EGOWEAVE: STRUCTURED MOTION SUPERVISION FROM CALIBRATED EGOCENTRIC VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric demonstrations could provide scalable human-motion supervision for robot learning, but camera projection and motion reference frames make recordings difficult to reuse across devices. We present EgoWeave, which converts heterogeneous egocentric video into structured motion supervision. A hand-centered virtual camera and crop-consistent decoding let one frozen hand estimator process monocular and stereo pinhole or fisheye video. Timestamped reference poses then place the recovered head, wrists, and hand landmarks in a common coordinate system. We organize this motion as head movement, future-head-relative wrists, and normalized wrist-local articulation, and train with a compositional landmark objective before robot-native finetuning. In controlled fixed-WiLoR experiments with annotated hand regions, hand-centered viewing reduces camera-space MPJPE by 52.9% on Aria and 29.8% on Quest 3 in the 40–60° off-axis range. In the evaluated short-budget setting, the complete EgoWeave recipe improves task success over direct finetuning by 5.20–20.50 percentage points across four robot suites; under structured supervision, replacing official annotations with EgoWeave's reconstructed motion further improves all four suites. Code is available at https://anonymous.4open.science/r/EgoWeave/.

1 INTRODUCTION

Human first-person recordings capture manipulation experience without requiring a robot to execute each interaction (Chi et al., 2024; Wang et al., 2024a; Xu et al., 2025; Hoque et al., 2026). They are not directly reusable supervision: pinhole and fisheye cameras expose different local geometry, and camera-relative trajectories mix viewpoint motion with hand motion.

Using these recordings for robot learning requires consistent hand reconstruction across camera models and an explicit representation of viewpoint and hand motion. EgoWeave adapts calibrated images to a frozen hand estimator and organizes the recovered motion into structured pretraining targets. Perspective-cropping methods provide the local-view construction (Yu et al., 2021; Zhou et al., 2023; Facebook Research, 2024); we use the same calibrated geometry when sampling those views and recovering predictions. The exported records make head, wrist, and articulation frames explicit, so head motion remains separate from hand motion (Hoque et al., 2026; Qiu et al., 2025; Yu et al., 2026).

EGOWEAVE follows this path from image to motion record (Figure 1). A global detector selects a hand ray, a hand-centered view feeds frozen WiLoR (Potamias et al., 2025), and matching decoding restores physical-camera coordinates. Tracked poses then place the prediction in a common metric reference. The same record format supports monocular and stereo inputs and both live and offline processing.

The common-coordinate record is not yet the final learning target. We separate head motion, wrist placement, and hand shape so that each part has a clear reference frame, then combine them in a 153D target. A compositional landmark objective couples these factors during human pretraining, followed by robot-native finetuning (Black et al., 2025). Reconstruction crops generate labels while the policy retains original RGB. We evaluate observation geometry in fixed-estimator controls, then compare annotation sources and supervision recipes in downstream robot learning.

Our contributions are:

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.

- **A calibrated observation-to-record interface.** We adapt observations rather than retrain the hand regressor: calibrated pinhole and fisheye images are mapped to a hand-centered perspective representation, and the same virtual-camera geometry governs sampling and prediction recovery for a shared frozen estimator.
- **Structured motion supervision for robot pretraining.** We combine head motion, future-head-relative wrist poses, and normalized wrist-local articulation in a structured pretraining target, with compositional landmark supervision before robot-native finetuning.
- **Evaluation of reconstruction and robot transfer.** Fixed-WiLoR controls reveal the effect of local observation direction in off-axis conditions. Source–recipe comparisons then show how pipeline-generated labels and structured supervision transfer across the four evaluated robot suites.

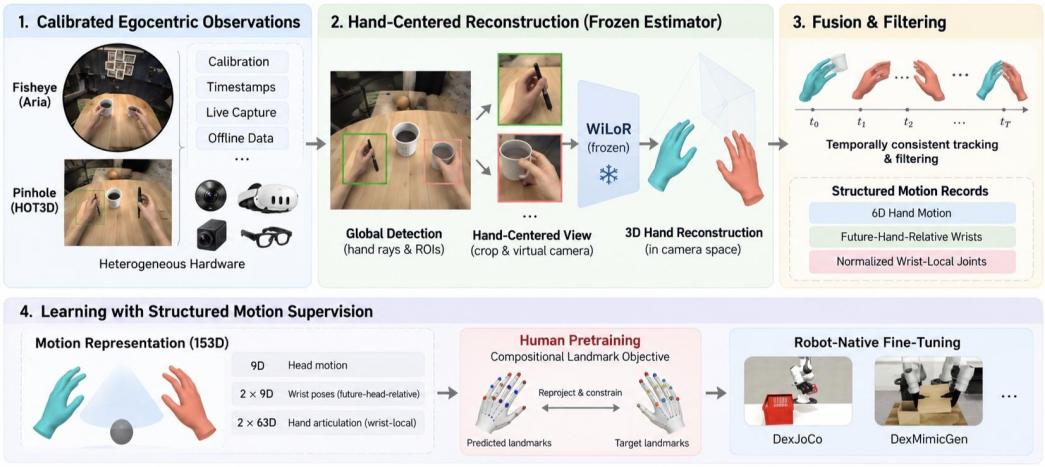


Figure 1: **EgoWeave overview.** Calibrated local views and matched decoding produce motion records for structured human pretraining and native robot adaptation. Official labels enter through the same record format. The policy receives original RGB, not reconstruction crops.

2 Related Work

Egocentric demonstrations and motion data. UMI, DexCap, and DexUMI collect manipulation demonstrations through handheld grippers, articulated-hand capture, and a hand exoskeleton, respectively (Chi et al., 2024; Wang et al., 2024a; Xu et al., 2025). These interfaces offer different combinations of portability and motion measurement. Public datasets provide complementary annotation sources: HOT3D supplies synchronized multi-view observations, reference poses, and calibrated crop utilities (Banerjee et al., 2025; Facebook Research, 2024); EgoDex pairs video with detailed skeletal annotations (Hoque et al., 2026); and EgoVerse integrates heterogeneous capture through shared processing (Punamiya et al., 2026). Such resources differ not only in image projection but also in reference points, landmark definitions, and availability. EgoWeave provides a common reconstruction and motion-record interface for these observations.

Local observation geometry and reconstruction. PCLs accounts for location-dependent perspective effects, JHands constructs hand-oriented virtual views, and the HOT3D toolkit maps calibrated source cameras to local perspective crops (Yu et al., 2021; Zhou et al., 2023; Facebook Research, 2024). Building on these cropping methods, EgoWeave pairs calibrated sampling with matching coordinate recovery to generate motion supervision for robot pretraining. At the estimator level, WiLoR combines hand localization and reconstruction (Potamias et al., 2025), HaWoR adds camera-motion recovery and motion completion (Zhang et al., 2025), FisheyeViT uses fisheye-aware features and body cues (Wang et al., 2024b), and EgoForce models hand–forearm geometry for metric camera-space recovery (Millerdurai et al., 2026).

Adaptation methods address a different part of the problem. V-HPOT predicts virtual-camera depth and performs self-supervised test-time optimization; S2DHand updates a pretrained single-view es-

timator using cross-view consensus and transformation constraints (Mucha et al., 2026; Liu et al., 2024). EgoWeave instead uses supplied calibration for deterministic transforms around fixed estimator weights. Consequently, our fixed-WiLoR controls isolate observation geometry, while complete-system comparisons retain each method's estimator, adaptation, and input requirements. They answer different questions and are reported separately.

Human-motion supervision and robot transfer. EgoMimic jointly trains on human and robot demonstrations (Kareer et al., 2025); LAPA learns latent actions before robot finetuning (Ye et al., 2025); EgoScale combines human pretraining with aligned human–robot mid-training (Zheng et al., 2026); and MimicDreamer aligns human and robot videos, viewpoints, and actions (Li et al., 2025). Action interfaces also differ: EgoVLA retains a shared wrist/MANO space with deployment-time retargeting, whereas H-RDT transfers parameters to robot-specific actions (Yang et al., 2025; Bi et al., 2026). Explicit targets include camera-frame wrists and fingertips in EgoDex (Hoque et al., 2026), joint head–wrist–fingertip motion in HAT (Qiu et al., 2025), and coordinated head–hand motion in EgoMI (Yu et al., 2026). Our recipe assigns wrists to the corresponding future head frame and articulation to a normalized wrist-local frame. A composed-landmark objective couples these predictions, while robot finetuning uses native actions. The 48D/153D experiments compare complete supervision recipes rather than the addition of coordinates alone.

3 Calibrated Egocentric Hand Reconstruction

Reusable motion supervision requires the observation geometry and coordinate recovery to agree. EgoWeave pairs hand-centered input sampling with crop-consistent decoding, then exports motion records with shared reference-frame conventions (Figure 2). Camera-specific processing uses deterministic transforms around one frozen estimator.

At time t, a calibrated central camera C supplies image $I_C(t)$ and projection Π_C, including its intrinsics and distortion parameters. The projection may be pinhole, radial–tangential, fisheye, or another calibrated central-camera model. Common-frame export additionally uses timestamped metric poses $T_{WC}(t)$ in reference frame W, where T_{AB} maps coordinates from B to A. Registered extrinsics relate cameras in a calibrated rig.

