run weights are one for scene geometry and camera-relative pose, ten for balanced mask BCE and the object-local

Table 6: Final five-source training mixture and the role of each source. The logical epoch length is 10,000.

Source	Count	Weight	Primary role
SeqMix / SyntheticObject	5,500	55%	exact synthetic geometry and diverse joint camera–object motion
NAVI	1,000	10%	clean real object-centric multi-view geometry
TACO	1,500	15%	real hand–object interaction and occlusion; target supervision is derived rather than treated as released GT
OPT3D	1,000	10%	real geometry with native target support
COPE119	1,000	10%	real joint camera–object motion closely matched to the target task
Total	10,000	100%	

point term, and, in Stage II, one for relative object pose and 0.1 for point–pose consistency. The latter two weights are zero in Stage I. The pose losses combine angular rotation with a Huber translation loss after detached sequence-scale calibration: camera translation has multiplier 100 and Huber delta 0.1; object translation has multiplier 1,000 and delta 0.01. The point–pose consistency loss uses smooth L1 with beta 0.01. Object-normal supervision has coefficient one; the structure terms below enter separately, so the coefficient ten on the local point term does not multiply the entire structure group. In the current implementation, object-local supervision uses balanced target-mask supervision together with local point/normal constraints and a structure group

$$\mathcal{L}_{\text{structure}} = 10\mathcal{L}_{\text{grid}} + 2\mathcal{L}_{\text{logz}} + 10\mathcal{L}_{\text{grad}} + \lambda_{\text{surf}}\mathcal{L}_{\text{surf}} + 10\mathcal{L}_{\text{edge}}. \tag{27}$$

The surface weight is $\lambda_{\text{surf}} = 5$ in Stage I and zero in Stage II. Here $\mathcal{L}_{\text{grad}}$ acts on smooth non-edge interior neighbors, $\mathcal{L}_{\text{surf}}$ acts on same-surface pairs that cross decoder patch boundaries, and $\mathcal{L}_{\text{edge}}$ preserves depth discontinuities. Each component enters the total loss once.

The patch-boundary term supervises local 3D differences rather than forcing neighboring points to coincide:

$$\mathcal{L}_{\text{surf}} = \left\| (\widehat{\mathbf{X}}_q - \widehat{\mathbf{X}}_p) - (\mathbf{X}_q - \mathbf{X}_p) \right\|_1, \tag{28}$$

computed on object, depth-valid, non-edge adjacent pairs that cross a 14×14 patch boundary. This construction targets patch-seam artifacts without imposing ordinary smoothness across genuine geometry edges.

C Extended Evaluation Protocol

Bold and underlined table entries denote the best and second-best results within each dataset; – indicates an unevaluated metric.

C.1 Pose metrics

For each eight-view sequence, all 28 pairs compare $\widehat{\mathbf{Q}}_i^{-1}\widehat{\mathbf{Q}}_j$ with $\mathbf{A}_i\mathbf{A}_j^{-1}$. Rotation and translation-axis errors are

$$e_R = \cos^{-1}\left[\text{clip}\left(\frac{\text{tr}(\widehat{\mathbf{R}}_{ij}(\mathbf{R}_{ij}^{\text{gt}})^\top) - 1}{2}, -1, 1\right)\right], \tag{29}$$

$$e_{T\text{-dir}} = \cos^{-1}\left[\text{clip}\left(\frac{|\widehat{\mathbf{t}}_{ij}^\top \mathbf{t}_{ij}^{\text{gt}}|}{\|\widehat{\mathbf{t}}_{ij}\|\,\|\mathbf{t}_{ij}^{\text{gt}}\|}, 0, 1\right)\right]. \tag{30}$$

RRA and RTA are the corresponding threshold recalls; translation-axis error is sign-insensitive. GT translations shorter than 10^{-5} m are excluded from direction metrics. With $r(u) = \Pr[\max(e_R, e_{T\text{-dir}}) \le u]$, the reported discrete AUC is

$$\text{AUC@}\tau = \frac{1}{\tau}\sum_{u=1}^{\tau} r(u), \qquad \tau \in \{5, 15, 30\}. \tag{31}$$

Means, P90 (90th percentiles), and recalls are computed within each sequence and then averaged equally across sequences. Rotation and translation-direction P90 errors are in degrees; recall and AUC are reported as percentages. OnePose uses 80 sequences, with all methods sharing eight-frame center stride-24 windows. Flagged failed pairs receive 180° rotation and 90° valid translation-axis error. Angular scores use raw rotations; HO3D ADD/ADD-S AUC up to 10 cm is evaluated separately with a first-view depth/pose anchor.

