EGOJEPA: BEYOND RETARGETING FOR DEXTEROUS MANIPULATION FROM EGOCENTRIC VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Egocentric human videos offer a scalable alternative to costly robot demonstrations, yet transferring their manipulation priors to robot control remains challenging. The commonly adopted retargeting-based approaches convert human wrist and finger poses into embodiment-specific action commands through hand-crafted geometric mappings. However, such mappings require meticulous recalibration across different embodiments and risk propagating hand-tracking errors into downstream robot control pipelines. To this end, we propose **EgoJEPA**, a three-stage framework that learns embodiment-agnostic representations from egocentric human videos without rule-based human-to-robot retargeting: 1) *Human-action pre-training* learns effect-aware latent actions through JEPA-style visual prediction conditioned on hand-motion differences; 2) *Latent-action-based world modeling* jointly predicts future visual dynamics and latent actions, which decouples human prior pre-training from robot-specific control parameterizations; 3) *Embodiment-specific tuning* adapts the pretrained model to diverse robots through lightweight action adapters. Experiments on five real-world manipulation tasks across three robot embodiments demonstrate the effectiveness and efficiency of EgoJEPA. It achieves a 74% average success rate under the out-of-domain evaluation, compared with 65% for the retargeting baseline, while reducing cumulative cross-embodiment wall-clock cost by 47.2%, from 743.2 to 392.7 hours.

1 INTRODUCTION

Despite recent advances in robotic manipulation policies (Chi et al., 2025; Zhao et al., 2023; Kim et al., 2024), scaling such systems remains constrained by their reliance on task-specific robot demonstrations, typically obtained through teleoperation or specialized capture systems (Zhao et al., 2023; Khazatsky et al., 2024; Chi et al., 2024). Collecting these demonstrations is costly, labor-intensive, and time-consuming, posing a significant bottleneck to scalable robot learning. In contrast, *egocentric human videos* provide a scalable supervision source for transferring manipulation skills from human hand-object interactions to robotic systems (Ma et al., 2026b; Feng et al., 2026).

Earlier *action-free* methods operate without explicit hand-pose annotations, instead learning latent action representations directly from inter-frame visual dynamics (Ye et al., 2025; Wang et al., 2026b; Bjorck et al., 2025; Gao et al., 2026; Bharadhwaj et al., 2024). However, the resulting representations may capture generic visual variations rather than action-induced physical effects, limiting their interpretability and transferability to robot control. To move beyond learning solely from visual supervision, recent approaches explicitly leverage human hand-pose annotations and adopt a *retargeting-based pipeline* (Kareer et al., 2025; Yang et al., 2025; Tao et al., 2025; Ma et al., 2026a; Wang et al., 2026c). Specifically, human hand motions, such as wrist and finger poses, are captured using specialized devices (*e.g.*, Leap Motion (Weichert et al., 2013) or Aria glasses (Engel et al., 2023)) and subsequently mapped to a target robot action space through hand-crafted rules for policy pre-training or human-robot co-training.

Despite the intuitive formulation, the retargeting pipeline suffers from two key limitations: 1) *Embodiment-specific mapping*. Retargeting relies on low-level geometric or kinematic mappings tailored to a particular robot embodiment, rather than learning transferable action semantics. Consequently, such mappings are tightly coupled to the robot morphology and typically require meticulous recalibration whenever either changes. As shown in Figure 1 (a), human–robot calibration alone

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.

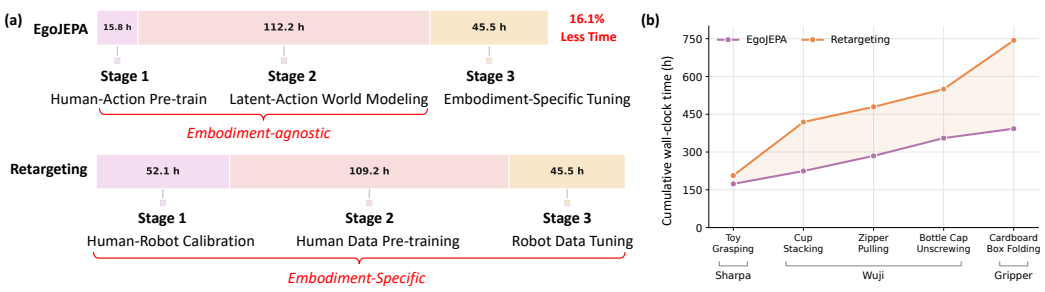

Figure 1: **Training efficiency and scalability of EgoJEPA *v.s.* retargeting.** **(a)** The retargeting pipeline requires embodiment-specific human-to-robot mappings and repeated calibration, whereas EgoJEPA learns an embodiment-agnostic latent action space in Stages 1–2 and performs embodiment-specific grounding only in Stage 3; **(b)** Cumulative training cost across robot embodiments. Reusing the embodiment-agnostic models from Stages 1–2 reduces the per-embodiment adaptation overhead of EgoJEPA relative to the retargeting pipeline.

takes 52.1 hours, accounting for 25.2% of the total pipeline time. Figure 1 (b) shows that this cost compounds when scaling to multiple robots, *i.e.*, retargeting must repeat the embodiment-specific human-robot calibration and pre-training whenever the robot morphology changes, which causes the cumulative wall-clock time to grow rapidly. 2) *Noise propagation*. Retargeting is highly sensitive to errors in estimated human motion. Hand-pose estimation noise caused by occlusions, viewpoint variations, or temporal jitter can propagate through the retargeting process, resulting in physically infeasible robot actions. As a concrete stress test, we inject Gaussian noise into human action annotations while keeping robot training unchanged. As shown in Figure 6 (a), under a mild per-axis position noise of $\sigma = 10$ mm, the retargeting pipeline on the task of `bottle cap unscrewing` drops drastically from 60% to 30% success rate, showing that annotation errors are directly amplified into downstream robot control failures.

To this end, we propose **EgoJEPA**, an **Ego**centric instantiation of the **J**oint **E**mbedding **P**redictive **A**rchitecture (Assran et al., 2023) that learns embodiment-agnostic latent action representations from human videos. As shown in Figure 1, EgoJEPA comprises three stages: 1) *Human-action pre-training* learns generic effect-aware latent actions through human-motion-conditioned visual prediction; 2) *Latent-action-based world modeling* jointly predicts future visual dynamics and the *latent* actions learned in Stage 1 via joint flow matching, rather than embodiment-specific robot commands; and 3) *Embodiment-specific tuning* adapts the pre-trained world model to robot action spaces through lightweight adapters. Evaluated on five real-world manipulation tasks across three robot embodiments (two dexterous hands and a gripper), EgoJEPA achieves the highest average success rate of 74% under the out-of-domain evaluation, outperforming the action-free method LAPA (Ye et al., 2025), the retargeting pipeline, and the prior egocentric human-video method DexWild (Tao et al., 2025). In addition, EgoJEPA significantly reduces the cost of scaling across robot embodiments. By reusing Stage 1–2 models and adapting only Stage 3 for each new robot, EgoJEPA reduces cumulative wall-clock cost from 743.2 to 392.7 hours over retargeting (*c.f.* Figure 1 (b)).

Our contributions are summarized as follows:

- We propose EgoJEPA, a JEPA-style framework that eliminates the need for rule-based human-to-robot retargeting by learning embodiment-agnostic representations from human videos.
- We introduce effect-aware latent actions as a transferable interface between human motion priors and robot control, enabling efficient policy adaptation across diverse embodiments via lightweight action adapters.
- Extensive experiments on five real-world manipulation tasks across three robot embodiments demonstrate the effectiveness and cross-embodiment scalability of EgoJEPA.

