Institute of Trustworthy Embodied AI
Fudan University

ActiveMimic: Egocentric Video Pretraining with Active Perception

Xingyao Lin[1,2] **Guojin Zhong**[1] **Tianyi Lu**[1]
Ziyi Ye[1] **Yichen Zhu**[3] **Zuxuan Wu**[1,2,4] **Yu-Gang Jiang**[1]
[1] Fudan University, [2] Shanghai Innovation Institute, [3] Current Robotics, [4] NeoteAI

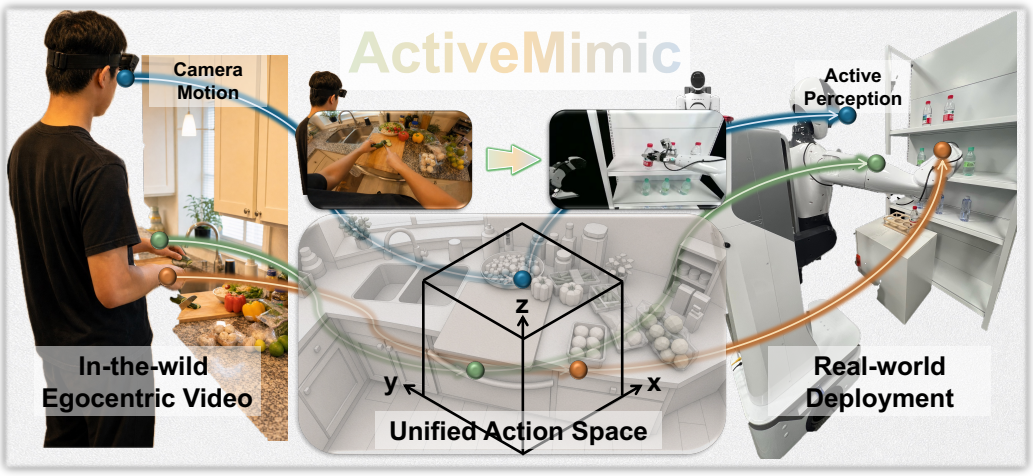

Figure 1: **ActiveMimic acquires active perception from in-the-wild egocentric human video and transfers it to real-world humanoid robots.** *Left to center:* egocentric camera motion and wrist action together form a 27-dimensional unified action representation that enables the model to jointly learn active perception and manipulation. *Center to right:* active perception is transferred to a humanoid robot, which repositions its viewpoint actively during task execution.

Abstract

Egocentric human video offers a scalable alternative to robot data for pretraining, yet models pretrained on such video consistently underperform those pretrained on robot data. We attribute this gap to a missing signal, the active perception behavior in egocentric videos, where humans continuously reposition their viewpoint during manipulation, inducing camera motion that standard pipelines treat as noise. To address this, we present **ActiveMimic**, a pretraining framework that recovers synchronized camera and wrist trajectories from a single body-worn RGB camera, models camera motion as a viewpoint action, and jointly learns active perception and manipulation from in-the-wild egocentric human video before adapting to a target robot. Empirically, real-world experiments across tasks with diverse active perception demands show that ActiveMimic consistently surpasses baselines pretrained on human video and matches state-of-the-art models pretrained on robot data. Further analysis provides evidence that active perception capability originates from egocentric human video pretraining rather than robot-specific fine-tuning, confirming active perception as the key to unlocking egocentric human video for robot pretraining.

Keywords: Robot Manipulation, Egocentric Human Video, Active Perception

Project Page: https://activemimic.github.io/

1 Introduction

Robot foundation models have become a central paradigm in robotic manipulation [1, 2, 3, 4, 5, 6]. A common training strategy combines a Vision-Language Model (VLM) with an action expert [7, 8], pretrains on large-scale robot data [9], and adapts to downstream tasks. However, robot data remains expensive to collect, difficult to scale, and limited in task diversity. Instead, egocentric human videos offer a scalable alternative, being cheaper to acquire, covering a broader range of daily activities, and are easy to scale. While appealing, models pretrained on egocentric human data consistently underperform those pretrained on robot data.

Existing studies attribute this gap to the absence of action supervision and focus on constructing proxy action labels, such as hand trajectories [10, 11], hand point clouds [12], or object motion signals [13]. These approaches, however, miss a key signal: during manipulation, humans continuously reposition their viewpoint through head and body movements, inducing substantial camera motion in egocentric videos that standard pipelines treat as noise. In this paper, we argue that explicitly modeling this active perception behavior [14, 15, 16] is key to unlocking egocentric human video for robot pretraining.

More specifically, modeling active perception requires recovering synchronized camera and wrist trajectories from egocentric human videos. However, wrist motion recovered from such videos inevitably conflates hand movement with camera rotation and translation, resulting in an inherent camera and hand coupling. Without resolving this coupling, a model cannot correctly learn either camera motion or hand motion. While existing methods that decouple camera motion and hand motion rely on dedicated capture hardware beyond a single body-worn RGB camera [17, 18, 19, 20], preventing them from scaling to in-the-wild video, our goal is to resolve this coupling without specialized hardware, computing synchronized camera and wrist trajectories using off-the-shelf vision models alone and producing a unified action representation that captures how perception and manipulation jointly evolve.