C.2 Trajectory diagnostics

Physical object trajectories combine predicted camera poses and calibrated object nodes, followed by one sequence-wide Sim(3) alignment. ATE summarizes position errors; RPE compares consecutive selected views. These SeqMix diagnostics use 50 training-overlapping scenes and are separate from the shared-frame angular scores.

C.3 Reconstruction metrics

Predicted poses fuse the predicted foreground local geometry. Scoring applies one first-view correspondence-based Sim(3) to the entire fusion, without ICP or per-frame alignment. The fit uses intersecting predicted/GT foreground with valid depth (32–6,000 correspondences, seed 3407). Both clouds are voxelized at 1 mm and capped at 60,000 points. CD is the mean of the two directed mean nearest-neighbor distances (cm); Chamfer-L2 averages squared distances (cm^2) (Mescheder et al., 2019). Precision and recall use distances strictly below δd_{GT}; F-score is their harmonic mean, with default $\delta = 0.02$. Unsigned normal consistency uses symmetric nearest-neighbor matches and 20-neighbor PCA normals (Gkioxari et al., 2019). All scores are sequence-macro averages. The main reconstruction populations are COPE119 (35), UOPE56 (56), HO3D (13), and NAVI (110); frame sampling follows Sec. 4.1.

D Additional Quantitative Results

The following comparisons use a 32×32 query grid and fusion threshold 0.75, with the original checkpoint and input groups. The controlled trajectory diagnostic retains its original configuration.

D.1 External pose transfer

Table 7 extends the six-method comparison to five exploratory subsets: HouseCat6D (10 bottle/cup targets across five official test scenes), H2O (six sequences from the prepared `pose_train` split), HILO and TransCG (six targets each), and synthetic REGRAD (six targets). The latter three use their locally prepared training partitions, which are external to our final five-source training. Each sample contains eight uniformly spaced views and one first-view box. All five datasets are excluded from our target-domain training and checkpoint selection.

Figure 5 shows joint recall across angular thresholds.

D.2 Full pose threshold curves

Table 8 expands the main-paper motion benchmarks with matched rotation and translation-axis means, recall at 15°, and discrete joint AUC at three thresholds. These quantities are aggregated from the same pair records and sequence populations as the headline scores. Figure 6 shows the underlying joint-recall curves.

D.3 Symmetry-aware rotation

We rescore the fixed predictions using the released category-level symmetry conventions:[1] rotations about the object y-axis for bottles, bowls, and cans in COPE119, and bottles in the evaluated HouseCat6D subset. Other targets retain identity symmetry, including mugs and handled cups. For

[1] COPE119 evaluation code; HouseCat6D evaluation code.

Table 7: Small-scale external pose transfer with frozen weights. All six methods use the same samples within each dataset. Angles are in degrees and AUC in percent; bold/underline indicate best/second-best values within each subset.

Method	$R_{\text{mean}} \downarrow$	$T^{\text{dir}}_{\text{mean}} \downarrow$	AUC@15↑
HouseCat6D ($n = 10$)			
π^3 + black-bbox	<u>7.66</u>	<u>10.64</u>	<u>39.69</u>
MapAnything	22.24	27.42	7.64
Fast3R	27.56	37.78	4.19
MUSt3R	18.24	25.10	9.33
CUT3R	18.99	26.13	10.05
ObjAligner	**2.24**	**3.08**	**79.52**
H2O ($n = 6$)			
π^3 + black-bbox	25.20	37.41	2.38
MapAnything	23.41	36.38	3.57
Fast3R	<u>20.06</u>	33.74	<u>8.29</u>
MUSt3R	21.18	<u>29.84</u>	6.07
CUT3R	39.14	31.61	3.02
ObjAligner	**16.02**	**21.67**	**14.09**
HILO ($n = 6$)			
π^3 + black-bbox	<u>61.81</u>	<u>30.39</u>	<u>15.32</u>
MapAnything	99.73	51.25	1.35
Fast3R	110.39	58.27	0.48
MUSt3R	107.83	55.56	0.00
CUT3R	93.56	45.91	1.90
ObjAligner	**3.37**	**2.81**	**77.74**
TransCG ($n = 6$)			
π^3 + black-bbox	<u>18.53</u>	<u>19.91</u>	<u>25.12</u>
MapAnything	44.20	45.54	3.57
Fast3R	46.26	49.79	0.56
MUSt3R	42.38	37.51	1.75
CUT3R	45.54	54.56	0.00
ObjAligner	**7.49**	**10.01**	**38.41**
REGRAD ($n = 6$)			
π^3 + black-bbox	<u>53.50</u>	<u>32.58</u>	<u>14.80</u>
MapAnything	82.70	42.32	3.89
Fast3R	96.84	48.13	2.26
MUSt3R	87.43	51.78	0.36
CUT3R	69.92	37.78	2.66
ObjAligner	**3.74**	**5.22**	**63.97**