2 Related Work

WAMs for Dexterous Manipulation. WAMs move beyond action-only policies by incorporating future state prediction into the action learning process, thereby modeling the coupled evolu-

tion of visual dynamics and action control (Wang et al., 2026a; Shen et al., 2026; Li et al., 2025). While early WAMs mainly target low-DoF gripper manipulation, recent works have extended this paradigm toward dexterous manipulation (Goswami et al., 2025; Kim et al., 2026; Jiang et al.). DexWM (Goswami et al., 2025) conditions latent-state prediction on fine-grained hand keypoint displacements for dexterous interaction modeling. DWM (Kim et al., 2026) simulates dexterous behaviors using scene-action-conditioned video diffusion without explicitly decoding robot actions. However, existing dexterous WAMs rely on embodiment-specific robot demonstrations, limiting cross-embodiment transfer and scalability. EgoJEPA instead learns transferable manipulation priors from egocentric human videos, enabling efficient adaptation to diverse robots.

Robot Learning from Human Videos. Due to the high cost of teleoperated robot demonstrations, recent work has increasingly turned to human-centric robot learning, which leverages abundant human videos as scalable semantic and physical priors (Ma et al., 2026b; Feng et al., 2026). Existing approaches can be broadly grouped into two categories. 1) *Latent-based methods* learn action-related representations from *action-free* videos by encoding inter-frame changes (Ye et al., 2025; Wang et al., 2026b; Bjorck et al., 2025; Gao et al., 2026) or future visual dynamics (Wu et al., 2024; Cheang et al., 2024; Bharadhwaj et al., 2024). However, lacking explicit action grounding, such latents may conflate controllable hand motion with camera motion and passive scene dynamics, making them less interpretable as action representations. 2) *Retargeting-based methods* recover explicit human motion, *e.g.*, wrist poses and finger poses captured by devices like Leap Motion or Aria glasses, and map them to robot-specific action spaces for pre-training or co-training (Kareer et al., 2025; Yang et al., 2025; Tao et al., 2025; Ma et al., 2026a; Wang et al., 2026c; Zheng et al., 2026). Although these approaches provide more structured supervision, they remain tied to embodiment-specific geometric mappings and are sensitive to hand-tracking noise. To this end, EgoJEPA encodes explicit human hand motions into effect-aware latent actions, enabling transfer to robot control without rule-based retargeting while improving robustness to hand-tracking noise. While concurrent work such as Being-H0 (Luo et al., 2025) adopts a similar spirit by discretizing MANO trajectories into motion tokens via a VQ-VAE structure, its motion tokens rely exclusively on isolated hand-pose sequences without grounding the tokenizer in the visual context. Consequently, such tokenizers are agnostic to visual interaction semantics, failing to capture the critical contact dynamics and spatial alignment between hands and target objects.

Joint-Embedding Predictive Architectures. JEPAs formulate latent world modeling by predicting the future evolution of a system in a compact representation space rather than reconstructing raw observations (LeCun et al., 2022). Early works apply JEPA to self-supervised representation learning, where models learn semantic representations by predicting latent embeddings of masked image or video patches (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Recent studies further explore JEPA for action-conditioned latent world modeling (Assran et al., 2025; Zhou et al., 2024). LeWorldModel (Maes et al., 2026) extends JEPA into an end-to-end action-conditioned world model via SIGReg regularization (Bardes et al., 2021). DINO-WM (Zhou et al., 2024) instead leverages pretrained visual representations to learn latent dynamics by predicting future semantic features. Following this spirit, EgoJEPA investigates how human motion can serve as a physical condition for latent world modeling, which enables adaptation to diverse robot embodiments.

3 EGOJEPA

We propose EgoJEPA, a framework that learns embodiment-agnostic action representations from egocentric human videos and efficiently adapts them to downstream robotic manipulation tasks. As illustrated in Figure 2, EgoJEPA comprises three key stages: human-action pre-training (Stage 1), latent-action-based world modeling (Stage 2), and embodiment-specific tuning (Stage 3).

3.1 STAGE 1: HUMAN-ACTION PRE-TRAINING

Instead of relying on rule-based, embodiment-specific retargeting pipelines, EgoJEPA first learns an effect-aware latent space grounded in *human-action*-conditioned visual prediction. Specifically, we define the "human action" as the difference between adjacent-frame hand annotations, such as 2D/3D keypoints, MANO parameters (Romero et al., 2022), or other hand-pose representations. This difference reflects the local hand motion responsible for visual state transitions, providing an embodiment-agnostic physical latent which can be later adapted to different robot embodiments.

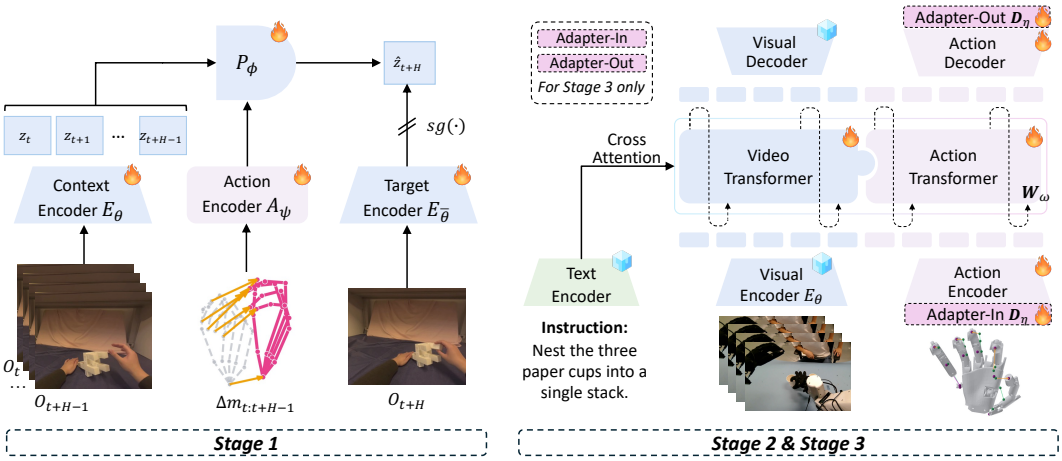

Figure 2: **Overview of the EgoJEPA framework. Stage 1**: An effect-aware latent action space is learned from ego-centric human video data without embodiment-specific action re-targeting. Specifically, the predictor $\boldsymbol{P}_\phi$ predicts the future representation $\hat{z}_{t+H}$ from the visual context encoded by $\boldsymbol{E}_\theta$ and the hand-motion differences $\Delta m_{t:t+H-1}$ encoded by $\boldsymbol{A}_\psi$; **Stage 2**: A world-action model $\boldsymbol{W}_\omega$ is trained on ego-centric human video data via joint flow matching to model future visual dynamics and latent human actions extracted by the frozen action encoder $\boldsymbol{A}_\psi$; **Stage 3**: Using robot demonstration data, the learned action representation is grounded in a target robot embodiment by inserting embodiment-specific input and output adapters $\boldsymbol{D}_{\eta_r}$ into the action branch, which map between relative robot actions and the shared latent action space.

Let $\mathcal{D}_{\text{ego}}$ denote an egocentric human video dataset, where each trajectory contains RGB observations $\{o_t\}_{t=1}^{T}$ and aligned hand annotations $\{m_t\}_{t=1}^{T}$. We instantiate the *human action* over a horizon H as adjacent-frame hand annotation differences.

$$\Delta m_{t:t+H-1} = \big(m_{i+1} - m_i\big)_{i=t}^{t+H-1}. \tag{1}$$

We then optimize a human-action-conditioned JEPA objective (Assran et al., 2023) that predicts the target visual representation from visual context and hand-motion differences, thereby grounding the learned action-effect latent in visual state transitions.

$$\underset{\phi,\psi,\theta}{\text{minimize}} \left\| \boldsymbol{P}_\phi\Big(\big(\boldsymbol{E}_\theta(o_i)\big)_{i=t}^{t+H-1}, \boldsymbol{A}_\psi(\Delta m_{t:t+H-1})\Big) - \text{sg}\big(\boldsymbol{E}_{\bar{\theta}}(o_{t+H})\big) \right\|_1, \tag{2}$$

where $\boldsymbol{E}_\theta$ and $\boldsymbol{E}_{\bar{\theta}}$ denote the online and target visual encoders, respectively, with $\bar{\theta}$ updated by an exponential moving average of θ. The action encoder $\boldsymbol{A}_\psi$ maps the human action chunk $\Delta m_{t:t+H-1}$ into the latent action embedding. The predictor $\boldsymbol{P}_\phi$ then combines this action condition with the context representations $(\boldsymbol{E}_\theta(o_i))_{i=t}^{t+H-1}$ to predict the target representation at time $t+H$. The stop-gradient operator $\text{sg}(\cdot)$ prevents gradients from flowing through the target branch.

