EGOXPERIENCE: LEARNING WORLD-GROUNDED EGOCENTRIC REPRESENTATIONS FROM HUMAN EXPERIENCE

Anonymous authors
Paper under double-blind review



Figure 1: **Overview of the EgoXperience Paradigm.** Task-driven capture and EgoX-Engine produce EgoX-Data-6M, supporting joint reconstruction with EgoX.

ABSTRACT

Reconstructing human activity from egocentric video requires more than recovering individual 3D components: the observer's motion, surrounding geometry, and hand-object interactions must fit together in a consistent world. Achieving this coherence requires both joint reconstruction and aligned supervision, yet existing methods and datasets offer limited support for either. We present **EGOXPERIENCE**, a full-stack paradigm that addresses these needs through data collection, automatic annotation, unified modeling, and benchmarking. Central to EgoXperience is an automatic annotation engine, **EGOX-Engine**, that couples geometrically verified object tracking with SLAM-based world alignment, yielding consistent supervision for the camera, scene, hands, and objects. This yields **EGOX-Data-6M**, a 50-hour dataset of task-driven manipulation and locomotion within and across indoor and outdoor environments, with second-level task captions. We carefully select one hour across these activity categories as **EGOX-Bench** to evaluate joint reconstruction. To translate this supervision into reconstruction, we introduce **EGOX**, trained with EGOX-Data-6M and complementary annotations from additional public datasets. Given monocular egocentric video and first-frame object masks, our model achieves joint metric reconstruction of the camera, scene, hands, and objects in a shared world coordinate frame with a single forward pass. Extensive experiments demonstrate state-of-the-art performance across scene estimation and world-space hand-object pose recovery. Against the strongest evaluated baselines, EgoX reduces camera, hand, and depth errors by 68.9%, 65.6%, and 42.9%, and improves object pose scores by 28.0%.

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.

1 INTRODUCTION

Egocentric video offers a natural perspective on how humans and embodied agents perceive and interact with the world, requiring only a wearable camera without external camera setups. By capturing the wearer's movement and close-range interactions, it provides visual cues for recovering camera pose, scene depth, hands, and objects. However, existing methods (Wang et al., 2025a; 2026a; Hu et al., 2026; Lyu et al., 2026; Zhu et al., 2026) typically recover only a subset of these tightly coupled components. Combining outputs from different methods often requires reconciling inconsistent scales and coordinate frames, complicating their use in downstream embodied AI applications. This motivates a unified feed-forward framework that jointly recovers the camera, scene, hands, and objects within a single model.

Training such a unified model, however, requires large-scale egocentric data with joint annotations of the camera, scene, hands, and objects. Although egocentric video is easy to record, obtaining reliable joint annotations is considerably more difficult due to frequent hand-object occlusions, rapid head motion, and motion blur. Existing datasets (Cao et al., 2026c; Banerjee et al., 2025; Liu et al., 2022; Chao et al., 2021; Wang et al., 2025b) have made progress toward this goal, often using calibrated multi-camera or motion-capture systems to support reliable annotations. However, these systems require substantial setup and calibration, making collection difficult to scale across diverse real-world environments. Scaling jointly annotated egocentric data therefore calls for a collection and annotation pipeline that reduces reliance on such specialized infrastructure.

To address these challenges, we present EGOXPERIENCE, a full-stack paradigm for egocentric 3D reconstruction, encompassing data collection, annotation, modeling, and benchmarking. Central to its annotation pipeline is **EGOX-Engine**, which tracks target objects across video frames and uses pre-scanned object models and observed depth to estimate and verify their 3D poses. SLAM-estimated camera poses place these object trajectories in a shared world coordinate system alongside scene geometry and hand annotations, providing geometrically aligned supervision. Applying this engine to 50 hours of task-driven recordings spanning workspace manipulation, cross-room interactions, and outdoor loco-manipulation yields **EGOX-Data-6M**, further enriched with task captions at second-level temporal resolution. To complement this dataset, we harmonize additional public interaction datasets (Banerjee et al., 2025; Liu et al., 2022; Kwon et al., 2021; Wang et al., 2025b; Hampali et al., 2022; Chao et al., 2021) under a common training formulation, aligning their geometric conventions while retaining their partial annotations. We also establish **EGOX-Bench** from one hour of held-out recordings, enabling joint evaluation of camera, scene depth, hand, and object reconstruction across all capture settings.

To translate this joint supervision into joint reconstruction, we introduce **EGOX**. Given monocular egocentric video and first-frame object masks, EGOX jointly reconstructs the camera, scene depth, hands, and objects at metric scale in a shared world coordinate system, all in a single forward pass. During training, we supervise each component and additionally use the predicted camera poses to transform object and hand predictions into normalized world coordinates, where we supervise their positions and orientations. This ties camera, hand, and object estimation together, encouraging geometrically consistent reconstruction. Evaluated on four benchmarks, EGOX achieves state-of-the-art performance across scene estimation and world-space hand–object pose recovery. It ranks first in depth, hand, and object metrics across all four benchmarks and achieves the best camera accuracy on HOI4D. Relative to the strongest competing methods, EgoX reduces camera nATE by 68.9% on HOI4D, hand W-MPJPE and depth AbsRel by 65.6% and 42.9% on EgoX-Bench, respectively, and improves object AUC-S_{50} by 28.0% on HOT3D.

2 RELATED WORK

2.1 SCENE AND ENTITY RECONSTRUCTION

DUSt3R (Wang et al., 2024) predicts pairwise pointmaps, while VGGT (Wang et al., 2025a) jointly estimates cameras and scene geometry. CUT3R (Wang et al., 2025c) introduces recurrent scene memory, and VGGT-Ω (Wang et al., 2026a) reduces inter-frame attention through scene registers. Dynamic-scene methods recover moving geometry through world-space point tracking or temporally conditioned point prediction (Zhang et al., 2025b; Feng et al., 2025; Sucar et al., 2025; 2026; Zhang et al., 2026; Cao et al., 2026a). Entity-aware models incorporate humans, instances, or hands

into reconstruction: Human3R (Chen et al., 2025) jointly estimates human meshes, cameras, and scenes; HSImul3R (Cao et al., 2026b) enables simulator-ready joint reconstruction of humans and scenes; IGGT (Li et al., 2026b) couples geometry with instance segmentation; and HGGT (Liu et al., 2026) estimates MANO parameters and cameras from uncalibrated views. EgoX extends this line of work to moving hands and manipulated objects, jointly reconstructing their motion, camera poses, and depth in a shared metric frame.

2.2 Hand-object reconstruction

HaMeR (Pavlakos et al., 2024) and WiLoR (Potamias et al., 2025) estimate MANO hand meshes from images. For world-space hand motion, HaWoR (Zhang et al., 2025a) combines hand reconstruction with SLAM and motion infilling, while Dyn-HaMR (Yu et al., 2025) uses camera tracking, motion priors, and test-time optimization. Hand3R (Hu et al., 2026), MINT (Zhu et al., 2026), and ReViV (Lyu et al., 2026) couple hand reconstruction with cameras or scene geometry; ReViV additionally predicts depth, body motion, and gaze. EgoX further incorporates manipulated-object motion into the same feed-forward reconstruction.

Keypoint Transformer (Hampali et al., 2022) and HOPformer (Bansal et al., 2026) estimate hand–object interactions without recovering a world-camera trajectory. HOLD (Fan et al., 2024) optimizes hand and object representations from video, while ForeHOI (Chen et al., 2026) completes object masks and shapes before geometric pose estimation. EgoGrasp (Fu et al., 2026) reconstructs hands and objects in separate stages, and WHOLE (Ye et al., 2026) requires object templates, a metric camera trajectory, and test-time guidance. Related object-pose systems use CAD models, reference images, or RGB-D observations (Labbé et al., 2022; Wen et al., 2024; Lee et al., 2025). In contrast, EgoX jointly estimates camera poses, depth, hand geometry, and rigid-object poses from RGB observations and optional first-frame object masks.

2.3 Egocentric datasets and supervision

HO-3D (Hampali et al., 2020), DexYCB (Chao et al., 2021), H2O (Kwon et al., 2021), H2O-3D (Hampali et al., 2022), HOI4D (Liu et al., 2022), ARCTIC (Fan et al., 2023), HO-Cap (Wang et al., 2025b), and HOT3D (Banerjee et al., 2025) provide hand-object annotations under different capture and geometric conventions. Recent datasets broaden scene coverage: SHOW3D (Rim et al., 2026) records indoor and outdoor interactions with a mobile ego–exo rig, Aria Digital Twin (Pan et al., 2023) provides scanned scenes and world-space trajectories, and HD-EPIC (Perrett et al., 2025) captures unscripted activities with digital twins. Large corpora such as Ego4D (Grauman et al., 2022), EPIC-KITCHENS (Damen et al., 2022), and EgoDex (Hoque et al., 2026) provide broader human experience but more limited aligned geometric supervision. ACE-Data-0 (Cao et al., 2026c) builds an ambient capture engine for calibrated and synchronized multimodal data collection in a controlled environment. EgoX-Data-6M complements these resources with aligned camera, depth, hand, and object annotations across workspace manipulation, cross-area interaction, and outdoor locomotion.