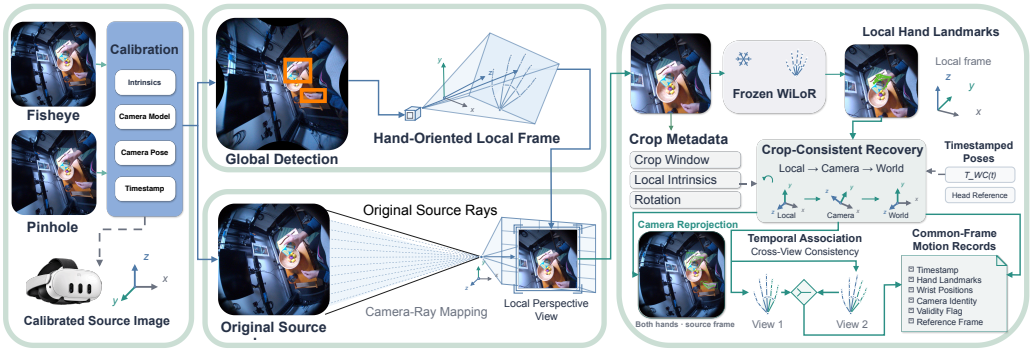


Figure 2: **Reconstruction and record export.** Global detection selects each local view; source pixels supply the crop. Matching crop metadata restores camera-space predictions, and tracked poses provide common-frame alignment. Monocular/stereo and live/offline execution share this per-view core.

3.1 Hand-Centered Observation Geometry

A broad view provides context for hand localization, while a hand-aligned perspective view places the reconstruction region near its own optical axis. We therefore separate global detection from local reconstruction. Both virtual cameras share the physical camera center, so their image mappings depend on viewing rays rather than scene depth.

A global pinhole view G with intrinsics K_G supplies the detector input. A detected box center $\mathbf{b}$ defines the source-camera ray

$$\mathbf{d} = \frac{R_{CG} K_G^{-1} \tilde{\mathbf{b}}}{\left\| R_{CG} K_G^{-1} \tilde{\mathbf{b}} \right\|_2}, \tag{1}$$

where R_{CG} rotates global-view axes into C and tildes denote homogeneous pixels; an axis-aligned global view uses $R_{CG} = I$. The detected ray $\mathbf{d}$ defines the optical axis of a local camera L that shares the physical center of C. This separates hand localization from hand reconstruction: the global view provides spatial context, whereas the local view places the detected hand near the center of the estimator's perspective domain.

For a local-camera pixel $\mathbf{q}_L$, the corresponding ray in the physical source-camera frame is

$$\mathbf{r}_C(\mathbf{q}_L) = R_{CL} \frac{K_L^{-1} \tilde{\mathbf{q}}_L}{\left\| K_L^{-1} \tilde{\mathbf{q}}_L \right\|_2}, \qquad I_L(\mathbf{q}_L) = \mathcal{S}(I_C, \Pi_C(\mathbf{r}_C(\mathbf{q}_L))), \tag{2}$$

where K_L is the local-camera intrinsic matrix, R_{CL} maps local axes to the source-camera frame, Π_C is the calibrated source projection, and $\mathcal{S}$ denotes bilinear sampling. The same ray-space construction therefore supports pinhole and fisheye sources; only Π_C changes with the camera model. Because the source pixels are sampled directly through Π_C, the local patch preserves the physical source rays while avoiding the need to materialize a full-frame intermediate image.

The geometric motivation follows from the differential scales of a radially symmetric projection $\rho = f(\theta)$, where θ is the angle from the camera optical axis:

$$s_r = |f'(\theta)|, \qquad s_t = \left| \frac{f(\theta)}{\sin\theta} \right|, \qquad \kappa(\theta) = \frac{\max(s_r, s_t)}{\min(s_r, s_t)}. \tag{3}$$

Here s_r and s_t measure radial and tangential differential scales on the viewing sphere, and κ measures local sampling anisotropy. For an equidistant fisheye, $f(\theta) = f_s\theta$ and $\kappa(\theta) = \theta/\sin\theta$. A local pinhole view has $\rho(\alpha) = f_p \tan\alpha$ and $\kappa(\alpha) = \sec\alpha$ around its own optical axis. Aligning that axis with the detected hand reduces the residual angle α, bringing the hand region into a more isotropic perspective parameterization. This predicts the strongest reconstruction gains for off-axis hands, which we evaluate under fixed WiLoR weights in Section 5.2. The local basis, crop metadata, and coordinate decoder are specified in the following subsection and in Appendix G.

3.2 Direct Source-to-Patch Reconstruction

Given a reference up vector $\mathbf{a}$, we construct the local-camera basis as

$$\mathbf{y} = \frac{\mathbf{a} - (\mathbf{a}^\top \mathbf{d})\mathbf{d}}{\|\mathbf{a} - (\mathbf{a}^\top \mathbf{d})\mathbf{d}\|_2}, \qquad \mathbf{x} = \mathbf{y} \times \mathbf{d}, \qquad R_{CL} = [\mathbf{x}\ \mathbf{y}\ \mathbf{d}]. \tag{4}$$

An orthogonal fallback is used when $\mathbf{a}$ is nearly parallel to $\mathbf{d}$. Reprojected detection-box corners determine the local hand extent. Let K_L denote local intrinsics and let $A \in \mathbb{R}^{3\times3}$ map final patch pixels to the local image plane, combining crop placement, resizing, and hand-side normalization. For patch pixel $\mathbf{q}$, the estimator input is

$$\mathbf{r}_L(\mathbf{q}) = \frac{K_L^{-1} A \tilde{\mathbf{q}}}{\left\| K_L^{-1} A \tilde{\mathbf{q}} \right\|_2}, \qquad I_L(\mathbf{q}) = \mathcal{S}(I_C, \Pi_C(R_{CL}\mathbf{r}_L(\mathbf{q}))). \tag{5}$$

where $\mathcal{S}$ is bilinear sampling through the calibrated source projection, with the finite-canvas validity mask and padding specified in Appendix G. Global detection selects the crop, but the crop draws its pixels directly from the original image I_C, not from the detector's rectified image.

Local orientation changes the viewing geometry, while direct sampling composes the projection and crop transforms into one resampling operation per patch. Global detection retains its own image preparation. Pinhole and fisheye sources use different Π_C but share the same estimator and decoding conventions.

3.3 Common-Frame Motion Records

For each detected hand, the frozen estimator E produces

$$\hat{m}_L = E(I_L). \tag{6}$$

The output adapter retains the crop extent, center, local focal length, and hand-side convention used to construct the input. It restores hand-side coordinates, converts the wrist basis to the record convention, and returns predictions to the physical camera:

$$\hat{m}_C = G_{L\to C}(\hat{m}_L). \tag{7}$$

The timestamped camera pose then places the geometric fields in the common frame:

$$\hat{m}_W(t) = T_{WC}(t)\, G_{L\to C}(\hat{m}_L(t)). \tag{8}$$

The rigid transform acts on geometry, not record metadata. Tracked poses establish the metric reference; hand depth and shape remain estimator predictions rather than quantities recovered from calibration alone.

Monocular estimates enter temporal tracking directly. Stereo candidates are time-associated and mapped to W before selection and fusion. Wrist-position and orientation consistency guide association; the registered synchronization and residual-translation handling are specified in Appendix G. The tracker distinguishes measured updates, short-gap predictions, and missing outputs. Offline formalization converts these states into field-level validity and supervision masks; complete predictions within confirmed tracks can remain eligible for training (Appendix C).

Each record contains a tracked head-reference pose, wrist poses (R_s, p_s), and 21 ordered landmarks per hand. For common-frame landmark $x_{s,j}$, wrist-local articulation is $\ell_{s,j} = R_s^\top (x_{s,j} - p_s)$. Exports specify reference points, axes, handedness, units, timestamps, calibration identity, and validity. Live capture and offline reannotation share this record schema, while scheduling and available temporal context differ. Section 4 constructs motion targets from the records; policies retain the original RGB observations.

4 Learning from Reconstructed Motion

Official and reconstructed records use the same target construction for $\pi_{0.5}$ human pretraining and native robot adaptation. The representation separates head movement, wrist placement, and hand shape so that each part has a clear reference frame before the components are combined.

4.1 Hierarchical Motion Targets

For future frame $f_k = t + k$, $k = 1, \ldots, 32$, let p_H, R_H denote the exported head-reference pose and p_s, R_s the wrist pose of hand $s \in \{L, R\}$. Head motion is $\Delta p_H = p_H(f_k) - p_H(t)$ and $\Delta R_H = R_H(t)^\top R_H(f_k)$. Future-head-relative wrists are $w_s = R_H(f_k)^\top [p_s(f_k) - p_H(f_k)]$ and $Q_s = R_H(f_k)^\top R_s(f_k)$. The 21 wrist-local landmarks are centered at joint zero and normalized by wrist-to-middle-MCP distance: $\bar{\ell}_{s,j} = (\ell_{s,j} - \ell_{s,0})/d_s$. With $\rho(R)$ concatenating the first two rotation columns (Zhou et al., 2019),

$$a_{t,k} = \left[\Delta p_H, \rho(\Delta R_H), w_L, \rho(Q_L), w_R, \rho(Q_R), \mathrm{vec}(\bar{\ell}_L), \mathrm{vec}(\bar{\ell}_R)\right] \in \mathbb{R}^{153}. \tag{9}$$

The $9 + 18 + 126$ channels encode head motion, wrist placement, and articulation. Head translation retains world axes. The current 153D state includes absolute head pose; three complete-recording progress features extend offline conditioning to 156D but are excluded from native robot states. Appendix A specifies the state, reference conventions, and normalization.