3.2 Stage 2: Latent-Action-Based World Modeling

Building on Stage 1, we train a *latent-action world model* to jointly predict future visual dynamics and the corresponding latent actions. Notably, the action branch is supervised by frozen JEPA action embeddings $\boldsymbol{A}_\psi(\Delta m)$, avoiding the reliance on retargeted coordinate-level robot actions.

Given an egocentric trajectory, we first construct the *clean* targets for the future visual dynamics and the latent action sequences as follows.

$$z_1 = \boldsymbol{E}_\theta(o_{t+1:t+H}), \qquad u_1 = \boldsymbol{A}_\psi(\Delta m_{t:t+H-1}), \tag{3}$$

where z_1 represents the visual dynamics, while u_1 denotes the corresponding latent action embedding. $\boldsymbol{E}_\theta$ and $\boldsymbol{A}_\psi$ denote the visual and action encoders adopted from the Stage 1 pre-trained model, respectively. We then formulate latent-action world modeling as a joint flow matching problem over future visual representations and latent actions. Specifically, for a flow timestep $\tau \in [0,1]$, we construct the noisy joint state by linear interpolation:

$$z_\tau = \tau z_1 + (1-\tau) z_0, \qquad u_\tau = \tau u_1 + (1-\tau) u_0, \tag{4}$$

where $z_0 \sim \mathcal{N}(0, I)$ and $u_0 \sim \mathcal{N}(0, I)$ are Gaussian noises with the same dimensionalities as z_1 and u_1. The world-action model $\boldsymbol{W}_\omega$ is trained to predict the velocity field from the noisy joint state to the clean video-action target $[z_1, u_1]$:

$$\mathcal{L}_{\text{Stage2}}(\omega) = \mathbb{E}_{z,u,\tau}\left[\left\|\boldsymbol{W}_\omega\left([z_\tau, u_\tau], q, \tau\right) - \left([z_1, u_1] - [z_0, u_0]\right)\right\|_2^2\right], \tag{5}$$

where q denotes the language instruction, and $[z_\tau, u_\tau]$ denotes the noisy output token sequence at flow timestep τ, consisting of future visual tokens z_τ and latent action tokens u_τ.

3.3 Stage 3: Embodiment-Specific Tuning

In Stage 3, EgoJEPA grounds the shared action representation into a target robot embodiment. Specifically, we initialize the world-action model $\boldsymbol{W}_\omega$ from Stage 2 and insert lightweight embodiment-specific adapters into its action branch. As shown in Figure 2 (right), the *input adapter* maps robot action tokens into the shared action-token space of the pretrained WAM, while the *output adapter* maps the WAM action-channel output back to the target robot action space.

To align the robot supervision with the relative hand-motion representations learned in Stage 1, we express robot actions as temporal differences. Given a sequence of absolute robot states $\{a_i\}_{i=t}^{t+H}$, the delta-action chunk is defined as follows.

$$\Delta a_{t:t+H-1} = \{a_{i+1} - a_i\}_{i=t}^{t+H-1}, \tag{6}$$

where H denotes the action prediction horizon. Let o_t denote the visual observation at time t, and let $o_{t+1:t+H}$ denote the subsequent RGB observation sequence over a horizon of H. We construct the clean future visual target and robot delta-action target as

$$z_1 = \boldsymbol{E}_\theta(o_{t+1:t+H}), \qquad \Delta a_1 = \Delta a_{t:t+H-1}, \tag{7}$$

where z_1 denotes the future visual dynamics of the robot trajectory, and Δa_1 denotes the corresponding executable relative robot action chunk. We then construct the noisy visual and action states through linear interpolation:

$$z_\tau = \tau z_1 + (1-\tau) z_0, \qquad \Delta a_\tau = \tau \Delta a_1 + (1-\tau)\Delta a_0, \tag{8}$$

where $z_0 \sim \mathcal{N}(0, I)$ and $\Delta a_0 \sim \mathcal{N}(0, I)$ are the sampled Gaussian noises.

Let $\boldsymbol{W}_\omega$ denote the pretrained world-action model from Stage 2, and let $\boldsymbol{D}_\eta$ denote the action adapter for the specific robot embodiment. The Stage 3 objective is formulated as follows.

$$\mathcal{L}_{\text{Stage3}}(\omega, \eta) = \mathbb{E}_{z,\Delta a,\tau}\left[\left\|\boldsymbol{D}_\eta \circ \boldsymbol{W}_\omega\left([z_\tau, \Delta a_\tau], q, \tau\right) - \left([z_1, \Delta a_1] - [z_0, \Delta a_0]\right)\right\|_2^2\right], \tag{9}$$

where q denotes the language instruction, and $[z_\tau, \Delta a_\tau]$ denotes the noisy output token sequence composed of future visual tokens and robot delta-action tokens. We use $\boldsymbol{D}_\eta^r \circ \boldsymbol{W}_\omega$ to denote the adapted world-action model, where the embodiment-specific adapter is applied to the action channel while the visual encoder $\boldsymbol{E}_\theta$ is kept unchanged.

4 Experiments

4.1 Experimental Setup

Human Pre-training Data. We use a commercially acquired egocentric human dataset for stages 1 and 2. The dataset contains 1,111,943 episodes from 62,518 recording sessions, amounting to approximately 20,106.6 hours and 2.17 billion video frames. It covers 64,686 task instructions and 608 scene labels across diverse household and industrial activities. Each episode consists of a monocular RGB video recorded at approximately 30Hz, together with camera calibration and motion annotations comprising wrist poses and 21 three-dimensional keypoints for each hand.

Robot Post-training Tasks. We evaluate EgoJEPA on five real-world manipulation tasks spanning three robot embodiments including a Sharpa dexterous hand (Sharpa), a Wuji dexterous hand (Wuji Technology), and a Xense Tac UMI gripper (Xense Robotics). Each task is associated with a task-specific robot demonstration dataset for Stage 3 post-training: 1) *Toy grasping* uses the Sharpa hand

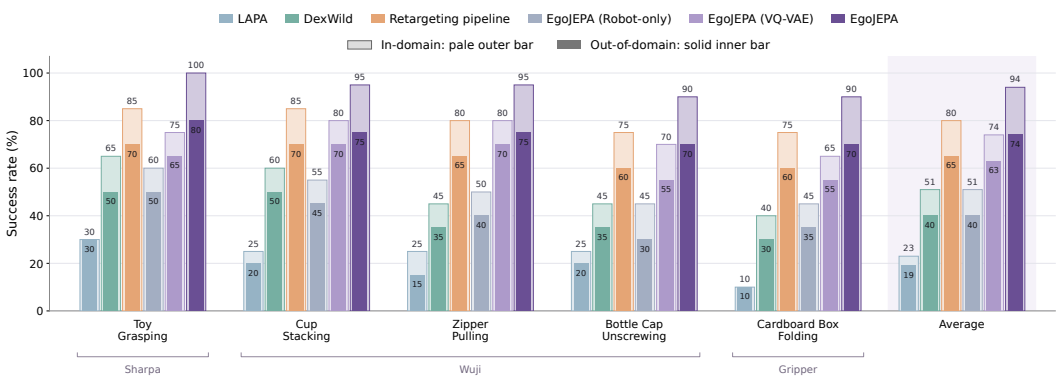

Figure 3: **In-domain and out-of-domain manipulation performance** across five tasks and three embodiments including Sharpa, Wuji, and Gripper. Bars report success rates over 20 trials per task, with the average success rate across tasks shown on the right.

to grasp a plush toy; 2) *Cup stacking* uses the Wuji hand to arrange cups into a target stack, requiring sequential manipulation and precise placement; 3) *Zipper pulling* uses the Wuji hand to pinch and slide a zipper pull; 4) *Bottle cap unscrewing* uses the Wuji hand to uncap a bottle through coordinated grasping; 5) *Cardboard box folding* uses the Xense Tac UMI gripper to fold a paper box into a target configuration, requiring multi-step manipulation. See Appendix A.1 for detailed specifications on both human pre-training and robot post-training datasets.