3 The EgoXperience Paradigm

Our EgoXperience unifies data capture, annotation and model learning for joint metric reconstruction of the camera, scene depth, hands, and objects from egocentric videos. As illustrated in Fig. 1, this paradigm comprises EgoX-Engine for data processing and annotation (Sec. 3.1), EgoX-Data-6M with 50 hours of task-driven recordings spanning manipulation and locomotion (Sec. 3.2), and EgoX for joint reconstruction from monocular video (Sec. 3.3).

3.1 EgoX-Engine

Learning to jointly reconstruct egocentric activity requires annotations that describe the camera, scene, hands, and objects in relation to one another. Obtaining these annotations is challenging because the observer and the interacting entities move simultaneously: a change in image position may arise from camera motion, entity motion, or both. EgoX-Engine addresses this ambiguity by using stereo depth to recover geometry at metric scale and SLAM-estimated camera poses to align

Table 1: **Comparison of Egocentric Datasets.** We compare data scale, capture settings, and annotation coverage. Obj. counts objects with 3D models.

			Capture setting			Annotations					
Dataset	Hours	Obj.	Single-area	Cross-area	Outdoor	3D Hand	6DoF	Mesh	Depth	Mask	Caption
H2O (Kwon et al., 2021)	1.1	8	✓	✗	✗	✓	✓	✓	✓	✗	✗
HOI4D (Liu et al., 2022)	22	800	✓	✗	✗	✓	✓	✓	✓	✓	✗
ARCTIC (Fan et al., 2023)	2.2	11	✓	✗	✗	✓	✓	✓	✗	✗	✗
Ego-Exo4D (Grauman et al., 2024)	221	0	✓	✗	✓	✓	✗	✗	✗	✗	✓
EgoDex (Hoque et al., 2026)	829	0	✓	✗	✗	✓	✗	✗	✗	✗	✓
HOT3D (Banerjee et al., 2025)	13.9	33	✓	✗	✗	✓	✓	✓	✗	✓	✗
HO-Cap (Wang et al., 2025b)	0.7	64	✓	✗	✗	✓	✓	✓	✓	✓	✗
EgoVerse (Punamiya et al., 2026)	1362	0	✓	✗	✗	✓	✗	✗	✗	✗	✓
SHOW3D (Rim et al., 2026)	20	21	✓	✓	✓	✓	✓	✓	✗	✗	✓
ACE-Data-0 (Cao et al., 2026c)	150	50	✓	✓	✗	✓	✓	✓	✗	✓	✓
EgoX-Data-6M	**50**	**101**	✓	✓	✓	✓	✓	✓	✓	✓	✓

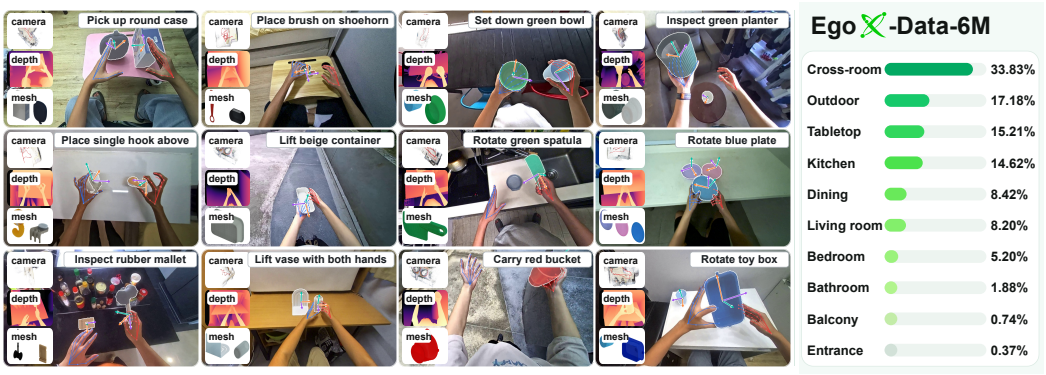

Figure 2: **Overview of EgoX-Data-6M.** Examples and statistics summarize capture settings, task coverage, and aligned multimodal annotations.

different annotations throughout each recording. This allows hand-object motion to be expressed relative to the surrounding scene as the observer moves. Fig. 3 summarizes the overall pipeline.

Capture and geometric reference. We record with the HOMIE Gen2 head-mounted rig (Ropedia R&D Team, 2026), using its associated processing pipeline to estimate metric depth from the synchronized, calibrated front stereo pair and recover camera trajectories via SLAM from its four calibrated fisheye cameras. For each recording, we use the SLAM coordinate system as a fixed session frame and transform camera-relative scene, hand, and object estimates into this frame.

Object geometry and motion. Object annotation requires both a geometric reference for each instance and its changing pose during interaction. We construct metric object meshes from 31 synchronized cameras mounted on a calibrated dome. SAM 3 (Carion et al., 2025) then extracts object masks from the captured views, which are used for mesh reconstruction with MV-SAM3D (Li et al., 2026a). To recover the object's physical dimensions, we put an AprilTag (Olson, 2011) board inside the dome to provide the scale reference.

Given these metric object references, the engine recovers each object's motion in the egocentric recording. Specifically, we first identify the target object using a first-frame mask and then specify the interaction interval over which to track it. Using the masked RGB image and its corresponding stereo depth, we estimate the initial pose of the pre-scanned object model in the camera frame. From this initial pose, we apply FoundationPose (Wen et al., 2024) to track the object throughout the specified interval, producing a camera-relative pose sequence. We verify this sequence by checking whether the object model placed at the estimated pose agrees with the observed depth. Sequences that pass verification are transformed through the SLAM-estimated camera poses into the session frame, aligning object motion with the scene and hand annotations.

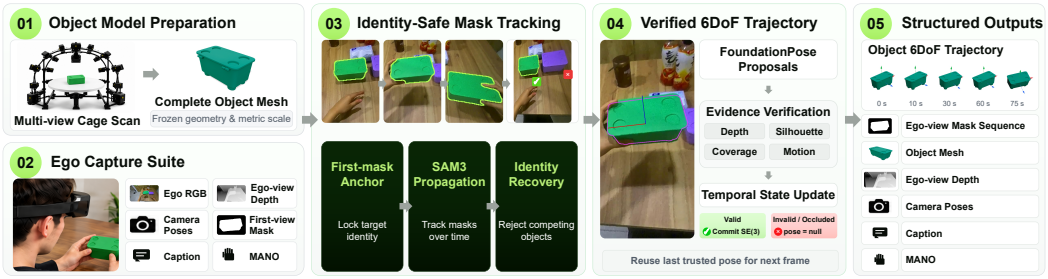

Figure 3: **Overview of EgoX-Engine.** The pipeline converts egocentric recordings into structured multimodal annotations, including aligned geometry and temporally localized task captions.

Scene and hand annotations. The scene and hand annotations used in this alignment are derived from the same calibrated stereo observations. Specifically, we reconstruct scene geometry by back-projecting stereo depth and recover 3D hand keypoints through cross-view triangulation. We use these keypoints to constrain MANO fitting, with additional reprojection, depth, and temporal terms to refine the estimated hand geometry and motion. Hand shape parameters are shared across frames within each recording, while hand pose, position, and orientation are estimated independently for each frame. Finally, we transform the fitted hands and scene points through the camera trajectory into the session frame, placing them in the same metric coordinate system as the tracked objects.

3.2 EgoX-Data-6M

Given EgoX-Engine, we apply it to annotate task-driven egocentric captures, constructing EgoX-Data-6M, which comprises 50 hours of video, 101 rigid objects, and 270 task scripts. The recordings cover activities ranging from manipulation within a single workspace to interactions across rooms and loco-manipulation in outdoor environments. The engine provides geometrically aligned camera, scene, hand, and object annotations across these settings, supporting joint reconstruction under both local manipulation and extended movement. Fig. 2 illustrates the activity coverage and accompanying annotations.

Task-driven collection. We organize data collection around task scripts, each specifying a manipulation goal, target objects, and a capture setting. Specifically, workspace scripts focus on local object manipulation, whereas cross-room and outdoor scripts incorporate movement between interaction locations.