4.2 Compositional Supervision and Robot Adaptation

One action expert predicts the 32-frame chunk from recorded RGB, task text, and current state tokens (Lipman et al., 2023; Black et al., 2025). Training-set statistics standardize the channels. For target z, noise ϵ, and noise time τ, $x_\tau = (1 - \tau)z + \tau\epsilon$; masked flow matching predicts $\epsilon - z$.

Missing channels are zero-filled after standardization and excluded from the loss (Appendix A). Unstandardizing $\hat{z} = x_\tau - \tau v_\theta$ gives $\hat{a}$. After decoding rotations, compose

$$F_{s,j}(a; s_t) = p_H(t) + \Delta p_H + R_H(t)\Delta R_H \left[w_s + Q_s \, d^* \bar{\ell}_{s,j} \right]. \tag{10}$$

Predictions and targets use the same current pose and fixed reference hand length d^*. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda_{\text{geo}} \mathcal{L}_{\text{geo}}, \qquad \mathcal{L}_{\text{geo}} = \mathbb{E}_{\text{valid}} \left[\frac{1}{63} \sum_{j=0}^{20} \| F_{s,j}(\hat{a}; s_t) - F_{s,j}(a; s_t) \|_2^2 \right]. \tag{11}$$

Geometric supervision requires all components of a hand and averages valid hands, steps, examples, and noise draws; flow supervision uses available channels. Appendix A gives weights and reductions. This couples component errors through the resulting hand geometry without discarding channel-level supervision.

Human pretraining adapts $\pi_{0.5}$'s projections to 153D and updates all parameters. Robot finetuning transfers compatible backbone and action-expert parameters, adapts projections and normalization to the native interface, and updates all parameters on robot demonstrations (Appendix D). Deployed policies directly predict native actions; Section 5.4 evaluates the complete recipes.

5 Evaluation

We evaluate EgoWeave from calibrated hand reconstruction to robot learning. We first isolate hand-centered observation geometry with a fixed estimator, then evaluate reconstruction across camera setups, followed by human-to-robot transfer and a physical-robot feasibility check.

5.1 Experimental Setup

We use the H2O test set and the Aria and Quest 3 subsets of HOT3D (Kwon et al., 2021; Banerjee et al., 2025) for reconstruction evaluation. H2O contains 23,391 perspective-corrected RGB frames and 46,742 labeled hand frames. The HOT3D split contains 16 development and 48 held-out clips, with 5,109 Aria and 10,581 Quest 3 fixed-ROI targets.

All reconstructions use a common 21-joint convention. MPJPE and wrist error are computed over finite predictions; Coverage uses the full target list, and all-target PCK@50 counts missing predictions as failures. Geometry experiments use frozen WiLoR weights, whereas external methods retain their own estimators, detectors, and adaptation procedures.

5.2 Effect of Hand-Centered Viewing

We isolate viewing direction by fixing WiLoR weights, annotated ROIs, hand side, input resolution, coordinate recovery, and source sampling. The global view (G) follows the physical optical axis, whereas the hand-centered view (L) follows the hand-center ray. Both paths use one direct source resampling, while all other estimator, crop, and decoding settings remain fixed.

The largest gains occur off axis (Table 1). In the 40–60° bins, MPJPE decreases from 99.00 to 46.66 mm on Aria and from 137.14 to 96.28 mm on Quest 3, corresponding to reductions of 52.9% and 29.8%. Near the optical axis, the difference is smaller; on Quest 3 at 0–20°, MPJPE changes from 48.79 to 51.76 mm.

Hand-centered sampling also reduces local sampling anisotropy in 363 of 375 views (Table 2). In the 40–70° bin, the median singular-value ratio decreases from 1.6674 to 1.1031, with improvement in all 158 views.

The fixed-ROI experiment isolates viewing geometry. With detected inputs, the EgoWeave local view reduces MPJPE from 94.05 to 83.88 mm and raises all-target PCK50 from 34.41% to 48.53% relative to whole-image undistortion followed by cropping (Table 3).

5.3 Reconstruction Across Camera Setups

On H2O without ground-truth ROIs or depth, EgoWeave obtains the lowest camera-space and root-aligned MPJPE, at 23.80 and 11.95 mm, with 99.96% coverage (Table 4). EgoForce achieves lower

Table 1: **Effect of viewing direction.** All controlled factors except virtual-camera orientation are fixed. Relative reduction is $(G - L)/G$.

Device	Angle	Targets	Camera-space MPJPE (mm) ↓ Global (G)	Hand-centered (L)	Reduction (%) ↑
Aria	0–20°	672	42.86	**40.68**	5.1
Aria	20–40°	3,103	49.50	**46.70**	5.7
Aria	40–60°	1,334	99.00	**46.66**	52.9
Quest 3	0–20°	1,398	**48.79**	51.76	−6.1
Quest 3	20–40°	5,819	58.41	**47.55**	18.6
Quest 3	40–60°	3,364	137.14	**96.28**	29.8



Figure 3: **Qualitative comparison of global-axis and hand-centered reconstruction.** Panels (a1, a2) show global-axis predictions (G, red), while (b1, b2) show hand-centered predictions (L, green). Yellow skeletons denote ground truth. Each pair (a_i, b_i) corresponds to the same Aria frame; indices $i = 1, 2$ denote two matched examples. All predictions use the same frozen WiLoR weights.

Table 2: **Sampling anisotropy at the hand-center ray.** Lower values indicate more isotropic sampling.

Angle	Views	Median anisotropy ↓ Global	Hand-centered	Improved (%) ↑
0–20°	27	1.0405	1.0095	81.5
20–40°	190	1.1704	1.0456	96.3
40–70°	158	1.6674	1.1031	100.0
All	375	–	–	96.8

Table 3: **Reconstruction with detected inputs.** MPJPE is computed over finite predictions, while Coverage and PCK@50 use the full target list and count missing outputs as failures.

Observation path	MPJPE (mm) ↓	Coverage (%) ↑	PCK@50 (%) ↑
Raw distorted image + crop	116.18	80.65	29.62
Whole-image undistortion + crop	94.05	75.65	34.41
Hand-centered local crop (EgoWeave)	**83.88**	75.64	**48.53**

wrist error and higher all-target PCK@50, while V-HPOT online TTO reduces its camera-space MPJPE from 154.72 to 62.66 mm.

Table 4: **Single-view reconstruction on H2O.** Coverage and PCK@50 use all 46,742 labeled hand frames.

Method	Camera MPJPE ↓	Wrist error ↓	Root-aligned MPJPE ↓	Coverage (%) ↑	PCK@50 (%) ↑
EgoForce (Millerdurai et al., 2026)	25.30	**21.74**	15.42	99.77	**92.83**
V-HPOT no-TTO (Mucha et al., 2026)	154.72	106.90	66.22	98.61	3.47
V-HPOT online TTO	62.66	60.95	50.69	99.82	45.94
EgoWeave (ours)	**23.80**	24.76	**11.95**	**99.96**	91.60

On Quest 3, we compare hand-shape reconstruction with S2DHand using ground-truth ROIs (Table 5). S2DHand uses calibrated stereo crops and either its source checkpoint or a checkpoint adapted for 10 epochs; EgoWeave is evaluated on each corresponding valid-target subset.

Table 5: **Hand-shape reconstruction on Quest 3 with GT ROIs.** Coverage and valid-target counts refer to S2DHand (Liu et al., 2024).

S2DHand configuration and valid outputs				Shape error (mm) ↓	
Checkpoint	Output	Coverage (%)	Valid targets	S2DHand	EgoWeave (single-view)
Source	Mono	95.01	10,053	54.86	**12.73**
Source	Dual	93.27	9,869	56.09	**12.75**
Adapted	Mono	95.01	10,053	52.15	**12.73**
Adapted	Dual	93.27	9,869	52.41	**12.75**

Shape error centers each 21-landmark output at landmark zero and applies the same wrist-to-middle-MCP scale normalization before averaging Euclidean landmark errors; values are reported in millimetres on the corresponding valid-target subsets.

S2DHand adaptation reduces shape error to 52.15/52.41 mm for mono/dual outputs; EgoWeave obtains 12.73/12.75 mm on the same valid subsets without camera-specific regressor adaptation.

5.4 Human-to-Robot Transfer

We evaluate transfer on DexJoCo(Wang et al., 2026), DexMimicGen(Jiang et al., 2025), and RoboTwin 2.0 (Chen et al., 2026). DexJoCo uses four tasks with 50 episodes per task. DexMimicGen A and B denote two robot embodiments, each evaluated on its corresponding fixed task suite with 50 episodes per task or condition. RoboTwin 2.0 uses five tasks with 50 fixed seeds per task.

Direct finetuning starts from the original $\pi_{0.5}$ checkpoint; all other methods first pretrain on human motion. Within each benchmark suite, robot-native finetuning and evaluation settings are shared.

To compare annotation sources, we apply the EgoWeave reconstruction pipeline to the same EgoDex RGB recordings used by the official labels. We refer to the resulting motion labels as *EgoWeave-reconstructed annotations*. Figure 4 shows matched examples from the two annotation sources.

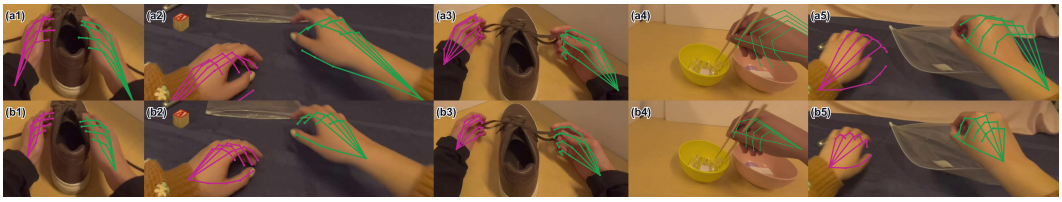


Figure 4: **Qualitative comparison between official EgoDex and EgoWeave-reconstructed annotations.** Panels (a1–a5) show the official EgoDex annotations, while (b1–b5) show the corresponding annotations reconstructed by EgoWeave. Each pair (a_i, b_i) uses the same RGB frame. Magenta and green skeletons denote the left and right hands, respectively.