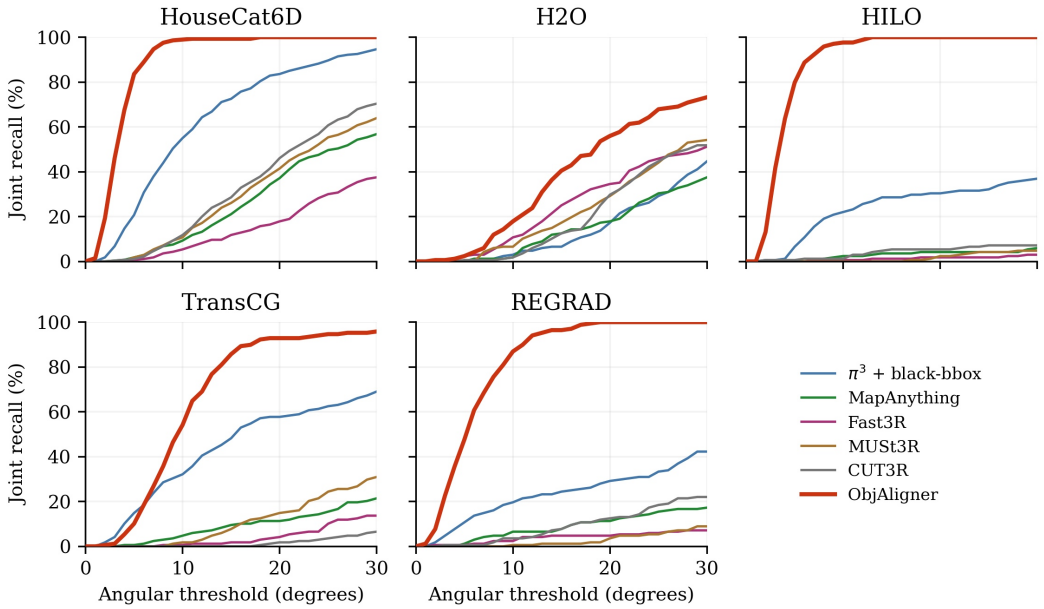

Figure 5: Joint pose recall on the five external subsets in Table 7. A pair succeeds when both angular errors are at most the threshold; curves are sequence-macro averages.

Table 8: Extended pose metrics on the complete original evaluation populations. Angles are in degrees; recall and AUC are percentages. ObjAligner uses the refined inference setting.

Method	$R_{\text{mean}} \downarrow$	$T^{\text{dir}}_{\text{mean}} \downarrow$	RRA@15↑	RTA@15↑	AUC@5↑	AUC@15↑	AUC@30↑
COPE119							
π^3 + black-bbox	17.08	31.36	48.16	24.08	0.16	4.78	21.50
MapAnything	16.63	34.17	52.76	26.63	1.24	7.63	22.68
Fast3R	21.26	39.85	40.92	20.41	1.35	6.76	17.36
MUSt3R	16.90	36.21	51.73	20.92	0.27	5.28	18.10
CUT3R	25.46	41.54	39.59	15.20	0.22	3.86	13.72
ObjAligner	**8.23**	**17.19**	**89.18**	**61.63**	**4.20**	**28.10**	**51.58**
UOPE56							
π^3 + black-bbox	18.49	36.22	49.36	24.36	0.69	7.38	20.63
MapAnything	17.42	37.21	55.23	23.21	0.64	7.64	21.01
Fast3R	18.74	36.12	52.23	25.96	0.85	9.29	23.04
MUSt3R	17.00	38.32	56.82	18.43	0.22	4.03	16.76
CUT3R	17.76	36.62	53.70	23.02	0.46	7.15	20.45
ObjAligner	**10.26**	**19.37**	**82.02**	**54.21**	**2.47**	**21.48**	**45.39**
HO3D							
π^3 + black-bbox	37.16	34.93	47.53	28.85	0.22	6.81	19.32
MapAnything	57.56	43.49	39.01	21.15	0.60	6.04	14.58
Fast3R	49.95	46.77	33.52	10.44	0.05	2.31	7.76
MUSt3R	42.58	43.34	49.73	17.03	0.00	2.14	11.32
CUT3R	46.02	42.16	42.58	18.96	0.33	4.69	13.98
ObjAligner	**30.64**	**29.66**	**54.40**	**40.38**	**0.66**	**10.57**	**27.08**

