EGOHEP: PREDICTIVE HAND–EYE PRETRAINING FOR ROBOT VISUAL REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Robotic manipulation requires visual representations that support the selection of relevant objects and regions for interaction. A central challenge is to learn from the way humans actively select visual information and coordinate it with subsequent hand movements. We introduce **EgoHEP**, a predictive hand-eye pretraining framework that learns this relationship from gaze and hand-wrist motion in egocentric recordings. EgoHEP combines hand-wrist reconstruction and forecasting with spatial attention predicted under gaze supervision, then aligns attention-weighted visual features with representations of future hand-wrist trajectories across multiple horizons, linking visual selection to manipulation dynamics. Experiments in simulation and on real robots demonstrate the transferability of visual representations learned through predictive hand–eye supervision, with EgoHEP achieving state-of-the-art performance on multiple downstream tasks.

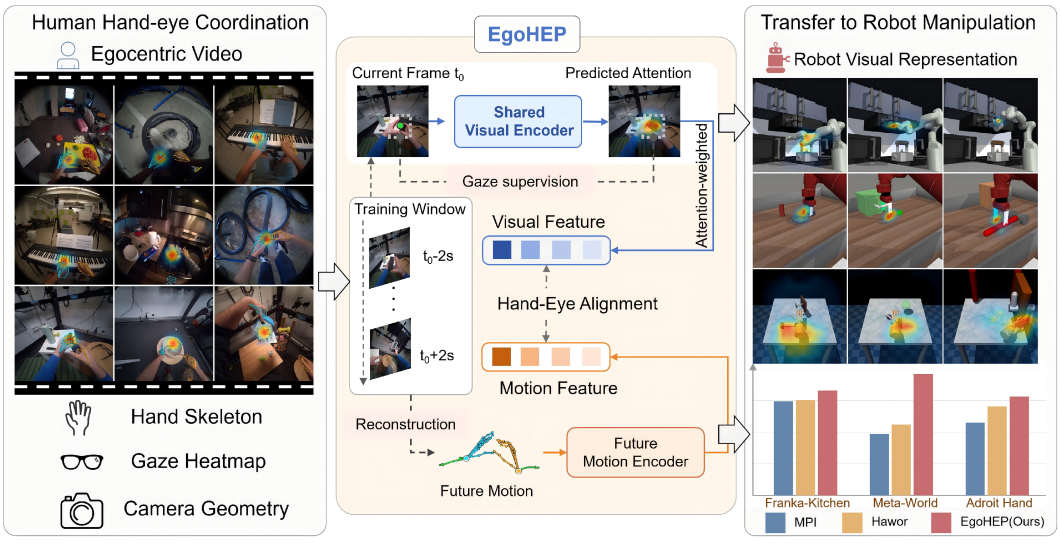

Figure 1: We propose EgoHEP, a predictive hand-eye pretraining framework that transfers human **hand-eye coordination** to robotic manipulation. Given egocentric human videos with corresponding hand poses, gaze, and camera parameters, EgoHEP learns temporal representations of future hand-wrist motion while using gaze supervision to focus visual attention on the interactive object. The resulting representations capture **human interaction dynamics** and improve performance across multiple downstream robotic manipulation tasks.

1 INTRODUCTION

Human videos provide useful supervision for learning visual representations for robotic manipulation. Existing methods learn from temporal structure, interaction prediction, and hand-object affordances Nair et al. (2022); Zeng et al. (2024); Srirama et al. (2024). However, these methods do not explicitly connect where humans look with how their hands move next. This connection is relevant

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.

to manipulation, which requires both locating an interaction region and executing an appropriate motion. Behavioral studies show that human gaze often reaches grasp sites and target locations before the hand or manipulated object arrives Johansson et al. (2001). This suggests that gaze and subsequent hand motion could jointly indicate which visual information matters for upcoming interaction. This motivates an essential question: *Can jointly modeling where humans look and how their hands move provide a better visual representation for robotic manipulation?*

To investigate this question, we turn to egocentric human recordings that allow visual observations to be studied together with gaze and hand motion. Ego-Exo4D Grauman et al. (2024), provides egocentric videos with gaze estimates and hand-pose annotations. Recent approaches such as EgoMimic Kareer et al. (2025) and EgoScale Zheng et al. (2026) further demonstrate the value of egocentric human demonstrations for robot manipulation through large-scale human pretraining. Building on these advances, we focus on how the relationship between visual selection and motion can serve as a visual pretraining objective. Gaze provides supervision about where a person looks, while subsequent hand and wrist motion describes how the interaction unfolds. Therefore, we introduce EgoHEP, a predictive hand-eye pretraining framework that connects these signals through a shared visual representation. Alongside hand-wrist reconstruction and forecasting, EgoHEP uses gaze to supervise spatial attention predicted from frame, and aligns the attention-weighted visual features for future hand-wrist motion.

We construct training data from HOT3D Banerjee et al. (2025) and EgoExo4D Grauman et al. (2024), combining official and estimated hand–wrist labels for motion pretraining, and using official hand–wrist with gaze labels for joint optimization. Historical visual–motor context supports hand–wrist reconstruction and prediction, capturing local hand and wrist displacement. Gaze supervises spatial attention predicted solely from current frame. Through a shared scene encoder, attention-weighted visual features are aligned with future motion encoded for each hand. This objective connects current visual selection with subsequent manipulation dynamics. We evaluate the pretrained visual encoder within MPI Zeng et al. (2024) and ACT Zhao et al. (2023a) policies using current visual observations. In EgoHEP, human signals serve only as pretraining supervision, and inference requires neither human gaze or hand pose annotations.

