MINT: A Unified Model for World-Space Camera and Hand Motion Estimation from Scalable Egocentric Pipeline Supervision

Zijie Zhu[1,3,4], Weiren Cai[3], Yizhou Wang[1,3], Zhenjie Yang[4], Yide Liu[3,5], Jiahao Chen[3,*], and Guanqi He[2,3,*]

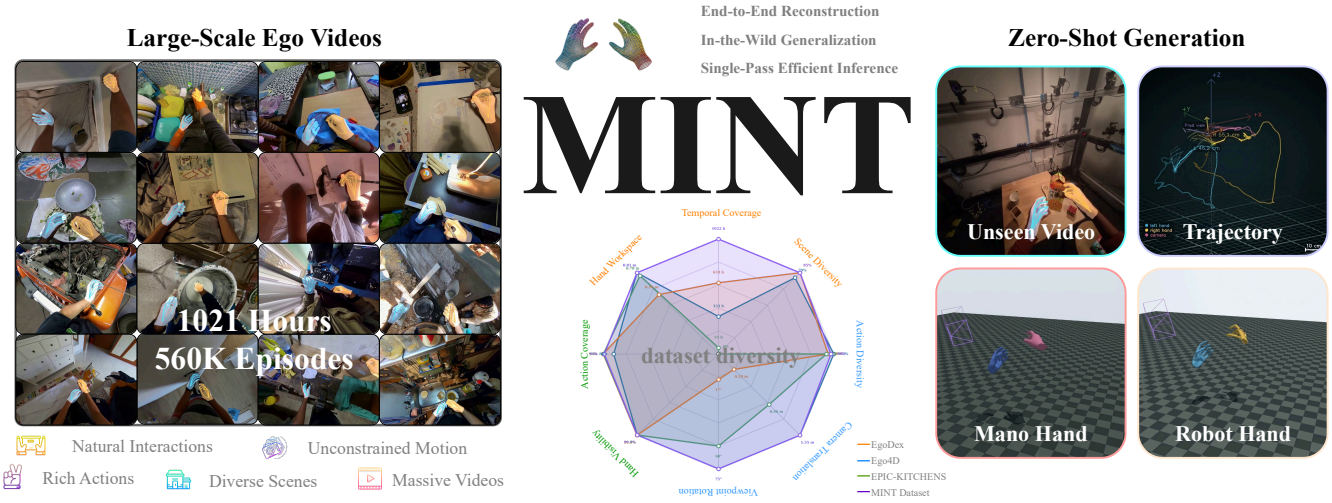

Fig. 1. MINT converts ordinary egocentric video into world-space motion supervision, and MINT recovers that motion with a single feed-forward model. Top: EgoPipeline labels 1,021 hours of public egocentric video drawn from Ego4D [8], EPIC-KITCHENS [4] and EgoDex [10] (center) with camera trajectories and bimanual MANO [21] states; each pair shows an input frame and the recovered hands reprojected onto it. Bottom: trained on that supervision, without any data from the evaluation domains, MINT transfers zero-shot to unseen domains.

***Abstract*— Recovering camera and hand motion in world coordinates from egocentric video is a key capability for activity understanding, robot learning, and augmented reality. Existing systems typically decompose this problem into separate stages for camera motion, depth estimation, hand reconstruction, and trajectory refinement, resulting in substantial computational overhead and preventing the joint modeling of camera and hand motion. We introduce MINT (Minting IN-the-Wild Trajectories), a foundation model for world-space hand motion reconstruction from ego-centric RGB video. From a single shared spatiotemporal video representation, MINT jointly predicts the camera trajectory, field of view (FoV), camera-frame hand states, and per-frame hand presence, and then produces world-space hand motion via explicit coordinate transformations. Training such a model at scale is challenging, since paired world-space camera and hand annotations are scarce. We therefore develop an open-source labeling EgoPipeline that converts large collections of public egocentric videos into structured camera-and-hand trajectory supervision. MINT is first pretrained on these large-scale pseudo-labels and then fine-tuned on a small set of high-quality camera-and-hand annotations. Across public benchmarks MINT approaches state-of-the-art accuracy without seeing either benchmark in training, reaching 23.62 mm MPJPE-p and 10.69 mm PA-MPJPE-p for camera-frame bimanual reconstruction on HOT3D, 4.69 mm RPE-T and 0.284° RPE-R for camera trajectory, and a $3.67\times$ end-to-end speedup over the labeling pipeline that supervises it. We release the model, training and inference code, labeling pipeline, and a curated 1,021-hour egocentric trajectory dataset.**

***Index Terms*— egocentric video; world-space hand reconstruction; camera trajectory; pipeline amortization; cross-domain generalization**

[1]ShanghaiTech University, [2]Tsinghua University, [3]Wuji Technology, [4]The University of Hong Kong, and [5]Zhejiang University.
[*]Corresponding authors.

I. Introduction

Egocentric video records rich human motion and hand–object interactions, and recovering their world-space motion provides an important source of motion supervision for embodied AI training [14]. Yet accurate world-space supervision remains scarce. Capturing such supervision typically requires specialized sensing, precise calibration, synchronized devices, or costly motion reconstruction, making it difficult to obtain large-scale, high-quality world-space supervision from publicly available egocentric video.

Existing reconstruction methods decompose egocentric motion into separate modules for hand detection, motion reconstruction, and camera trajectory estimation. Such pipelines can recover individual sequences, but typically require complex optimization procedures and substantial engineering effort. Hand estimation depends on a detector and can fail under severe occlusion, while errors in cascaded reconstruction propagate across stages, and representations are difficult to share across tasks. Meanwhile, large-scale 3D reconstruction models provide a basis for jointly modeling camera and hand motion. Their pretraining captures long-

range temporal correspondence, camera motion, and scene geometry within a shared spatio-temporal representation. We therefore ask whether such a unified geometric representation can support both camera and hand motion reconstruction.

We present MINT, a unified model for recovering world-space camera and bimanual hand motion from egocentric RGB video. MINT adapts a pretrained geometric reconstruction model to egocentric data and predicts camera extrinsics, horizontal and vertical fields of view, left- and right-hand MANO [21] states, and per-frame hand observability from a shared representation. It requires no external hand detector, motion infiller, or test-time optimization, and no dense depth or point-map intermediate. For each 32-frame window, all outputs are obtained in a single encoder forward pass, while videos of arbitrary length are covered by chaining overlapping windows (Sec. IV-B), enabling unified motion reconstruction.