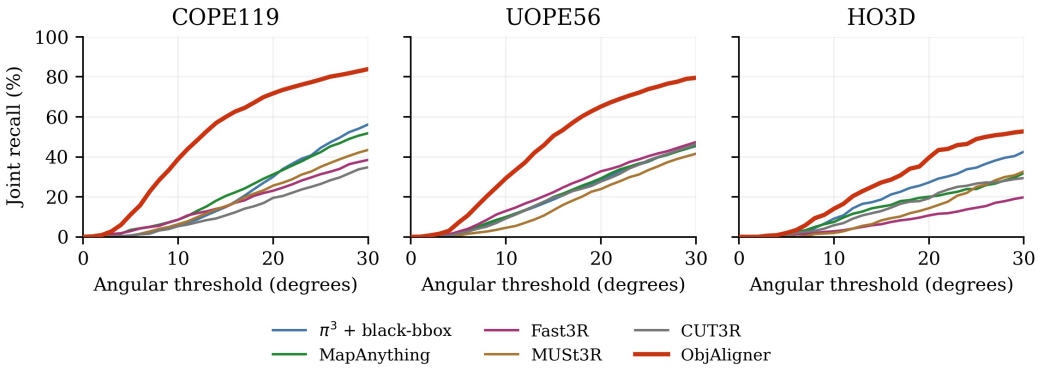

Figure 6: Joint pose recall versus angular threshold for the main motion benchmarks. A pair succeeds when both rotation and translation-axis errors are at most the threshold. Curves are sequence-averaged; the main AUC@15 is the mean recall at integer thresholds 1°–15°.

the object symmetry group G_o, we compute

$$e_{R,ij}^{\text{sym}} = \min_{\mathbf{S}\in G_o} d_{SO(3)}\left(\widehat{\mathbf{R}}_{ij}, \mathbf{R}_i^{\text{gt}}\mathbf{S}(\mathbf{R}_j^{\text{gt}})^\top\right), \tag{32}$$

where $\widehat{\mathbf{R}}_{ij}$ is the rotation of $\widehat{\mathbf{Q}}_i^{-1}\widehat{\mathbf{Q}}_j$ and $\mathbf{R}_i^{\text{gt}}$ is the rotation of $\mathbf{A}_i$. The continuous-axis minimum is evaluated analytically. All methods share the same symmetry rules; Table 9 averages over all sequences, including asymmetric targets. This auxiliary scoring leaves predictions and the raw main-table metrics unchanged.

Table 9: Raw and symmetry-aware rotation on COPE119 Test35 (21 symmetric-category sequences out of 35) and the existing HouseCat6D test subset (5 out of 10 targets). R is the sequence-macro mean pairwise error in degrees; RRA@15 is recall at 15° (%). Raw/Sym use identical predictions and populations.

Method	COPE119 Test35				HouseCat6D (10 targets)			
	$R_{\text{raw}}\downarrow$	$R_{\text{sym}}\downarrow$	$\text{RRA}_{\text{raw}}\uparrow$	$\text{RRA}_{\text{sym}}\uparrow$	$R_{\text{raw}}\downarrow$	$R_{\text{sym}}\downarrow$	$\text{RRA}_{\text{raw}}\uparrow$	$\text{RRA}_{\text{sym}}\uparrow$
π^3 + black-bbox	17.08	14.34	48.16	61.22	7.66	6.01	87.14	92.14
MapAnything	16.63	12.96	52.76	65.92	22.24	17.92	36.43	50.00
Fast3R	21.26	17.49	40.92	54.39	27.56	19.90	27.14	48.57
MUSt3R	16.90	12.09	51.73	69.39	18.24	14.67	51.43	65.36
CUT3R	25.46	19.05	39.59	56.12	18.99	16.25	53.93	65.71
ObjAligner	**8.23**	**5.00**	**89.18**	**98.67**	**2.24**	**1.94**	**100.00**	**100.00**

D.4 Controlled trajectory diagnostics

Table 10: Auxiliary physical-trajectory diagnostics on 50 training-overlapping SeqMix scenes. One sequence-wide object-frame calibration and one sequence-wide Sim(3) are used by the native trajectory evaluator; no per-frame alignment is applied. These are distinct from the shared-frame angular scores. †: inference-only removal.

Method	ATE RMSE↓ (cm)	ATE mean↓ (cm)	RPE-R RMSE↓ ($^\circ$)	RPE-T RMSE↓ (cm)
Adapted π^3	5.78	5.20	55.56	11.86
Direct (frozen base)	4.67	4.23	37.04	11.01
w/o Local Support†	6.31	5.61	92.42	12.80
Full ObjAligner	**2.14**	**1.97**	**9.18**	**4.61**

Table 10 summarizes trajectory errors; Figure 7 shows their cumulative distributions over all 50 sequences.