We evaluate EgoHEP on 21 downstream tasks across simulation benchmarks and real-world experiments, covering single-arm manipulation in Franka-Kitchen Gupta et al. (2019) and Meta-World Yu et al. (2020), dexterous manipulation in Adroit Hand Rajeswaran et al. (2017), bimanual manipulation in RoboTwin Chen et al. (2025), and real-world bimanual tasks. Our results demonstrate the transferability of EgoHEP across diverse robot embodiments and policy architectures, outperforming MPI and HaWoR Zhang et al. (2025). As shown in Fig. 1, EgoHEP improves performance over MPI by 16.5% on Adroit Hand, 38.0% on Meta-World, and 6.8% on Franka-Kitchen.

To summarize, we make the following contributions:

- We propose EgoHEP, a predictive hand-eye pretraining framework that integrates hand-wrist motion and gaze to learn human interaction dynamic representations for robotic manipulation.
- We develop a gaze-supervised alignment objective that connects scene and attention-weighted visual features with future hand-wrist motion representations across multiple horizons, linking current visual selection to subsequent manipulation dynamics.
- We provide an empirical analysis across diverse robot embodiments. Our results demonstrate the transferability of EgoHEP, achieving state-of-the-art performance on multiple downstream tasks.

2 Related Work

Ego Vision Understanding: Egocentric vision research now extends beyond early action recognition to large-scale, multimodal, and task-oriented behavior understanding. EPIC-KITCHENS Damen et al. (2018) and Ego4D Grauman et al. (2022) establish large-scale video resources covering daily activities, while EgoVLP Lin et al. (2022) and LaViLa Zhao et al. (2023b) further connect manipulation processes with semantic descriptions through joint video and language learning. Recent datasets record finer-grained interaction signals. Ego-Exo4D Grauman et al. (2024) and HOT3D Banerjee et al. (2025) provide three-dimensional hand motion and camera information,

and the portions captured with Project Aria Engel et al. (2023) also include synchronized gaze. EgoM2P Li et al. (2025) and EgoGazeVQA Peng et al. (2026) use these signals for cross-modal prediction and gaze-guided video understanding. OCT Liu et al. (2022) further predicts future hand trajectories and object interaction hotspots.

Hand and Eye Representation: Hand motion and gaze describe the execution of an action and the operator's visual selection, respectively, and recent work has begun to convert them into structured supervision. HaWoR Zhang et al. (2025) reconstructs hand motion in world coordinates from egocentric videos, whereas HRP Srirama et al. (2024) uses future hand positions, contact regions, and interaction objects as affordance supervision for visual encoders. VRB Bahl et al. (2023) transfers human-video affordances to robotics, while Li et al. Li et al. (2018) jointly model gaze and actions for video recognition. For gaze modeling, GazeVLA Li et al. (2026) uses gaze as an intent representation before action generation. These methods improve the modeling of action states, interaction locations, or task intentions, but hand motion and gaze generally serve different learning roles and are not jointly used to constrain a scene representation for transfer to robots.

Visual Representation Learning for Robotics: Visual representation learning for robotics is moving from generic visual pretraining toward explicit modeling of temporal changes and interaction structure. MVP Xiao et al. (2022) and VC-1 Majumdar et al. (2023) use masked reconstruction to learn fine-grained visual information, R3M Nair et al. (2022) and Voltron Karamcheti et al. (2023) incorporate temporal relations and language semantics, and RSP Jang et al. (2024) learns visual dynamics through future-frame prediction. VIP Ma et al. (2022) and LIV Ma et al. (2023) learn control-oriented representations and rewards from human videos, with LIV additionally incorporating language. HRP Srirama et al. (2024) and MPI Zeng et al. (2024) further incorporate hand, contact, interaction-object, or future-interaction information into their pretraining objectives. Meanwhile, EgoEngine Liu et al. (2026) directly converts egocentric human videos into robot observations and action demonstrations, expanding the data available for policy learning while introducing the need to address action retargeting, scene generation, and embodiment mismatch. Overall, existing methods do not jointly use future hand motion and visual attention in egocentric data to shape a transferable scene encoder. EgoHEP aligns gaze-supervised visual features with future hand-wrist motion for transfer to robot control.

3 METHOD

3.1 EGOCENTRIC SUPERVISION CONSTRUCTION

Egocentric observations. Let $\mathcal{V} = \{I_t\}_{t=1}^{T}$ denote an egocentric video sequence with available hand-wrist and gaze supervision. These signals describe local hand , hand movement through space, and visual attention, respectively. We construct training windows from HOT3D Banerjee et al. (2025) and EgoExo4D Grauman et al. (2024), with ground-truth annotations from both datasets. The hand-wrist pretraining stage additionally uses EgoExo4D windows with pseudo hand and wrist labels that pass the data qualification criteria. We distinguish these pseudo-labels from ground-truth supervision and account for missing or invalid targets using validity masks.

Hand-wrist labels. For each hand $s \in \{L, R\}$, let $\mathbf{x}_{t,j}^{s} \in \mathbb{R}^3$ denote joint positions in a common anchor camera frame within each window, with $j = 0$ denoting the wrist and $j = 1, \ldots, J$ the $J = 20$ non-root joints. We define

$$\mathbf{w}_t^s = \mathbf{x}_{t,0}^s, \qquad \mathbf{p}_{t,j}^s = \mathbf{x}_{t,j}^s - \mathbf{w}_t^s, \quad j = 1, \ldots, J. \tag{1}$$

Stacking the local offsets gives the joint configuration $\mathbf{q}_t^s \in \mathbb{R}^{3J}$. The full skeleton is recovered by adding the wrist position to these offsets and transformed to the evaluation camera frame before computing errors over all 21 joints.

The binary mask $m_{t,j}^s$ indicates label validity. A local offset requires both its joint and wrist labels to be valid. Supervised losses exclude invalid targets; temporal differences require all contributing states to be valid.