With this in mind, we introduce ActiveMimic, a pretraining framework that models viewpoint and wrist motion so as to perceive and act in an active manner. In particular, ActiveMimic derives a unified action representation encoding the viewpoint motion of the camera alongside the bimanual wrist motion, all expressed in a common reference frame, allowing the model to learn their relationships through a single flow matching objective. We compute this unified action space on Ego4D [21], a large-scale egocentric dataset covering diverse daily activities and hand-object manipulation. It is worth noting that the approach is general and can be extended to any in-the-wild egocentric data. Once camera and wrist actions are aligned, we pretrain the model to predict both camera and wrist actions from egocentric observations, learning active perception jointly with manipulation. Finally, the pretrained model is adapted to the target robotic embodiment using robot-specific data, transferring the active perception capability acquired during pretraining.

Real-world experiments on tasks spanning diverse active perception demands show that ActiveMimic consistently surpasses baselines pretrained on human video and matches state-of-the-art models pretrained on robot data. Our analysis further reveals that active perception originates from egocentric video pretraining rather than robot-specific fine-tuning, and that camera motion supervision facilitates representational transfer from human perception to robot control.

In summary, our contributions are threefold. **(a) ActiveMimic: an active-perception-aware pretraining framework for in-the-wild egocentric video.** We extract synchronized camera and wrist trajectories from egocentric human video and jointly model active perception with manipulation, enabling scalable pretraining without dedicated capture hardware. **(b) Active perception is the key to unlocking egocentric human video for robot pretraining.** Real-world experiments demonstrate that camera motion supervision consistently improves success rate across tasks with diverse active perception demands. **(c) Active perception originates from pretraining and transfers from human to robot.** We show that active perception is acquired during egocentric pretraining rather than robot fine-tuning, and that camera motion supervision facilitates representational transfer from human perception to robot control.

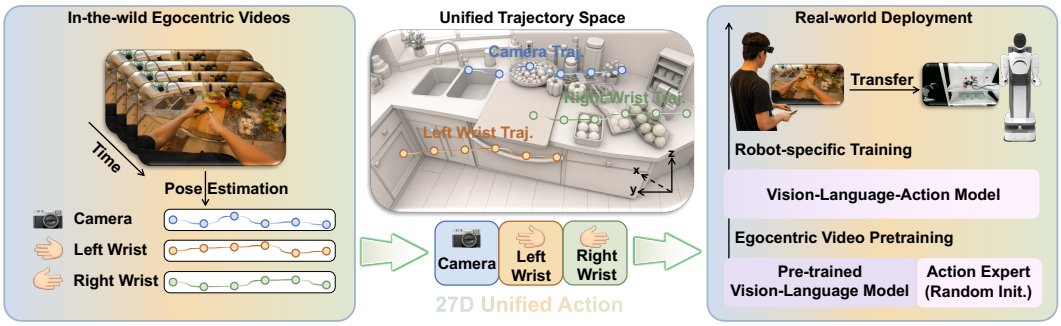

Figure 2: **Overview of ActiveMimic.** *Left:* recovering synchronized camera and wrist trajectories from a single body-worn RGB camera. *Middle:* resolving camera-wrist coupling and encoding as a unified 27D action. *Right:* pretraining on the 27D action to jointly model active perception and manipulation, then adapting to the target robot.

2 Related Work

Learning from human videos Human videos offer a cheaper, more scalable, and more diverse alternative to robot data for pretraining. One line of work estimates proxy action labels from such videos, including hand trajectories [10, 11], hand point clouds [12], and object motion signals [13], but supervises only hand or object motion and offers no signal about viewpoint action. A complementary line reads both viewpoint and hand actions directly from dedicated capture hardware [17, 18], restoring action supervision at the cost of additional cameras [22] or wearable sensors [19, 20, 23] beyond a single body-worn RGB camera, which restricts its applicability to in-the-wild egocentric human video.

Active perception A long-standing problem in robotics and computer vision is active perception [14, 15, 16], where an agent actively controls its viewpoint to reduce perceptual uncertainty rather than passively receiving images. Classically, it has been studied as Next-Best-View planning [24, 25, 26, 27, 28, 29], where viewpoint selection is optimized independently of downstream manipulation. Recent work instead models camera motion and manipulation actions jointly within a shared action space, learning perception and action end-to-end [30, 31, 32, 33, 34, 35, 36]. This new paradigm, however, relies on dedicated data collection with human-operated capture rigs, ranging from VR headsets and controllers for robot teleoperation [32, 30] to wearable devices that record head and hand poses [19], and therefore cannot leverage in-the-wild egocentric human videos as a web-scale, organically produced corpus, analogous to the readily available web data that has driven the rapid scaling of modern VLM and LLM pretraining.

Learning active perception from egocentric human videos Among models trained on egocentric human videos [17, 37, 38], the dominant line supervises only proxy hand or object labels and leaves active perception unmodeled, while work that learns active perception from human data relies on additional cameras [22] or wearable sensors [19, 20] beyond a single body-worn RGB camera rather than on in-the-wild egocentric video. In contrast, ActiveMimic introduces a purely vision-based approach that recovers camera and wrist trajectories jointly from a single body-worn RGB camera, striking a balance between fidelity and scalability that enables active perception to be trained together with manipulation in a unified action space on in-the-wild egocentric videos.