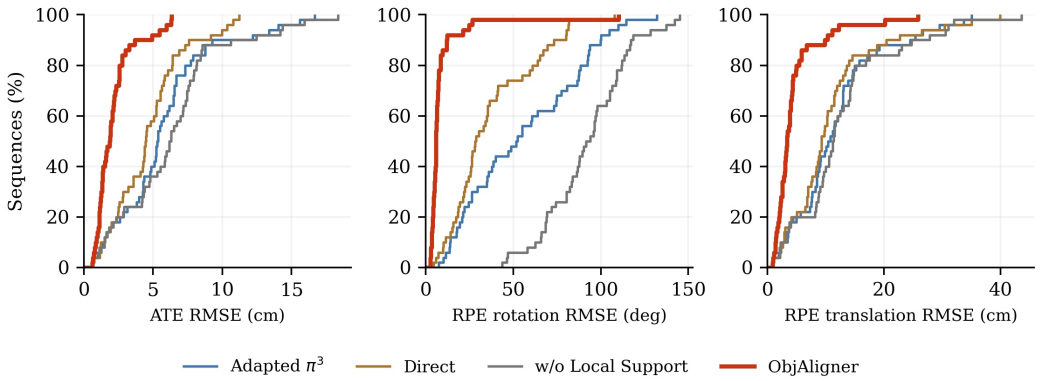

Figure 7: Cumulative trajectory-error distributions on 50 training-overlapping SeqMix scenes under the original inference configuration. Each curve reports the fraction of sequences below an error threshold after one sequence-wide Sim(3). LS denotes local support; its removal is inference-only.

D.5 Additional reconstruction metrics

Tables 11 and 12 report complementary reconstruction metrics side by side; Table 13 reports normal consistency. Figure 8 shows their threshold dependence. ObjAligner has the lowest Chamfer-L2 among the compared methods on COPE119, UOPE56, and NAVI.

Table 11: Reconstruction on NAVI and HO3D. CD/L2: cm/cm^2; P/R/F: precision, recall, and F-score (%) at 2% of the GT diagonal.

Method	CD↓	L2↓	P↑	R↑	F↑
NAVI					
π^3 + black-bbox	**0.148**	0.083	87.62	**95.71**	91.24
MapAnything	0.305	0.333	62.89	83.17	70.97
Fast3R	0.250	0.326	76.27	88.31	81.41
DA3	0.185	0.528	86.71	93.98	89.92
FreeSplatter-O	0.369	0.390	61.09	76.60	67.42
MUSt3R	0.275	0.211	62.62	88.59	72.61
CUT3R	0.798	4.913	36.23	65.17	45.57
ObjAligner	0.157	**0.065**	**88.22**	95.06	**91.25**
HO3D					
π^3 + black-bbox	1.256	4.600	29.46	60.90	38.61
MapAnything	1.130	5.938	35.21	67.66	44.77
Fast3R	0.972	4.679	**43.26**	62.17	**49.63**
DA3	2.020	12.938	22.53	52.90	29.37
FreeSplatter-O	1.295	4.725	28.63	47.27	34.35
MUSt3R	**0.817**	**1.711**	33.57	**68.71**	44.38
CUT3R	3.466	98.082	23.24	58.75	32.43
ObjAligner	0.826	1.838	39.05	66.59	48.23

Table 12: Reconstruction on COPE119 and UOPE56. Units and P/R/F definitions follow Table 11. All four datasets use their complete evaluation populations.

Method	CD↓	L2↓	P↑	R↑	F↑
COPE119					
π^3 + black-bbox	2.015	46.121	21.04	53.83	29.34
MapAnything	1.338	41.145	35.47	61.31	43.77
Fast3R	1.498	42.701	40.11	50.74	43.56
DA3	4.810	93.043	11.06	34.91	15.85
FreeSplatter-O	1.862	45.016	29.26	42.32	33.10
MUSt3R	2.567	1.09×10^5	24.39	54.30	31.25
CUT3R	3.094	74.788	16.28	41.40	22.22
ObjAligner	**1.221**	**38.254**	**46.59**	**64.51**	**52.05**
UOPE56					
π^3 + black-bbox	1.550	39.557	30.44	61.89	39.99
MapAnything	**1.223**	38.004	40.49	**67.26**	49.64
Fast3R	1.247	39.937	**49.80**	57.18	**52.51**
DA3	4.140	59.217	13.29	41.86	18.34
FreeSplatter-O	1.598	41.976	41.87	53.78	45.54
MUSt3R	1.658	1.87×10^4	30.30	63.24	39.57
CUT3R	2.240	46.583	21.01	54.10	29.48
ObjAligner	1.258	**37.067**	48.37	59.80	52.44

E Inference Sensitivity

E.1 Fusion-mask threshold sensitivity

Table 14 varies the fusion-mask threshold on six COPE119 sequences (bottle_02, bowl_04, camera_02, can_02, laptop_01, and mug_02), keeping poses, local geometry, and the first-view alignment fixed. The query grid is 12×12. Higher thresholds improve precision with a small recall trade-off; 0.75 yields the highest F-score in this sweep.