Gaze labels. We use dataset-provided Project Aria eye-gaze estimates. Gaze is associated with selected frames using device timestamps. After projection and the same geometric transformations applied to frames, gaze is represented by normalized image coordinates $g_t = (u_t, v_t) \in [0, 1)^2$. The mask m_t^g indicates valid temporal association and finite, in-bounds coordinates.

Each valid current-frame gaze point defines a Gaussian soft target Q_t on a 32×32 grid. We mask invalid image regions and normalize the remaining probability mass to one. The Gaussian width is specified in the implementation details.

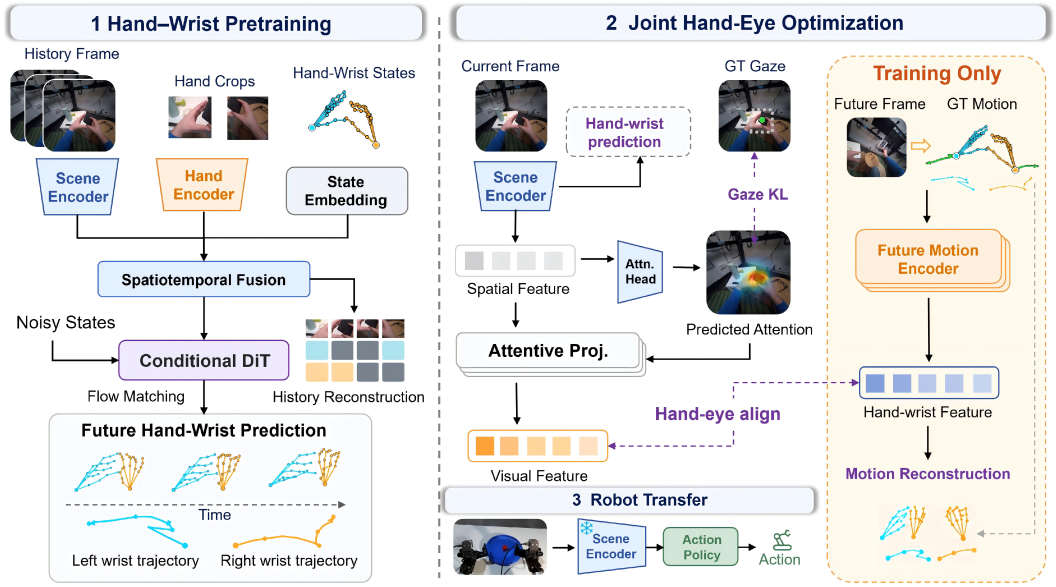

Figure 2: **Overview of EgoHEP.** Given egocentric video with hand crops, hand-wrist states, and gaze, EgoHEP first learns hand-wrist dynamics through reconstruction and future prediction.It then aligns gaze-supervised visual features with future hand-wrist motion, and transfers the learned representation to downstream robot manipulation tasks.

3.2 Hand–Wrist Pretraining (Stage I)

Motion representation. Human hand motion combines local joint configuration and global wrist translation. We represent each hand $s \in \{L, R\}$ by its wrist-relative joint configuration $\mathbf{q}_t^s = \{\mathbf{p}_{t,j}^s\}_{j=1}^J \in \mathbb{R}^{3J}$ and wrist position $\mathbf{w}_t^s \in \mathbb{R}^3$ in the anchor camera frame.

We model future motion as innovations relative to the most recent valid historical state. Let $\bar{\mathbf{q}}_t^s$ and $\bar{\mathbf{w}}_t^s$ denote the last valid joint and wrist states within the observation window, selected using their corresponding validity masks. For a future offset $k > 0$, the targets are

$$\Delta\mathbf{q}_{t+k}^s = \mathbf{q}_{t+k}^s - \bar{\mathbf{q}}_t^s, \qquad \Delta\mathbf{w}_{t+k}^s = \mathbf{w}_{t+k}^s - \bar{\mathbf{w}}_t^s. \tag{2}$$

These targets describe joint changes and wrist displacements relative to the observed history. Adding the reference states back to the predicted innovations recovers the future hand-wrist states. We retain the validity masks $m_{t,j}^s \in \{0, 1\}$: missing or invalid joints are masked.

Visual-motor context encoding. We jointly encode historical full-frame RGB, bilateral hand crops, local joint configurations, wrist states, and validity masks. Temporal information, camera intrinsics, and historical-to-anchor camera transforms provide additional context. These inputs are fused into a visual-motor representation $\mathbf{c}_t$.

Conditioned on the visual-motor context $\mathbf{c}_t$, we model future hand-wrist innovations using the conditional flow field $\mathbf{v}_\psi(\mathbf{z}_\tau, \tau \mid \mathbf{c}_t)$, where $\mathbf{z}_\tau$ is an intermediate flow state and $\tau \in [0, 1]$ denotes flow time. Adding the historical reference states to the predicted innovations recovers the future hand-wrist states. We optimize

$$\mathcal{L}_{\text{HW}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}}, \tag{3}$$

Here, $\mathcal{L}_{\text{flow}}$ is the flow-matching loss, $\mathcal{L}_{\text{rec}}$ supervises reconstructed states, and $\mathcal{L}_{\text{dyn}}$ combines velocity, acceleration, and bone-consistency losses. All loss weights are provided in the implementation details.

3.3 GAZE-SUPERVISED ATTENTION LEARNING

Predicted spatial attention with gaze supervision. Given the current RGB frame, the scene encoder produces spatial tokens on a 28×28 grid. An attention predictor maps these features to 32×32 spatial logits and applies a masked softmax over the valid image region to obtain the attention distribution P_t. We supervise P_t using the soft gaze target Q_t. The gaze objective $\mathcal{L}_{\text{gaze}}$ is the KL divergence $D_{\text{KL}}(Q_t \| P_t)$, averaged over samples with valid current-frame gaze.