3 Method

ActiveMimic reframes egocentric human video pretraining around the coupled evolution of active perception and manipulation. Rather than treating egocentric camera motion as incidental noise, we interpret it as a viewpoint action that reflects how humans actively position their viewpoint during task execution. Starting from raw egocentric videos, we recover temporally aligned camera and wrist trajectories and represent them in a unified trajectory space, enabling joint modeling of

active perception and manipulation (Sec. 3.1). The model is then trained on this structured signal to predict both camera and wrist action from egocentric observations, enabling it to acquire transferable perceptual representations prior to adaptation to the target robotic embodiment (Sec. 3.2).

3.1 From Egocentric Video to Unified Action Space

From a single body-worn RGB camera, we recover synchronized camera and wrist trajectories that jointly describe active perception and manipulation, without requiring additional sensors or controlled capture conditions. This involves three steps: recovering camera and wrist trajectories from RGB frames using off-the-shelf vision models, resolving the camera and wrist coupling by re-expressing all poses in a common reference frame, and encoding the result as a unified 27-dimensional action representation. In particular, we consider Ego4D [21], a large-scale egocentric dataset covering diverse daily activities; details of dataset filtering, temporal segmentation, and instruction annotation are provided in Sec. A.2.

Recovering camera and wrist trajectories. For each egocentric video, we estimate three synchronized pose trajectories using off-the-shelf vision models: the egocentric camera trajectory and the left and right wrist trajectories. We denote by $T_{\text{ref}}^{\text{tgt}} \in SE(3)$ the rigid transformation of the target frame expressed in the reference frame. For each frame $k \in \{1, \dots, K\}$ in an episode of K frames, we estimate the egocentric camera pose $T_{\text{cam}_1}^{\text{cam}_k}$, expressed in the coordinate system of the camera at the first frame of the episode, together with the left and right wrist poses $T_{\text{cam}_k}^{\text{wrist}_k^L}$ and $T_{\text{cam}_k}^{\text{wrist}_k^R}$, expressed in the coordinate system of the current-frame camera. The egocentric camera trajectory serves as the operational realization of the viewpoint action introduced earlier; rigidly attached to the wearer, it encodes active perception independent of the mounting configuration (head-, chest-, or glasses-mounted). Wrist poses are estimated by SAM-3D-Body [39]. The camera trajectory is recovered by VGGT [40] as a scale-normalized path $\tilde{T}_{\text{cam}_1}^{\text{cam}_k}$, whose translational component is determined only up to a global scale factor. To recover the metric scale, we align the per-pixel depth maps from VGGT with metric depth estimates from UniDepth [41] via a median depth ratio, aggregated into an episode-level scale factor λ. The metric camera trajectory $T_{\text{cam}_1}^{\text{cam}_k}$ is obtained by scaling the translational component of $\tilde{T}_{\text{cam}_1}^{\text{cam}_k}$ by λ while keeping its rotation unchanged. Details of the scale recovery procedure are provided in Sec. A.1.

Resolving camera and wrist coupling. The recovered camera and wrist trajectories are coupled: wrist poses are expressed in the current-frame camera coordinate system while the camera trajectory is anchored to the first frame, so any displacement in the wrist poses between frames reflects both actual wrist movement and the camera's own rotation and translation; using these wrist poses directly as action supervision would therefore conflate wrist movement with camera motion. We resolve this coupling by re-expressing all poses in a common chunk-relative reference frame. Since the policy operates on fixed-length temporal chunks rather than full episodes, we re-center all poses for each chunk. For a chunk of length H, let i denote its start frame in the episode-level index and let $\tau \in \{0, \dots, H-1\}$ denote the chunk-local offset. The corresponding episode-level frame index is $k = i + \tau$. We re-center all camera poses in this chunk to the coordinate system cam_i of the chunk's first frame. The chunk-relative camera pose at offset τ is

$$T_{\text{cam}_i}^{\text{cam}_{i+\tau}} = \left(T_{\text{cam}_1}^{\text{cam}_i}\right)^{-1} T_{\text{cam}_1}^{\text{cam}_{i+\tau}}, \tag{1}$$

and the wrist poses are followed by composing the chunk-relative camera pose at offset τ with the current-frame wrist estimates,

$$T_{\text{cam}_i}^{\text{wrist}_{i+\tau}^L} = T_{\text{cam}_i}^{\text{cam}_{i+\tau}}\, T_{\text{cam}_{i+\tau}}^{\text{wrist}_{i+\tau}^L}, \tag{2}$$

and analogously for the right wrist. This construction decouples camera and wrist motions by placing them in a single spatial reference frame cam_i.