Annotation coverage. We provide geometric annotations in a session-fixed metric frame, allowing camera motion, scene structure, and hand-object interactions to be related throughout each recording. Additionally, we include second-level task captions to describe the ongoing activity. Tab. 1 compares our dataset with existing egocentric resources. Compared with existing datasets, our EgoX-Data-6M brings together workspace manipulation, cross-room interactions, and outdoor locomotion with camera, scene, hand, and object annotations in a shared metric frame, supporting joint reconstruction across these activity settings.

Benchmark and complementary training data. To evaluate all reconstruction components on the same recordings, we reserve a quality-controlled, one-hour subset spanning the three activity settings as EgoX-Bench. This subset is used exclusively for joint evaluation of camera, scene, hand, and object reconstruction.

To broaden the training data beyond our own recordings, we incorporate additional public datasets. Using these datasets together requires accounting for differences in both geometric conventions and available annotations. We therefore convert their geometric labels to a common coordinate and scale convention and supervise the model using the labels available in each dataset. This allows recordings with partial annotations to contribute to training without requiring every dataset to provide camera, scene, hand, and object labels. We also clean erroneous annotations and align depth with hand and

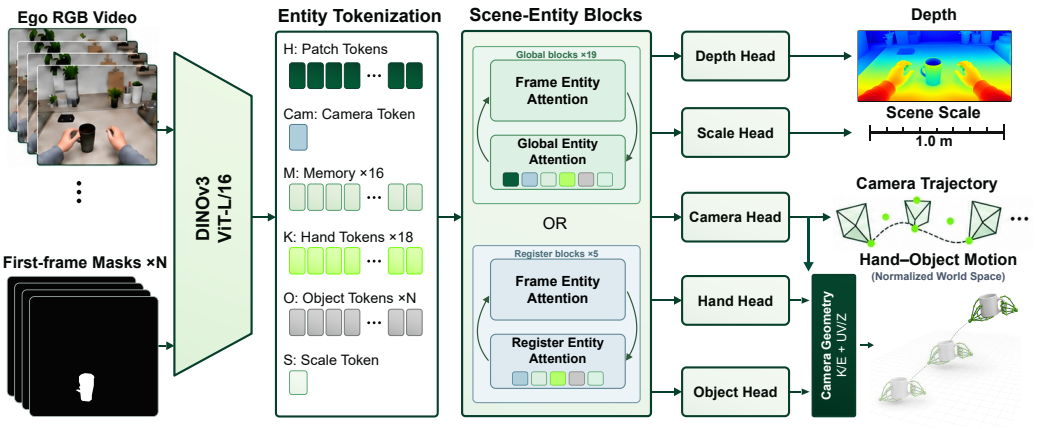

Figure 4: **Overview of EgoX.** Shared scene and entity features support joint reconstruction, while predicted cameras and scale place all outputs in a common metric world frame.

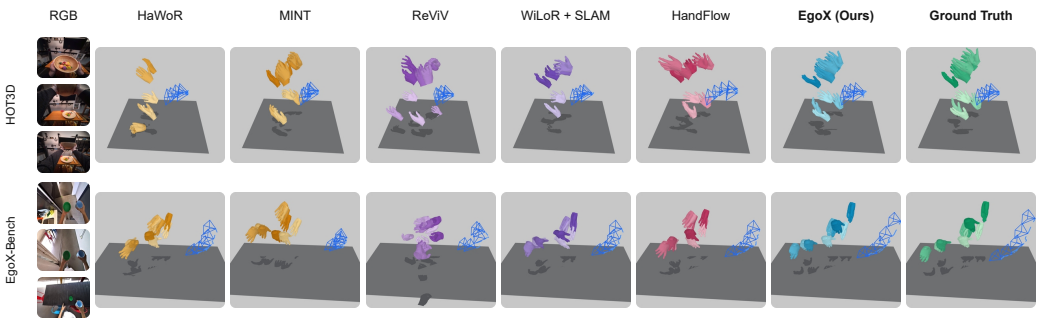

Figure 5: **World-Space Hand and Camera Reconstruction.** Qualitative comparisons on HOT3D and EgoX-Bench. Predictions and ground truth are shown in a common world frame.

object geometry to reduce conflicting supervision (Appendix B). Sec. 3.3 explains how the model learns from this combined supervision; the data mixture is detailed in Appendix G.

3.3 EGOX

To translate the aligned supervision into joint reconstruction, we propose EGOX, which is capable of predicting the camera, scene, and motions of both hands and objects within a single model. The central challenge to achieve this goal is that the predictions must agree geometrically as both the viewpoint and interaction change. As illustrated in Fig. 4, we address this through a joint scene and entity representation, followed by explicit composition through the predicted camera poses. Supervision on the resulting hand and object motions then ties their reconstruction to camera estimation.

Problem formulation. Given monocular observations $\mathcal{I} = \{I_t\}_{t=1}^{T}$ and optional first-frame object masks $\mathcal{B} = \{B_{1,j}\}_{j=1}^{N}$ identifying the target objects, EGOX jointly predicts camera intrinsics and poses, dense depth, two-hand MANO states, object origins and orientations, and a positive metric scale in a single forward pass. We express all reconstructed outputs in $\mathcal{W}$, the coordinate system of the first-frame camera. For those complementary datasets without ground-truth object masks, we adopt the pipeline to reconstruct only cameras, scene geometry, and hands.

Joint scene and entity representation. To jointly estimate cameras, scene geometry, and hand-object motion, we encode them with a shared Transformer. Starting from image patch features extracted by DINOv3 (Siméoni et al., 2025), we introduce camera, hand, and object tokens in each view, using distinct left/right embeddings to identify the hands and first-frame mask descriptors

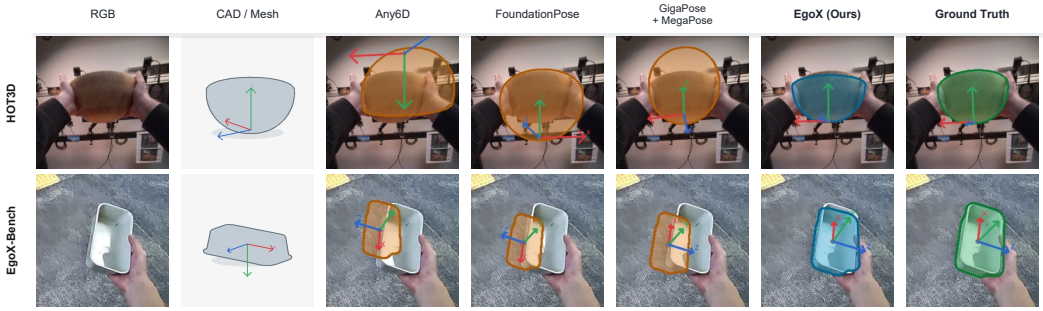

Figure 6: **Camera-Frame Object Pose Reconstruction.** Qualitative comparisons of object 6-DoF predictions with ground truth on HOT3D and EgoX-Bench.

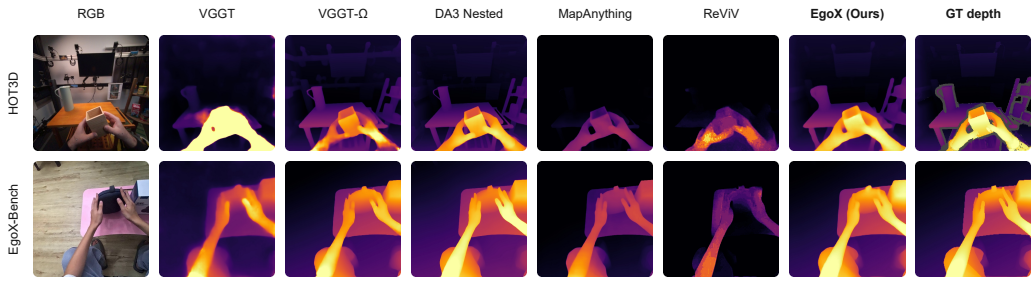

Figure 7: **Depth Reconstruction.** GT: refined official depth. VGGT/VGGT-Ω: one scale fitted to DA3 Nested depth per sequence; others: predicted metric scale.

to associate target objects across views. These tokens gather image evidence through within-view attention and exchange information across observations through cross-view attention, forming a joint representation for reconstruction. Architectural details are provided in Appendix G.

Reconstruction in a shared metric frame. From the joint representation, we reconstruct scene geometry and hand-object motion in a common reference frame $\mathcal{W}$, defined by the first-frame camera coordinate system. To combine predictions across views, we estimate camera intrinsics $\hat{K}_t$ and extrinsics $(\hat{R}_t, \hat{\tau}_t)$ mapping $\mathcal{W}$ to camera t. These parameters allow us to recover scene points, wrist positions, and object origins from their image coordinates and predicted depths. Specifically, for a homogeneous image position $\tilde{u} = (u_x, u_y, 1)^\top$ and normalized depth z, its position in $\mathcal{W}$ is:

$$\mathcal{U}_t(\tilde{u}, z) = \hat{R}_t^\top \left(z\hat{K}_t^{-1}\tilde{u} - \hat{\tau}_t \right). \tag{1}$$

This shared mapping ties scene and entity positions to the same camera estimates. For hand and object orientations, we similarly use inverse camera rotation $\hat{R}_t^\top$ to transform camera-frame hand and object orientations into $\mathcal{W}$, keeping MANO joint rotations local and expressing object rotations relative to their first-frame orientations.