The paired examples illustrate differences in landmark placement and hand geometry; their downstream effect is evaluated below.

The comparison crosses two factors: annotation source and supervision formulation. *EgoDex-style pretraining* uses official EgoDex annotations with wrist–fingertip supervision. *EgoWeave w/ official labels* uses official annotations with the structured target. *EgoWeave w/ wrist–fingertip supervision* uses reconstructed annotations with the EgoDex-style target. Full *EgoWeave* combines EgoWeave-reconstructed annotations with structured head–wrist–hand supervision.

EgoWeave exceeds direct finetuning by 5.20–20.50 percentage points across the four suites. With structured supervision fixed, reconstructed annotations improve over official annotations by 4.33–8.00 points. With reconstructed annotations fixed, structured supervision improves over wrist–

Table 6: **Human-to-robot transfer.** Success rate (%) after robot-native finetuning. DMG-A and DMG-B denote two DexMimicGen robot embodiments.

Method	DexJoCo	DMG A	DMG B	RoboTwin 2.0
Direct finetuning	42.00	27.33	65.00	40.40
EgoDex-style pretraining	43.50	30.67	66.00	32.80
EgoWeave w/ official labels	54.50	28.67	75.00	39.60
EgoWeave w/ wrist–fingertip sup.	55.00	29.33	66.00	44.40
EgoWeave (ours)	**62.50**	**35.33**	**79.33**	**45.60**

fingertip supervision by 1.20–13.33 points. With official annotations, the structured target improves three suites but decreases DexMimicGen A from 30.67% to 28.67%.

5.5 Real-World Experiments

We finally evaluate the transfer performance on a Franka robot with a gripper using the *Bread transfer* task, where the robot must move bread from a toaster to a cutting board. Each method is evaluated for 40 trials under the same success criterion.

Table 7: **Real-world Bread-transfer results on Franka.** Each method is evaluated over 40 trials. Bold marks the best result.

Method	Successes / trials	Success (%) ↑
Direct finetuning	8/40	20.0
EgoDex-style pretraining	15/40	37.5
EgoWeave w/ official labels	8/40	20.0
EgoWeave w/ wrist–fingertip sup.	10/40	25.0
EgoWeave (ours)	**21/40**	**52.5**

EgoWeave achieves the highest success rate of 52.5%, improving over direct finetuning by 32.5 percentage points (Table 7). Human-motion pretraining alone also improves performance, with EgoDex-style pretraining reaching 37.5%. The physical-robot result follows the direction of the simulation transfer results while using a separate task and success predicate. It therefore checks whether the reconstructed-motion interface remains useful when the policy must execute native actions under real sensing and control latency, rather than only matching a simulated trajectory.

The result separates the roles of the two pretraining ingredients: human data provide broad manipulation variation, while the structured record makes that variation consistent across cameras and maps it to a robot action space. The additional gain from the complete recipe shows that annotation quality and target factorization still matter after the backbone has learned from human motion.

6 Discussion and Conclusion

EgoWeave connects calibrated observations through a unified reconstruction interface to structured motion supervision. Controlled reconstruction studies, four-suite simulation transfer, and the Franka Bread-transfer evaluation support the full pipeline and its downstream utility. These conclusions are scoped to the evaluated datasets, budgets, and task protocols.

Future work. Future work will investigate calibration uncertainty, cross-device generalization, and larger-scale deployment. We plan to evaluate how intrinsic and extrinsic calibration errors affect hand-centered sampling, coordinate recovery, and the resulting motion records, and to develop confidence estimates for identifying unreliable frames. We will also extend the evaluation to additional camera layouts, robot embodiments, longer-horizon tasks, and more diverse bimanual interactions. Finally, we will study lightweight online adaptation and richer supervision for contacts and object relations while preserving the interpretable head–wrist–articulation factorization. These experiments will clarify when deterministic geometric processing is sufficient and when learned adaptation is needed for robust real-world transfer.

AI USE STATEMENT

AI tools assisted with literature organization, experiment design discussion, manuscript drafting and revision, LaTeX preparation, code inspection, and diagram preparation. The authors remain responsible for the data provenance, implementation, analyses, reported numbers, references, and final manuscript. Method and interface diagrams are schematic; quantitative claims refer to the reported measurements.

REFERENCES

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*, pp. 7061–7071, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Banerjee_HOT3D_Hand_and_Object_Tracking_in_3D_from_Egocentric_Multi-View_CVPR_2025_paper.html.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-RDT: Human manipulation enhanced bimanual robotic manipulation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(22):18135–18143, March 2026. doi: 10.1609/aaai.v40i22.38875. URL https://doi.org/10.1609/aaai.v40i22.38875.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, brian ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/black25a.html.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62192.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal Manipulation Interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. URL https://www.roboticsproceedings.org/rss20/p045.html.

Facebook Research. Hand tracking toolkit: Camera warping and hand-crop utilities. Official source repository, 2024. URL https://github.com/facebookresearch/hand_tracking_toolkit/blob/55e97a82cd6d510030dac24b71c7206e979caf28/hand_tracking_toolkit/dataset.py. Initial public revision, June 14, 2024; accessed September 16, 2026.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, pp. 4218–4237, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/07fcc6e2b89439d3ee5ab60939aaa6a0-Abstract-Conference.html.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Jim Fan, and Yuke Zhu. DexMimicGen: Automated data generation for bimanual dexterous manipulation via

imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16923–16930. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127809. URL https://doi.org/10.1109/ICRA55743.2025.11127809.

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)*, pp. 13226–13233. IEEE, 2025. doi: 10.1109/ICRA55743.2025.11127989. URL https://doi.org/10.1109/ICRA55743.2025.11127989.

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*, pp. 10138–10148, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Kwon_H2O_Two_Hands_Manipulating_Objects_for_First_Person_Interaction_Recognition_ICCV_2021_paper.html.

Haoyun Li, Ivan Zhang, Runqi Ouyang, Xiaofeng Wang, Zheng Zhu, Zhiqin Yang, Zhentao Zhang, Boyuan Wang, Chaojun Ni, Wenkang Qin, Xinze Chen, Yun Ye, Guan Huang, Zhenbo Song, and Xingang Wang. MimicDreamer: Aligning human and robot demonstrations for scalable VLA training, 2025. URL https://arxiv.org/abs/2509.22199.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, and Yoichi Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 677–686, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Liu_Single-to-Dual-View_Adaptation_for_Egocentric_3D_Hand_Pose_Estimation_CVPR_2024_paper.pdf.

Christen Millerdurai, Shaoxiang Wang, Yaxu Xie, Vladislav Golyanik, Didier Stricker, and Alain Pagani. EgoForce: Forearm-guided camera-space 3D hand pose from a monocular egocentric camera. In *ACM SIGGRAPH 2026 Conference Papers*, pp. 1–12. ACM, 2026. doi: 10.1145/3799902.3811047. URL https://doi.org/10.1145/3799902.3811047.

Wiktor Mucha, Michael Wray, and Martin Kampel. Towards egocentric 3d hand pose estimation in unseen domains. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026. URL https://openaccess.thecvf.com/content/WACV2026/papers/Mucha_Towards_Egocentric_3D_Hand_Pose_Estimation_in_Unseen_Domains_WACV_2026_paper.pdf.

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*, pp. 12242–12254, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Potamias_WiLoR_End-to-end_3D_Hand_Localization_and_Reconstruction_in-the-wild_CVPR_2025_paper.html.

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.

Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J. Yoon, Ryan Hoque, Lars Paulsen, Ge Yang, Jian Zhang, Sha Yi, Guanya Shi, and Xiaolong Wang. Humanoid policy $\sim$ human policy. In *Proceedings of The 9th Conference on Robot Learning*,

volume 305 of *Proceedings of Machine Learning Research*, pp. 2888–2906. PMLR, 2025. URL https://proceedings.mlr.press/v305/qiu25a.html.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Proceedings of Robotics: Science and Systems*, 2024a. URL https://www.roboticsproceedings.org/rss20/p043.html.

Hanwen Wang, Weizhi Zhao, Xiangyu Wang, Siyuan Huang, He Lin, Boyuan Zheng, Rongtao Xu, Gang Wang, Yao Mu, He Wang, Lue Fan, Hongsheng Li, Zhaoxiang Zhang, and Tieniu Tan. Dexjoco: A benchmark and toolkit for task-oriented dexterous manipulation on mujoco, 2026. URL https://arxiv.org/abs/2605.16257.

Jian Wang, Zhe Cao, Diogo Luvizon, Lingjie Liu, Kripasindhu Sarkar, Danhang Tang, Thabo Beeler, and Christian Theobalt. Egocentric whole-body motion capture with FisheyeViT and diffusion-based motion refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 777–787, 2024b. doi: 10.1109/CVPR52733.2024.00080. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wang_Egocentric_Whole-Body_Motion_Capture_with_FisheyeViT_and_Diffusion-Based_Motion_Refinement_CVPR_2024_paper.html.

Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 437–459. PMLR, 2025. URL https://proceedings.mlr.press/v305/xu25b.html.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, Hongxu Yin, Sifei Liu, Song Han, Yao Lu, and Xiaolong Wang. EgoVLA: Learning vision-language-action models from egocentric human videos, 2025. URL https://arxiv.org/abs/2507.12440v3.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 28213–28239, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/45d74e190008c7bff2845ffc8e3facd3-Paper-Conference.pdf.

Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. PCLs: Geometry-aware neural reconstruction of 3d pose with perspective crop layers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9064–9073, 2021. doi: 10.1109/CVPR46437.2021.00895. URL https://openaccess.thecvf.com/content/CVPR2021/html/Yu_PCLs_Geometry-Aware_Neural_Reconstruction_of_3D_Pose_With_Perspective_Crop_CVPR_2021_paper.html.

Justin Yu, Yide Shentu, Di Wu, Pieter Abbeel, Ken Goldberg, and Philipp Wu. EgoMI: Learning active vision and whole-body manipulation from egocentric human demonstrations, 2026. URL https://arxiv.org/abs/2511.00153v2.

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*, pp. 1805–1815, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_HaWoR_World-Space_Hand_Motion_Reconstruction_from_Egocentric_Videos_CVPR_2025_paper.html.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, Trevor Darrell, Furong Huang, Yuke Zhu, Danfei Xu, and Linxi Fan. Egoscale: Scaling dexterous manipulation with diverse egocentric human data, 2026. URL https://arxiv.org/abs/2602.16710.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5745–5753, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html.

Zhishan Zhou, Zhi Lv, Shihao Zhou, Minqiang Zou, Tong Wu, Mochen Yu, Yao Tang, and Jiajun Liang. 1st place solution of egocentric 3d hand pose estimation challenge 2023 technical report: A concise pipeline for egocentric hand pose reconstruction, 2023. URL https://arxiv.org/abs/2310.04769.

A Representation Contracts

Head, wrist, and local-articulation definitions. Let $p_H(t), R_H(t)$ denote the world position and orientation of the head-motion reference at observation time t. Each export records its reference point and axes together with the recording trajectory. The equations below apply to the reference selected by each export; matched annotation-source comparisons use the same registered reference convention. The transfer configurations in Table 6 use these registered reference conventions. Hand $s \in \{L, R\}$ has wrist pose p_s, R_s and wrist-local landmarks $\ell_{s,j}$, $j = 0, \ldots, 20$, after producer-to-training conversion. For future chunk frame $f_k = t + k$, $k = 1, \ldots, 32$,

$$\Delta p_H(t, f_k) = p_H(f_k) - p_H(t), \qquad \Delta R_H(t, f_k) = R_H(t)^\top R_H(f_k). \tag{12}$$

Head translation remains in world axes, with the current head pose supplied in the geometric state. Wrists describe hand placement relative to the head at the corresponding future frame:

$$w_s(f_k) = R_H(f_k)^\top [p_s(f_k) - p_H(f_k)], \qquad Q_s(f_k) = R_H(f_k)^\top R_s(f_k). \tag{13}$$

Each hand and frame uses its own wrist-to-middle-metacarpophalangeal distance to normalize local articulation:

$$d_s(f_k) = \|\ell_{s,9}(f_k) - \ell_{s,0}(f_k)\|_2, \qquad \bar{\ell}_{s,j}(f_k) = \frac{\ell_{s,j}(f_k) - \ell_{s,0}(f_k)}{d_s(f_k)}. \tag{14}$$

Head and wrist translations retain metre units; local articulation is dimensionless. The encoded target composition places the recentered landmark zero at the recorded wrist origin. It preserves relative articulation at the reference hand scale, while the reconstruction record retains any estimator-predicted offset between that origin and landmark zero.

153D serialization. Table 8 fixes the ordering of Equation 9. The six coordinates corresponding to $\bar{\ell}_{s,0}$ are structurally zero because landmark zero is the local origin. These coordinates remain in the target and receive the local-articulation flow weight of 0.3 when their hand labels are valid. The raw record retains metric landmarks and provenance, while the normalized local target retains articulation proportions.

Table 8: Zero-based, half-open 153D action slices. Rotation channels concatenate matrix columns.

Slice	Quantity	Frame or anchor
$[0, 3)$	Head translation displacement	World axes; current anchor
$[3, 9)$	Head relative rotation, 6D	Future head relative to current
$[9, 12)$	Left wrist position	Future head
$[12, 18)$	Left wrist rotation, 6D	Future head
$[18, 21)$	Right wrist position	Future head
$[21, 27)$	Right wrist rotation, 6D	Future head
$[27, 90)$	Left normalized local21	Wrist-local
$[90, 153)$	Right normalized local21	Wrist-local

Geometry and numerical conventions. The encoder centers producer-supplied wrist-local landmarks at landmark zero. When the wrist-to-middle-MCP distance is at most 10^{-6}, its scale fallback is the mean centered landmark distance exceeding 10^{-6}, or the numerical value 1 if none exists. Label-validity flags determine supervision availability separately from this numerical fallback. For 6D rotation columns a_1, a_2, the decoder D computes

$$\begin{aligned} b_1 &= \frac{a_1}{\max(\|a_1\|_2, 10^{-6})}, & u_2 &= a_2 - (a_2^\top b_1) b_1, \\ b_2 &= \frac{u_2}{\max(\|u_2\|_2, 10^{-6})}, & D(a_1, a_2) &= [b_1 \;\; b_2 \;\; b_1 \times b_2]. \end{aligned} \tag{15}$$

These denominator safeguards define the numerical handling of small or collinear columns. Equation 18 composes normalized articulation at a fixed reference scale; the raw record preserves the original metric landmarks.

48D reference convention. Let C_t, c_t denote current camera-to-world rotation and position, p_s, R_s the wrist pose, and $x_{s,j}$ world landmarks. The wrist–fingertip reference uses the current-camera wrist pose and five fingertips as state. Each hand contributes the following 24D future action:

$$\left[C_t^\top (p_s(f) - p_s(t)), \rho(C_t^\top R_s(f)), \{C_t^\top (x_{s,j}(f) - x_{s,j}(t))\}_{j\in\{4,8,12,16,20\}}\right]. \tag{16}$$

Positions are corresponding-point displacements; wrist rotation is a future orientation in the current camera frame. Fingertip motion includes wrist translation. This formula defines the 48D wrist–fingertip export used in the reference comparison, while the official annotations contain additional fields. Structured motion supervision additionally changes head supervision, intermediate joints, scale normalization, and geometric loss.

State, progress, and statistics. The current 153D geometric state uses absolute world head position and 6D orientation, current head-relative wrists, and current normalized local landmarks. Appending $[r_t, \sin(2\pi r_t), \cos(2\pi r_t)]$, with $r_t = t/(T-1)$ for an episode of T frames, gives the 156D conditioning state for offline human-pretraining windows. These progress features are not used by the deployed reconstruction path or by the native robot state. The future action remains 153D and contains frames $t+1$ through $t+32$.

Human states and actions use channel-wise standardization $(x-\mu)/(\sigma+10^{-6})$. In the partial-annotation export, statistics are fitted over valid training values, and standard deviations below 10^{-6} are replaced by 1. Missing channels are set to zero after standardization. The standardized state is discretized into text tokens for policy conditioning; its continuous values are also retained for geometric supervision and recovered by the inverse standardization. Text identifies unavailable tracking for the current hands. Future availability supplies target preprocessing and loss masks. Robot fine-tuning uses its own state schema and normalization.

Flow-matching objective. For standardized target actions z with missing channels filled by zero, Gaussian noise ϵ, and noise time τ, structured motion supervision uses $x_\tau = (1-\tau)z + \tau\epsilon$ and target velocity $\epsilon - z$. One noise time is sampled per example and shared across its chunk, with $\tau = 0.001 + 0.999\, b$, $b \sim \mathrm{Beta}(1.5, 1)$. Let $m_{k,c} \in \{0,1\}$ denote target availability, $A_k = \sum_c m_{k,c}\alpha_c$, and $e_{k,c} = v_\theta(x_\tau, o_t, \tau)_{k,c} - (\epsilon - z)_{k,c}$. Then

$$\ell_{\mathrm{FM},k} = \begin{cases} \dfrac{\sum_c m_{k,c}\alpha_c e_{k,c}^2}{A_k}, & A_k > 0, \\ 0, & A_k = 0, \end{cases} \qquad \mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\frac{\sum_k \ell_{\mathrm{FM},k}}{\max(1, \sum_k \mathbf{1}[A_k > 0])}\right]. \tag{17}$$

The expectation averages over examples and noise draws. The per-channel weights for head translation, head rotation, wrist position, wrist rotation, and local articulation are respectively 3, 1.5, 20, 4, and 0.3. For partial annotations, head-motion channels require current and future reference validity, wrist channels require the corresponding future head and hand, and local channels require the corresponding future hand. Complete labels set all availability masks to one.

Compositional geometric supervision. The clean-action estimate $\hat{z} = x_\tau - \tau v_\theta(x_\tau, o_t, \tau)$ is unnormalized before composition. The continuous current state is likewise unnormalized. Predicted reference-scale landmarks are

$$\begin{aligned} &\hat{p}_H = p_H(t) + \widehat{\Delta p}_H, && \hat{R}_H = R_H(t) D(\widehat{\rho(\Delta R_H)}), \\ &\hat{\tilde{x}}_{s,j} = \hat{p}_H + \hat{R}_H \left[\hat{w}_s + D(\widehat{\rho(Q_s)}) d^* \hat{\tilde{\ell}}_{s,j}\right]. \end{aligned} \tag{18}$$

Target landmarks $\tilde{x}_{s,j}$ use their own target head, wrist, and local-articulation components in the same composition. Both hands use the fixed configuration value $d^* = 0.09564778953790665\,\mathrm{m}$ for predictions and targets.