27D action representation. The decoupled chunk-relative poses are encoded into a unified 27-dimensional action vector that jointly captures viewpoint action and bimanual manipulation. Each chunk-relative pose, written in homogeneous form as

$$T = \begin{bmatrix} R & t \\ \mathbf{0}^\top & 1 \end{bmatrix}, \qquad R \in SO(3),\ t \in \mathbb{R}^3, \tag{3}$$

is encoded by its translation and a continuous 6D rotation representation [42]:

$$p = t \in \mathbb{R}^3, \qquad r_{6D} = \left[R_{:,1};\, R_{:,2}\right] \in \mathbb{R}^6, \tag{4}$$

where $R_{:,j}$ denotes the j-th column of R. Concatenating the camera and both wrist encodings yields a unified chunk-relative action vector for each chunk start i and offset τ:

$$a_{i,\tau} = \big[\underbrace{p^{\text{cam}}_{i,\tau},\, r^{\text{cam}}_{6D,i,\tau}}_{\text{camera (9D)}},\, \underbrace{p^{\text{wrist}^L}_{i,\tau},\, r^{\text{wrist}^L}_{6D,i,\tau}}_{\text{left wrist (9D)}},\, \underbrace{p^{\text{wrist}^R}_{i,\tau},\, r^{\text{wrist}^R}_{6D,i,\tau}}_{\text{right wrist (9D)}}\big] \in \mathbb{R}^{27}. \tag{5}$$

This unified 27D action space enables the model to jointly learn the coupled dynamics of camera and wrist motion within a single prediction objective.

3.2 Architecture and Training Strategy

With the decoupled camera and wrist action from Sec. 3.1, we introduce a two-stage training strategy that injects active perception capability into the model. We first describe the model architecture, then detail the training strategy.

Architecture and training objective. The architecture of ActiveMimic adopts a mix-of-transformers design [43, 44, 3] that combines a visual-language prefix with an action-expert suffix. The visual-language prefix encodes images and a tokenized prompt into a multimodal context, onto which the action expert attends together with the current state and a continuous time variable to predict a chunk of future continuous actions. The policy is trained with a conditional flow-matching objective. The loss is defined as

$$\mathcal{L} = \mathbb{E}_{a,\epsilon,t}\left\|v_t(a_t, o) - (\epsilon - a)\right\|_2^2, \tag{6}$$

where $a_t = t\epsilon + (1-t)a$ is a noisy sample of the clean action chunk a, Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, time step $t \sim \mathcal{U}(0, 1)$, and o denotes the overall conditioning context. The action chunk a refers to the 27D unified action defined in Sec. 3.1 during egocentric human video pretraining and to the robot action chunk during robot-specific fine-tuning. At inference, the prefix representation is encoded once and cached, and the action chunk is recovered by initializing from Gaussian noise and iteratively denoising via Euler integration along the learned velocity field.

Two-stage training. Training follows a two-stage recipe: an egocentric human video pretraining stage on the dataset constructed in Sec. 3.1, followed by a robot-specific training stage that adapts the pretrained policy to the target robotic embodiment. During pretraining, the visual-language prefix is initialized from a pretrained VLM checkpoint [45] while the action expert is initialized at random, and the policy is supervised with the chunk-relative camera and wrist targets, so that it learns to model active perception jointly with manipulation from large-scale egocentric human video. The subsequent robot-specific training stage retains the same architecture and is initialized entirely from the pretrained weights, training on robot-specific data to transfer the active perception capability acquired during pretraining to the robotic embodiment.

4 Experiments

We structure our evaluation around four questions that together assess whether active perception is the key to unlocking egocentric human video for robot pretraining. **Q1** (Sec. 4.2). Does camera motion supervision improve real-world task performance? **Q2** (Sec. 4.3). Do the camera and wrist trajectories recovered from egocentric video carry effective pretraining signals? **Q3** (Sec. 4.4). Does active perception come from egocentric pretraining, and how does the model use it? **Q4** (Sec. 4.5). Does camera motion supervision enable human-to-robot representational transfer?

4.1 Experimental Setup

Robot platform. We conduct all real-world experiments on a humanoid upper-body robot (AGIBOT G1) equipped with a 2-DoF head, a 2-DoF waist, and two 7-DoF arms with parallel-jaw grippers. The robot observes through three RGB cameras: one head-mounted and two wrist-mounted. The head camera, together with the head and waist joints, forms the active perception subsystem that enables viewpoint repositioning during task execution.