The reconstruction in $\mathcal{W}$ obtained above still uses normalized depths and translations. To recover metric scale, we apply a shared ScaleHead to predict one positive factor $\hat{s}$ for the sequence, scaling scene points and camera, wrist, and object positions consistently: $\hat{p}^{\mathcal{W}} = \hat{s}\mathcal{U}_t(\tilde{u}, z)$. For hands, this adjusts only the wrist positions while preserving the metric dimensions of the MANO meshes.

Geometrically coupled supervision. Building on this shared reconstruction, we supervise not only individual predictions but also the resulting hand-object geometry in $\mathcal{W}$. To compute this geometric supervision, we transform hand and object predictions through the estimated cameras and compare the reconstructed geometry with ground truth. The resulting loss backpropagates through the camera transformations, updating both entity and camera predictions. We combine these supervision terms into the training objective:

$$\mathcal{L} = \mathcal{L}_{\text{local}} + \mathcal{L}_{\text{dense}} + \mathcal{L}_{\text{coupled}} + \mathcal{L}_{\text{scale}}, \tag{2}$$

Table 2: **Hand Reconstruction on Four Benchmarks.** PA-, WA-, and W-MPJPE are reported in millimetres. World-space errors use each method's estimated camera trajectory.

	HOT3D			EgoX-Bench			HOI4D			H2O		
Method	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓	PA-MPJPE ↓	WA-MPJPE ↓	W-MPJPE ↓
HaWoR	6.06	29.80	56.30	10.98	125.73	363.47	18.61	50.01	110.96	5.68	25.83	64.62
MINT	9.94	37.76	85.85	10.19	149.91	467.65	17.94	43.64	108.76	5.74	28.38	50.04
ReViV	14.01	49.90	109.02	23.89	111.48	501.29	25.18	83.40	354.39	8.69	23.40	45.22
WiLoR + SLAM	5.90	25.52	58.55	8.51	63.03	328.81	19.41	47.66	132.94	5.11	21.78	74.95
HandFlow	5.54	59.01	119.48	11.14	75.76	375.11	19.91	57.02	120.26	7.07	19.00	34.63
EgoX (Ours)	**3.70**	**15.54**	**38.89**	**4.64**	**53.81**	**113.24**	**3.32**	**15.12**	**42.67**	**4.97**	**11.41**	**23.95**

Table 3: **Camera-Frame Object Pose Reconstruction.** ADD/ADD-S: mm; AUC: %. Extra inputs: $\hat{D}/\hat{K}$, DA3 Nested depth/intrinsics; M, per-frame masks; A, GT metric CAD.

	HOT3D			EgoX-Bench			HOI4D			H2O		
Method	ADD ↓	ADD-S ↓	AUC-S_{50} ↑	ADD ↓	ADD-S ↓	AUC-S_{50} ↑	ADD ↓	ADD-S ↓	AUC-S_{50} ↑	ADD ↓	ADD-S ↓	AUC-S_{50} ↑
Any6D$^{\hat{D},\hat{K},M}$	67.06	44.05	42.99	88.25	41.41	38.39	67.01	53.26	34.61	72.75	36.57	38.37
FoundationPose$^{\hat{D},\hat{K},A}$	92.49	59.87	23.32	85.26	55.04	31.99	94.76	73.02	13.39	91.91	30.97	46.03
GigaPose + MegaPose$^{\hat{K},A,M}$	71.92	32.42	57.59	69.96	26.87	59.80	97.94	88.68	3.91	89.08	53.08	29.14
EgoX (Ours)	**28.00**	**14.03**	**73.71**	**27.11**	**12.41**	**76.59**	**44.55**	**21.11**	**59.08**	**22.10**	**10.29**	**80.42**

Table 4: **Camera and Depth Reconstruction.** RRE: degrees; $\delta_{1.25}$: %. VGGT/VGGT-Ω: one scale fitted to DA3 Nested depth per sequence; others: predicted metric scale.

	HOT3D				EgoX-Bench				HOI4D				H2O			
	Camera		Depth†		Camera		Depth†		Camera		Depth†		Camera		Depth†	
Method	nATE ↓	RRE (°) ↓	AbsRel ↓	$\delta_{1.25}$ ↑	nATE ↓	RRE (°) ↓	AbsRel ↓	$\delta_{1.25}$ ↑	nATE ↓	RRE (°) ↓	AbsRel ↓	$\delta_{1.25}$ ↑	nATE ↓	RRE (°) ↓	AbsRel ↓	$\delta_{1.25}$ ↑
VGGT	0.0064	0.64	0.2182	69.40	0.0684	2.86	0.2877	66.53	0.0310	1.50	0.0676	98.60	0.0036	0.27	0.0726	94.36
VGGT-Ω	**0.0018**	**0.16**	0.1031	92.20	**0.0071**	**0.61**	0.0806	91.57	0.0212	1.56	0.0558	99.68	0.0030	0.27	0.0784	95.46
CUT3R	0.0449	3.75	0.2748	69.19	0.0756	6.29	0.4317	54.07	0.0455	5.08	0.0647	97.05	0.0120	1.56	0.0887	93.09
Any4D	0.0267	2.24	0.2899	70.23	0.0604	4.55	0.4161	48.15	0.0273	1.91	0.0757	95.80	0.0083	0.67	0.1075	87.26
DA3 Nested	0.0036	0.23	0.0799	94.40	0.0135	0.92	0.1785	82.29	0.0307	1.54	0.0746	97.34	**0.0026**	**0.26**	0.0823	92.64
MapAnything	0.0321	19.40	0.3182	64.20	0.0283	2.11	0.1429	87.12	0.0220	1.63	0.0727	95.36	0.0071	0.51	0.1009	92.01
ReViV	0.0285	1.98	0.2558	61.19	0.0350	3.81	0.3972	53.60	0.0231	2.06	0.2075	60.42	0.0072	0.84	0.2562	53.07
EgoX (Ours)	0.0059	0.88	**0.0604**	**96.21**	0.0295	2.21	**0.0460**	**97.37**	**0.0066**	**0.78**	**0.0312**	**99.98**	0.0035	0.33	**0.0700**	**99.70**

where $\mathcal{L}_{\text{local}}$ supervises camera and view-local entity predictions, $\mathcal{L}_{\text{dense}}$ supervises depth and scene points, $\mathcal{L}_{\text{coupled}}$ supervises the composed hand-object geometry, and $\mathcal{L}_{\text{scale}}$ supervises metric scale. To accommodate varying annotation coverage across datasets, we apply each loss only where its required labels are available, allowing fully and partially annotated data to contribute to the same objective. Loss definitions and training settings are provided in Appendix G.

4 Experiments

We evaluate camera, depth, hand, and object reconstruction on HOT3D, EgoX-Bench, H2O, and HOI4D using identical 150-observation clips across methods. Evaluation sequences are excluded from EgoX training. EgoX uses RGB and first-frame object masks; object baselines may additionally require estimated depth, intrinsics, per-frame masks, or metric CAD assets. World-space hand metrics use estimated cameras; object poses are evaluated in camera coordinates. No method receives GT camera poses or scale in the main comparisons. Appendix C details the evaluation protocols.

4.1 Main results

World-space hands. EgoX achieves the lowest PA-, WA-, and W-MPJPE on all four benchmarks (Table 2). On HOT3D, it reduces W-MPJPE from 56.30 to 38.89 mm relative to HaWoR, the strongest baseline on this metric, despite HaWoR's adaptive DROID-SLAM and Metric3D-based scale (Zhang et al., 2025a). On EgoX-Bench, W-MPJPE falls from 328.81 mm for WiLoR + SLAM to 113.24 mm. Figure 5 illustrates improved world-space placement: baseline hands recover plau-

Table 5: **ScaleHead Ablation.** Results on HOT3D with predicted (P) or oracle (G) scale.

Setting	s	W ↓	AUC-S$_{50}$ ↑	AbsRel ↓
w/ head	P	**38.89**	**73.71**	**0.0604**
w/o head	G	40.67	60.87	0.0620

Table 6: **DepthHead Ablation.** Results on HOT3D with predicted scale in both settings.