Let $g_{k,s}$ be one when all head-motion, wrist-pose, and local-landmark target channels needed for hand s at time k are valid. Define $B_{k,s} = \frac{1}{63}\sum_{j=0}^{20} \left\| \hat{\tilde{x}}_{k,s,j} - \tilde{x}_{k,s,j} \right\|_2^2$. The equal-hand-weight geometric objective is

$$\ell_{\mathrm{geo},k} = \frac{\sum_s g_{k,s} B_{k,s}}{\max(1, \sum_s g_{k,s})}, \qquad \mathcal{L}_{\mathrm{geo}} = \mathbb{E}\left[\frac{\sum_k \ell_{\mathrm{geo},k}}{\max(1, \sum_k \mathbf{1}[\sum_s g_{k,s} > 0])}\right]. \tag{19}$$

The total objective is $\mathcal{L}_{153} = \mathcal{L}_{\mathrm{FM}} + 100\,\mathcal{L}_{\mathrm{geo}}$. Each valid time step first averages its valid hands, each example averages its valid time steps, and the expectation averages examples and noise draws. The geometric term propagates landmark error through the predicted head, wrist, and local-articulation components at the common reference scale.

B Cross-Source Reconstruction Study

The reconstruction study separates projection geometry, estimator accuracy, and detected-input reconstruction quality. The fixed-region-of-interest (ROI) geometry study operates on annotated regions and fixed 256-pixel patches with the same WiLoR weights, hand side, coordinate recovery, and output convention. The detected-input study uses the shared detector and compares complete observation paths. All target lists are frozen before evaluation; conditional mean per-joint position error (MPJPE) uses finite outputs, whereas Coverage and the percentage of correct keypoints within 50 mm (PCK@50) retain every target in the denominator.

Fixed-ROI geometry controls. The three paths in Table 9 distinguish global-axis viewing, global remapping, and direct hand-centered viewing. B first creates a global perspective image and then crops a hand box. G samples a global-axis patch directly from the source image. L samples a hand-centered local camera directly from the source image.

Table 9: Fixed-ROI geometry controls on 5,109 Aria and 10,581 Quest 3 targets. All rows share WiLoR weights, annotated ROI, hand side, and coordinate recovery. Errors are in mm; all targets produce finite predictions in this controlled study.

Input path	Aria MPJPE	Quest 3 MPJPE	Aria aligned	Quest 3 aligned	Aria PCK@50	Quest 3 PCK@50
B: global remap + crop	61.34	81.02	36.01	53.34	47.95%	37.84%
G: global-axis direct	61.55	82.17	35.99	53.69	47.47%	37.64%
L: local direct	45.89	63.60	23.62	18.78	63.99%	57.55%

Relative to G, L lowers MPJPE by 25.4% on Aria and 22.6% on Quest 3. These controls isolate the contribution of hand-centered viewing under a shared direct source-sampling and decoding pipeline.

The complete qualitative global/local comparison is shown in Figure 5. Each pair uses the same source frame and hand region; the left panel is the global-axis view and the right panel is the hand-centered local view. The quantitative comparisons are reported in Table 9 and Table 2.

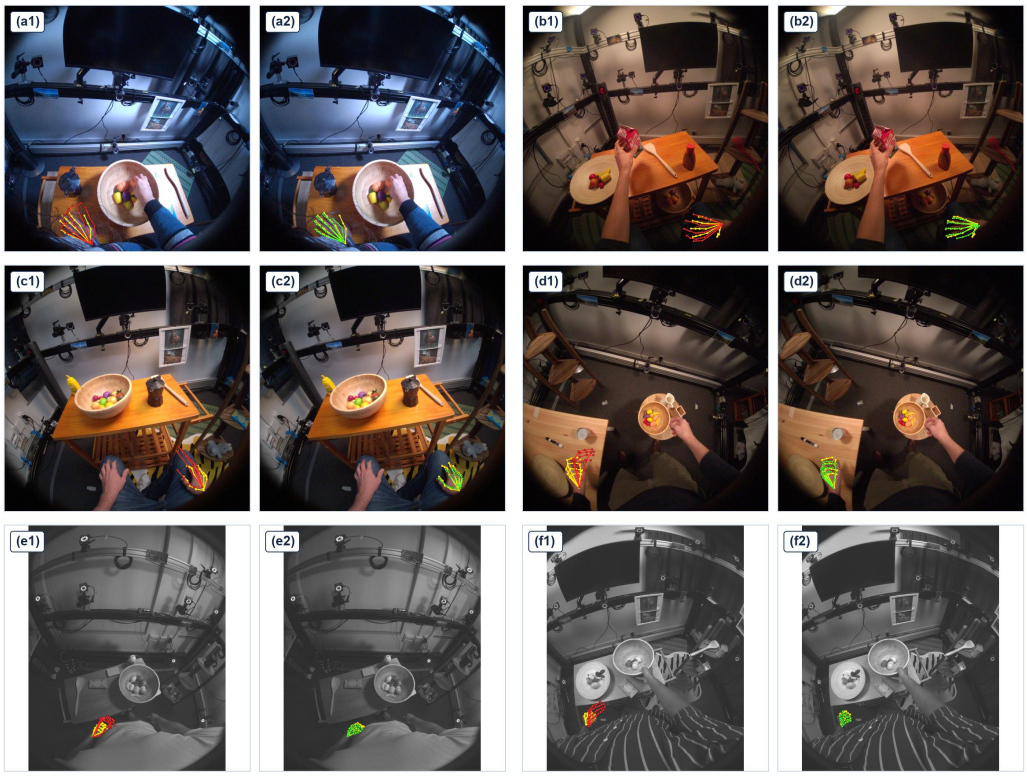


Figure 5: **Global-axis and hand-centered local views.** Each pair uses the same source frame and hand region. The left panel shows the global-axis view and the right panel the hand-centered local view, with the corresponding prediction overlays.

C Training Availability from Reconstruction Status

Formal hand labels. The producer audit records a hand as measured when a fused state has contributing views, predicted when it has a state without a new observation, and missing when no state is available. The offline formalizer requires a complete finite set of 21 landmarks and a valid wrist rotation. Within each tracking epoch, it marks complete outputs from the configured confirmation point onward as valid, including subsequent short-gap predictions. The inspected export uses confirmation length one; producer-side cross-side cleanup precedes formalization.

Canonical landmark and wrist streams expose `valid`/`fail` status, a timestamp, and either a finite payload or a failure reason with a null payload. Measured/predicted identity remains in audit metadata. The training reader verifies agreement of wrist and landmark timestamps and status, then derives per-hand availability from `status=valid`. Consequently, a complete prediction retained within a confirmed track can supply the same channel supervision as a measured update. Unconfirmed, missing, or cleared outputs supply no hand supervision. Label availability therefore measures usable exported coverage, while measured-update coverage must be counted separately.

Window admission and channel masks. Let C_t denote current camera validity, H_u head-reference validity at frame u, and $V_{s,u}$ formal hand validity after the producer's export and cleanup. In the partial-label training adapter, anchors have stride two, require $C_t \wedge H_t$, and are admitted when at least one hand label is valid in the next 32 frames:

$$C_t \wedge H_t \wedge \left(\bigvee_{k=1}^{32} \bigvee_{s \in \{L,R\}} V_{s,t+k} \right). \tag{20}$$

Current hand availability and complete bimanual coverage are not prerequisites. Table 10 specifies the current-state and future-action masks; the three progress features have valid current-state entries

and no action entries. Missing channels are zero-filled after normalization. A missing current hand can still have valid future targets.

Table 10: Channel availability for the structured motion target with partial annotations, for future frame $f = t + k$. Each condition applies to all channels of the indicated component.

Component	Current-state mask	Future-action mask
Head pose / motion (9D)	H_t	$H_t \wedge H_f$
Hand s wrist pose (9D)	$H_t \wedge V_{s,t}$	$H_f \wedge V_{s,f}$
Hand s local landmarks (63D)	$V_{s,t}$	$V_{s,f}$

The flow objective uses these per-channel masks with the fixed weights in Appendix A. A hand contributes to geometric supervision only when its head-motion, wrist-pose, and local-landmark channels are all valid at that future step. The losses do not directly consume the producer's measured/predicted identity or audit-quality metadata. Head-reference construction and cleanup rules are versioned with each export, while the channel-mask logic operates on the resulting reference and hand availability.

D Training and Evaluation Registry

The five main configurations are Direct finetuning, EgoDex-style pretraining, EgoWeave w/ official labels, EgoWeave w/ wrist–fingertip supervision, and EgoWeave (ours). Direct finetuning starts from the original $\pi_{0.5}$ checkpoint and proceeds directly to native robot finetuning. EgoDex-style pretraining uses official annotations with wrist-and-fingertip supervision. The two structured supervision configurations use the structured head–wrist–hand target, with official annotations for the first and EgoWeave-reconstructed annotations for the second. Human pretraining uses human batches; robot batches enter during native finetuning.

Within each suite, the evaluation implementation, task partition, action sampling, replanning horizon, controller rate, retry rule, and episode denominator are registered with the checkpoint. The transfer table compares complete recipes: label provenance, target contents, normalization, and auxiliary geometric supervision may differ between rows.

Human pretraining. The structured target contains 9D viewpoint motion, 18D head-relative wrist pose, and 126D normalized local articulation, with a 156D conditioning state and 32-frame future chunks. The policy uses one recorded RGB observation and task/state tokens, a single action expert, and full-parameter optimization from the $\pi_{0.5}$ base checkpoint. Initialization copies shape-compatible parameters and initializes the dimension-mismatched action input kernel, output kernel, and output bias for the 153D interface. The action input bias is retained. The 48D reference follows Appendix A; its target layout and training configuration are recorded separately.