Real-world egocentric videos contain diverse scenes and hand motion, but lack accurate world-space hand supervision. We therefore build EgoPipeline to produce motion supervision at scale from ordinary video. We process 1,729 hours from Ego4D [8], EPIC-KITCHENS [4], and EgoDex [10] into 1,021 hours of world-space camera and hand motion supervision (Fig. 1). Because the camera trajectories in this supervision inherit the scale ambiguity of monocular reconstruction, we introduce a second training stage that adapts the camera head alone using a human egocentric dataset captured with high-precision devices. The pseudo-labels provide broad coverage, while this dataset provides precise absolute scale.

Neither HOT3D [1] nor ARCTIC [7] is included in the training data; both are therefore evaluated in the zero-shot setting. On HOT3D, MINT approaches state-of-the-art accuracy on all eight metrics we report, remaining close to a model trained on that benchmark. For camera trajectory estimation, MINT achieves the lowest relative translation error on ARCTIC.

Our contributions are:

- **Unified egocentric reconstruction.** MINT jointly recovers camera state, field of view, bimanual hand state, and hand observability from a unified spatio-temporal geometric representation. Each 32-frame window requires a single encoder forward pass, with no detector, motion infiller, or test-time optimization.
- **Scalable egocentric data production.** EgoPipeline converts large-scale ordinary egocentric video into structured world-space motion supervision and combines broad, automatically reconstructed data with a small amount of precise geometric supervision.
- **Large-scale structured egocentric data.** We provide a broad corpus of reconstructed egocentric motion containing camera trajectories, bimanual hand state, world-space motion, hand observability, and action descriptions.

II. Related Work

A. Egocentric Video Understanding

Large-scale egocentric video datasets, including Ego4D [8], EPIC-KITCHENS [4], and EgoDex [10], provide important resources for learning human actions and visual-language representations from diverse real-world activities. ViTRA [14] further demonstrates the value of large-scale real-life human videos for pretraining, covering diverse tasks, objects, and environments and improving zero-shot action prediction. Extracting 3D hand-motion supervision from such videos typically requires multiple stages, including camera calibration, depth estimation, visual SLAM, hand reconstruction, and trajectory processing, resulting in a complex data production pipeline with substantial engineering effort for optimization and deployment.

B. Video-Based Hand Pose Estimation

MANO [21] provides a generic low-dimensional parameterization from images to hand pose and shape. Monocular hand reconstruction methods mainly recover hand pose and shape by estimating MANO [21] parameters from RGB input. HaMeR [18] employs a pretrained ViT with large-scale training data to improve generalization to real-world images, while WiLoR [19] combines hand localization and 3D reconstruction for single-frame hand-state estimation. Video-based methods further introduce temporal modeling. DynHaMR [37] incorporates a two-hand interaction motion prior with test-time optimization, while HaWoR [39] combines camera-frame hand estimation, egocentric SLAM trajectory estimation, and motion infilling based on monocular metric depth. ViDiHand [33] exploits the spatiotemporal prior of a pretrained video diffusion model to jointly estimate hand presence and temporally stable 3D hand pose without an explicit detector or test-time optimization, while the reconstructed hand states remain in the camera frame.

The 3D supervision used by these methods is primarily obtained from dedicated sensing systems and staged laboratory captures, including the multi-depth-camera setup of DexYCB [2], the head-mounted system of HOT3D [1], and the multi-view marker-based motion capture system of ARCTIC [7]. These systems provide accurate 3D annotations but rely on specialized acquisition setups, with limited coverage of hand motions, object interactions, and real-world scenes.

C. Feed-Forward 3D Reconstruction

Camera and scene reconstruction has increasingly moved from per-video optimization toward learned feed-forward estimation. DROID-SLAM [24] combines deep networks with differentiable optimization to jointly estimate camera poses and dense geometry, while MegaSaM [13] incorporates learned geometric representations into SLAM for dynamic and casually captured videos.

Recent feed-forward methods directly predict camera and scene geometry from images or videos. VGGT [32] uses cross-view attention to estimate camera parameters and dense geometry from multiple frames, while π^3 [31] employs permutation-equivariant modeling to predict affine-invariant