Setting	s	W ↓	AUC-S$_{50}$ ↑	AbsRel ↓
w/ head	P	**38.89**	**73.71**	0.0604
w/o head	P	67.53	41.09	*N/A*

sible local hand poses but exhibit large deviations from the ground-truth positions in world coordinates, whereas EgoX closely matches the ground truth.

Object 6-DoF. EgoX achieves the best ADD, ADD-S, and AUC-S$_{50}$ on all four benchmarks (Table 3). On EgoX-Bench, its AUC-S$_{50}$ is 76.59, compared with 59.80 for GigaPose + MegaPose, the strongest competing result, although that pipeline additionally uses metric CAD assets and per-frame masks. EgoX also outperforms Any6D on HOI4D, where Any6D receives estimated depth, intrinsics, and per-frame masks. These comparisons show more accurate camera-frame poses with fewer auxiliary inputs (Appendix C). Figure 6 shows larger translation or rotation offsets for the object baselines, while EgoX more closely matches the ground-truth pose.

Camera motion. Compared with ReViV, which also reconstructs cameras, depth, and hands, EgoX achieves lower nATE and RRE on all four benchmarks (Table 4). On HOI4D, EgoX also leads the geometry baselines: nATE is 0.0066 versus 0.0212 for VGGT-Ω, and RRE is 0.78° versus 1.50° for VGGT. VGGT-Ω leads on HOT3D and EgoX-Bench, while DA3 Nested leads on H2O. VGGT-Ω's camera advantage may reflect its large-scale geometric training on diverse static and dynamic scenes (Wang et al., 2026a).

Depth reconstruction. EgoX achieves the lowest AbsRel and highest $\delta_{1.25}$ on all four benchmarks (Table 4). On EgoX-Bench, AbsRel is 0.0460 versus 0.3972 for ReViV and 0.0806 for VGGT-Ω; $\delta_{1.25}$ is 97.37 versus 91.57 for VGGT-Ω. EgoX obtains these results with its own predicted metric scale, while VGGT-Ω uses scale estimated from DA3 Nested depth (Section C). The results support accurate metric depth alongside hand and object reconstruction. Figure 7 shows that EgoX better matches the refined ground-truth depth.

4.2 Ablation Studies

Metric scale prediction. We remove ScaleHead and use oracle scale (G) to convert the ablated model's predictions to metric units. Despite this advantage, W-MPJPE increases from 38.89 to 40.67 mm and AUC-S$_{50}$ decreases from 73.71 to 60.87 (Table 5), showing that ScaleHead benefits joint reconstruction beyond final rescaling.

Dense-depth prediction. We remove DepthHead while retaining the remaining architecture and predicted scale. W-MPJPE increases to 67.53 mm and AUC-S$_{50}$ decreases to 41.09 (Table 6), showing that explicit depth prediction improves hand and object reconstruction through the shared representation.

5 Conclusion

We present EgoXperience, a unified approach to learning world-grounded egocentric representations through task-driven capture, geometrically aligned annotation, and joint reconstruction. EgoX-Engine provides aligned multimodal annotations for the 50-hour EgoX-Data-6M, from which we curate the held-out one-hour EgoX-Bench. Trained on these annotations and harmonized public datasets, EgoX jointly reconstructs cameras, dense depth, hands, and rigid-object 6-DoF poses in a common metric world frame. Experiments on EgoX-Bench and public benchmarks demonstrate strong performance in hand and object reconstruction and depth estimation. Our current formulation assumes rigid objects and does not predict their meshes. Future work will explore object mesh prediction and multiview inputs.

AI USE STATEMENT

We used generative AI tools to improve the language and readability of the manuscript. The authors reviewed and edited the AI-assisted text and take full responsibility for the final manuscript.

ETHICS STATEMENT

All participants consented to the collection and research use of their recordings and authorized their public release. The collected data contain no sensitive personal information. We plan to publicly release the dataset and benchmark to support research on egocentric reconstruction.

REPRODUCIBILITY STATEMENT

We describe public-dataset cleaning and geometric alignment in Appendix B, model and training details in Appendix G, and evaluation protocols in Appendix C. Runtime measurement settings are provided in Appendix D. We will publicly release our dataset, benchmark, model code, and pretrained weights to facilitate reproducibility.

REFERENCES

AgiBot World Team. Agibot world 2026. https://huggingface.co/datasets/agibot-world/AgiBotWorld2026, 2026. 21

Armen Avetisyan, Christopher Xie, Henry Howard-Jenkins, Tsun-Yi Yang, Samir Aroudj, Suvam Patra, Fuyang Zhang, Duncan Frost, Luke Holland, Campbell Orme, Jakob Engel, Edward Miller, Richard Newcombe, and Vasileios Balntas. Scenescript: Reconstructing scenes with an autoregressive structured language model, 2024. URL https://arxiv.org/abs/2403.13064. 21

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071, June 2025. 2, 3, 4, 18, 21

Siddhant Bansal, Zhifan Zhu, Shashank Tripathi, Jiahe Zhao, Michael J. Black, and Dima Damen. Towards in-the-wild egocentric 3d hand-object pose estimation. In *European Conference on Computer Vision (ECCV)*, 2026. 3, 17, 18

Yukang Cao, Xinying Guo, Mingyuan Zhang, Haozhe Xie, Chenyang Gu, and Ziwei Liu. Crowdmogen: Event-driven collective human motion generation. *International Journal of Computer Vision*, 134(1):29, 2026a. 2

Yukang Cao, Haozhe Xie, Fangzhou Hong, Long Zhuo, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Hsimul3r: Physics-in-the-loop reconstruction of simulation-ready human–scene interactions. In *European Conference on Computer Vision*, pp. 509–528. Springer, 2026b. 3

Yukang Cao, Haozhe Xie, Beichen Wen, Runmao Yao, Yinghao Liu, Yue Huang, Zhichao Liao, Yunxiang Wang, Haiheng Liu, Xingshun Tian, Dawei Su, Long Zhuo, Dacheng Tao, Xiaogang Wang, Liang Pan, and Ziwei Liu. Ace-data-0: Human-centric ambient capture as embodied data engine. *arXiv preprint arXiv:2607.28625*, 2026c. 2, 3, 4, 21

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. *arXiv preprint arXiv:2507.12465*, 2025a. 17

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-anything: Simulation-ready physical 3d assets from single image. *arXiv preprint arXiv:2511.13648*, 2025b. 17

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-omni: Unified simulation-ready physical 3d generation for rigid, deformable, and articulated objects, 2026d. URL https://arxiv.org/abs/2605.21572. 17

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. URL https://arxiv.org/abs/2511.16719. 4

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 18, 21

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, 2026. URL https://arxiv.org/abs/2602.06226. 3, 17

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Pons-Moll Gerard. Human3r: Everyone everywhere all at once. *arXiv preprint arXiv:2510.06219*, 2025. 3, 17

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *International Journal of Computer Vision*, 2022. 3, 18

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3, 4, 18

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 494–504, 2024. 3, 17

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*, 2025. 2, 17

Hongming Fu, Wenjia Wang, Xiaozhen Qiao, Rolandos Alexandros Potamias, Taku Komura, Shuo Yang, Zheng Liu, and Bo Zhao. Egograsp: World-space hand-object interaction estimation from egocentric videos, 2026. URL https://arxiv.org/abs/2601.01050. 3, 17

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19012, 2022. 3, 18

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard

Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19383–19400, 2024. 4

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 18

Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *IEEE Computer Vision and Pattern Recognition Conference*, 2022. 2, 3, 17, 18, 21

Ryan Hoque, Peide Huang, David J. Yoon, Mouli sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openresix.net/pdf/d4fea01944202c75aa13dc34757e154478eb8df7.pdf. 3, 4, 18

Wendi Hu, Haonan Zhou, Wenhao Hu, and Gaoang Wang. Hand3r: Online 4d hand-scene reconstruction in the wild, 2026. URL https://arxiv.org/abs/2602.03200. 2, 3, 17

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. Vipe: Video pose engine for 3d geometric perception, 2025. URL https://arxiv.org/abs/2508.10934. 17

Zhi Jing, Siyuan Yang, Jicong Ao, Ting Xiao, Yu-Gang Jiang, and Chenjia Bai. Humanoidgen: Data generation for bimanual dexterous manipulation via LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=Mk9ykil8eP. 17

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. *CVPR*, 2023. 21

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)*. IEEE, 2026. 17

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10138–10148, October 2021. 2, 3, 4, 18, 21

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022. 3, 17

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 Computer Vision and Pattern Recognition Conference (CVPR)*, 2025. 3, 17

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026a. 4

Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, Hang Yin, Michael Lingelbach, Minjune Hwang, Ayano Hiranaka, Sujay Garlanka, Arman Aydin, Sharon Lee, Jiankai Sun, Mona Anvari, Manasi Sharma, Dhruva Bansal, Samuel Hunter, Kyu-Young Kim,