Implementation Details. In Stage 1, the context and target visual encoders are initialized from the pretrained Wan2.2 VAE encoder (Wan et al., 2025), and the predictor is a 24-layer Transformer with frame-causal attention. In Stage 2, the video Transformer is initialized from Wan2.2-TI2V-5B (Wan et al., 2025), while the Stage 1 action embeddings u_1 are average-pooled along the feature dimension to obtain 128-dimensional latent action targets for stable training. In Stage 3, the input and output action adapters each comprise a two-layer MLP with a ReLU activation between the layers. Full training configurations are provided in Table A3.

Evaluation Protocol. We evaluate all policies under both in-domain and out-of-domain settings across five tasks and three robot embodiments. In-domain evaluation uses the same object instances and spatial layouts as Stage 3 training. Out-of-domain evaluation considers two forms of distribution shift: 1) *object shifts*, using unseen object instances with different shapes, sizes, or materials while preserving task requirements; 2) *spatial shifts*, placing objects at positions and orientations outside the Stage 3 training distribution but within the robot's reachable workspace. All policies are evaluated without additional fine-tuning under matched test configurations. We conduct 20 trials per task for each setting and define success as full task completion.

Compared Methods. We evaluate EgoJEPA against five competitive methods and architectural variants: 1) *LAPA* (Ye et al., 2025), an action-free method that learns discrete latent actions directly from inter-frame visual changes for policy pre-training; 2) *DexWild* (Tao et al., 2025), an egocentric human-video method that co-trains dexterous manipulation policies on in-the-wild videos and robot demonstrations; 3) *Retargeting-based pipeline*, which shares our architecture but replaces the learned latent actions with embodiment-specific robot targets derived via geometric retargeting, followed by fine-tuning on robot data; 4) *EgoJEPA (Robot-only)*, an ablation that omits Stage 1–2 human-video pre-training and is trained strictly on robot demonstrations; and 5) *EgoJEPA (VQ-VAE)*, a variant that replaces Stage 1 JEPA pre-training with a VQ-VAE reconstruction objective (Van Den Oord et al., 2017) over hand-motion annotations, while maintaining identical Stage 2–3 protocols. Refer to Appendix A.2 for more implementation details.

4.2 Real-World Performance

We present in-domain and out-of-domain manipulation results in Figure 3, highlighting three key findings: 1) *State-of-the-art performance.* EgoJEPA attains the highest success rate on all five tasks, with average success rates of 94% and 74% under in- and out-of-domain settings, respectively, sur-

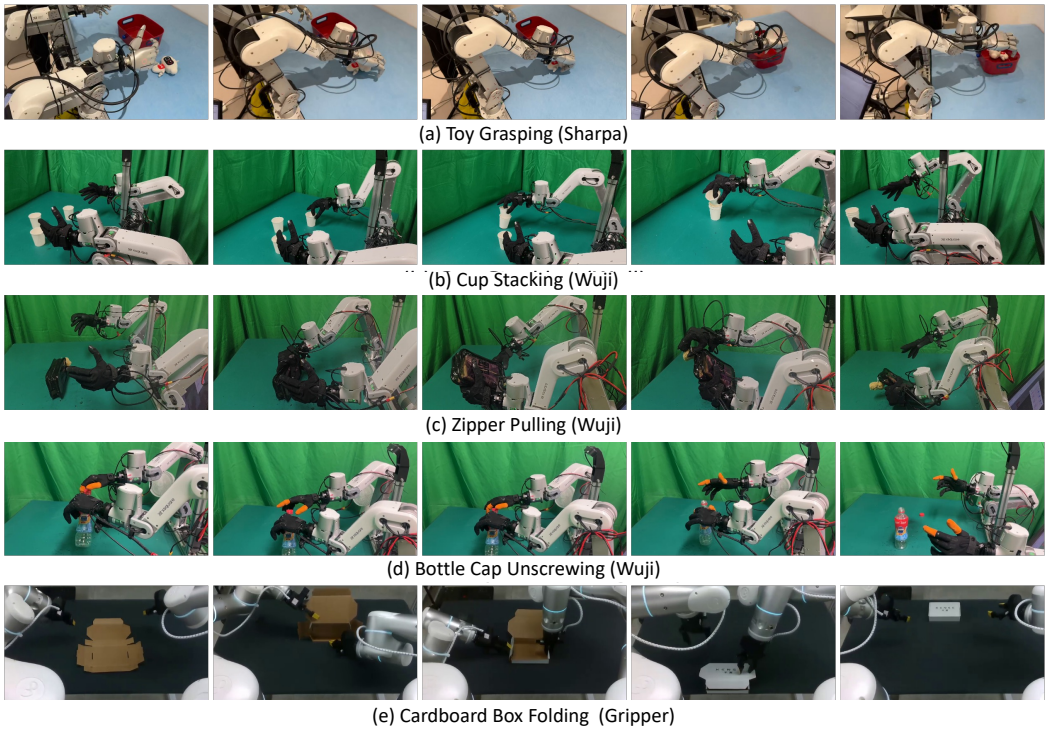

Figure 4: **Qualitative results** on five real-world robotic manipulation tasks.

passing the retargeting pipeline by 14% and 9% absolute margins. 2) *Generalization gap.* All methods exhibit lower average success rates under out-of-domain evaluation, highlighting the challenges posed by unseen object instances and spatial layouts. Nevertheless, EgoJEPA achieves the highest out-of-domain success rate of 74%, compared with 65% for retargeting and 40% for DexWild. 3) *Task-level variation.* `Bottle cap unscrewing` and `cardboard box folding` yield EgoJEPA's lowest out-of-domain success rates. The former requires coordinated bimanual control to stabilize and rotate the bottle, while the latter involves multi-step operations with strict geometric alignments. These constraints increase sensitivity to contact noise and error accumulation, making reliable completion challenging.

4.3 Ablation Studies

We conduct ablation studies to evaluate EgoJEPA's architectural design and data efficiency. All policy evaluations follow the out-of-domain setting.

Benefits of Human Pre-training. To assess the contribution of human pre-training to downstream manipulation, we compare EgoJEPA with a robot-only variant, *i.e.*, EgoJEPA (Robot-only) in Figure 3, which retains the same architecture but omits human-data training in Stages 1–2 and is trained directly on robot demonstrations. As shown, human pre-training increases the average out-of-domain success rate from 40% to 74%, with improvements across all five tasks. The largest gain occurs on `bottle cap unscrewing`, with success increasing from 30% to 70%. This task requires coordinated bimanual control to stabilize the bottle while rotating the cap, suggesting the value of human manipulation priors for learning coordinated actions. We further examine Stage 3 training dynamics on `bottle cap unscrewing` by comparing action loss over training steps between EgoJEPA and its robot-only variant. As shown in Figure 5 (left), EgoJEPA converges faster and achieves substantially lower training loss, suggesting that priors learned from egocentric human videos facilitate embodiment-specific adaptation under limited robot supervision.

JEPA *vs.* VQ-VAE. To assess the benefit of JEPA-based predictive learning, we compare EgoJEPA with EgoJEPA (VQ-VAE), which replaces Stage 1 predictive learning with VQ-VAE-based action reconstruction (Van Den Oord et al., 2017), using the same data and downstream training protocol.

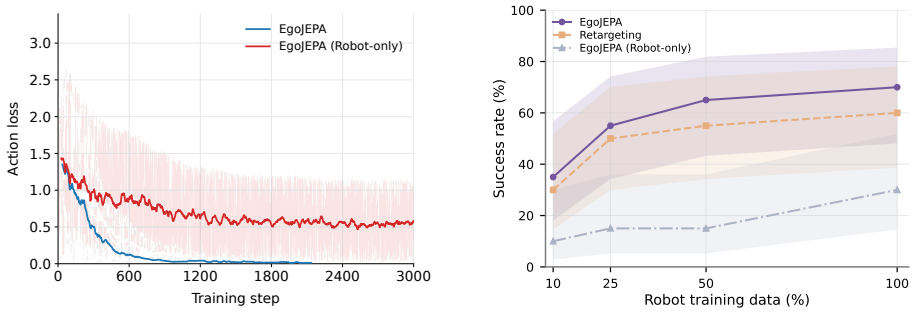

Figure 5: **Left: Training convergence comparisons.** Action loss during Stage 3 training on robot data for EgoJEPA and its robot-only variant. **Right: Robot data efficiency analysis.** Success rates on the task of `bottle cap unscrewing` with varying fractions of robot demonstrations. Shaded regions indicate 95% Wilson confidence intervals based on 20 evaluation trials per setting.