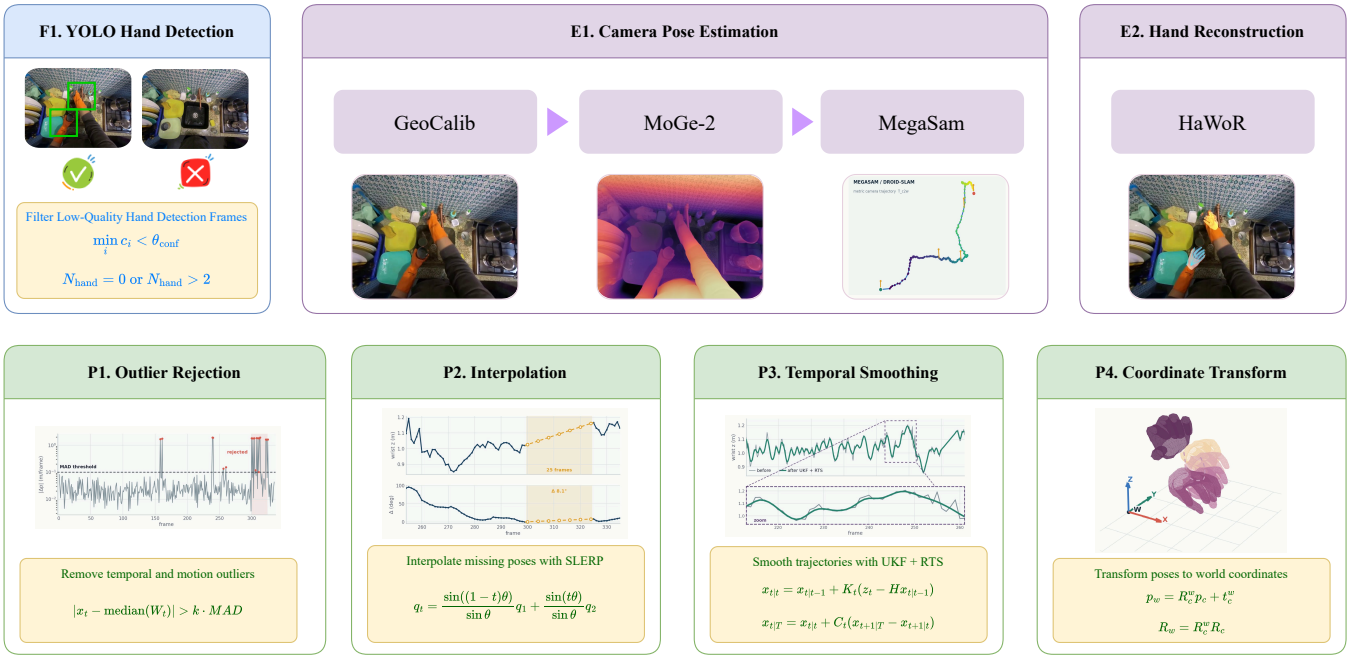

Fig. 2. Overview of EgoPipeline. Egocentric videos are first pre-filtered by hand detection. The retained clips are then processed by GeoCalib for camera intrinsics, MoGe-2 for monocular depth, MegaSaM for camera pose, and HaWoR for bimanual hand reconstruction. A final post-processing stage stabilizes the recovered trajectories and composes the camera-frame hand states with the camera trajectory to produce world-space bimanual motion supervision.

camera poses and scale-invariant local point maps. LingBot-Map [3] extends feed-forward reconstruction to long streaming videos by aggregating historical observations into a spatiotemporal geometric representation, enabling continuous estimation of camera trajectories and scene geometry over long sequences. These works establish a feed-forward paradigm for geometric reconstruction from long video sequences. Our work extends this paradigm to egocentric hand reconstruction by jointly modeling camera and bimanual motion and recovering temporally consistent 3D hand motion in a unified world coordinate system.

III. EgoPipeline Structured Supervision

Existing egocentric hand reconstruction and hand–object interaction datasets are predominantly collected in controlled environments with predefined tasks, limiting the diversity of bimanual interactions and real-world conditions they capture [1], [2], [7]. In contrast, publicly available egocentric videos cover diverse activities, objects, environments, and interactions. To exploit this diversity for large-scale 3D motion learning, we collect 1,729 hours of egocentric video spanning open-ended daily activities, fine-grained bimanual manipulation on tabletops, and structured kitchen interactions. These videos contain visual and motion variations that are difficult to reproduce in controlled capture, including complex backgrounds, dynamic object interactions, severe hand occlusions, variations in hand scale, and illumination conditions.

To extract 3D motion supervision from these videos, we develop a first-person video processing pipeline inspired by the data construction paradigm of ViTRA [14], illustrated in Fig. 2. We first compress and encode the videos for improved I/O throughput and use high-confidence YOLO detections to filter clips with no detected hands for more than 1 s or more than two detected hands. For the remaining clips, we estimate camera intrinsics with GeoCalib [26], undistort the frames, and inject the monocular depth prior from MoGe-2 [30] into MegaSaM [13] or DROID-SLAM [24] to recover metric-scale camera trajectories. In parallel, HaWoR [39] reconstructs both hands and estimates per-frame motion states in the camera coordinate system, with each hand represented by a 6D wrist pose and MANO [21] parameters. We then combine the camera-space hand states with the recovered camera trajectory to obtain world-space bimanual motion.

Because HaWoR relies on YOLO-based hand detections, reconstruction failures and intermittent detections can produce outliers, abrupt trajectory changes, and short missing intervals. We therefore apply unified post-processing with outlier removal, missing-interval interpolation, and UKF temporal filtering to stabilize the recovered trajectories and provide more reliable motion supervision. Finally, we construct 1,021 hours of effective data with world-coordinate 3D bimanual motion supervision.

Despite enabling large-scale supervision generation, the pipeline relies on multiple heterogeneous models for hand detection, camera calibration, depth estimation, SLAM, and hand reconstruction. This design incurs substantial computational and engineering overhead, while errors from different stages can propagate through the reconstruction process. We therefore propose MINT, a unified feed-forward egocentric 3D reconstruction framework that jointly models camera and bimanual motion directly from video. MINT learns their coupled geometry and predicts camera and hand motion in

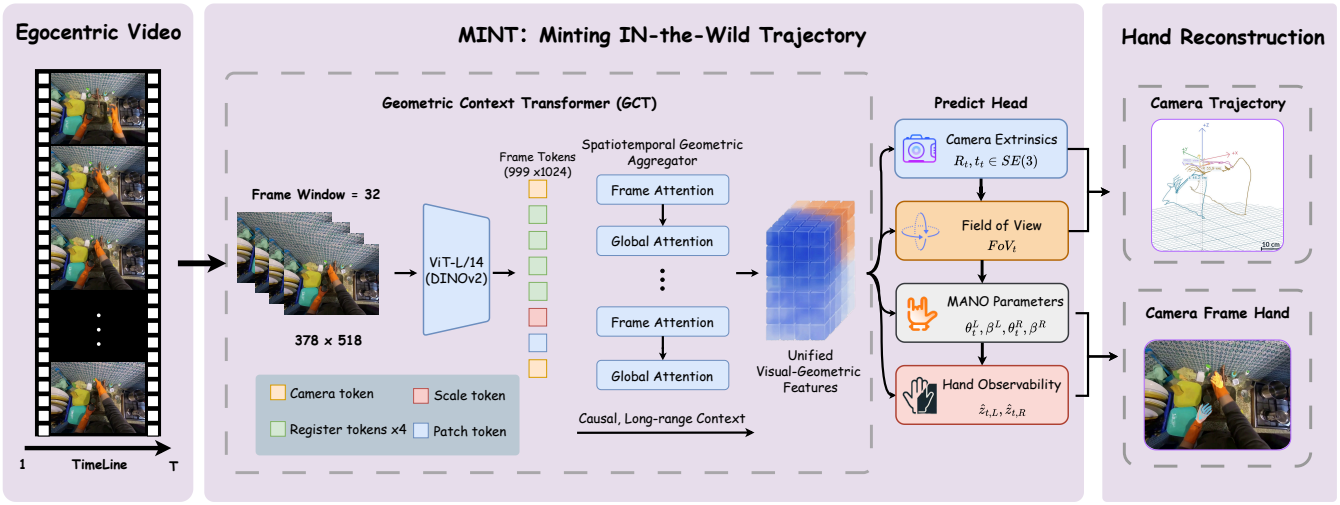

Fig. 3. Architecture of MINT. Each window of 32 egocentric frames is encoded at 378×518 by a DINOv2 ViT-L/14 patch embedding, and the resulting frame tokens are processed by the spatiotemporal geometric aggregator of GCT, which alternates frame attention and global attention so that every frame of the window is encoded with bidirectional temporal context. The unified visual-geometric features carry camera, register, scale, and patch tokens, and feed four prediction heads for camera extrinsics, field of view, bimanual MANO parameters, and hand observability. Composing the predicted camera trajectory with the camera-frame hand states yields world-space bimanual reconstruction.

a common world coordinate system.