Alan Lou, Caleb R Matthews, Ivan Villa-Renteria, Jerry Huayang Tang, Claire Tang, Fei Xia, Yunzhu Li, Silvio Savarese, Hyowon Gweon, C. Karen Liu, Jiajun Wu, and Li Fei-Fei. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 21

Hao Li, Zhengyu Zou, Fangfu Liu, Xuanyang Zhang, Fangzhou Hong, Yukang Cao, Yushi LAN, Manyuan Zhang, Gang YU, Dingwen Zhang, and Ziwei Liu. IGGT: Instance-grounded geometry transformer for semantic 3d reconstruction. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=swiL18PmUV. 3, 17

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast and Robust Structure and Motion from Casual Dynamic Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 17

Yumeng Liu, Xiao-Xiao Long, Marc Habermann, Xuanze Yang, Cheng Lin, Yuan Liu, Yuexin Ma, and Ligang Liu. Hggt: Robust and flexible 3d hand mesh reconstruction from uncalibrated images. *arXiv preprint arXiv:2603.23997*, 2026. 3, 17

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*, 2024. 18

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21013–21022, June 2022. 2, 3, 4, 18, 21

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. Reviv: Reconstructing the viewer and the view in 4d from monocular egocentric video. In *European Conference on Computer Vision (ECCV)*, 2026. 2, 3, 17

John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J. Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 21

Edwin Olson. Apriltag: A robust and flexible visual fiducial system. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3400–3407, 2011. 4

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception, 2023. URL https://arxiv.org/abs/2306.06362. 3, 18

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 17

Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. Hd-epic: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. 3, 18

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 17

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 17

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Wenkai Xuan, Chenyu Yang, Xi Wang, James Fort, Richard Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. Egoverse: An egocentric human dataset for robot learning from around the world, 2026. URL https://arxiv.org/abs/2604.07607. 4

Patrick Rim, Kevin Harris, Braden Copple, Shangchen Han, Xu Xie, Ivan Shugurov, Sizhe An, He Wen, Alex Wong, Tomas Hodan, and Kun He. SHOW3D: Capturing scenes of 3d hands and objects in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3, 4, 18

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017. 17

Ropedia. Xperience-10M: A large-scale egocentric multimodal dataset with structured 3D/4D annotations. Hugging Face, 2026. URL https://huggingface.co/datasets/ropedia-ai/xperience-10m. Dataset. 21

Ropedia R&D Team. HOMIE Gen2: Scale Real-World Data for Physical AI. https://ropedia.com/homie, 2026. Accessed 2026-09-19. 4

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104. 6

Edgar Sucar, Zihang Lai, Eldar Insafutdinov, and Andrea Vedaldi. Dynamic point maps: A versatile representation for dynamic 3d reconstruction, 2025. URL https://arxiv.org/abs/2503.16318. 2, 17

Edgar Sucar, Eldar Insafutdinov, Zihang Lai, and Andrea Vedaldi. V-DPM: 4d video reconstruction with dynamic point maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2, 17

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*, 2025a. 2, 17

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. 2, 9, 17

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025b. URL https://openreview.net/forum?id=hpu6r8oLw9. 2, 3, 4, 18, 21

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)*, 2025c. 2, 17

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 17

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. Arthoi: Taming foundation models for monocular 4d reconstruction of hand-articulated-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026b. 17

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*, 2024. 3, 4, 17

Yufei Ye, Jiaman Li, Ryan Rong, and C. Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *CVPR Findings*, 2026. 3, 17

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 17

Zhao-Heng Yin, Changhao Wang, Luis Pineda, Francois Hogan, Krishna Bodduluri, Akash Sharma, Patrick Lancaster, Ishita Prasad, Mrinal Kalakrishnan, Jitendra Malik, Mike Lambeta, Tingfan Wu, Pieter Abbeel, and Mustafa Mukadam. Dexteritygen: Foundation controller for unprecedented dexterity, 2025. URL https://arxiv.org/abs/2502.04307. 17

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025. 3, 17

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 *CVPR*, 2026. 2, 17

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. 3, 8, 17, 18

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 *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=lJpqxFgWCM. 2, 17

Zijie Zhu, Weiren Cai, Yizhou Wang, Zhenjie Yang, Yide Liu, Jiahao Chen, and Guanqi He. Mint: A unified model for world-space camera and hand motion estimation from scalable egocentric pipeline supervision, 2026. URL https://arxiv.org/abs/2609.04958. 2, 3, 17, 18

A Extended Related Work

A.1 Joint geometry and entity reconstruction

DUSt3R (Wang et al., 2024) predicts pairwise pointmaps, while VGGT (Wang et al., 2025a) jointly estimates cameras and geometry. CUT3R (Wang et al., 2025c) uses recurrent scene memory, and VGGT-Ω (Wang et al., 2026a) reduces inter-frame attention costs with scene registers. EgoX combines register-mediated and global attention with hand, object, and scale tokens. For dynamic scenes, MonST3R (Zhang et al., 2025b) reconstructs moving content, and St4RTrack (Feng et al., 2025) tracks 3D points in a common world frame. Dynamic Point Maps (Sucar et al., 2025) recovers rigid-object motion through masked point alignment, V-DPM (Sucar et al., 2026) extends pointmaps to video, and D4RT (Zhang et al., 2026) predicts 3D points conditioned on source, target, and camera-reference times. EgoX directly predicts articulated hands and rigid object poses alongside cameras and depth. Human3R (Chen et al., 2025) combines human meshes with scene and camera reconstruction, while IGGT (Li et al., 2026b) jointly learns geometry and instance segmentation. HGGT (Liu et al., 2026) uses hand tokens to estimate MANO parameters and cameras from uncalibrated views of a fixed hand configuration. Unlike these methods, EgoX jointly reconstructs the camera, depth, and moving hands and objects over time.

A.2 Hand and hand–object reconstruction

HaMeR (Pavlakos et al., 2024) regresses MANO (Romero et al., 2017) parameters from hand crops, while WiLoR (Potamias et al., 2025) pairs hand detection with mesh estimation. For world-space motion, HaWoR (Zhang et al., 2025a) combines hand reconstruction with SLAM and motion infilling; Dyn-HaMR (Yu et al., 2025) uses camera tracking, motion priors, and optimization. Hand3R (Hu et al., 2026) couples a hand estimator with recurrent scene modeling, while MINT (Zhu et al., 2026) and ReViV (Lyu et al., 2026) jointly estimate hand and camera motion, with ReViV also modeling depth, body motion, and gaze. EgoX additionally predicts manipulated-object motion.

Simulation-ready objects (Cao et al., 2025a;b; 2026d) and synthetic motion (Jing et al., 2025; Yin et al., 2025) enable interaction data at scale, but may not capture natural human manipulation. Keypoint Transformer (Hampali et al., 2022) and HOPformer (Bansal et al., 2026) estimate hand–object interactions without a world-camera trajectory. HOLD (Fan et al., 2024) optimizes hand and object representations from video; ForeHOI (Chen et al., 2026) completes masks and shapes before geometrically estimating object poses, without requiring separate input masks. ArtHOI (Wang et al., 2026b) handles articulated objects, whereas EgoX models rigid objects.

For world-space interactions, EgoGrasp (Fu et al., 2026) reconstructs hands and objects in separate stages. WHOLE (Ye et al., 2026) uses a motion prior and test-time guidance but requires object templates and a metric camera trajectory. EgoX jointly predicts cameras, depth, hands, and objects from RGB video and initial object masks, without either external input. Its reconstruction losses evaluate hand, object, and depth predictions through the predicted camera, providing training signals for camera estimation.

A.3 Camera, depth, and object-pose pipelines

Modular reconstruction also provides a source of training annotations. MegaSaM (Li et al., 2025) and ViPE (Huang et al., 2025) combine learned geometry with camera and depth optimization. Metric3D (Yin et al., 2023) and UniDepth (Piccinelli et al., 2024) predict metric depth from single images, while MapAnything (Keetha et al., 2026) predicts metric geometry from multiple views. HaWoR uses metric depth to scale its SLAM trajectory. EgoX instead learns a scale per window shared by depth and camera, hand, and object translations, though shared scale alone does not ensure geometric consistency.

MegaPose (Labbé et al., 2022) requires a CAD model, FoundationPose (Wen et al., 2024) accepts a model or reference images, and Any6D (Lee et al., 2025) uses an RGB-D reference. Their camera-relative poses can be mapped to world coordinates using a camera trajectory. EgoX likewise transforms predictions into a common frame, but learns camera and object estimates jointly. These methods also differ in their requirements for depth, pose initialization, scale, and masks.