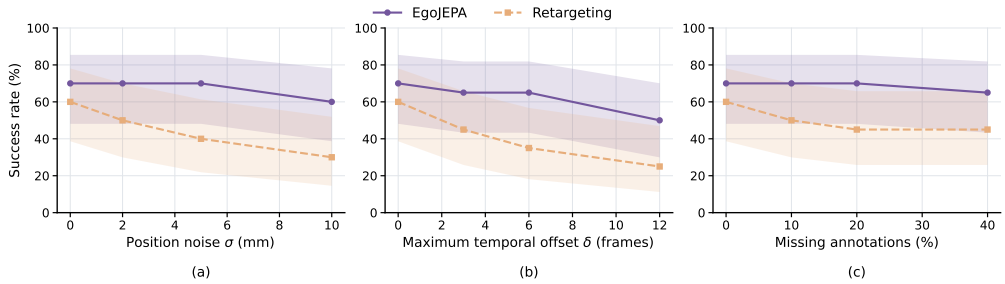

Figure 6: **Robustness analysis to human action annotation errors.** The success rate on `bottle cap unscrewing` with increasing (a) position noise standard deviation σ, (b) maximum temporal offset δ and (c) missing-annotation fraction p in human pre-training data (stage 1 & 2).

As shown in Figure 3, EgoJEPA increases the average out-of-domain success rate from 63% to 74%, with gains across all five tasks. These results suggest that learning the visual consequences of human motion yields more effective representations for downstream control than reconstructing motion alone. Refer to Appendix A.2 for more details.

Robot Data Efficiency. We evaluate data efficiency on `bottle cap unscrewing` using 10%, 25%, 50%, and 100% of the robot demonstration dataset during adaptation. As shown in Figure 5 (right), EgoJEPA consistently outperforms both robot-only training (*i.e.*, EgoJEPA (Robot-only)) and the retargeting baseline across all data budgets. Notably, with only 10% of the demonstrations, EgoJEPA achieves a 35% success rate, surpassing the 30% attained by the robot-only model trained on the full dataset (100%). Similarly, using just 50% of the demonstrations, EgoJEPA reaches a 65% success rate, outperforming retargeting using all demonstrations (60%). These results demonstrate that pre-trained effect-aware representations enable substantially more data-efficient policy adaptation, significantly reducing reliance on costly robot teleoperation data.

Robustness to Human Annotation Errors. Hand-motion annotations in human videos are susceptible to localization errors, temporal misalignment, and tracking failures, which can propagate through retargeting into robot action supervision. To evaluate robustness against noisy human pre-training data, we independently introduce three types of perturbation to hand annotations during Stages 1–2: 1) *Position noise*: additive Gaussian noise with per-axis standard deviation $\sigma \in \{0, 2, 5, 10\}$ mm; 2) *Temporal jitter*: random frame shifts relative to original video frames, sampled uniformly from $\{-\delta, \ldots, \delta\}$ with maximum temporal offset $\delta \in \{0, 3, 6, 12\}$ frames; and 3) *Annotation dropout*: masking contiguous annotation segments covering a fraction $p \in \{0\%, 10\%, 20\%, 40\%\}$ of total frames, with missing keypoints linearly interpolated. All variants are evaluated on `bottle cap unscrewing`, keeping Stage 3 robot demonstration data and training settings identical across experiments. See Appendix A.3 for further implementation details.

As illustrated in Figure 6, EgoJEPA demonstrates significantly higher resilience to all three annotation error types compared with the retargeting pipeline. Under the maximum position noise ($\sigma = 10$ mm), EgoJEPA experiences a modest 10% success-rate drop (70% to 60%), whereas retargeting suffers a drastic 30% decline (60% to 30%). Among the evaluated perturbations, temporal

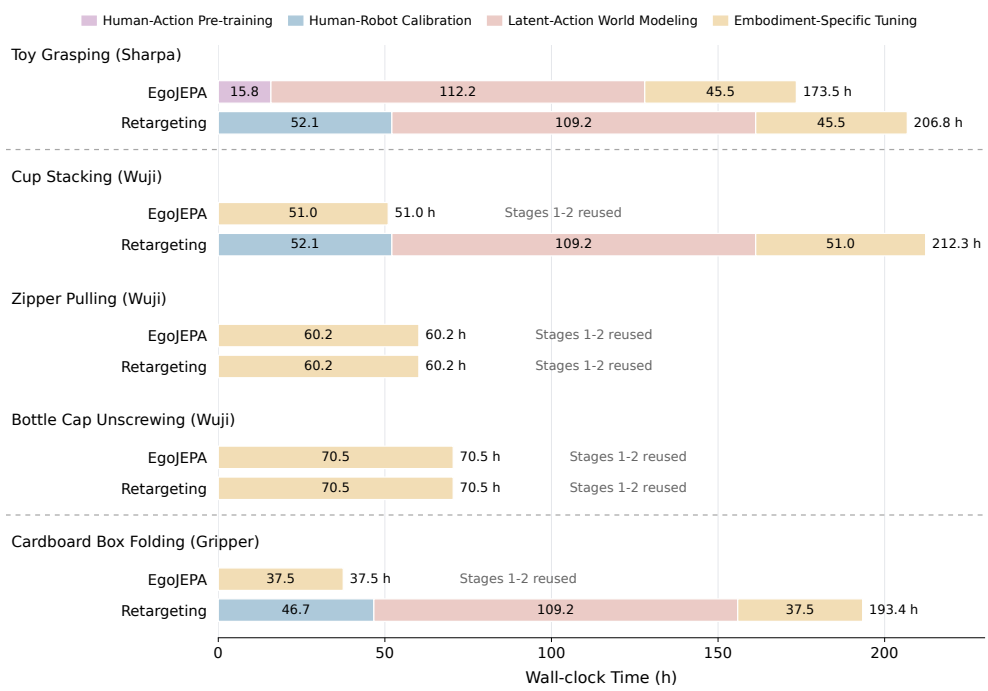

Figure 7: **Training efficiency across five tasks and three robot embodiments.** Stacked bars show the wall-clock time required for each task, with colors denoting training stages. EgoJEPA achieves greater training efficiency by reusing human pre-training (Stages 1–2) across robot embodiments.

jitter induces the most severe performance decay for both models, while annotation dropout exhibits the mildest impact. Specifically, at the highest perturbation levels, EgoJEPA's success rate degrades to 50% under temporal jitter ($\delta = 12$) but maintains 65% under annotation dropout ($p = 40\%$).

Cross-Embodiment Efficiency. To further evaluate the computational efficiency, we record the wall-clock adaptation costs across five tasks and three distinct robot embodiments. Wall-clock time is reported rather than GPU hours because the human-robot calibration step in the retargeting baseline is executed entirely on CPUs, making real-time wall-clock duration a more consistent and fair measure of end-to-end deployment cost. As illustrated in Figure 7, EgoJEPA seamlessly reuses pre-trained human action representations (Stages 1–2) across all downstream embodiments. In contrast, the retargeting pipeline requires re-running expensive calibration and latent world modeling whenever a new robot embodiment is introduced. For instance, when adapting to the Xense Tac UMI gripper for `cardboard box folding`, EgoJEPA requires only 37.5 hours of Stage 3 tuning, compared with 193.4 hours for retargeting (including 46.7 hours of CPU-based calibration and 109.2 hours of GPU-based world modeling). Across all five tasks, EgoJEPA reduces cumulative wall-clock training time from 743.2 hours to 392.7 hours, achieving a 47.2% overall cost reduction. These results highlight EgoJEPA's substantial time and compute savings when transferring shared, embodiment-agnostic action priors to new robotic platforms.