Attention-weighted visual representation. We resample the predicted attention to the 28×28 visual token grid, apply the valid-patch mask, and renormalize the weights. A weighted sum of the spatial tokens then yields the attention-weighted representation $\mathbf{z}_t^{\text{att}}$. Pooling uses model-predicted attention, while dataset-provided gaze supplies the supervision target. Both attention and the pooled representation are therefore computed from RGB without gaze input at inference.

3.4 JOINT HAND-EYE PREDICTIVE OPTIMIZATION (STAGE II)

Predictive visual-motion alignment. We retain the hand-wrist objective using official ground-truth supervision only, denoted by $\mathcal{L}_{\text{HW}}^{\text{GT}}$. For each hand and horizon $h \in \{4, 8, 16\}$, a Transformer encodes the complete future motion prefix into a normalized embedding through masked temporal pooling. An auxiliary loss $\mathcal{L}_{\text{mot}}$ reconstructs the prefix from this embedding to preserve trajectory information.

The visual query is obtained by projecting the predicted-attention-pooled feature $\mathbf{z}_t^{\text{att}}$ together with the prediction time interval and hand-side indicator, followed by normalization. Both hands share the predicted attention, and the projection is shared across hands and horizons. A unidirectional contrastive loss $\mathcal{L}_{\text{InfoNCE}}$ matches each visual query to the future motion embedding of the same sample, hand, and horizon, using valid gaze-motion pairs.

Joint hand-eye optimization. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{HW}}^{\text{GT}} + \lambda_{\text{g}} \mathcal{L}_{\text{gaze}} + \lambda_{\text{m}} \mathcal{L}_{\text{mot}} + \lambda_{\text{a}} \mathcal{L}_{\text{InfoNCE}}. \tag{4}$$

We cycle through the three prediction horizons, using one per training step, and gradually increase the alignment weight after an initial stabilization period. The hand-wrist, gaze, and contrastive objectives jointly optimize the shared trainable scene encoder, while motion reconstruction regularizes the motion embedding. The gaze-attention pathway does not directly enter the hand-wrist forecasting decoder. Loss weights and scheduling details are provided in the implementation details.

4 EXPERIMENTS

4.1 HAND-EYE PRETRAINING DATA

Dataset composition. We use Project Aria egocentric recordings from HOT3D and EgoExo4D. Official hand-wrist annotations from both datasets are supplemented with HaWoR-generated EgoExo4D pseudo-labels for dynamics pretraining. Pseudo-labeling applies to hand and wrist supervision; gaze labels come from official eye-gaze estimation.

Table 1 summarizes the training pools. The mixed-supervision dynamics pool contains 34,843 windows from 1,193 recordings. Joint hand-eye optimiation uses a subset of 8,248 windows, restricted to official hand-wrist annotations and gaze.

Table 1: Statistics of hand-eye pretraining data. Training data are deduplicated within each stage.

Training stage	HOT3D GT	EgoExo4D GT	EgoExo4D Pseudo	Total
Hand-wrist pretraining (Stage I)	9,163	1,166	24,514	34,843
Joint hand-eye optimization (Stage II)	7,082	1,166	0	8,248

label quality filtering. To control pseudo-label quality, we assign discrete weights based on prediction validity, finite joint coordinates, and wrist anchoring availability. Where valid official body-wrist observations are available, they provide positional anchors for translating the predicted hand.

Table 2: **Evaluation on Franka-Kitchen and Meta-World.** We report success rates (%) over 50 evaluation episodes per model and task. Average is the unweighted mean across five tasks within each benchmark. The highest and second-highest distinct values in each column within each benchmark are **bolded** and underlined, respectively.

Franka-Kitchen						
Method	Turn Knob	Open Microwave	Slide Door	Light	Hinge Door	Average
MPI-Base	54.0	<u>34.0</u>	**100.0**	**82.0**	28.0	59.6
HaWoR	72.0	26.0	94.0	70.0	**40.0**	60.4
Hand-Wrist Pretraining	<u>78.0</u>	30.0	<u>98.0</u>	70.0	<u>36.0</u>	<u>62.4</u>
EgoHEP (ours)	**88.0**	**40.0**	96.0	<u>76.0</u>	32.0	**66.4**
Meta-World						
Method	Assembly	Bin Picking	Button Press	Drawer Open	Hammer	Average
MPI-Base	0.0	<u>4.0</u>	<u>78.0</u>	<u>96.0</u>	16.0	38.8
HaWoR	22.0	0.0	**100.0**	80.0	22.0	44.8
Hand-Wrist Pretraining	<u>40.0</u>	0.0	**100.0**	94.0	<u>28.0</u>	<u>52.4</u>
EgoHEP (ours)	**74.0**	**36.0**	**100.0**	**98.0**	**76.0**	**76.8**

Wrist agreement provides a positional reliability cue without establishing the accuracy of finger joint configurations.

We further filter temporal discontinuities. Large wrist jumps between consecutive valid frames invalidate the affected hand-wrist supervision slots, after which window-level hand coverage is re-evaluated. Quality thresholds and masking statistics are provided in the appendix.

4.2 Experimental Setup

Benchmarks and Policy integration. We evaluate visual representations on 18 simulation tasks and three real-world manipulation tasks. The simulation benchmarks comprise five tasks each from Franka-Kitchen and Meta-World for general manipulation, four Adroit Hand tasks for dexterous manipulation, and four RoboTwin tasks for bimanual manipulation. The real-world tasks further evaluate transfer to physical robot execution.

We integrate the visual representations into established policy. For Franka-Kitchen and Meta-World, we adopt the policy architecture used by MPI Zeng et al. (2024). For Adroit Hand, RoboTwin, and the real-world tasks, we use ACT Zhao et al. (2023a). Within each benchmark, the policy architecture is held fixed across representation comparisons.