A.4 Datasets and geometric evaluation

HO-3D (Hampali et al., 2020), DexYCB (Chao et al., 2021), H2O (Kwon et al., 2021), H2O-3D (Hampali et al., 2022), HOI4D (Liu et al., 2022), ARCTIC (Fan et al., 2023), TACO (Liu et al., 2024), HO-Cap (Wang et al., 2025b), and HOT3D (Banerjee et al., 2025) support hand–object reconstruction under varied capture protocols. ARCTIC includes articulated objects, and HOT3D provides motion-capture reference poses. However, capture volumes or clip protocols often limit evaluation of continuous camera travel between workspaces.

More recent datasets broaden the settings. SHOW3D (Rim et al., 2026) captures indoor and outdoor interactions with a mobile ego–exo rig, but defines annotations relative to the moving rig. EPIC-Contact (Bansal et al., 2026) provides fitted meshes and contact correspondences in kitchen video. Aria Digital Twin (Pan et al., 2023) provides camera and object trajectories, scanned scenes, and rendered depth; HD-EPIC (Perrett et al., 2025) provides unscripted kitchen recordings with digital twins and 3D object motion, though its labels serve a different purpose from dense hand and object pose trajectories.

Large collections such as Ego4D (Grauman et al., 2022) and EPIC-KITCHENS (Damen et al., 2022) support learned geometric supervision. EgoDex (Hoque et al., 2026) scales manipulation data, while MINT (Zhu et al., 2026) builds a camera-and-hand corpus through automatic annotation. Our data spans workspace manipulation, travel between indoor workspaces, and outdoor locomotion, with hand, object, and scene annotations. This enables evaluation across camera displacement, capture setting, and sequence duration using camera, depth, and world-space interaction errors.

B Public Dataset Cleaning and Geometric Alignment

We systematically clean HOT3D, HOI4D, H2O, H2O-3D, HO-Cap, and DexYCB before joint training. We discard frames with erroneous hand annotations and remove spurious points from depth-derived point clouds. We further align depth-derived geometry, MANO hand meshes, and object meshes in a common 3D coordinate system, correcting inconsistencies in their relative placement and scale. These steps reduce erroneous labels and conflicting geometric supervision across reconstruction tasks.

C Evaluation Protocol

Datasets and splits. We evaluate on four sequence-disjoint benchmarks: HaWoR's validation split of HOT3D (Zhang et al., 2025a), the held-out one-hour EgoX-Bench, the official H2O test split, and our sequence-level HOI4D validation split. Evaluation sequences are excluded from EgoX training. All methods use identical sequence lists and matched 150-observation clips.

Inputs and matched protocol. EgoX receives RGB and optional first-frame object masks. Object-pose baselines additionally use DA3 Nested depth and intrinsics, per-frame masks, or metric CAD assets as required; their runtimes include any required DA3 preprocessing. Any6D receives DA3 Nested depth and intrinsics together with per-frame object masks, but no CAD asset. FoundationPose receives DA3 Nested depth and intrinsics together with the metric CAD asset. GigaPose + MegaPose receives DA3 Nested intrinsics, the metric CAD asset, and per-frame object masks. Ground-truth camera poses and scale are not supplied as reconstruction inputs in the main comparisons.

For HandFlow, we follow the released right-hand-only protocol by mirroring left-hand inputs and mapping recovered joints back to the original camera frame. HandFlow and WiLoR are paired with the same DROID-SLAM camera trajectory; neither receives ground-truth camera poses, depth, or hand annotations. Results are averaged over matched clips after applying common annotation-validity masks.

Metrics and coordinates. Hands are evaluated using PA-, W-, and WA-MPJPE. Cameras use normalized absolute trajectory error (nATE) and relative rotation error (RRE); depth uses absolute relative error (AbsRel) and threshold accuracy $\delta_{1.25}$; object poses use ADD, ADD-S, and AUC-S_{50}.

World-space hand metrics use each pipeline's estimated camera trajectory, whereas object pose is evaluated in camera coordinates.

Scale. EgoX and the native metric baselines retain their estimated metric scale. VGGT and VGGT-Ω use one scale coefficient fitted from DA3 Nested depth per input sequence. Alignment, aggregation, validity masks, and failure handling are held fixed across methods.

D Runtime Analysis

We measure method-level inference time for 150 frames on one NVIDIA A800 (Table 7). Models and reusable assets are preloaded; data loading, visualization, and disk I/O are excluded. Runtime for object-pose baselines includes DA3 Nested whenever estimated depth or intrinsics are required.

Table 7: **Inference Efficiency.** Method-level runtime for 150 frames on one NVIDIA A800. Object baselines include DA3 Nested preprocessing when required.

Hand reconstruction			*Object 6D pose*		
Method	Time ↓	FPS ↑	Method	Time ↓	FPS ↑
HaWoR	25.37 s	5.91	Any6D	166.54 s	0.90
MINT	**8.30 s**	**18.08**	FoundationPose	31.72 s	4.73
ReViV	14.91 s	10.06	GigaPose + MegaPose	134.03 s	1.12
EgoX	14.60 s	10.28	**EgoX**	**14.60 s**	**10.28**

E Reconstruction Capabilities

Reconstruction scope. Existing methods reconstruct different subsets of the egocentric scene and often rely on multi-stage inference. Table 8 compares their output coverage and inference structure. EgoX is the only end-to-end method in this comparison that jointly reconstructs cameras, depth, hands, and object 6-DoF poses in a shared world frame.

Table 8: **Reconstruction Outputs and Inference Structure.** Checkmarks denote supported outputs; inference is classified as end-to-end or multi-stage.

Method	Outputs					Inference
	Cam.	Depth	Hands	Obj. 6DoF	World-space	
VGGT-Ω	✓	✓	✗	✗	✗	End-to-end
MapAnything	✓	✓	✗	✗	✓	End-to-end
DA3 Nested	✓	✓	✗	✗	✓	End-to-end
HaWoR	✓	✗	✓	✗	✓	Multi-stage
Hand3R	✓	✓	✓	✗	✓	End-to-end
MINT	✓	✗	✓	✗	✓	End-to-end
ReViV	✓	✓	✓	✗	✓	End-to-end
HandFlow	✓	✗	✓	✗	✓	End-to-end
FoundationPose	✗	✗	✗	✓	✗	Multi-stage
Any6D	✗	✗	✗	✓	✗	Multi-stage
GigaPose + MegaPose	✗	✗	✗	✓	✗	Multi-stage
HOPformer	✗	✗	✓	✓	✗	End-to-end
WHOLE	✗	✗	✓	✓	✓	Multi-stage
EgoX	✓	✓	✓	✓	✓	End-to-end

F Additional Qualitative Results

Figures 8, 9, and 10 extend the main qualitative comparisons with additional HOI4D and H2O sequences for world-space hand and camera reconstruction, camera-frame object pose, and depth, respectively.

F.1 Additional hand and camera comparisons

Figure 8 supplements Figure 5 with additional world-space hand and camera reconstructions.

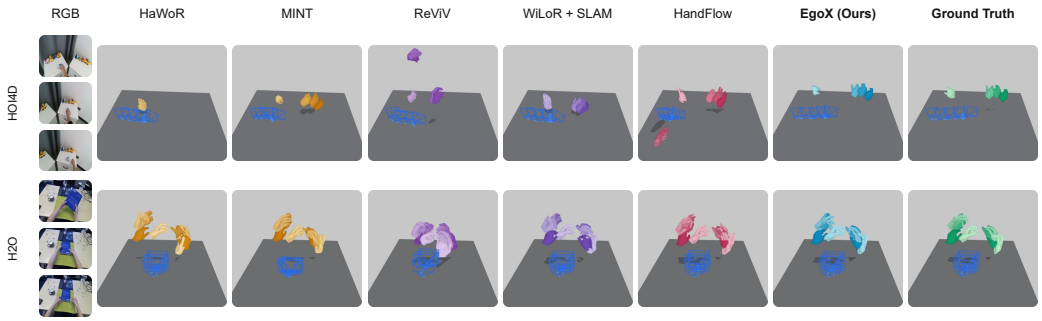

Figure 8: **Additional World-Space Hand and Camera Reconstruction.** Qualitative comparisons on HOI4D and H2O, supplementing Figure 5.

F.2 Additional object-pose comparisons

Figure 9 supplements Figure 6 with additional camera-frame object-pose estimates.