IV. Method

A. Overview

Monocular hand reconstruction aims to recover 3D hand pose and motion from egocentric videos. Existing methods typically rely on a pipeline of hand detection, camera tracking, hand fitting, and post-processing, which requires substantial engineering effort, accumulates errors across stages, and limits overall processing throughput. We therefore leverage a pretrained Geometric Context Transformer (GCT), the geometric encoder of LingBot-Map [3] trained for long-range 3D reconstruction from video, to learn a unified visual-geometric representation for hand reconstruction. GCT encodes a window of frames in a single bidirectional spatiotemporal pass, and four task heads regress camera extrinsics, field of view (FoV), MANO parameters for both hands, and hand observability, as shown in Fig. 3. MINT recovers camera and bimanual hand motion in the world coordinate frame from egocentric video with a single feed-forward model, without a hand detector, motion infiller, or test-time optimization. A window is processed in one pass, and videos longer than the window are covered by chaining windows as described in Sec. IV-B.

We adopt a two-stage training strategy. In the first stage, the visual-geometric representation and task heads are jointly trained on diverse egocentric datasets to establish coupled camera and hand reconstruction. In the second stage, while preserving the learned geometric representation and hand reconstruction capability, we adapt only the camera extrinsics head on a small in-house corpus whose camera trajectories are captured with high-precision devices. This curriculum transfers long-range 3D reconstruction priors to egocentric hand motion recovery while specializing camera estimation for accurate world-space annotation.

B. Visual Geometry Grounded Backbone

Egocentric video interleaves articulated hand motion, hand–object contact, camera egomotion, and the surrounding 3D layout in a single stream, and none of them is fully recoverable from an isolated frame. To obtain temporally consistent geometric features, MINT adopts the pretrained geometric context transformer (GCT) of LingBot-Map [3], a visual geometry grounded transformer [32], as its shared spatiotemporal geometric encoder. Reconstruction training over a broad distribution of scenes has established in that model robust geometric priors for scene structure, viewpoint change, and trajectory consistency, and its aggregation of features along a video sequence also matches the temporal input of egocentric recordings more closely than models built for unordered image sets.

a) Geometric feature extraction: We retain the original visual encoder and geometric aggregator so that these pretrained priors are inherited. Each frame is resized to 378×518 and encoded by a ViT-L/14 encoder, initialized from DINOv2 [17] and pretrained jointly with the aggregator, into 999 tokens of dimension 1024. One camera token, four register tokens, and one scale token are prepended to every frame, so that frame-level global quantities are carried in dedicated slots. The sequence then alternates between frame attention and global attention,

$$\boldsymbol{X}_{1:T}^{l} = G_{\text{global}}^{l}\left(G_{\text{frame}}^{l}(\boldsymbol{X}_{1:T}^{l-1})\right), \tag{1}$$

where a frame module models the spatial structure inside a single image and a global module propagates geometric context along time, so that egomotion is inferred from cross-frame evidence rather than from single-frame visual cues. Features taken from several depths are aggregated into shared

prediction features, which the camera and bimanual heads of Sec. IV-C read from the same forward pass.

b) Long-sequence inference: The global attention of GCT operates on a clip of fixed length and is trained at $T = 32$ frames, so extending attention across a complete recording would let computation and memory grow with sequence length. We therefore segment a video of arbitrary length into overlapping sliding windows. Each window is encoded independently by the same shared network, the camera trajectories of adjacent windows are aligned and chained in $\mathrm{SE}(3)$, and the hand parameters and observability scores of the frames two windows share are fused. This design bounds the computation of a single pass to one fixed window while keeping the total cost approximately linear in the length of the video, which supports continuous geometric estimation over long sequences.

C. Prediction Heads

Through large-scale geometric reconstruction pretraining, GCT acquires strong priors for scene geometry, viewpoint change, and temporal motion, allowing its features to encode scene structure, viewpoint changes, and cross-frame motion. However, these generic geometric features do not directly specify the structured state required for hand reconstruction. Unlike rigid scene geometry, egocentric hand reconstruction must jointly recover wrist pose, articulated joint motion, and hand shape while handling occlusion and partial observability during hand–object interaction. We therefore introduce dedicated hand and camera prediction heads on top of these geometric features, mapping them to structured bimanual MANO [21] states and camera states. The hand state is recovered in the camera frame and coupled with camera motion through a differentiable camera-to-world transformation, while a hand observability head suppresses unreliable supervision from partially observed hands.

For each hand $s \in \{L, R\}$, we parameterize its state as

$$\hat{\boldsymbol{h}}_{t,s} = [\hat{\boldsymbol{\Lambda}}_{t,s}, \hat{\boldsymbol{\Gamma}}_{t,s}, \hat{\boldsymbol{\Theta}}_{t,s}, \hat{\boldsymbol{\beta}}_{t,s}] \in \mathbb{R}^{109}, \tag{2}$$

where $\boldsymbol{\Lambda}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\beta}$ denote wrist translation, wrist rotation, fifteen joint rotations, and hand shape, respectively. All rotations use the continuous six-dimensional representation, resulting in 218 hand parameters for both hands at each frame.