Evaluation protocol. We compare EgoHEP with MPI-Base, which uses the same visual backbone, the original pretrained HaWoR scene encoder transferred to downstream policies, and Hand-Wrist Pretraining. Hand-Wrist Pretraining denotes our Stage I checkpoint, trained with ground-truth and pseudo hand–wrist labels without gaze supervision or predictive alignment.

In contrast, GT-HW in Table 5 is a Stage II variant trained using only ground-truth hand-wrist motion supervision.For all downstream tasks, we freeze the visual encoder and train only the policy-side parameters. Within each task, methods use matched demonstration sets and policy-training budgets. We evaluate each model on 50 rollouts per simulation task and report task-wise success rates. Real-world comparisons use a fixed number of trials per task across methods. Demonstration counts and policy-training hyperparameters are provided in the implementation details.

4.3 Simulation Manipulation Results

Our experiments are evaluated across manipulation settings with diverse robot embodiments and interaction requirements. For downstream policy learning, we freeze the scene encoder and train the action policy using task-specific training data. We adopt the same policy architecture as MPI for Franka-Kitchen and Meta-World, and use ACT for Adroit Hand and RoboTwin. Each method is evaluated over 50 episodes per task, and we report task-wise success rates.

General Single-Arm Manipulation: Franka-Kitchen and Meta-World. Franka-Kitchen mainly involves interactions with household objects, where the manipulated object remains fixed in the scene and the main challenge lies in approaching the correct part and executing an appropriate local motion. EgoHEP reaches 66.4% success, with its clearest gains appearing on *turn knob* and *open microwave*. Both tasks contain a relatively localized interaction point: the knob or the handle, followed by a constrained motion. This structure resembles the hand-eye relation modeled during pretraining: visual attention first concentrates on an interaction region, after which the hand follows a structured trajectory. It is therefore reasonable that these tasks benefit from preserving information about both interaction location and subsequent motion. The remaining Kitchen tasks reveal a different picture. *Slide door* is already solved at a very high rate by nearly all representations. *Hinge door*, by contrast, remains less stable. Opening a large hinged object involves not only finding the handle but maintaining a suitable interaction direction over a larger displacement.

Meta-World contains a broader range of interaction patterns, and the differences between representations become more pronounced. In *assembly*, the policy must reason about the relative configuration between an object and a target rather than simply approaching a fixed handle. EgoHEP increases success from 40% to 74% over the hand-wrist pretraining model. One interpretation is that the attended region must encode not only "where the hand goes" but also the spatial relationship that determines where the manipulated object should be placed. The largest improvement occurs on *hammer*. Hammering requires locating the tool, acquiring it with an appropriate grasp, and subsequently using it toward another interaction target. In contrast, *button press* is already saturated and *drawer open* is close to saturation.

Table 3: **Dexterous manipulation on Adroit Hand.** Success rates (%) on four tasks.

Method	Door	Hammer	Pen	Relocate	Average
MPI-Base	<u>92.0</u>	52.0	32.0	8.0	46.0
HaWoR	<u>92.0</u>	<u>80.0</u>	42.0	12.0	<u>56.5</u>
Hand-Wrist Pretraining	88.0	66.0	**56.0**	<u>14.0</u>	56.0
EgoHEP (ours)	**96.0**	**82.0**	<u>52.0</u>	**20.0**	**62.5**

Dexterous Manipulation: Adroit Hand. Adroit Hand introduces a substantially different embodiment. The policy controls a high-dimensional dexterous hand, while EgoHEP is pretrained from human egocentric observations without mapping human joint trajectories into the robotic action space. The resulting 62.5% average success therefore provides an interesting test of whether the learned visual representation transfers independently of direct embodiment correspondence.

Among the four tasks, *Hammer* again shows a relatively large improvement. Although the Adroit hand differs greatly from a human hand in actuation and control, the perceptual problem still contains a familiar structure: identify a tool, establish a task-compatible grasp, and maintain information about the subsequent manipulation target. The fact that both Meta-World and Adroit hammering show strong gains suggests that tool-mediated interaction may be a particularly useful setting for studying predictive visual representations, because the visually relevant object at the current moment has a clear relationship with a later action sequence.

Relocate provides a different challenge. All methods remain relatively weak, including EgoHEP. Here, success depends on both object acquisition and maintaining the object under dexterous control while moving it to a target location. The modest improvement of EgoHEP may indicate that better visual initialization helps with the early phases of interaction, while much of the remaining difficulty lies in closed-loop state estimation and control after the object has already been grasped.

Bimanual Manipulation: RoboTwin. RoboTwin evaluates whether EgoHEP transfers from human hand-eye pretraining to coordinated dual-arm manipulation. EgoHEP achieves the highest average success rate of 76.5%, compared with 66.5% for MPI-Base, 73.0% for HaWoR, and 71.5% for hand-wrist pretraining. The largest gain appears on *Bell Press*, where EgoHEP reaches 54%, substantially higher than the hand-wrist model. This task requires locating a compact functional region before contact, which is closely related to the spatial attention supervision introduced by gaze.

The remaining tasks reveal that this advantage is not uniform. On *Stapler Press*, EgoHEP is slightly below HaWoR and hand-wrist pretraining, suggesting that identifying the interaction region alone

Table 4: **Evaluation on RoboTwin.** We report success rates (%) over 50 evaluation episodes per model and task. Average is the mean across the four bimanual manipulation tasks.