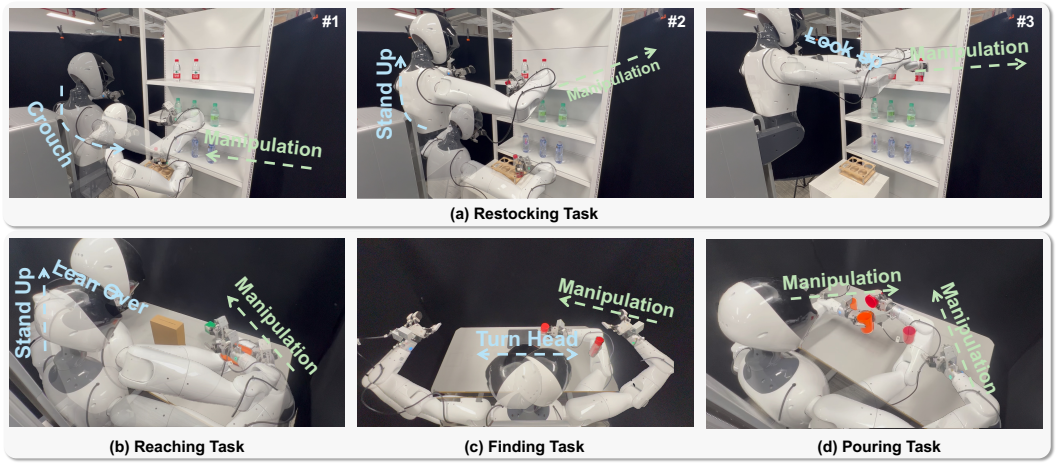

Figure 3: **Real-world tasks.** *(a) Restocking*: the robot crouches to pick up a water bottle from the table, then stands and looks up to scan the shelf for an empty slot and places it. *(b) Reaching*: the robot stands up and leans over an obstacle to reach the target object behind it. *(c) Finding*: the robot turns its head left or right to locate a yogurt and grasps it with the corresponding arm. *(d) Pouring*: the robot uses both hands to transfer liquid from a source container to a receiving container.

Tasks. We evaluate ActiveMimic on four real-world tasks spanning the active perception spectrum (Fig. 3). *(a) Restocking* is the most demanding: the robot crouches to pick up a water bottle from the table, stands and looks up to scan the shelf for an empty slot, then places the bottle. The shelf has three tiers at 70, 100, and 130 cm; we award one point for pickup and one for placement. *(b) Reaching* requires standing up and leaning over a 24 cm obstacle to grasp a target object initialized in a 20×20 cm region behind it. *(c) Finding* requires active search: the target yogurt is initialized in one of two 15×25 cm regions on the left or right side of the table, and the robot turns its head to locate and grasp it with the corresponding arm. *(d) Pouring* requires bimanual coordination to transfer liquid between two containers initialized in separate 20×20 cm regions. We train the four tasks on 270, 30, 60, and 90 teleoperated demonstrations and evaluate over 81, 18, 36, and 45 trials, respectively. We report end-to-end success rate as the primary metric and additionally score Restocking by average points per trial.

Pretraining data. We build our pretraining corpus from the Hands and Objects subset of Ego4D, which already targets egocentric hand-object manipulation. We further filter this subset to remove clips unsuitable for active perception supervision, yielding 2,561 episodes that amount to roughly 10 hours of video at 10 fps and an average of 130 frames per episode. Details of the additional filtering procedure are provided in Sec. A.2.

Baselines. We compare ActiveMimic against four baselines. (i) π_0 [3], initialized from the publicly released checkpoint and fine-tuned on our robot-specific data. (ii) *MotoVLA* [12], a state-of-the-art model pretrained on human video whose pretraining corpus mixes robot data with RH20T [46] human video, serving as the strongest available representative of pretraining on human video. (iii) *ActiveMimic*$_{\textit{wrist-only}}$ shares the Ego4D [21] corpus and architecture of ActiveMimic but is supervised only with the 18D wrist action, isolating the contribution of camera motion supervision. (iv) *ActiveMimic*$_{\textit{sft-only}}$ skips egocentric pretraining entirely and trains only on robot-specific data, isolating the contribution of egocentric human video pretraining.

4.2 Comparison with Baselines

Fig. 4 shows that ActiveMimic surpasses all baselines on all four tasks, achieving success rates of 90.1% on Restocking, 88.9% on Reaching, 91.7% on Finding, and 93.3% on Pouring. Among the ActiveMimic variants, both ActiveMimic$_{\text{wrist-only}}$ and ActiveMimic$_{\text{sft-only}}$ fall behind ActiveMimic

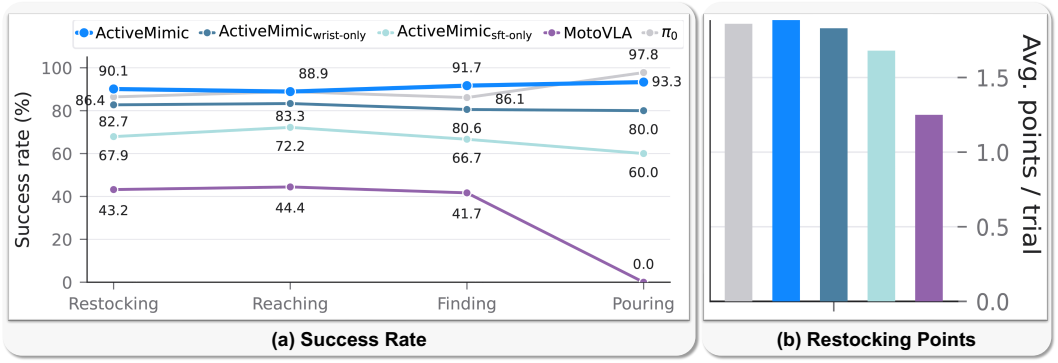

Figure 4: **Real-world results.** *(a) Success rate*: end-to-end success rate (%) on the four real-world tasks. *(b) Restocking points*: average points per trial on Restocking, with one point awarded for picking up the bottle and one for placing it on the shelf.