Native robot parameter transfer. The native-policy suites transfer the vision-language backbone and internal action-expert parameters, use robot-specific state/action schemas and normalization statistics, and update all model parameters. Their objectives supervise native robot actions; the head–wrist–local21 geometric objective belongs to the human-pretraining stage.

DexJoCo retains parameter leaves with matching names and shapes from the human-pretrained checkpoint, including the action input bias. Dimension-mismatched action input/output projection leaves use the target model's initialization. The single-arm interface has action/state widths $22/23$. The retained bimanual capture/replay interface has widths $44/46$; these interface dimensions are registry metadata and are not an additional column in the four-suite transfer snapshot.

DexMimicGen transfers parameters outside the action projections from the human-pretrained checkpoint. Its action input/output projection kernels and biases are initialized from the original $\pi_{0.5}$ base checkpoint, cropping the action-axis dimensions to the native width. DexMimicGen A uses action/state widths $24/38$, and DexMimicGen B uses $30/36$. The robot recipes use their own percentile-normalization statistics, distinct from the human-stage mean/std normalization.

Evaluation provenance. The evaluation registry defines fields for rollout sampling uncertainty, independently trained checkpoint variation, checkpoint identities, source manifests, normalization

files, parameter transfer/initialization reports, training-window counts, optimizer updates, updated-parameter counts, and per-task outcomes. The current draft reports aggregate snapshot values; paired rollout intervals and independent-checkpoint variation are not reported here. The registry identifies the fields needed to report these quantities for each contrast.

E Real-Robot Experimental Setup

This appendix specifies the physical Bread-transfer protocol used for the Franka evaluation in Section 5.5. The task instruction was "Take the bread out of the toaster and place it on the cutting board." All evaluated checkpoints used the same nominal workspace arrangement, reset procedure, observation interface, controller, and success criterion. Each method contributed 40 total attempts. The denominator in the main table is therefore 40 attempts per method; the reported success rate is the number of successful attempts divided by 40. Technical invalids remain identified separately in the trial record and are not reclassified as policy failures.

Robot and visual observations. The platform was a Franka Research 3 (FR3) arm equipped with a Franka Hand parallel gripper. Three Intel RealSense RGB cameras provided the observations: an external D435i (serial number 042222071168), an external D435 (serial number 401622073871), and a wrist-mounted D405 (serial number 260522273410). Each camera produced 640×480 RGB frames at 30 Hz. Exposure was fixed at 8000. Frames were resized to 224×224 without cropping before being passed to the policy. The robot state consisted of the seven measured joint positions and the measured gripper width, giving an 8-dimensional state vector.

Table 11: Camera configuration for the real-robot Bread-transfer evaluation.

View	Camera	Serial number	Resolution	Rate
External view 1	RealSense D435i	042222071168	640×480	30 Hz
External view 2	RealSense D435i	401622073871	640×480	30 Hz
Wrist view	RealSense D405	260522273410	640×480	30 Hz

Policy output and chunk execution. The Pi0.5 policy emitted a 50-step action sequence. Each action contained seven absolute joint-position targets and one gripper-width target. The inference adapter validated the output shape and exposed the first 30 actions to the robot controller. A new observation was captured after each 30-step chunk, whose nominal duration was 1 s. Each trial allowed at most 60 policy calls. Continuous or unbounded execution was disabled, and output recording was enabled. While inference was running, the arm held its latest safe target.

Arm control and reset. Absolute joint-position commands were tracked by a joint-impedance controller at 30 Hz. Every transmitted command was checked against a maximum joint-reference velocity of 0.60 rad/s, a maximum joint-reference acceleration of 2.0 rad/s^2, and a maximum target-to-measured joint difference of 0.25 rad. Command state was preserved across policy chunks; if a control deadline was missed, the scheduler resynchronized to the current time instead of sending a catch-up burst. The nominal initial joint configuration was

$$\begin{bmatrix} -0.04675926,\ 0.02376588,\ 0.04375975,\ -2.12153602, \\ 0.01218572,\ 2.15952754,\ 0.77104223 \end{bmatrix} \text{rad}.$$

Before every counted attempt, the arm returned to this configuration and the gripper was fully opened. Execution was rejected when the reported maximum gripper opening was outside 70–90 mm.

Gripper control. The Franka Hand was controlled asynchronously so that blocking gripper communication did not interrupt the 30 Hz arm loop. Gripper predictions were applied at their corresponding action indices. The closing/opening speed was 0.10 m/s and the grasping force was 20 N. Predicted widths below 30 mm triggered closing, while widths above 60 mm triggered opening. The inner and outer grasp tolerances were 5 mm and 80 mm, respectively. During a physical gripper transition, the arm held its last target until the Hand reported completion or an error.

Recorded signals and outcome definition. Robot states and all camera streams were recorded at approximately 30 Hz. Each trial stored joint positions, velocities, torques, collision/contact states, Cartesian state, gripper width and grasp status, three synchronized MP4 streams, policy observations and predicted action chunks, timestamped controller/gripper/inference/safety events, and checkpoint and trial metadata.

A trial was successful only when the robot grasped one intended bread slice, removed it completely without dragging or substantially moving the toaster, retained it during transport, moved it over the cutting board, released it onto the board, and completed the sequence without operator intervention or a robot safety stop. Grasping both slices, slipping or dropping the bread, dragging or tipping the toaster, collisions, stopping before release, releasing outside the board, an operator emergency stop, or exhausting the trial budget counted as failures. Camera disconnections, camera-busy errors, invalid gripper calibration, communication faults, and unrelated hardware failures were recorded as technical invalids with their reason. They remained in the 40-attempt denominator and were not labeled policy failures.

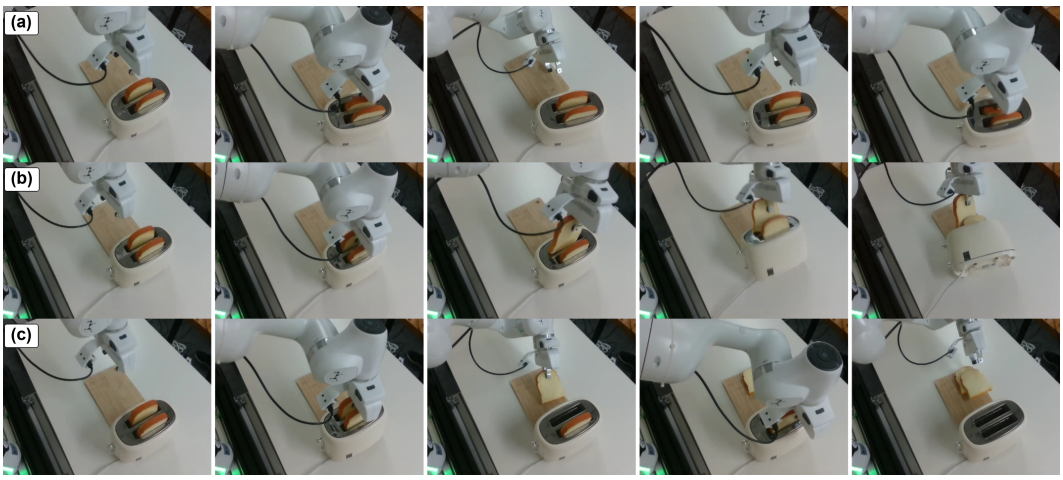


Figure 6: **Representative real-robot Bread-transfer rollouts.** Rows (a)–(c) correspond to Direct finetuning from the original $\pi_{0.5}$ checkpoint, EgoDex-style pretraining with official EgoDex labels and wrist–fingertip supervision, and EgoWeave (ours) with reconstructed labels and structured motion supervision, respectively. Each row shows a temporally ordered sequence from the toaster interaction to the subsequent bread-placement stage.

Implementation records. The controller, policy adapter, robot/camera configuration, and recording configuration were versioned with the trial bundle. The archived rollout directory therefore provides the checkpoint identity, camera mapping, controller limits, event timestamps, and outcome record needed to reproduce the protocol independently of the aggregate success table.

F Per-task Transfer Results and Evaluation Protocol

The following tables report the task-level success rates underlying the short-budget simulation transfer study. The first four rows follow the baseline order in Table 6; the geometric-loss ablation is shown immediately before the complete EgoWeave recipe. Bold values identify the highest rate for each task, with ties retained. Rates are percentages; task sets and episode counts are given in Section 5.4.

Suite success is the unweighted mean of the displayed task rates. The five main-text configurations reproduce Table 6, including 79.33% for EgoWeave on DexMimicGen B. Without the geometric loss, suite success is 54.50%, 24.67%, 65.00%, and 42.40% on DexJoCo, DexMimicGen A, DexMimicGen B, and RoboTwin, respectively. The complete recipe has higher suite averages in all four cases, although the difference varies by task: removing the loss improves Pouring and Pick dual bottles and leaves Handover block unchanged.

Table 12: Per-task success rates (%) on DexJoCo single-arm. Bold values mark the highest rate for each task; ties are retained.

Method	Hammer nail	Pick bucket	Pinch tongs	Water plant
Direct finetuning	28	42	42	56
EgoDex-style pretraining	**38**	48	30	58
EgoWeave w/ official labels	28	**66**	54	70
EgoWeave w/ wrist–fingertip supervision	**38**	58	56	68
EgoWeave w/o geometric loss	34	60	56	68
EgoWeave (ours)	**38**	**66**	**68**	**78**

Table 13: Per-task success rates (%) on DexMimicGen A. Bold values mark the highest rate for each task; ties are retained.