Method	Bell Press	Stapler Press	Pot Lift	Microphone Handover	Average
MPI-Base	34.0	44.0	<u>96.0</u>	92.0	66.5
HaWoR	<u>42.0</u>	**64.0**	92.0	<u>94.0</u>	<u>73.0</u>
Hand-Wrist Pretraining	26.0	<u>62.0</u>	**100.0**	**98.0**	71.5
EgoHEP (ours)	**54.0**	58.0	**100.0**	<u>94.0</u>	**76.5**

may be insufficient when success also depends strongly on contact geometry and execution direction. *Pot Lift* and *Microphone Handover* are already solved at high rates by multiple representations, leaving limited room to distinguish the effect of additional hand-eye supervision.

Together with the single-arm and dexterous results, RoboTwin suggests that the usefulness of gaze supervision depends more on the structure of the interaction than on the robot embodiment itself. The strongest gains tend to occur when a visually localized region is closely connected to the subsequent manipulation, whereas tasks dominated by precise contact geometry, proprioceptive feedback, or already-saturated execution show smaller differences.

4.4 Real-World Manipulation

The real-world tasks in Fig. 3 evaluate object handover, coordinated placement, and bimanual lifting. As reported in Fig 4, EgoHEP achieves the highest success rate on all three tasks, averaging 65.0%, compared with 51.7% for hand-wrist pretraining and 25.0% for MPI. With the visual encoders frozen and the same ACT policy setup, these results support the transfer of human hand-eye pretraining to physical robot manipulation.

The task-wise results suggest complementary benefits from motion prediction and joint hand-eye supervision. On *Pass Banana*, hand-wrist pretraining already raises success from 5% to 55%, while EgoHEP further improves it to 65%. The largest additional gain occurs on *Lift Plate*, where EgoHEP reaches 70%, exceeding hand-wrist pretraining by 20 percentage points; *Place Pen* improves from 50% to 60%. These tasks involve relating object locations to coordinated actions by both hands, making the improvements consistent with the intended role of linking visual attention to future motion.

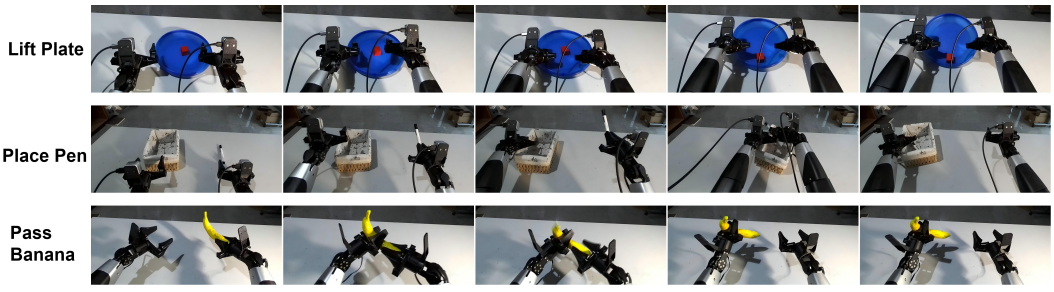

Figure 3: Real-world manipulation. Example execution sequences for *Lift Plate*, *Place Pen*, and *Pass Banana*, covering coordinated lifting, object placement, and handover between grippers. Frames are ordered temporally from left to right.

Fig. 5 illustrates different trajectories executed by policies using MPI and EgoHEP during real-world deployment. Interestingly, the EgoHEP policy follows a trajectory that more closely resembles how humans approach and grasp a plate, whereas the MPI policy first aligns its grippers before grasping. This qualitative observation suggests that hand-eye pretraining helps capture human interaction patterns and preferences, which can inform downstream robot behavior.

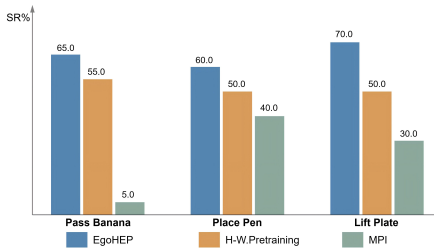

Figure 4: Result on real-world bimanual tasks, success rates (SR%) .

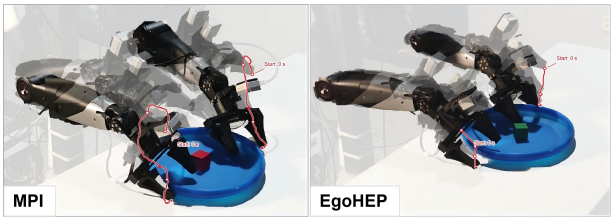

Figure 5: **Trajectory comparison.** MPI and EgoHEP follow different grasp trajectories in the *Lift Plate* task.

Table 5: **Ablation of hand-eye supervision.** All variants are evaluated after Stage II training. GT-HW uses only ground-truth hand-wrist motion supervision; Gaze adds gaze supervision; EgoHEP further adds predictive visual-motion alignment (Align.). The highest success rate in each task and average column within each benchmark is **bolded**.

Franka-Kitchen									
Method	Hand-wrist	Gaze	Align.	Turn Knob	Open Microwave	Slide Door	Light	Hinge Door	Avg.
GT-HW	✓	–	–	66.0	40.0	**96.0**	60.0	**38.0**	60.0
Gaze	✓	✓	–	84.0	**42.0**	94.0	70.0	28.0	63.6
EgoHEP (ours)	✓	✓	✓	**88.0**	40.0	**96.0**	**76.0**	32.0	**66.4**
Meta-World									
Method	Hand-wrist	Gaze	Align.	Assembly	Bin Picking	Button Press	Drawer Open	Hammer	Avg.
GT-HW	✓	–	–	2.0	18.0	**100.0**	**100.0**	**94.0**	62.8
Gaze	✓	✓	–	40.0	0.0	**100.0**	**100.0**	84.0	64.8
EgoHEP (ours)	✓	✓	✓	**74.0**	**36.0**	**100.0**	98.0	76.0	**76.8**

4.5 Ablation of hand-eye supervision

Table 5 separates the contributions of gaze supervision and predictive visual-motion alignment. Adding gaze supervision alone yields only a modest improvement in the suite averages, from 60.0% to 63.6% on Franka Kitchen and from 62.8% to 64.8% on Meta-World, with mixed task-level effects. In contrast, introducing predictive alignment produces a substantially larger gain, raising performance to 66.4% and 76.8%, respectively. The improvement is significant on Meta-World tasks such as *assembly* and *bin picking*, where aligning gaze-conditioned visual features with future hand-wrist motion appears more informative than gaze supervision alone. These results suggest that the main benefit of EgoHEP comes not simply from predicting where humans look, but from coupling the attended visual representation with the dynamics of what the hands do next.