The camera state is parameterized as

$$\hat{\boldsymbol{c}}_t = [\hat{\boldsymbol{t}}_t, \hat{\boldsymbol{q}}_t, \hat{\boldsymbol{f}}_t] \in \mathbb{R}^9, \tag{3}$$

comprising the camera translation $\boldsymbol{t}_t$, a unit quaternion $\boldsymbol{q}_t$ with rotation matrix $\boldsymbol{R}_t$ under the world-to-camera convention, and the vertical and horizontal fields of view $\boldsymbol{f}_t$.

a) Structured Hand Motion Prediction: Egocentric hand motion follows a structured kinematic model, and frame-wise features alone are insufficient to reliably recover the full MANO state, particularly under rapid motion and frequent occlusion. We therefore directly decode the structured states of both hands from the spatio-temporal geometric features.

Each hand carries four learnable queries corresponding to wrist translation, wrist rotation, joint rotations, and hand shape. Let $\boldsymbol{X}_t$ denote the final-layer spatio-temporal tokens of frame t. Each query interacts with these tokens through

$$\boldsymbol{Q}^s_{t,k} = \mathrm{CrossAttn}(\boldsymbol{q}^s_k, \boldsymbol{X}_t, \boldsymbol{X}_t), \quad k \in \{\Lambda, \Gamma, \Theta, \beta\}. \tag{4}$$

The four resulting features are jointly mapped to the MANO state.

Frame-wise decoding alone does not guarantee temporal consistency, since rapid motion and occlusion can cause discontinuities across neighboring frames. We therefore condition the hand queries on the previous prediction and perform two refinement iterations. The first iteration predicts absolute rotations, while subsequent iterations predict a bounded axis-angle increment in the current local frame:

$$\tilde{\boldsymbol{\omega}} = \frac{\boldsymbol{\omega}}{\|\boldsymbol{\omega}\|}\, \omega_{\max} \tanh\!\left(\frac{\|\boldsymbol{\omega}\|}{\omega_{\max}}\right), \qquad \omega_{\max} = 30^\circ. \tag{5}$$

The update is then applied as $\boldsymbol{R}^{(k+1)} = \boldsymbol{R}^{(k)} \exp([\tilde{\boldsymbol{\omega}}^{(k)}]_\times)$, and mapped back to the continuous six-dimensional representation. The residual layers are zero-initialized, making the refinement module an identity mapping at the beginning of training.

b) Camera State and Hand–Camera Geometric Coupling: The hand state is recovered in the camera frame, whereas the final hand motion must be expressed in a common world frame. The camera state therefore determines both camera motion and the mapping of the hand trajectory into world space.

We adopt the LingBot-Map [3] camera head, which predicts camera pose through iterative refinement. The previous prediction is embedded and injected into a causal temporal trunk through adaptive layer normalization, which predicts the current update. We retain four refinement iterations and move the field of view out of the feedback path, reducing the iterative state from nine to seven dimensions.

The field of view is predicted by an independent branch:

$$\hat{\boldsymbol{f}}_t = \mathrm{Softplus}\big(H_{\mathrm{FoV}}(\boldsymbol{X}^{\mathrm{cam}}_t)\big), \tag{6}$$

where the camera token is first projected to 512 dimensions and processed by two temporal blocks. The predicted extrinsics and fields of view are then concatenated into the camera state in Eq. (3).

Let $\boldsymbol{p}^c_{t,s}$ and $\boldsymbol{Q}^c_{t,s}$ denote the wrist translation and rotation matrix of hand s in the camera frame. They are mapped to the world frame as

$$\boldsymbol{p}^w_{t,s} = \boldsymbol{R}_t^\top (\boldsymbol{p}^c_{t,s} - \boldsymbol{t}_t), \quad \boldsymbol{Q}^w_{t,s} = \boldsymbol{R}_t^\top \boldsymbol{Q}^c_{t,s}. \tag{7}$$

Joint rotations and hand shape remain invariant under this transformation. Since the transformation is differentiable, the world-frame loss can jointly constrain the camera and hand predictions, explicitly coupling their geometry.

c) Hand Observability: Hands in egocentric videos are frequently affected by object interaction, self-occlusion, and image-boundary truncation, making some frames unreliable for geometric supervision. We therefore introduce a lightweight hand observability head to determine whether

Table 1. Camera-frame bimanual reconstruction on ARCTIC [7] and HOT3D [1]. All rows are scored under the coverage-aware protocol of Sec. V-A, in which a hand a method declines to report is charged the error of a canonical MANO [21] placeholder. MPJPE-p and PA-MPJPE-p in mm, GO-p in degrees, CT-p in m, Jitter in mm/frame^2. ViDiHand [33] (marked *) is trained on most of both benchmarks and is therefore a reference row rather than a comparable one, while MINT is zero-shot. The MINT + UKF row applies the filter of Sec. III at inference time and changes nothing else. Bold marks the best value among the comparable rows, that is excluding the reference row.

		Detection			3D pose		Orient. & pos.		Temporal
	Method	FAcc ↑	Recall ↑	F1 ↑	MPJPE-p ↓	PA-MPJPE-p ↓	GO-p ↓	CT-p ↓	Jitter ↓
ARCTIC	InterWild [16]	0.878	0.943	0.959	30.82	15.95	25.39	0.097	46.58
	HaMeR [18]	0.875	0.943	0.957	29.20	14.60	24.91	0.095	18.28
	Hamba [5]	0.833	0.912	0.941	31.23	17.17	27.82	0.110	15.36
	WildHands [20]	0.879	0.946	0.960	25.70	13.94	22.32	**0.058**	12.97
	OmniHands [15]	0.866	0.949	0.954	29.67	14.20	24.58	0.087	45.31
	WiLoR [19]	**0.919**	0.951	0.974	**22.01**	**11.87**	**17.36**	0.075	24.09
	Dyn-HaMR [37]	0.842	0.918	0.951	27.90	17.02	25.95	0.121	12.84
	HaWoR [39]	0.700	0.817	0.895	45.36	26.38	43.33	0.149	19.79
	ViDiHand [33]*	0.997	0.999	0.999	21.67	9.82	14.64	0.047	3.18
	MINT	0.916	**0.957**	**0.978**	51.03	27.71	24.19	0.140	12.26
	MINT + UKF	0.916	0.957	0.978	51.09	27.70	24.22	0.140	**2.54**
HOT3D	InterWild [16]	0.669	0.881	0.868	77.17	24.81	58.50	0.213	101.16
	HaMeR [18]	0.692	0.904	0.883	68.31	21.46	49.64	0.102	23.63
	Hamba [5]	0.632	0.828	0.853	71.73	29.62	56.53	0.128	18.51
	WildHands [20]	0.655	0.863	0.844	52.79	28.95	53.93	0.157	22.89
	OmniHands [15]	0.649	0.895	0.868	63.28	22.68	49.12	0.133	69.51
	WiLoR [19]	0.827	0.897	0.937	30.97	19.98	25.75	0.098	17.98
	Dyn-HaMR [37]	0.614	0.811	0.802	74.21	38.20	43.85	0.571	44.94
	HaWoR [39]	0.348	0.499	0.654	71.40	66.03	79.35	0.262	23.87
	ViDiHand [33]*	0.948	0.974	0.983	21.51	11.38	15.83	0.040	3.74
	MINT	**0.940**	**0.977**	**0.950**	**23.61**	10.70	16.78	**0.073**	11.52
	MINT + UKF	0.940	0.977	0.950	23.62	**10.69**	**16.77**	0.073	**2.39**

each hand in the current frame can provide reliable supervision.