5 Conclusion

We introduced EgoJEPA, a three-stage framework that transfers manipulation priors from egocentric human videos to robot control without rule-based retargeting. By grounding latent actions in visual effects and integrating them into world modeling, EgoJEPA decouples human pre-training from embodiment-specific control, enabling shared pretrained models to be reused across robots through lightweight action adapters. Experiments across five real-world tasks and three embodiments demonstrate that EgoJEPA outperforms the retargeting pipeline by an absolute margin of 9% under out-of-domain evaluation while reducing cumulative cross-embodiment training time by 47.2%. In addition, EgoJEPA exhibits greater robustness to human annotation noise, supporting scalable learning from imperfectly annotated egocentric videos.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature searches, manuscript polishing, and experimental design. We have not used generative AI tools to generate synthetic datasets or implement the proposed method. Developing theoretical models, formulating mathematical claims, and writing proofs are not applicable to this work. We reviewed and revised the AI-assisted contents against experimental records, and verified references against their original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2024.

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

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. *arXiv preprint arXiv:2402.10329*, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

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.

Zhiyuan Feng, Qixiu Li, Huizhi Liang, Rushuai Yang, Yichao Shen, Zhiying Du, Zhaowei Zhang, Yu Deng, Li Zhao, Hao Zhao, et al. From human videos to robot manipulation: A survey on scalable vision-language-action learning with human-centric data. *arXiv preprint arXiv:2606.00054*, 2026.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models for learning dexterous hand-object interactions from human videos. *arXiv preprint arXiv:2512.13644*, 2025.

Taoran Jiang, Liqian Ma, Yixuan Guan, Jiaojiao Meng, Weihang Chen, Zecui Zeng, Lusong Li, Dan Wu, Jing Xu, and Rui Chen. Dexsim2real2: Building explicit world model for precise articulated object dexterous manipulation, 2024. *URL https://arxiv. org/abs/2409.08750*.

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.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29663–29673, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

Juncheng Ma, Jianxin Bi, Yufan Deng, Xuanran Zhai, Kewei Zhang, Ye Huang, Bo Liang, Shukai Gong, Jiankai Tu, Xiaotian Tang, et al. Humanscale: Egocentric human video can outperform real-robot data for embodied pretraining. *arXiv preprint arXiv:2606.20521*, 2026a.

Junyi Ma, Erhang Zhang, Haoran Yang, Ditao Li, Chenyang Xu, Guangming Wang, and Hesheng Wang. Robot learning from human videos: A survey. *arXiv preprint arXiv:2604.27621*, 2026b.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

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

Sharpa. Sharpa Wave. `https://www.sharpa.com/pages/wave`.

Qiuhong Shen, Shihua Zhang, Yue Liao, Qi Li, Zhenxiong Tan, Shizun Wang, Shuicheng Yan, and Xinchao Wang. World action models: A survey. *arXiv preprint arXiv:2606.20781*, 2026.

Tony Tao, Mohan Kumar Srirama, Jason Jingzhou Liu, Kenneth Shaw, and Deepak Pathak. Dexwild: Dexterous human interactions for in-the-wild robot policies. *arXiv preprint arXiv:2505.07813*, 2025.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, et al. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*, 2026a.

Yuqi Wang, Xinghang Li, Wenxuan Wang, Junbo Zhang, Yingyan Li, Yuntao Chen, Xinlong Wang, and Zhaoxiang Zhang. Unified vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 80929–80944, 2026b.

Zhi Wang, Botao He, Kelin Yu, Seungjae Lee, Ruohan Gao, Furong Huang, and Yiannis Aloimonos. Humanego: Zero-shot robot learning from minutes of human egocentric videos. *arXiv preprint arXiv:2605.24934*, 2026c.

Frank Weichert, Daniel Bachmann, Bartholomäus Rudak, and Denis Fisseler. Analysis of the accuracy and robustness of the leap motion controller. *Sensors*, 13(5):6380–6393, 2013.

Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, volume 2024, pp. 10641–10662, 2024.

Wuji Technology. Wuji Hand. `https://wuji.tech/en/hand`.

Xense Robotics. XTac-UMI G1. `https://docs.xenserobotics.com/en/docs/latest/xtac/`.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pp. 28213–28239, 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*, 2023.

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.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

Table A1: **Summary statistics of the human pre-training dataset.** Dataset scale, recording specifications, task and scene coverage, and hand annotations.

Statistic	Value
Episodes	1,111,943
Video Frames	2,169,344,852
Total Duration	≈20, 106.6 hours
Frame Rate	≈29.97 Hz
Camera Views	1
Mean Episode Duration	≈65 s
Median Episode Duration	≈30 s
Distinct Task Instructions	64,686
Scene Labels	608
3D Keypoints per Hand	21

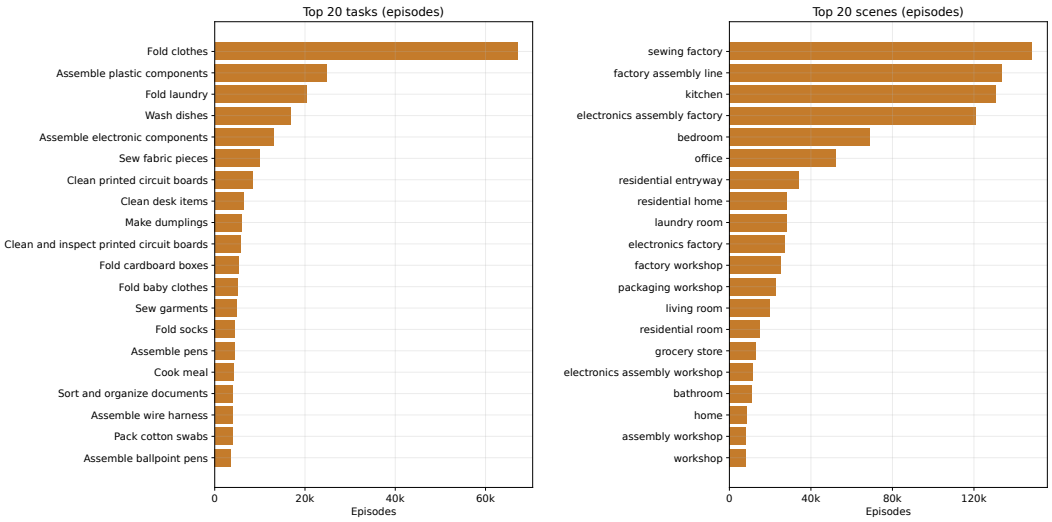

Figure A1: **Task and scene distributions of the human pre-training dataset.** The 20 most frequent tasks (left) and scenes (right), ranked by episode count.

A Appendix

A.1 Dataset Details

Human Pre-training Data Statistics. Table A1 summarizes the dataset scale, recording specifications, and annotation coverage. Figure A1 presents the 20 most frequent tasks and scenes, covering household activities and industrial operations such as clothes folding, dish washing, and component assembly. Figure A2 shows the range of episode durations and frame counts, encompassing both brief interactions and extended manipulation sequences. Figure A3 further illustrates representative egocentric video sequences with aligned hand-keypoint annotations, providing a qualitative view of the human pre-training data.

Robot Demonstration Statistics. Table A2 summarizes the embodiment, action dimensionality, episode count, and average duration for each task. The datasets cover five manipulation tasks using the Sharpa hand, Wuji hand, and Xense Tac UMI gripper, with action dimensions of 56, 52, and 14, respectively. Demonstration durations vary with task demands, ranging from brief grasping interactions to longer sequences involving coordinated manipulation.