Table 13: Estimated-normal consistency (NC, %). Normals use 20-neighbor PCA and symmetric unsigned nearest-neighbor agreement. NAVI: 110 groups; HO3D: 13 groups.

Method	NAVI NC↑	HO3D NC↑
π^3 + black-bbox	**82.97**	70.70
MapAnything	77.02	<u>72.48</u>
Fast3R	74.52	68.41
DA3	<u>82.80</u>	71.10
FreeSplatter-O	72.21	61.49
MUSt3R	71.18	66.35
CUT3R	69.96	67.48
ObjAligner	82.17	**73.99**

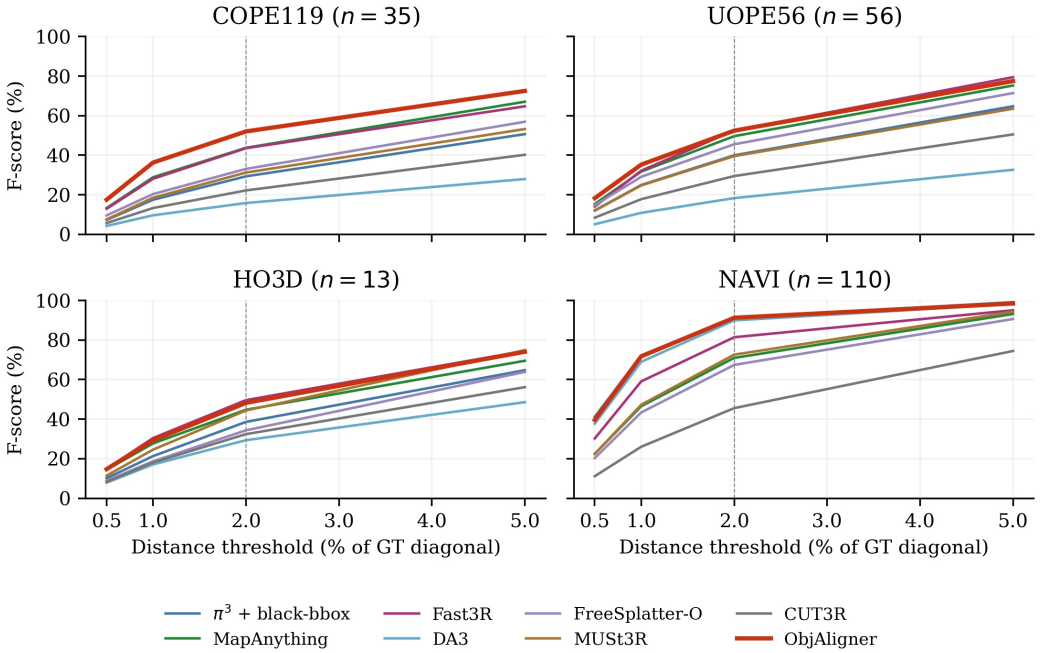

Figure 8: Visible-surface F-score versus distance threshold (percentage of the GT bounding-box diagonal). Each point is the sequence-macro average on the same population, alignment, and point samples as the reconstruction tables. The vertical guide marks the default 2% threshold.

Table 14: Fusion-mask sensitivity on six fixed sequences. CD is in centimeters; precision, recall, and F-score are percentages. Calibration support remains at probability 0.5.

Threshold	CD↓	Precision↑	Recall↑	F-score↑
0.25	0.969	50.37	**60.97**	54.82
0.3	0.969	50.59	60.95	54.93
0.5	0.966	51.32	60.76	55.29
0.7	0.963	51.94	60.52	55.55
0.75	**0.962**	**52.14**	60.46	**55.63**

E.2 Inference-time shared-query density

Table 15 extends the query grid from 8×8 to 48×48 on the same six sequences. The checkpoint, 168×168 crop, and 0.5 fusion threshold are fixed. Joint AUC reaches its maximum at 32×32; 40×40 and 48×48 offer no further gain in this metric. We use 32×32 for the refined inference evaluation.