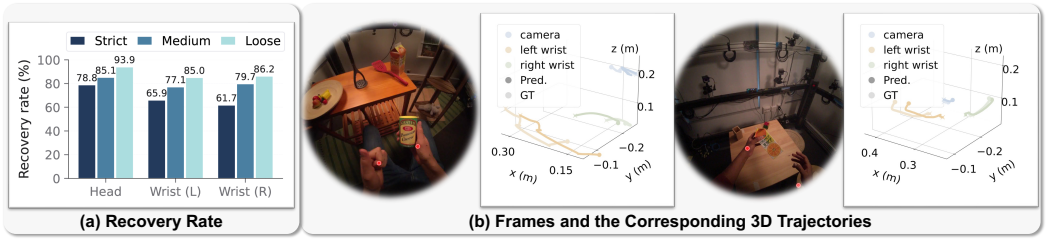

Figure 5: **Dataset characterization.** *Left*: recovery rates of predicted head and wrist poses on HOT3D at three tolerance tiers. *Right*: for two HOT3D videos, predicted wrist projections on a sampled frame and 3D chunk trajectories starting from that frame.

across the board, confirming that camera motion supervision during egocentric pretraining is the key differentiating factor. MotoVLA, which leverages a large mixed corpus of robot and human data, also falls behind ActiveMimic by a substantial margin on all tasks. Beyond these baselines, ActiveMimic achieves comparable or higher success rates than π_0 on all four tasks, showing that egocentric video pretraining matches a state-of-the-art model pretrained on robot data. On Restocking and Finding, the two tasks with the highest active perception demands, ActiveMimic clearly surpasses π_0 (90.1% vs. 86.4% and 91.7% vs. 86.1%), indicating that egocentric video provides active perception advantages that robot data alone does not capture. We further investigate where this capability originates (Sec. 4.4) and whether it transfers from human to robot (Sec. 4.5).

4.3 Egocentric Video Yields Effective Pretraining Labels

The 27D action labels constructed in Sec. 3.1 are designed to expose the coupling structure between active perception and manipulation from in-the-wild egocentric video. Validating this design requires an egocentric dataset with ground-truth head and wrist pose annotations, which Ego4D itself does not provide. We therefore evaluate our approach on HOT3D [47], an external egocentric dataset that supplies such annotations. We quantify label fidelity through the recovery rate on a randomly sampled 10% subset of HOT3D videos: for each sampled frame, the estimated head and wrist poses are compared against ground-truth annotations, and a frame is considered recovered when both the translational error and the rot6d L2 error fall within a specified tolerance. Under the strict tier (pos $\leq$ 0.8 m, rot6d L2 $\leq$ 0.6), head recovery reaches 78.82%, with left and right wrist recovery at 65.93% and 61.72%, respectively; under the loose tier, all three body parts exceed 85% (Fig. 5a). The approach operates purely from RGB video, without motion capture, inertial sensors, or calibrated multi-camera rigs, a deliberate fidelity-vs-scalability design choice that allows it to scale to arbitrary in-the-wild egocentric video. In addition, qualitative results show that estimated

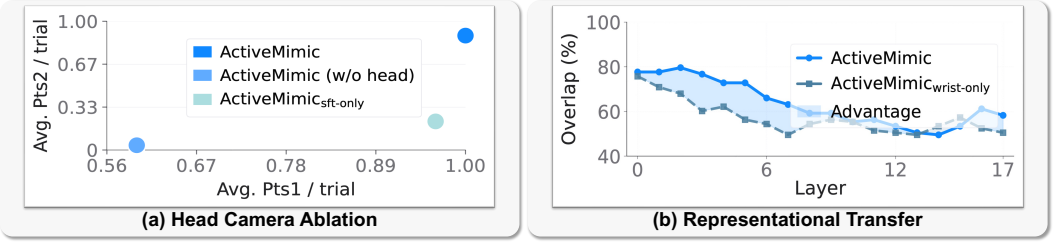

Figure 6: **Analysis experiments.** *(a)* Scores on Restocking for crouching to grasp the bottle (Pts1) and looking up to place it (Pts2). *(b)* Per-layer overlap (%) of the top-10% activated units under head-view vs. full-view inputs for ActiveMimic and ActiveMimic$_{\text{wrist-only}}$.

trajectories closely follow ground-truth trends on sampled HOT3D episodes (Fig. 5b). Together, these results (Fig. 5) confirm that the labels carry effective pretraining signals.

4.4 The Head Camera Enables Pretrained Active Perception

To pinpoint whether active perception comes from egocentric pretraining and how the model deploys it, we ablate Restocking under three inference conditions (Fig. 6a): ActiveMimic with all three cameras, ActiveMimic with the head camera zeroed out (w/o head), and ActiveMimic$_{\text{sft-only}}$ with all cameras. Notably, all three conditions reliably complete the pickup point, but the placement point reveals a stark separation. ActiveMimic scores 24 out of 27 on placement, whereas ActiveMimic$_{\text{sft-only}}$ achieves only 6 out of 27. This fourfold gap indicates that active perception capability is acquired during egocentric pretraining rather than robot-specific fine-tuning. Removing the head camera from the pretrained model collapses placement further to 1 out of 27, confirming that the model realizes this capability through the head camera. Together, egocentric pretraining provides active perception capability, and the head camera is how the model uses it.