Method	Box cleanup	Drawer cleanup	Lift tray
Direct finetuning	32	50	0
EgoDex-style pretraining	36	46	10
EgoWeave w/ official labels	24	**54**	8
EgoWeave w/ wrist–fingertip supervision	28	**54**	6
EgoWeave w/o geometric loss	36	32	6
EgoWeave (ours)	**40**	46	**20**

Table 14: Per-task success rates (%) on DexMimicGen B. Bold values mark the highest rate for each task; ties are retained.

Method	Coffee	Pouring	Can sort
Direct finetuning	28	**94**	73
EgoDex-style pretraining	18	**94**	86
EgoWeave w/ official labels	42	92	91
EgoWeave w/ wrist–fingertip supervision	28	86	84
EgoWeave w/o geometric loss	22	90	83
EgoWeave (ours)	**62**	82	**94**

Table 15: Per-task success rates (%) on RoboTwin 2.0. Bold values mark the highest rate for each task; ties are retained.

Method	Stamp seal	Turn switch	Move can pot	Handover block	Pick dual bottles
Direct finetuning	38	24	46	**40**	54
EgoDex-style pretraining	**40**	20	32	24	48
EgoWeave w/ official labels	32	32	44	24	66
EgoWeave w/ wrist–fingertip supervision	32	50	46	28	66
EgoWeave w/o geometric loss	30	32	48	34	**68**
EgoWeave (ours)	32	**52**	**52**	34	58

The additional row isolates the compositional geometric loss within the structured motion supervision configuration. The wrist–fingertip comparison also changes target contents and normalization.

Figure 7 shows representative matched-seed rollouts for three of the simulation tasks.

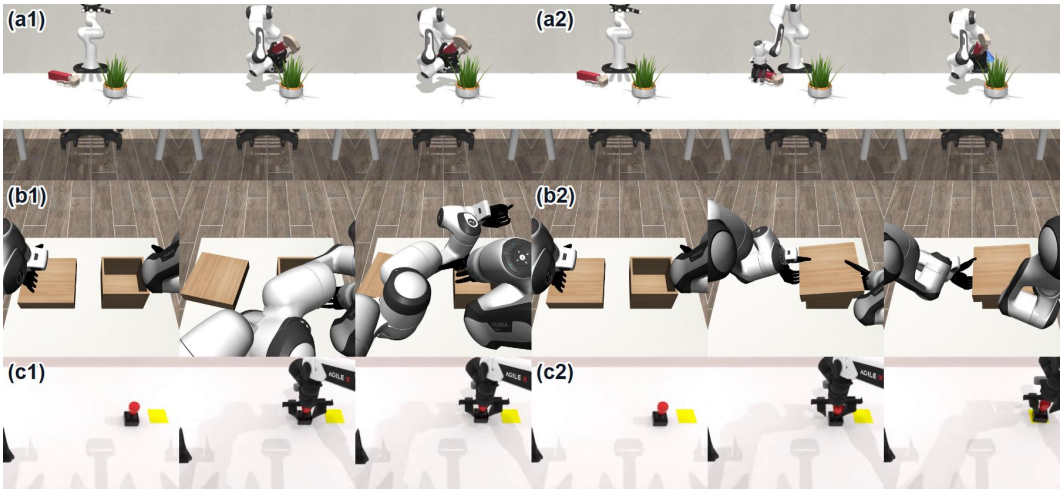


Figure 7: **Representative simulation rollouts.** Rows (a)–(c) show DexJoCo Water Plant, DexMimicGen A Box Cleanup, and RoboTwin 2.0 Stamp Seal, respectively. In each row, the left sequence shows chronological frames from Direct finetuning of the original $\pi_{0.5}$ checkpoint, and the right sequence shows EgoWeave (ours) with reconstructed labels and structured motion supervision. The two rollouts in each row use the same evaluation seed.

G Geometric Adapter and Motion-Record Details

Camera, crop, and sampling geometry. The projection interface supports pinhole/radial–tangential and fisheye models. Global and local virtual cameras share the physical optical center; source calibration, image dimensions, units, and the physical pose reference are retained with each record. Common-frame camera poses establish the reference for predicted hand geometry.

The global detector view uses calibrated intrinsics K_G. Fisheye inputs are mapped to this perspective view through the source projection. For the central hand ray d in Equation 1, choose a reference axis a and construct

$$x = \frac{a \times d}{\|a \times d\|_2}, \qquad y = d \times x, \qquad R_{CL} = [x\ y\ d]. \tag{21}$$

This specifies the local viewing direction and roll. For a global detection-box corner v_k, its local projection is

$$u_k = \pi\big(K_L R_{CL}^{\top} R_{CG} K_G^{-1} \tilde{v}_k\big)\,, \tag{22}$$

where π denotes perspective division. The axis-aligned extent of these corners defines the crop center, width, and height. With side sign $m = +1$ for right and -1 for left, the crop affine is

$$A_m = \begin{bmatrix} ms & 0 & c_x - mb/2 \\ 0 & s & c_y - b/2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{23}$$

Here b and s are the crop side length and pixel scale determined by the detected extent and the chosen patch resolution. Local coordinates outside the source canvas are masked; valid coordinates use bilinear interpolation.

Coordinate decoding and hand-side conventions. Let (σ, τ_x, τ_y) be WiLoR's predicted crop-camera parameters, (p_x, p_y) the principal point of K_L, and $\bar{f} = (K_{L,11} + K_{L,22})/2$ its mean focal length. The adapter recomputes camera translation from the preserved local crop metadata:

$$D = b\sigma + \varepsilon, \qquad t_L = \begin{bmatrix} 2(c_x - p_x)/D + m\tau_x \\ 2(c_y - p_y)/D + \tau_y \\ 2\bar{f}/D \end{bmatrix}. \tag{24}$$

where ε is a numerical safeguard. For a predicted hand-centered landmark $\hat{j}_j$ and $M_m = \mathrm{diag}(m, 1, 1)$, the recovered landmark is

$$\hat{x}_{L,j} = M_m \hat{j}_j + t_L, \qquad \hat{x}_{C,j} = R_{CL} \hat{x}_{L,j}. \tag{25}$$

The same reflection and translation apply to mesh vertices. The learning encoder subsequently centers local landmarks at joint zero (Appendix A); tracked camera poses place the decoded geometry in the common reference frame.

Let Q be the predicted MANO global orientation, and define

$$B = \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \qquad M = \mathrm{diag}(-1, 1, 1). \tag{26}$$

The standardized local wrist orientation is $R_{Lw} = QB$ for a right hand and $R_{Lw} = MQBM$ for a left hand. Reflecting both the camera and wrist bases preserves a proper rotation for the left hand. Physical-camera and common-frame orientations are $R_{Cw} = R_{CL} R_{Lw}$ and $R_{Ww} = R_{WC} R_{Cw}$, respectively.

Effect of local orientation. For an ideal equidistant fisheye with focal length f_s and a fixed-axis pinhole view with focal length f_g,

$$r_s = f_s \theta, \qquad r_g = f_g \tan\theta, \qquad \frac{\mathrm{d}r_g}{\mathrm{d}r_s} = \frac{f_g}{f_s} \sec^2\theta, \quad 0 \le \theta < \frac{\pi}{2}. \tag{27}$$

Here θ is measured from the shared axis of these two views. The expression characterizes radial magnification in this ideal model. The hand-centered local camera instead places the detected hand ray on its own optical axis. Composing source projection and crop transforms is a separate operation that avoids intermediate image resampling on the patch-generation path; global detection retains its own image preparation.

Temporal association and candidate selection. For synchronized multi-view observations, timestamps and calibrated camera poses place candidates in a common frame. Root translation can be associated with the preceding state when a small timestamp offset remains; orientation and articulation are not extrapolated across that offset. For each hand, candidates are filtered by position and orientation consistency, and at most one candidate per camera is retained before clustering. Other candidates join the seed when

$$\|\tilde{p}_i - \tilde{p}_j\|_2 \le \tau_p, \qquad \angle(R_i^\top R_j) \le \tau_R, \tag{28}$$

where j is the seed and $\angle$ is the relative rotation angle. A single candidate can form a cluster. Camera identifiers and box coordinates resolve remaining ties.

Equal-weight fusion and temporal filtering. For selected set $\mathcal{C}$, positions use the arithmetic mean $\bar{p} = |\mathcal{C}|^{-1} \sum_{i \in \mathcal{C}} \tilde{p}_i$. If $\sum_{i \in \mathcal{C}} R_i = U \Sigma V^\top$, the fused orientation is

$$\bar{R} = U \,\mathrm{diag}\big(1, 1, \det(UV^\top)\big)\, V^\top. \tag{29}$$

Detector scores influence selection, while accepted views receive equal fusion weight. A constant-velocity Kalman filter updates the root position, followed by adaptive position smoothing; orientation uses a spherical-interpolation exponential moving average. Common-frame root-relative landmark offsets are averaged across accepted views. When at least two views are available, bone-length estimates receive an exponential update. Offsets are then projected toward the available bone-length estimates, temporally smoothed, and added to the filtered root. Monocular observations follow the same temporal update with a single-view measurement.

Track lifecycle and offline processing. The tracker distinguishes measured updates, short-gap predictions, and missing outputs. Offline formalization then applies the validity rules in Appendix C; offline annotation may use neighboring frames when assigning export status, while the resulting training masks retain the distinction between available and missing channels.