For each hand, a learnable query cross-attends to the current-frame patch tokens. The tokens are projected to 256 dimensions, and the head predicts the left- and right-hand observability logits $\hat{z}_{t,L}$ and $\hat{z}_{t,R}$. The corresponding probabilities $\hat{a}_{t,s}$ determine whether that hand contributes to the hand reconstruction loss and world-frame loss.

D. Training Objectives

The egocentric reconstruction pipeline provides high-quality hand motion from large-scale videos, including wrist trajectories, joint motion, and hand–object interactions, while covering a broad range of tasks, scenes, and action variations. These data provide large-scale supervision for learning hand motion and geometric relationships in egocentric videos.

However, the corresponding camera trajectories typically rely on monocular depth and SLAM. Although monocular SLAM captures relative frame-to-frame motion well, its scale ambiguity can introduce systematic errors in the absolute trajectory scale. We therefore use two training stages: the first learns stable hand–camera geometry from large-scale and diverse data, while the second uses camera motion captured with high-precision devices to correct the remaining scale and absolute pose errors.

a) Stage 1: Learning Hand–Camera Coupled Geometry: The first stage does not directly fit an absolutely accurate camera trajectory. Instead, it exploits rich hand-motion supervision from large-scale egocentric data to learn a stable hand–camera geometric relationship. To improve generalization across tasks, actions, and scenes, we select clips with high scene and action diversity from Ego4D [8], EgoDex [10], and EPIC-KITCHENS [4], and initialize the geometric encoder with the pretrained GCT from LingBot-Map [3].

We jointly optimize four objectives: the camera motion loss $\mathcal{L}_{\text{cam}}$, bimanual MANO loss $\mathcal{L}_{\text{mano}}$, hand observability loss $\mathcal{L}_{\text{obs}}$, and world-frame consistency loss $\mathcal{L}_{\text{world}}$:

$$\mathcal{L}_{\text{stage1}} = \lambda_{\text{cam}}\mathcal{L}_{\text{cam}} + \lambda_{\text{mano}}\mathcal{L}_{\text{mano}} + \lambda_{\text{obs}}\mathcal{L}_{\text{obs}} + \lambda_{\text{world}}\mathcal{L}_{\text{world}}. \tag{8}$$

Through the differentiable transformation in Eq. (7), the world-frame loss jointly constrains the camera and hand predictions, allowing reliable hand-motion supervision to inform their geometric relationship.

The field of view and camera trajectory are predicted by separate heads, preserving the pretrained field-of-view capability while preventing camera-trajectory adaptation from interfering with it.

b) Stage 2: Camera Trajectory Correction: The first stage establishes stable hand–camera geometry, but monocular SLAM can still leave errors in absolute scale and pose. We therefore perform a second stage using camera trajectories captured with high-precision devices.

We freeze the learned geometric encoder, hand, observability, and field-of-view modules, and optimize only the camera trajectory prediction. The objective consists of a translation term $\mathcal{L}_t$, a rotation term $\mathcal{L}_R$, and a temporal consistency term $\mathcal{L}_{\text{temp}}$:

$$\mathcal{L}_{\text{stage2}} = \lambda_t\mathcal{L}_t + \lambda_R\mathcal{L}_R + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \tag{9}$$

Since hand reconstruction and hand–camera geometry are established in Stage 1, a small amount of high-precision camera trajectory data is sufficient to correct the absolute scale and pose while preserving the learned hand motion, hand–camera geometry, and field-of-view estimation.

V. Experiments

A. Setup

Datasets. Stage 1 uses the 1,021 hours of EgoPipeline output from Sec. III, which contains no manually annotated hands and no ground-truth camera poses. Stage 2 trains the camera head alone on a small in-house corpus recorded with a single rig, using the camera trajectories that rig provides; no hand supervision is used in this stage. Neither stage sees HOT3D [1] or ARCTIC [7], on which all results below are therefore zero-shot.

Metrics. Standard hand pose metrics typically evaluate only successfully matched hands and therefore do not account for missed detections. We therefore use a coverage-aware evaluation in which missed hands are penalized rather than excluded from pose evaluation [33], and we group the metrics by what they measure. Detection is measured by frame accuracy (FAcc), recall and F1. Camera-frame hand reconstruction is measured by MPJPE-p and PA-MPJPE-p for articulated pose, by GO-p for wrist orientation and by CT-p for hand placement, all computed under the coverage-aware protocol. For camera trajectories we report ATE and ATE% for global accuracy, RPE-T and RPE-R for relative motion accuracy, and the arc-length ratio for scale deviation. Trajectories are evaluated over the full sequence and aligned with $SE(3)$ without scale fitting, so the reported errors retain the metric scale the model recovers on its own. ATE and RPE are computed per sequence as an RMSE and summarized by a mean and a median. Temporal quality is assessed by RPE for the camera and by Jitter for the hands. RPE is particularly relevant to MINT because the model is trained with inter-frame motion increments, which makes it a direct measure of local motion consistency.

Implementation. MINT is trained on 32-frame clips at 378×518 resolution using AdamW with BF16 mixed precision. Stage 1 uses learning rates of 5×10^{-5}, 1×10^{-4} and 5×10^{-4} for the pretrained backbone, the camera and field-of-view heads and the hand and observability heads respectively, with an effective batch size of 1. In stage 2 the aggregator, the field-of-view head and the hand and observability heads are frozen, and only the camera head is optimized at a learning rate of 5×10^{-5}, with an effective batch size of 4. Both stages use a weight decay of 0.05 and a $2{,}000$-step warmup.