Data Licensing and Privacy Safeguards. To ensure strict regulatory compliance and ethical standards, all third-party and commercially purchased datasets utilized in our experiments are managed under rigorous licensing and privacy protocols:

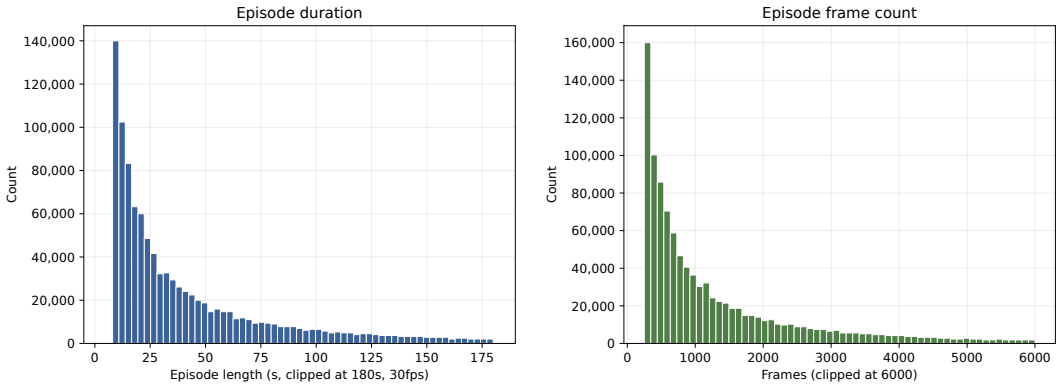

Figure A2: **Episode length distributions of the human pre-training dataset.** Histograms show episode durations in seconds (left) and frame counts (right).

Table A2: **Task-wise statistics** of the robot demonstration datasets.

Task	Embodiment	Action Dim.	# Episode	Avg. Length (s)
Plush Toy Grasping	Sharpa Hand	56	750	6.9
Cup Stacking	Wuji Hand	52	916	13.6
Zipper Manipulation	Wuji Hand	52	623	15.5
Bottle Cap Unscrewing	Wuji Hand	52	635	23.8
Paper-Box Folding	Xense Tac UMI gripper	14	843	297.6

- *Commercial Licensing & Usage Scope:* All commercially acquired datasets were legally licensed from standard commercial vendors under explicit agreements that grant authorization for non-commercial academic research and algorithmic evaluation. The usage fully complies with the intellectual property rights and End User License Agreements (EULA) specified by the vendors.
- *Privacy Safeguards & Anonymization:* Prior to model training and representation learning, all datasets underwent thorough privacy verification. Human subjects appearing in egocentric videos and annotations are strictly anonymized, with personally identifiable information (PII), *e.g.* facial features, license plates, and sensitive background information, either blurred or removed by the data providers in accordance with global privacy regulations (*e.g.*, GDPR).
- *Data Security & Responsible AI:* All purchased assets are securely stored on restricted-access servers with role-based access control to prevent unauthorized redistribution or leakages. No biometric identifiers or sensitive personal characteristics were collected, inferred, or processed.

A.2 Implementation Details

Training Configurations of EgoJEPA. Table A3 summarizes the data, model initialization, and optimization settings for all three stages. Stages 1–2 use egocentric human data, while Stage 3 initializes from the Stage 2 checkpoint and jointly tunes the video and action Transformers with embodiment-specific adapters on robot data. Stage 3 additionally incorporates three camera views and proprioceptive conditioning. All stages use AdamW, with stage-specific learning rates, schedules, and batch sizes as detailed in the table.

Implementations of EgoJEPA (VQ-VAE). We detail the implementation and training procedure of EgoJEPA (VQ-VAE), which replaces Stage 1 JEPA pre-training with VQ-VAE-based hand-motion reconstruction (Van Den Oord et al., 2017). Let $\mathcal{D}_{\text{ego}}$ denote an egocentric human video dataset, where each trajectory contains RGB observations $\{o_t\}_{t=1}^{T}$ and aligned hand annotations $\{m_t\}_{t=1}^{T}$. A VQ-VAE encoder f_{enc} maps each hand annotation m_t to a continuous representation $\mathbf{h}_t$, which is quantized using a learnable codebook $\mathcal{C} = \{\mathbf{e}_k\}_{k=1}^{K}$ as follows:

$$\mathbf{q}_t = \mathbf{h}_t + \text{sg}(\mathbf{e}_{k_t} - \mathbf{h}_t), \quad s.t. \quad k_t = \arg\min_{k \in \{1,\dots,K\}} \|\mathbf{h}_t - \mathbf{e}_k\|_2^2, \tag{A1}$$

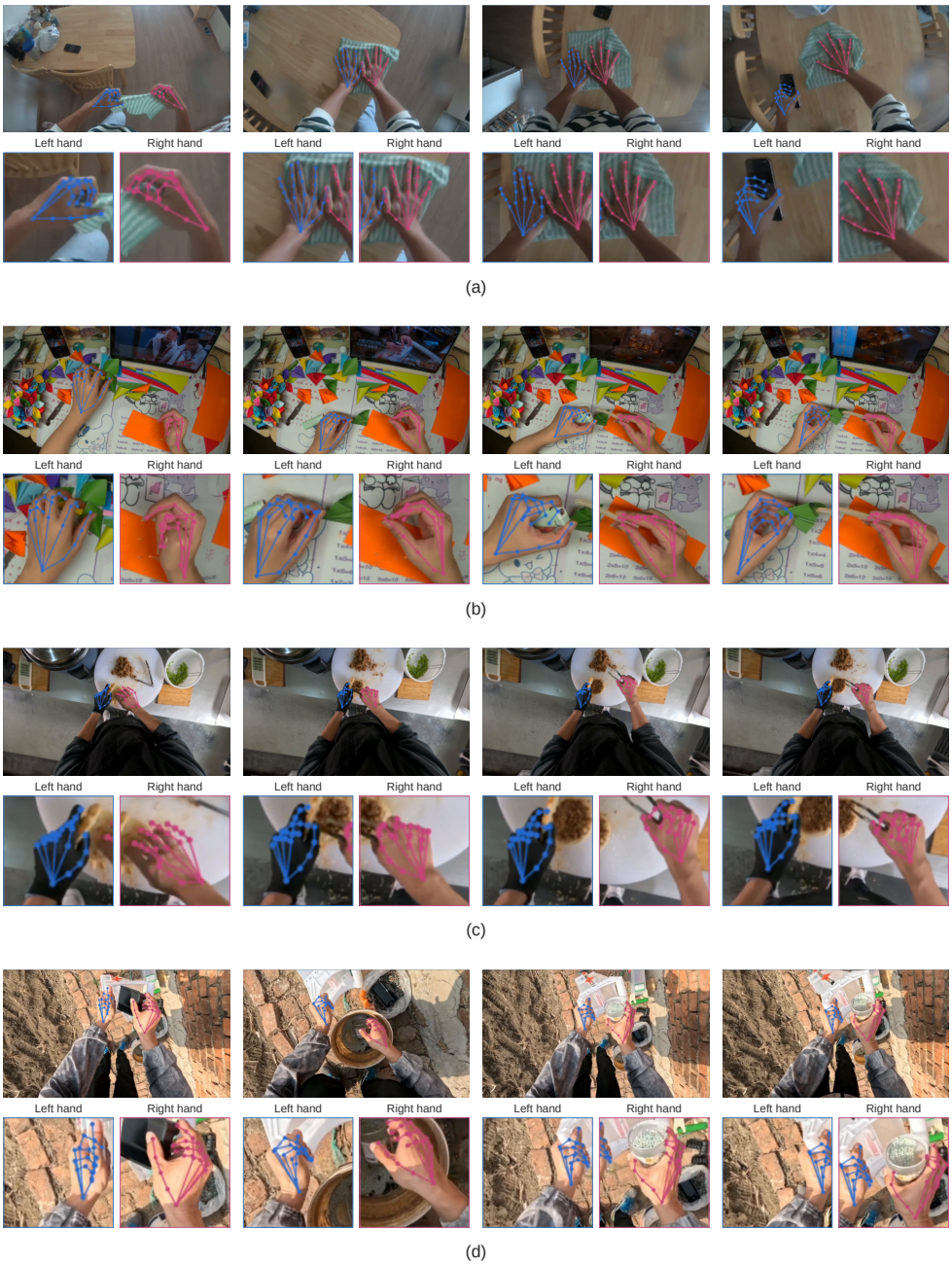

Figure A3: **Examples of egocentric human videos and hand annotations.** Each sequence shows sampled frames with enlarged hand crops, where blue and pink keypoints denote the left and right hands, respectively.

where $\mathrm{sg}(\cdot)$ is a straight-through estimator for encoder gradients, which preserves its input during the forward pass and blocks gradients during backpropagation.