4.5 Human-to-Robot Representational Transfer via Camera Motion Supervision

To investigate how camera motion supervision facilitates human-to-robot transfer, we compare ActiveMimic and ActiveMimic$_{\text{wrist-only}}$ under two inference conditions: full-view, where the model receives all three cameras, and head-view, where it receives only the head camera, approximating the single egocentric viewpoint in pretraining. Specifically, for each layer in the action expert, we identify the top-$K\%$ most activated units under each view and compute their overlap. The resulting overlap between head-view and full-view thus measures how much of the egocentric representational structure is preserved under the robot's multi-camera observation. Because this egocentric structure is learned from human video pretraining, higher preservation directly reflects stronger human-to-robot transfer. As shown in Fig. 6b, ActiveMimic maintains consistently higher overlap than ActiveMimic$_{\text{wrist-only}}$ across the early-to-mid layers (layers 0 through 11), where perceptual representations are encoded [48]. The higher overlap in perceptual layers indicates that camera motion supervision produces representations more robust to the observation modality shift, providing representational-level evidence that camera motion supervision strengthens human-to-robot transfer. We report $K{=}10$ and show in Sec. C.4 that the conclusion is robust to the choice of K.

5 Conclusion

We introduce ActiveMimic, an active-perception-aware pretraining framework for in-the-wild egocentric video. Across real-world tasks, ActiveMimic consistently surpasses baselines pretrained on human video, confirming active perception as the key to unlocking egocentric human video for robot pretraining. We further provide evidence that active perception originates from egocentric pretraining and that camera motion supervision facilitates representational transfer from human perception to robot control. Limitations are discussed in Sec. D.

References

[1] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023.

[2] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. P. Foster, P. R. Sanketi, Q. Vuong, et al. Openvla: An open-source vision-language-action model. In *CoRL*, 2025.

[3] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[4] J. Bjorck, F. Castañeda, N. Cherniadev, X. Da, R. Ding, L. Fan, Y. Fang, D. Fox, F. Hu, S. Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[5] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. In *ICLR*, 2025.

[6] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[7] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[8] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *ICLR*, 2023.

[9] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *ICRA*, 2024.

[10] Y. Liu, W. C. Shin, Y. Han, Z. Chen, H. Ravichandar, and D. Xu. Immimic: Cross-domain imitation from human videos via mapping and interpolation. In *CoRL*, 2025.

[11] X. Cai, R.-Z. Qiu, G. Chen, L. Wei, I. Liu, T. Huang, X. Cheng, and X. Wang. In-n-on: Scaling egocentric manipulation with in-the-wild and on-task data. *arXiv preprint arXiv:2511.15704*, 2025.

[12] A. Spiridonov, J.-N. Zaech, N. Nikolov, L. Van Gool, and D. P. Paudel. Generalist robot manipulation beyond action labeled data. In *CoRL*, 2025.

[13] T. Yoshida, S. Kurita, T. Nishimura, and S. Mori. Developing vision-language-action model from egocentric videos. *arXiv preprint arXiv:2509.21986*, 2025.

[14] R. Bajcsy. Active perception. *Proceedings of the IEEE*, 1988.

[15] R. Bajcsy, Y. Aloimonos, and J. K. Tsotsos. Revisiting active perception. *Autonomous Robots*, 2018.

[16] J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *IJCV*, 1988.

[17] R. Yang, Q. Yu, Y. Wu, R. Yan, B. Li, A.-C. Cheng, X. Zou, Y. Fang, X. Cheng, R.-Z. Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

[18] L. Y. Zhu, P. Kuppili, R. Punamiya, P. Aphiwetsa, D. Patel, S. Kareer, S. Ha, and D. Xu. Emma: Scaling mobile manipulation via egocentric human data. *RAL*, 2026.

[19] M. Shi, S. Peng, J. Chen, H. Jiang, Y. Li, D. Huang, P. Luo, H. Li, and L. Chen. Egohumanoid: Unlocking in-the-wild loco-manipulation with robot-free egocentric demonstration. *arXiv preprint arXiv:2602.10106*, 2026.

[20] R.-Z. Qiu, S. Yang, X. Cheng, C. Chawla, J. Li, T. He, G. Yan, D. J. Yoon, R. Hoque, L. Paulsen, et al. Humanoid policy $\sim$ human policy. In *CoRL*, 2025.

[21] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

[22] H. Luo, Y. Wang, W. Zhang, S. Zheng, Z. Xi, C. Xu, H. Xu, H. Yuan, C. Zhang, Y. Wang, et al. Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.

[23] S. Kareer, D. Patel, R. Punamiya, P. Mathur, S. Cheng, C. Wang, J. Hoffman, and D. Xu. Egomimic: Scaling imitation learning via egocentric video. In *ICRA*, 2025.

[24] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart. Receding horizon" next-best-view" planner for 3d exploration. In *ICRA*, 2016.

[25] M. Breyer, L. Ott, R. Siegwart, and J. J. Chung. Closed-loop next-best-view planning for target-driven grasping. In *IROS*, 2022.

[26] C. Connolly. The determination of next best views. In *ICRA*, 1985.

[27] M. Krainin, B. Curless, and D. Fox. Autonomous generation of complete 3d object models using next best view manipulation planning. In *ICRA*, 2011.

[28] M. Naazare, F. G. Rosas, and D. Schulz. Online next-best-view planner for 3d-exploration and inspection with a mobile manipulator robot. *RAL*, 2022.