B. Camera-Frame Hands

Detection and pose. In the zero-shot setting MINT remains close to the in-domain ViDiHand and ahead of the zero-shot baselines on HOT3D (Table 1). Predicting both hands on every frame with an explicit observability score keeps detection stable under egocentric occlusion. MINT reaches an FAcc of 0.940 on HOT3D, against 0.827 for the zero-shot WiLoR and 0.948 for the in-domain ViDiHand, and on ARCTIC it stays on par with WiLoR. Pose accuracy follows the same pattern, with 23.61 mm MPJPE-p and 10.70 mm PA-MPJPE-p on HOT3D, within 2.1 mm of ViDiHand on the former and below it on the latter, so a single feed-forward student stays close to the multi-stage systems that produced its supervision. On ARCTIC the two error sources separate. Wrist orientation stays accurate at $24.19°$ GO-p, whereas CT-p reaches 0.140 m, close to the 0.149 m of the HaWoR branch that generates our supervision. The residual error is therefore dominated by camera-space translation rather than by articulated pose, which the gap between MPJPE-p and PA-MPJPE-p confirms.

Temporal smoothness. Without a motion prior or test-time optimization, MINT is already smoother than the optimization-based baselines on both benchmarks. Applying the UKF of Sec. III at inference reduces Jitter by 79% on both benchmarks, below the in-domain ViDiHand, while MPJPE-p and PA-MPJPE-p change by less than 0.1 mm.

C. Camera Trajectory

On camera trajectory (Table 2) MINT gives the best relative translation accuracy on ARCTIC, 3.39 mm RPE-T against 8.73 mm for MegaSaM, and on HOT3D its 4.69 mm RPE-T is second only to MegaSaM, which completes 24 of 27 sequences. Absolute error is where the dedicated systems stay ahead, since a feed-forward pass over 32-frame windows has neither loop closure nor global optimization and accumulates 181.7 mm ATE on HOT3D against 49.1 mm for DROID-SLAM.

The second stage acts on scale rather than on shape. Without it the model reaches an arc-length ratio of 0.466 on HOT3D, so the recovered path is less than half the length of the true one, the signature of pseudo-labels whose scale comes from monocular depth. Stage 2 moves that ratio to 1.094 and cuts ATE from 524.7 to 181.7 mm and RPE-T from 8.75 to 4.69 mm. On ARCTIC the same correction overshoots, taking the ratio from 0.755 to 1.412, so ATE rises from 63.7 to 81.9 mm even though RPE-T still improves from 3.53 to 3.39 mm. Stage 2 therefore removes the systematic under-scaling that dominates on HOT3D, at the price of some over-scaling on a benchmark whose motion is already closer to metric.

D. Ablations

Data mixture. Table 3 varies only the source composition and scores camera-frame hand reconstruction on HOT3D, and each single source behaves as its capture conditions predict. Ego4D [8], which covers the widest range of everyday human tasks, is the stronger single source at 25.32 mm MPJPE-p and the smoothest of all rows at 8.58 Jitter. EgoDex [10], tabletop manipulation whose viewpoint moves little, holds detection at 0.949 F1 but loses ground on where the hand is, $20.69°$ GO-p and 0.083 m CT-p. The mixture is best in five of the six columns, including 23.61 mm MPJPE-p against 25.32 mm for the best single source, and is second on Jitter. Since zero-shot transfer is the target, we train on the mixture, which generalizes across the viewpoints, motions and intrinsics the benchmarks present rather than fitting the capture conditions of one source.

Initialization. Table 4 replaces the LingBot-Map [3] backbone with random initialization. The randomly initialized run is trained on a shorter schedule than the full model, so

Table 2. World-frame camera trajectory results on the HOT3D [1] and ARCTIC [7] validation sets. All sequences are evaluated over their *full* trajectories using $SE(3)$-only alignment without scale fitting, thereby preserving the original scale error, while the trajectory arc-length ratio measures scale consistency. MINT w/o stage 2 never sees metric ground truth.

	Method	ATE ↓ (mm)		RPE-T ↓ (mm)		RPE-R ↓ (deg)		Arc len.	ATE
		mean	med.	mean	med.	mean	med.	ratio → 1	% ↓
HOT3D	DROID-SLAM [24]	**49.1**	**39.6**	5.36	3.52	0.227	0.146	0.778	**0.36**
	HaWoR [39]	200.3	179.2	8.60	7.33	1.098	0.928	**0.950**	1.28
	InfiniteVGGT [38]	124.8	100.0	13.52	9.67	1.492	0.392	0.556	0.80
	LingBot-Map [3]	85.5	51.0	7.56	6.18	0.684	0.253	0.712	0.55
	MegaSaM† [13]	94.4	65.3	**3.18**	**2.13**	**0.082**	**0.063**	0.716	0.69
	MINT w/o stage 2	524.7	434.6	8.75	8.37	0.234	0.229	0.466	3.29
	MINT	181.7	155.5	4.69	4.78	0.284	0.259	1.094	1.15
ARCTIC	DROID-SLAM [24]	181.5	49.6	33.84	14.25	1.006	0.423	**0.964**	8.07
	HaWoR [39]	66.2	**29.4**	24.08	4.16	0.298	**0.151**	0.759	2.87
	InfiniteVGGT [38]	79.0	69.1	16.21	12.63	1.265	0.655	0.284	3.23
	LingBot-Map [3]	59.6	62.0	9.17	8.47	0.980	0.717	0.591	2.46
	MegaSaM† [13]	**51.4**	50.4	8.73	5.58	0.779	0.725	1.956	**2.15**
	MINT w/o stage 2	63.7	59.3	3.53	3.62	**0.251**	0.243	0.755	2.63
	MINT	81.9	83.2	**3.39**	**3.37**	0.256	0.251	1.412	3.40

Table 3. Data mixture, scored as camera-frame hand reconstruction on HOT3D [1] under the protocol of Table 1; only the source composition changes. EPIC-KITCHENS [4] and the single-rig in-house corpus are reported only within the mixture.