5 Conclusion

We introduced EgoHEP, a predictive hand-eye pretraining framework for learning transferable robot visual representations from egocentric human videos. EgoHEP combines hand-wrist dynamics prediction with gaze-supervised attention learning and aligns the attended visual features with future motion. Experiments on simulated and real-world manipulation tasks show improved over state-of-the-art baselines across multiple benchmarks, robot embodiments, and policy architectures. These findings support the value of jointly modeling human visual attention and future motion for robot representation learning. We believe EgoHEP represents an important step toward proactive robot perception through learning from human videos. In future work, we will explore representations to guide proactive perception, enabling robots to select where to observe based on interaction needs.

AI USE STATEMENT

In this work, we used AI tools to assist with the execution of experimental evaluations and data preprocessing and cleaning. We did not use generative AI tools to formulate the research ideas or design the core algorithms. Additionally, we used AI tools to assist with manuscript polishing and to search for relevant literature. We have reviewed all AI-assisted work and finalized the manuscript based on our own judgment. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 01–13. IEEE, 2023.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, 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)*, pp. 7061–7071. IEEE, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: the dataset. In *European conference on computer vision*, pp. 753–771. Springer, 2018.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, 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*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, 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*, pp. 19383–19400, 2024.

Abhishek Gupta, Vikash Kumar, Corey Lynch, Sergey Levine, and Karol Hausman. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning. *arXiv preprint arXiv:1910.11956*, 2019.

Huiwon Jang, Dongyoung Kim, Junsu Kim, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Visual representation learning with stochastic frame prediction. *arXiv preprint arXiv:2406.07398*, 2024.

Roland S Johansson, Göran Westling, Anders Bäckström, and J Randall Flanagan. Eye–hand coordination in object manipulation. *the Journal of Neuroscience*, 21(17):6917–6932, 2001.

Siddharth Karamcheti, Suraj Nair, Annie S Chen, Thomas Kollar, Chelsea Finn, Dorsa Sadigh, and Percy Liang. Language-driven representation learning for robotics. *arXiv preprint arXiv:2302.12766*, 2023.

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.

Chengyang Li, Kaiyi Xiong, Yuan Xu, Lei Qian, Yizhou Wang, and Wentao Zhu. Gazevla: Learning human intention for robotic manipulation. *arXiv preprint arXiv:2604.22615*, 2026.

Gen Li, Yutong Chen, Yiqian Wu, Kaifeng Zhao, Marc Pollefeys, and Siyu Tang. Egom2p: Egocentric multimodal multitask pretraining. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10830–10843. IEEE, 2025.

Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *European Conference on Computer Vision*, pp. 639–655. Springer, 2018.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022.

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3272–3282. IEEE, 2022.

Yangcen Liu, Shuo Cheng, Xinchen Yin, Woo Chul Shin, Alfred Cueva, Yiran Yang, Zhenyang Chen, Chuye Zhang, and Danfei Xu. Egoengine: From egocentric human videos to high-fidelity dexterous robot demonstrations. *arXiv preprint arXiv:2606.12604*, 2026.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.

Yecheng Jason Ma, Vikash Kumar, Amy Zhang, Osbert Bastani, and Dinesh Jayaraman. Liv: Language-image representations and rewards for robotic control. In *International Conference on Machine Learning*, pp. 23301–23320. PMLR, 2023.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Advances in Neural Information Processing Systems*, 36:655–677, 2023.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38, 2026.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.

Mohan Kumar Srirama, Sudeep Dasari, Shikhar Bahl, and Abhinav Gupta. Hrp: Human affordances for robotic pre-training. *arXiv preprint arXiv:2407.18911*, 2024.

Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, et al. Learning manipulation by predicting interaction. *arXiv preprint arXiv:2406.00439*, 2024.

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)*, pp. 1805–1815. IEEE, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023a.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6586–6597. IEEE, 2023b.

Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

A Implementation Details

A.1 Pretraining Data and Supervision

We use Aria egocentric recordings from HOT3D and EgoExo4D. Each window contains 16 historical observations and 16 future motion labels sampled at 8 Hz. Scene images and hand crops are resized to 448×448 and 256×256, respectively.

Table 6: Pretraining data pools. Counts refer to windows, not videos.

Data pool	Hand-wrist pretraining	EgoHEP
HOT3D official	9,163	7,082
EgoExo4D official	1,166	1,166
EgoExo4D pseudo-labeled	24,514	—
Total	34,843	8,248

Label sources. For HOT3D, hand joints are obtained from the official MANO annotations, and wrist labels correspond to their root joints. For the EgoExo4D official pool, both hand and wrist labels come from the released 3D hand annotations. In the pseudo-labeled pool, hand joints are predicted by HaWoR, and wrist positions may be anchored using official body-wrist annotations; some frames are replaced by complete official hand annotations. Gaze labels come from the datasets' released eye-tracking estimates: projected gaze for HOT3D and released 2D gaze coordinates for EgoExo4D. They are not generated by HaWoR.