Then, a VQ-VAE decoder f_{dec} maps the quantized representation $\mathbf{q}_t$ back to the hand annotation space, yielding $\hat{m}_t = f_{\mathrm{dec}}(\mathbf{q}_t)$. The overall training objective is as follows:

$$\mathcal{L}_{\mathrm{VQ}} = \mathbb{E}_{\tau \sim \mathcal{D}_{\mathrm{ego}},\, t} \left[\underbrace{\|m_t - \hat{m}_t\|_2^2}_{\mathcal{L}_{\mathrm{rec}}} + \underbrace{\|\,\mathrm{sg}(\mathbf{h}_t) - \mathbf{e}_{k_t}\|_2^2}_{\mathcal{L}_{\mathrm{codebook}}} + \underbrace{\|\mathbf{h}_t - \mathrm{sg}(\mathbf{e}_{k_t})\|_2^2}_{\mathcal{L}_{\mathrm{commit}}} \right], \tag{A2}$$

Table A3: **Training configurations** for the three stages of EgoJEPA.

Configuration	Stage 1	Stage 2	Stage 3
Training Data	Ego Data	Ego Data	Robot Data
Initialization	Wan2.2 VAE	Wan2.2-TI2V-5B	Stage 2 Checkpoint
Trainable Modules	All Modules	Video DiT + Action DiT	Video DiT + Action DiT + Adapters
Input Resolution	256×256	192×224	192×224
Camera Views	1	1	3
Proprioceptive Conditioning	No	No	Yes
Action Prediction Horizon	–	32	32
Latent Action Dimension	1024	128	128
Robot Data Dimension	–	–	56/52/14
Number of GPUs	32	64	64
Per-device Batch Size	128	128	64
Steps	3,500	6,000	3,000
Optimizer	AdamW	AdamW	AdamW
AdamW $\boldsymbol{\beta}$	$(0.9, 0.999)$	$(0.95, 0.999)$	$(0.95, 0.999)$
Learning Rate	4.25×10^{-4}	3×10^{-5}	1×10^{-4}
Minimum Learning Rate	0	1×10^{-6}	1×10^{-6}
LR Schedule	Linear	Cosine	Cosine
Weight Decay	0.04	0.01	0.01
Maximum Gradient Norm	–	1.0	1.0

This VQ-VAE objective learns discrete representations of per-frame hand annotations without visual-effect supervision. It replaces only Stage 1 human-action pre-training, while the training protocols for Stages 2 and 3 remain unchanged.

Implementations of Retargeting Pipeline. We convert human wrist poses and 3D hand keypoints into embodiment-specific robot action targets. For the Sharpa and Wuji hands, we use their respective official retargeting implementations[1][2], adapting the input keypoint ordering and coordinate conventions to each interface. Wrist poses are mapped to robot end-effector poses through calibrated coordinate transformations and tool offsets. For the Xense Tac UMI gripper, the thumb–index fingertip distance is mapped to the gripper opening using a calibrated, range-constrained mapping. Retargeted trajectories are aligned with the video frames and converted into relative actions consistent with the robot demonstrations.

A.3 Robustness Evaluation Protocol

We detail the experimental settings for evaluating robustness to human annotation errors. We independently introduce position noise, temporal jitter, and annotation dropout into human pre-training annotations, using identical perturbations for EgoJEPA and the retargeting pipeline. Stage 3 training remains unchanged, and both methods are evaluated on `bottle cap unscrewing` under the out-of-domain setting.

1) Position Noise. To evaluate robustness to localization errors in human hand-motion annotations, we perturb the annotated 3D positions with additive Gaussian noise. For each frame t, a translation vector is sampled independently and applied consistently to all keypoints of that hand:

$$\widetilde{\mathbf{p}}_{t,j} = \mathbf{p}_{t,j} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}\big(\mathbf{0}, \sigma^2 \mathbf{I}_3\big)\,, \tag{A3}$$

where $\mathbf{p}_{t,j} \in \mathbb{R}^3$ denotes the position of keypoint j, and σ is the noise standard deviation along each coordinate axis. The same translation is applied to the wrist position, preserving the hand geometry and leaving rotation annotations unchanged. We set $\sigma \in \{0, 2, 5, 10\}$ mm, with $\sigma = 0$ denoting clean annotations. Perturbations are introduced before computing inter-frame motion differences or performing retargeting. Both methods are pretrained using the same perturbed annotations and subsequently adapted and evaluated using clean robot data.

2) Temporal Jitter. To evaluate robustness to temporal misalignment between human videos and hand-motion annotations, we randomly shift the annotation sequence relative to the video. For each

[1] `https://github.com/sharpa-robotics/sharpa-vision-retargeting-sdk`
[2] `https://github.com/wuji-technology/wuji-retargeting`

training clip, an integer offset δ is sampled uniformly:

$$\widetilde{\mathbf{m}}_t = \mathbf{m}_{t+\delta}, \qquad \delta \sim \mathrm{Uniform}\{-k, \ldots, k\}, \tag{A4}$$

where $\mathbf{m}_t$ denotes the hand-motion annotation at frame t, and k controls the maximum temporal offset. The offset is shared across both hands and held constant within each clip, preserving motion continuity while disrupting visual–motion alignment. We set $k \in \{0, 3, 6, 12\}$ frames, corresponding to maximum offsets of approximately $\{0, 100, 200, 400\}$ ms at 30 Hz. Perturbations are applied before temporal subsampling and action extraction, using only clips with sufficient surrounding frames to accommodate the shifts. Both methods use the same perturbed annotations for human pre-training, followed by adaptation and evaluation on clean robot data.

3) Annotation Dropout. To evaluate robustness to missing hand-motion annotations caused by occlusion or tracking failures, we randomly mask contiguous annotation segments within each training clip. Let $\mathcal{D}$ denote the set of masked frames. The perturbed annotations are obtained as

$$\widetilde{\mathbf{m}}_t = \begin{cases} \mathbf{m}_t, & t \notin \mathcal{D}, \\ \mathrm{Interp}(\mathbf{m}_{t^-}, \mathbf{m}_{t^+}; t), & t \in \mathcal{D}, \end{cases} \tag{A5}$$

where t^- and t^+ are the nearest valid frames before and after the missing segment. Positions are reconstructed by linear interpolation, and rotations by spherical linear interpolation. Both the retargeting pipeline and our EgoJPEA use identical masks and interpolation rules for human pre-training, followed by adaptation and evaluation on clean robot data.

A.4 Broader Impact and Limitations

Broader Impact. This work advances generalist robot learning by presenting an embodiment-agnostic framework that effectively leverages large-scale, low-cost human egocentric videos. By bypassing tedious manual geometric retargeting and repeated human-robot calibrations, EgoJEPA significantly lowers the computational and labor barriers to deploying multi-embodiment dexterous robotic systems in real-world environments, such as domestic assistance and industrial automation. However, scaling learning from uncurated video datasets carries the risk of inheriting or amplifying biases present in human demonstration environments. Furthermore, as autonomous dexterous manipulation systems become more capable in physical deployment, safety, alignment, and strict physical containment protocols must be rigorously established to prevent unintended hardware collisions or operational hazards.

Limitations. Although EgoJEPA demonstrates strong generalization and cross-embodiment efficiency for dexterous manipulation, several limitations remain: *(1) Performance bounds under severe action quality issues:* While EgoJEPA exhibits noise robustness, extreme degradation in egocentric video actions (*e.g.*, severe temporal misalignments or prolonged occlusions) can still distort latent representations and bottleneck downstream task performance. *(2) Contact-intensive dynamics:* Performance drops on out-of-domain tasks that require fine-grained contact dynamics (*e.g.*, `bottle cap unscrewing`). Integrating explicit interaction modeling or force-feedback priors into the latent action space remains a key direction for future work.