E.3 Inference-time local crop resolution

Table 16 varies the crop size with a fixed 12×12 query grid and 0.5 fusion threshold. The default 168×168 crop matches the highest joint AUC and retains the highest F-score among the tested sizes and is kept for refined inference.

Table 15: Query-density sensitivity. Angles: degrees; CD: cm; AUC/F-score: %.

Query grid	$R_{\text{mean}} \downarrow$	$T^{\text{dir}}_{\text{mean}} \downarrow$	AUC@15↑	CD↓	F-score↑
8×8	9.58	18.74	22.42	**0.961**	54.83
12×12	9.43	18.75	22.10	0.966	55.29
16×16	9.41	**18.65**	22.74	0.962	55.55
20×20	9.43	18.76	22.46	0.962	55.55
24×24	9.40	18.71	22.74	0.963	**55.58**
32×32	9.40	18.71	**22.86**	0.962	55.55
40×40	**9.39**	18.69	22.82	0.963	55.48
48×48	9.39	18.70	**22.86**	0.962	55.52

Table 16: Crop-resolution sensitivity on the same six sequences. Units follow Table 15.

Crop size	$R_{\text{mean}} \downarrow$	$T^{\text{dir}}_{\text{mean}} \downarrow$	AUC@15↑	CD↓	F-score↑
112×112	9.47	**18.69**	**22.10**	**0.965**	55.28
168×168	**9.43**	18.75	**22.10**	0.966	**55.29**
224×224	9.46	18.79	21.83	0.966	55.28

F Additional Qualitative Results

Figures 9–11 show additional COPE119 and NAVI examples from the original inference setting. Each block presents GT, π^3, MapAnything, and ObjAligner, with RGB, pose overlays, and fused surfaces. Dashed teal and solid red boxes indicate GT and predicted poses, respectively.

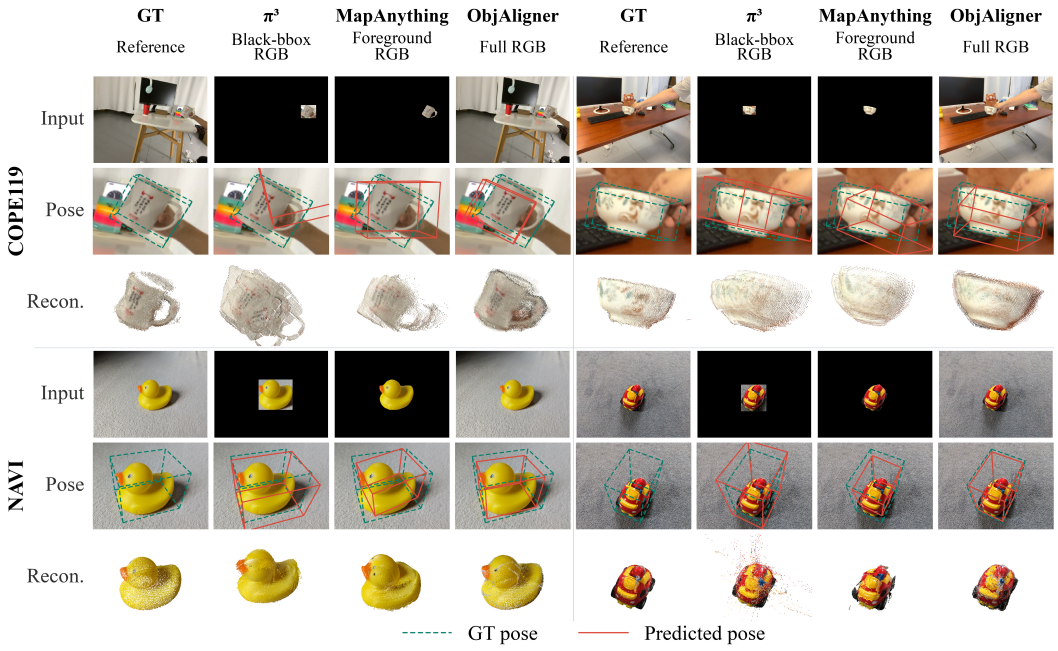

Figure 9: Pose and visible-surface comparisons on COPE119 (top) and NAVI (bottom).