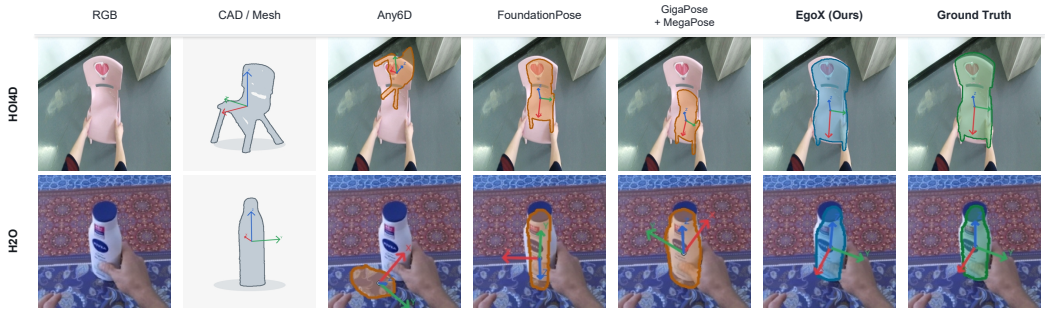

Figure 9: **Additional Camera-Frame Object Pose Reconstruction.** Qualitative comparisons on HOI4D and H2O, supplementing Figure 6.

F.3 Additional depth comparisons

Figure 10 supplements Figure 7 with additional depth predictions.

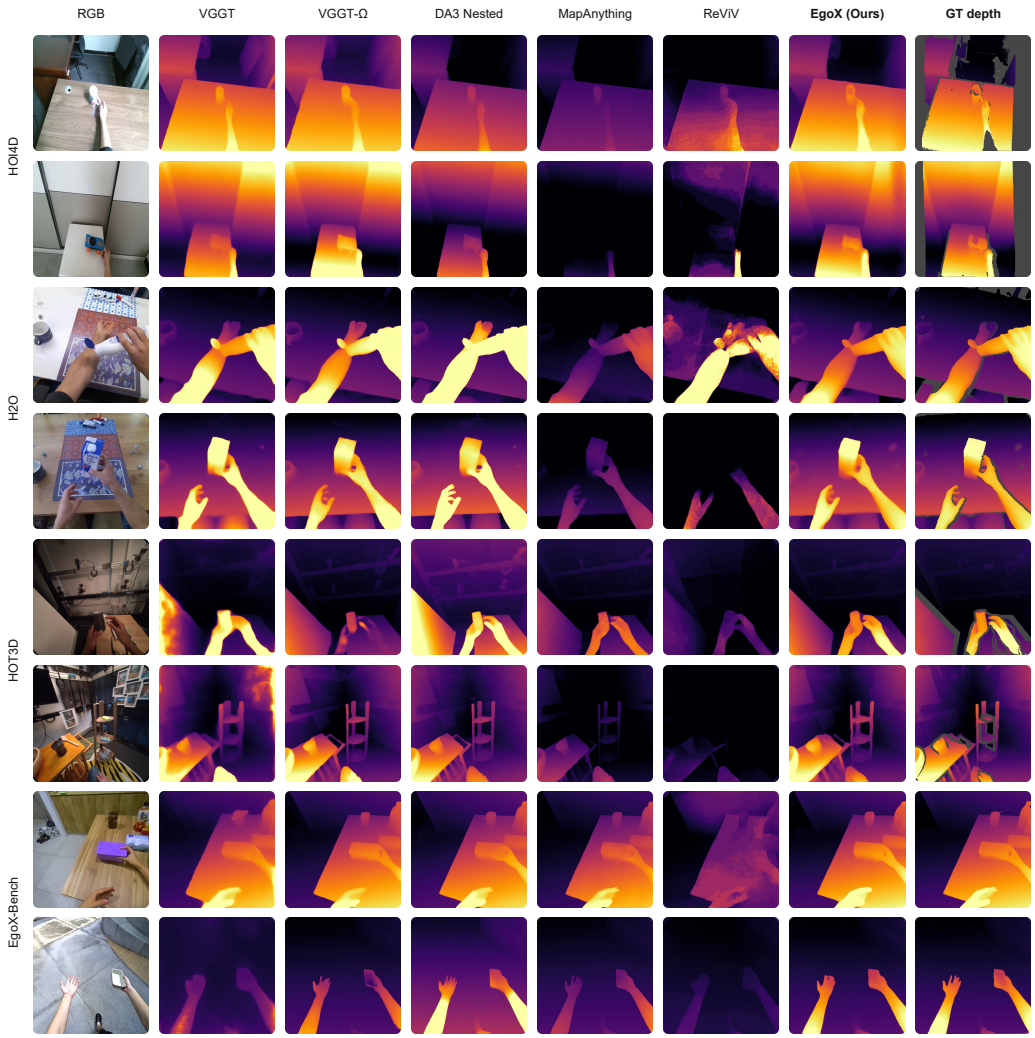

Figure 10: **Additional Depth Reconstruction Results.** Comparisons use the same reference-depth processing and scale protocol as Figure 7.

G Training and Compute

Training supervision. We combine the EgoX-Data-6M training split with six independently sourced public interaction datasets: HOT3D (Banerjee et al., 2025), HOI4D (Liu et al., 2022), H2O (Kwon et al., 2021), HO-Cap (Wang et al., 2025b), H2O-3D (Hampali et al., 2022), and DexYCB (Chao et al., 2021). We reprocess them so that depth, hand geometry, and object geometry agree under consistent metric camera conventions, avoiding contradictory supervision across tasks. Additional real-world sources are Xperience-10M (Ropedia, 2026), ACE-Data-0 (Cao et al., 2026c), and AgiBot World 2026 (AgiBot World Team, 2026); synthetic sources are Aria Synthetic Environments (Avetisyan et al., 2024), BEHAVIOR-1K (Li et al., 2024), SceneNet RGB-D (McCormac et al., 2017), and Dynamic Replica (Karaev et al., 2023). The pool contains approximately 280 hours. Training first mixes real and synthetic sources and then continues on an equal mixture of the retained real interaction sources. We sample $T \in \{1, \ldots, 24\}$ observations per input set.

Architecture details. DINOv3 ViT-L/16 provides patch features to a 24-block, 1024-dimensional reconstruction transformer. Nineteen blocks exchange the full token set across views and five use the compact register/entity set. Camera and depth heads read fused image features; hand and ob-

ject heads read their corresponding view-specific entity states. The model uses learned spatial and reference embeddings and no temporal encoding.

Coordinate and scale conventions. Training uses a common normalization factor z_{gt}, defined as the mean Euclidean distance of valid metric depth points from the reference-camera origin after transformation into $\mathcal{W}$. Metric depth and translations are divided by z_{gt}; rotations and intrinsics are unchanged. For wrist-centred MANO joints $\hat{J}^m_{\mathrm{rel},t}$ and reconstructed wrist $\hat{w}^n_t$, training-space and metric joints are

$$\hat{J}^n_t = \frac{\hat{J}^m_{\mathrm{rel},t}}{z_{\mathrm{gt}}} + \hat{w}^n_t, \qquad \hat{J}^m_t = \hat{J}^m_{\mathrm{rel},t} + \hat{s}\,\hat{w}^n_t. \tag{3}$$

Object translations use the annotated object origin. Object orientation is represented relative to the reference view as $\Delta A^o_{t,j} = A^o_{t,j}(A^o_{r,j})^\top$; the corresponding camera-conditioned target is $Q^{o,*}_{t,j} = R^*_t \Delta A^{o,*}_{t,j}$.

Detailed objective. The parameter term applies ℓ_1 losses to normalized camera translation, standardized unit quaternions, field of view, and normalized principal point; geodesic losses to hand and object rotations; Smooth L1 losses to MANO shape, image positions, and log-depths; and class-weighted binary cross-entropy to in-frame predictions. Image-position residuals are normalized by the longer image dimension. Ground-truth cameras define the view-local targets.

The dense losses $\mathcal{L}_D$ and $\mathcal{L}_P$ follow VGGT-Ω, combining confidence-weighted residuals, inverse-depth weighting, confidence regularization, and multiscale gradient penalties. Hand and object reconstruction use Smooth L1 penalties after composition through predicted cameras. Object rotation uses $d_{\mathrm{sym}}(A, B) = \min_{G\in\mathcal{S}_r} d(A, BG)$, where d is geodesic distance and $\mathcal{S}_r$ is the object's symmetry group in the reference orientation. The reference rotation is supervised toward identity without symmetry relaxation; the coupled term applies to non-reference views. All reductions use annotation-validity and in-frame masks, and object terms additionally require a bound reference mask. Missing annotations are not treated as negative labels.

Sampling and optimization. The source composition and two-stage mixture are summarized above. Each per-GPU batch samples $T \in \{1, \ldots, 24\}$ and uses a T-dependent batch size to maintain an approximately constant image budget. The image encoder is initialized from pretrained DINOv3 weights and the remaining modules randomly. All modules are optimized jointly using AdamW with linear warm-up, cosine decay, mixed precision, and module-wise gradient clipping.

The final release will report the resolved configuration of the evaluated checkpoint, including loss weights, source manifests, sampling probabilities, optimizer schedule, seed, precision, hardware, training time, and checkpoint-selection protocol.