Sources	F1 ↑	MPJPE-p ↓	PA-MPJPE-p ↓	GO-p ↓	CT-p ↓	Jitter ↓
Ego4D only	0.947	25.32	11.48	17.33	0.075	**8.58**
EgoDex only	0.949	26.47	11.94	20.69	0.083	15.29
Full mixture	**0.950**	**23.61**	**10.70**	**16.78**	**0.073**	11.52

Table 4. Backbone initialization, scored as camera-frame hand reconstruction on HOT3D [1] under the protocol of Table 1; the randomly initialized run is trained under a shorter schedule.

Init.	F1 ↑	MPJPE-p ↓	PA-MPJPE-p ↓	GO-p ↓	CT-p ↓	Jitter ↓
Random	**0.954**	27.00	12.96	17.77	0.095	18.44
LingBot-Map [3]	0.950	**23.61**	**10.70**	**16.78**	**0.073**	**11.52**

this comparison indicates a trend rather than a converged gap. Random initialization is worse on MPJPE-p, PA-MPJPE-p, GO-p and CT-p, and Jitter rises from 11.52 to 18.44; F1 is marginally higher at 0.954 against 0.950. Geometric pretraining thus matters for where the hand is and how smoothly it moves, not for whether it is found at all.

E. Inference Throughput

We evaluate inference efficiency at 512×384 and 30 fps under identical conditions. MINT encodes each video window in a single bidirectional spatiotemporal pass. On a single GPU, its per-frame latency is 72.4 ms, compared with 105.1 ms for HaWoR [39] and 1260.0 ms for VITRA [14].

Since the windows are mutually independent, extended egocentric recordings admit parallel processing across GPUs without inter-window coordination, which reduces the mean per-frame latency to 22.7 ms. EgoPipeline, although it parallelizes an originally serial cascade, retains a fixed inter-stage dependency whose slowest stage bounds its attainable throughput, and requires 83.4 ms per frame. MINT therefore achieves a $3.67\times$ end-to-end speedup over EgoPipeline, and $12.5\times$ over our reproduction of VITRA, which requires 283.3 ms.

Table 5. Inference efficiency comparison under identical conditions. VITRA [14] is reproduced for reference.

Method	Time (ms/fr.) ↓	fps ↑	Speedup ↑
One GPU			
VITRA [14]	1260.0	0.8	–
HaWoR [39]	105.1	9.5	12.0×
MINT	**72.4**	**13.8**	**17.4**×
Four GPUs			
VITRA [14]	283.3	3.5	–
EgoPipeline	83.4	12.0	3.4×
MINT	**22.7**	**44.1**	**12.5**×

VI. Conclusion

We presented MINT, a unified feed-forward model that recovers world-space camera and bimanual hand motion from egocentric video. No hand label in its training data is human-annotated: MINT is trained on $1{,}021$ hours of motion supervision that EgoPipeline extracts automatically from publicly available egocentric video, and a short second stage uses device-captured camera trajectories from a small in-house corpus to correct absolute scale. The resulting model transfers to unseen domains and is competitive zero-shot on both camera-frame hand reconstruction and camera trajectory estimation. These results indicate that automatically reconstructed motion supervision from public video is sufficient to learn generalizable egocentric motion reconstruction.

a) Limitations and future work: MINT remains affected by the scale drift inherited from monocular reconstruction in the supervision pipeline, and the 32-frame training window leaves long-horizon drift open. Future work will explore more accurate and diverse metric supervision to reduce scale errors and improve long-range trajectory reconstruction.

References

[1] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7061–7071, 2025.

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021.

[3] Lin-Zhuo Chen, Jian Gao, Yihang Chen, et al. Geometric context transformer for streaming 3d reconstruction. *arXiv preprint arXiv:2604.14141*, 2026.

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, et al. Scaling egocentric vision: the dataset. In *European conference on computer vision*, pages 753–771, 2018.

[5] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. *arXiv preprint arXiv:2407.09646*, 2024.

[6] Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6341–6351, 2024.

[7] 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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12943–12954, 2023.

[8] Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022.

[9] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. 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*, pages 19383–19400, 2024.

[10] Ryan Hoque, Peide Huang, David Yoon, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, pages 4218–4237, 2026.

[11] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233, 2025.

[12] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91, 2024.

[13] 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10486–10496, 2025.

[14] Qixiu Li, Yu Deng, Yaobo Liang, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

[15] Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Wei Jing, Qi Yan, Qianying Wang, Yebin Liu, and Hongwen Zhang. Omnihands: Towards robust 4d hand mesh recovery via a versatile transformer. *arXiv preprint arXiv:2405.20330*, 2024.

[16] Gyeongsik Moon. Bringing inputs to shared domains for 3D interacting hands recovery in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17028–17037, 2023.

[17] Maxime Oquab, Timothée Darcet, Théo Moutakanni, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[18] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024.

[19] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12242–12254, 2025.

[20] Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *European Conference on Computer Vision*, pages 183–202, 2024.

[21] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[22] Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024.

[23] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2070–2080, 2024.

[24] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

[25] Eugene Valassakis and Guillermo Garcia-Hernando. Handdgp: Camera-space hand mesh prediction with differentiable global positioning. In *European Conference on Computer Vision*, pages 479–496, 2024.

[26] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pages 1–20, 2024.

[27] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024.

[28] Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. Tram: Global trajectory and motion of 3d humans from in-the-wild videos. In *European Conference on Computer Vision*, pages 467–487, 2024.

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

[30] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *Advances in Neural Information Processing Systems*, 38:35928–35959, 2026.

[31] 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*, pages 10481–10497, 2026.

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

[33] Yuxi Wang, Chengkai Jin, Yufei Liu, Wenqi Ouyang, Tianyi Wei, Zhiwei Zeng, Siyuan Huang, Zhiqi Shen, and Xingang Pan. The Surprising Effectiveness of Video Diffusion Models for Hand Motion Reconstruction. *arXiv preprint arXiv:2606.30308*, 2026.

[34] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21222–21232, 2023.

[35] Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7072–7084, 2025.

[36] 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 *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9009–9019, 2023.

[37] Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27716–27726, 2025.

[38] Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams. *arXiv preprint arXiv:2601.02281*, 2026.

[39] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1805–1815, 2025.

[40] Junyi Zhang, Charles Herrmann, Junhwa Hur, et al. Monst3r: A simple approach for estimating geometry in the presence of motion. In *International Conference on Learning Representations*, pages 82863–82886, 2025.