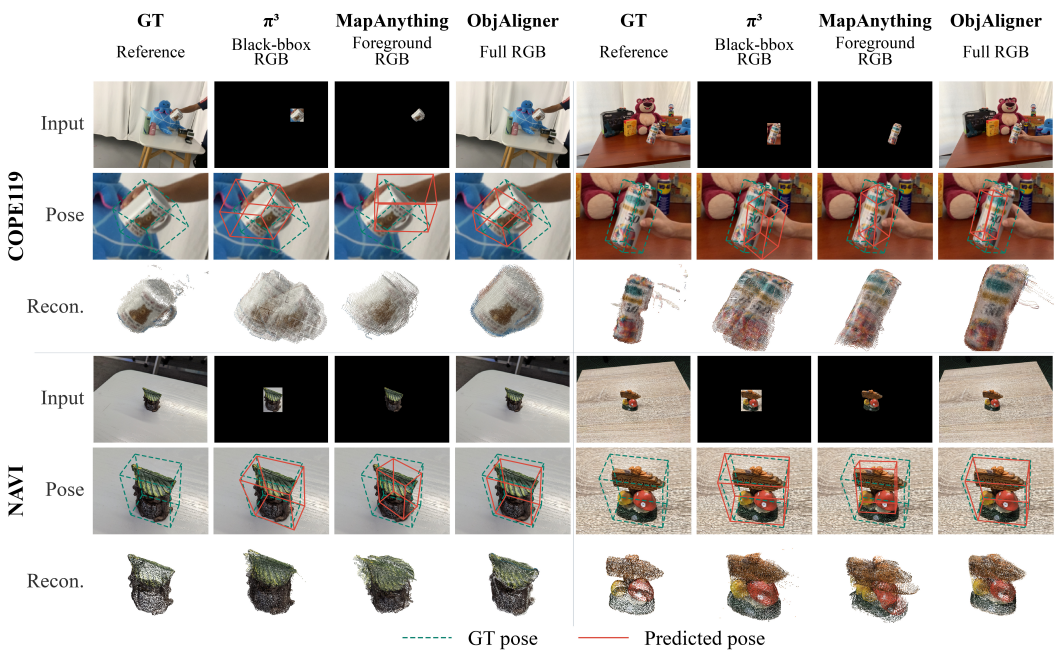

Figure 10: Further COPE119 and NAVI examples; layout follows Figure 9.

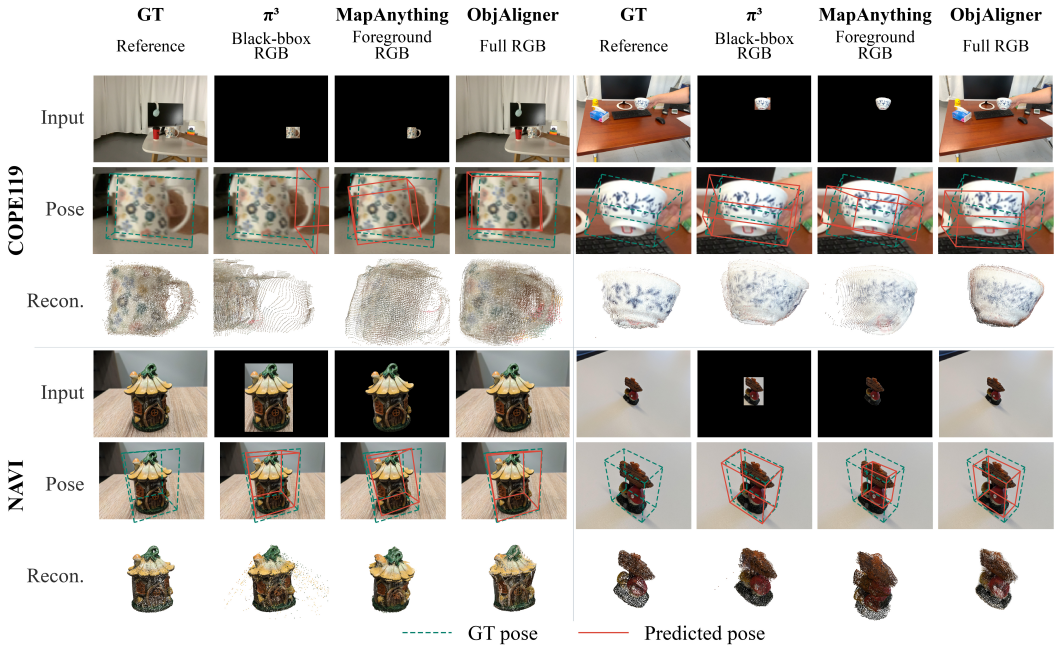

Figure 11: Additional pose overlays and fused surfaces; layout follows Figure 9.

G LIMITATIONS AND FUTURE WORK

ObjAligner jointly processes the supplied views and reconstructs the observed object surfaces. Its current formulation does not maintain a persistent object representation across successive input windows or explicitly complete unobserved geometry. A natural next step is streaming reconstruction: incrementally updating the shared object representation as new frames arrive, while preserving a consistent object coordinate frame with bounded memory. Incorporating shape priors to recover unobserved surfaces is another direction for future work.