Hand-wrist pretraining uses 34,843 windows from 1,193 unique takes/videos. EgoHEP uses 8,248 official hand–wrist windows from 100 unique takes/videos, with a long-run HOT3D:EgoExo4D sampling ratio of 4:1 and no HaWoR pseudo-labels. Pseudo-label quality weights are heuristic scores based on detection confidence and wrist association; wrist agreement does not establish finger-joint accuracy. Invalid supervision is masked. At runtime, adjacent wrist displacements exceeding 0.5 m in the fixed reference camera frame invalidate the corresponding supervision without smoothing trajectories or changing timestamps.

The EgoExo4D validation set contains 343 windows. Twenty HOT3D takes are excluded from Stage II training, with 400 windows used for mechanism diagnostics. These takes were exposed during Hand-wrist pretraining and therefore are not unseen test data for the complete pretraining pipeline.

A.2 Two-Stage Training and Checkpoint Selection

Shared encoders and prediction pathways are inherited, and the EMA future denoiser is promoted to the trainable prediction branch. Attention, visual projection, and motion encoding modules are newly initialized. The final Stage II checkpoint is used for robot transfer; robot success rates are not used for checkpoint selection.

Both stages retain the following hand-wrist loss terms and coefficients:

$$\begin{aligned}\mathcal{L}_{\mathrm{HW}} = \mathcal{L}_{\mathrm{flow}} + 0.5\mathcal{L}_{x_0} + 0.15\mathcal{L}_{\Delta} + 0.05\mathcal{L}_{\Delta^2} \\ + 0.05\mathcal{L}_{\mathrm{bone}} + 0.05\mathcal{L}_{\mathrm{history}}.\end{aligned} \tag{5}$$

Here, $\mathcal{L}_{\mathrm{flow}}$ is the conditional flow-matching objective, and $\mathcal{L}_{\mathrm{rec}} \equiv \mathcal{L}_{x_0}$ supervises the reconstructed clean hand–wrist states. The dynamics term combines velocity, acceleration, and bone consistency:

$$\mathcal{L}_{\mathrm{dyn}} = \lambda_{\mathrm{v}}\mathcal{L}_{\mathrm{vel}} + \lambda_{\mathrm{a}}\mathcal{L}_{\mathrm{acc}} + \lambda_{\mathrm{b}}\mathcal{L}_{\mathrm{bone}}. \tag{6}$$

Here, $\mathcal{L}_{\mathrm{vel}} \equiv \mathcal{L}_{\Delta}$ and $\mathcal{L}_{\mathrm{acc}} \equiv \mathcal{L}_{\Delta^2}$. Their effective coefficients in the full hand–wrist objective are 0.15, 0.05, and 0.05 for the first-order, second-order, and bone-consistency terms, respectively. The first- and second-order constraints use frame differences rather than time-normalized physical velocity and acceleration. Hand and wrist terms are combined with weights 0.6 and 0.4. Hand-wrist pretraining assigns a coefficient of 0.30 to the pseudo-labeled loss.

EgoHEP adds gaze KL divergence, motion reconstruction, and visual-to-motion InfoNCE to the GT hand–wrist objective. Gaze and motion reconstruction each have weight 0.10. The alignment

weight is zero for the first 200 updates, increases linearly to 0.10 over the next 200 updates, and remains fixed thereafter. Gaze targets are 32×32 Gaussian heatmaps with standard deviation 3; visual pooling always uses predicted attention. H4/H8/H16 correspond nominally to 0.5/1/2 s, with one horizon selected cyclically per update. The InfoNCE temperature is 0.07.

Table 7: Stage II optimization settings. LN denotes LayerNorm.

Setting	Value
Global batch	48
Precision	BF16; FP32 loss reduction
Optimizer	AdamW, $\beta = (0.9, 0.95)$
Weight decay / clipping	0.05 / 1.0
HW-head learning rate	7.5×10^{-5}
Fusion/projection/motion LR	5×10^{-5}
Attention learning rate	10^{-4}
Backbone learning rate	2×10^{-6}
Schedule	100-update warmup, then cosine decay to 10% of the initial rate
Denoiser EMA	0.999

The ablations preserve the Stage II initialization, data, and training budget. *GT-HW* disables all auxiliary objectives, whereas *Gaze* retains only gaze KL with weight 0.10.

A.3 Downstream Simulation Protocol

We evaluate five Franka-Kitchen tasks (sliding door, hinged door, microwave, light, and knob) and five Meta-World tasks (Assembly, Bin-picking, Button-press-topdown, Drawer-open, and Hammer). Each task uses the first 25 complete demonstrations. Online inputs are limited to the current RGB image and robot proprioception; ground-truth gaze, future motion, and object states are not policy inputs.

Only the scene encoder is frozen. The attention head, MAP adapter, proprioception projector, and BN-linear action head are trained on robot demonstrations. Global visual features, attention-pooled local features, and proprioception are concatenated for action prediction.

Table 8: Simulation training and evaluation settings.

Setting	Franka-Kitchen	Meta-World
Batch size	32	32
Training updates	20,007	20,280
Episode horizon	50 steps	500 steps
Success criterion	`solved_steps`> 5	Final-step `info.success`

Both suites use a left camera view and independent resets. The action-head learning rate is 10^{-3}; attention, MAP, and proprioception modules use 10^{-5}. We use AdamW with $\beta = (0.9, 0.999)$, weight decay 0.01, and gradient clipping at 1.0. Final BN statistics are recomputed using training demonstrations only, and only final policies are evaluated.

A.4 Real-Robot Setup

Real-world experiments use an **AgileX Dual Nero** dual-arm robot with a **ZED Mini** as the primary-view camera. We evaluate three tasks: *Lift Plate*, *Place Pen*, and *Pass Banana*. Each task uses 50 training demonstrations and 20 evaluation episodes. Task-wise success rates are computed as the number of successful episodes divided by 20.