[29] X. Zhang, D. Wang, S. Han, W. Li, B. Zhao, Z. Wang, X. Duan, C. Fang, X. Li, and J. He. Affordance-driven next-best-view planning for robotic grasping. In *CoRL*, 2023.

[30] J. Yu, Y. Shentu, D. Wu, P. Abbeel, K. Goldberg, and P. Wu. Egomi: Learning active vision and whole-body manipulation from egocentric human demonstrations. *arXiv preprint arXiv:2511.00153*, 2025.

[31] H. Xiong, X. Xu, J. Wu, Y. Hou, J. Bohg, and S. Song. Vision in action: Learning active perception from human demonstrations. In *CoRL*, 2025.

[32] Q. Zeng, C. Li, J. S. John, Z. Zhou, J. Wen, G. Feng, Y. Zhu, and Y. Xu. Activeumi: Robotic manipulation with active perception from robot-free human demonstrations. *arXiv preprint arXiv:2510.01607*, 2025.

[33] I. Chuang, A. Lee, D. Gao, M.-M. Naddaf-Sh, and I. Soltani. Active vision might be all you need: Exploring active vision in bimanual robotic manipulation. In *ICRA*, 2025.

[34] J. Kerr, K. Hari, E. Weber, C. M. Kim, B. Yi, K. Goldberg, A. Kanazawa, et al. Eye, robot: Learning to look to act with a bc-rl perception-action loop. In *CoRL*, 2025.

[35] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. In *CoRL*, 2025.

[36] M. Liu, E. Zhou, C. Chi, Y. Han, S. Rong, L. Chen, P. Wang, Z. Wang, and S. Zhang. Sapave: Towards active perception and manipulation in vision-language-action models for robotics. *arXiv preprint arXiv:2603.12193*, 2026.

[37] S. Kareer, K. Pertsch, J. Darpinian, J. Hoffman, D. Xu, S. Levine, C. Finn, and S. Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025.

[38] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, C. Yao, L. Zeng, Z. Feng, H. Liang, S. Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

[39] X. Yang, D. Kukreja, D. Pinkus, A. Sagar, T. Fan, J. Park, S. Shin, J. Cao, J. Liu, N. Ugrinovic, et al. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

[40] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer. In *CVPR*, 2025.

[41] L. Piccinelli, Y.-H. Yang, C. Sakaridis, M. Segu, S. Li, L. Van Gool, and F. Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024.

[42] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

[43] W. Liang, L. Yu, L. Luo, S. Iyer, N. Dong, C. Zhou, G. Ghosh, M. Lewis, W.-t. Yih, L. Zettlemoyer, et al. Mixture-of-transformers: A sparse and scalable architecture for multi-modal foundation models. *arXiv preprint arXiv:2411.04996*, 2024.

[44] S. Zhao, X. Zhang, J. Guo, J. Hu, L. Duan, M. Fu, Y. X. Chng, G.-H. Wang, Q.-G. Chen, Z. Xu, et al. Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*, 2025.

[45] L. Beyer, A. Steiner, A. S. Pinto, A. Kolesnikov, X. Wang, D. Salz, M. Neumann, I. Alabdulmohsin, M. Tschannen, E. Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

[46] H.-S. Fang, H. Fang, Z. Tang, J. Liu, C. Wang, J. Wang, H. Zhu, and C. Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

[47] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *CVPR*, 2025.

[48] Y. Luo, H. Chen, Z. Wu, B. Sui, J. Liu, C. Gu, Z. Liu, Q. Feng, J. Yu, S. Gu, et al. Look before acting: Enhancing vision foundation representations for vision-language-action models. *arXiv preprint arXiv:2603.15618*, 2026.

[49] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[50] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[51] K. Grauman, A. Westbury, L. Torresani, K. Kitani, J. Malik, T. Afouras, K. Ashutosh, V. Baiyya, S. Bansal, B. Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVPR*, 2024.

[52] S. Yin, Y. Ze, H.-X. Yu, C. K. Liu, and J. Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation. *arXiv preprint arXiv:2509.20322*, 2025.

[53] S. Wei, H. Jing, B. Li, Z. Zhao, J. Mao, Z. Ni, S. He, J. Liu, X. Liu, K. Kang, et al. ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.

[54] H. Jiang, J. Chen, Q. Bu, L. Chen, M. Shi, Y. Zhang, D. Li, C. Suo, C. Wang, Z. Peng, et al. Wholebodyvla: Towards unified latent vla for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025.

[55] X. Wang, T. Kwon, M. Rad, B. Pan, I. Chakraborty, S. Andrist, D. Bohus, A. Feniello, B. Tekin, F. V. Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *ICCV*, 2023.

[56] L. Xu, C. Yang, Z. Lin, F. Xu, Y. Liu, C. Xu, Y. Zhang, J. Qin, X. Sheng, Y. Liu, et al. Perceiving and acting in first-person: A dataset and benchmark for egocentric human-object-human interactions. In *ICCV*, 2025.

[57] X. Lin, X. Zhu, T. Lu, S. Xie, H. Zhang, X. Qiu, Z. Wu, and Y.-G. Jiang. Ask-to-clarify: Resolving instruction ambiguity through multi-turn dialogue. *arXiv preprint arXiv:2509.15061*, 2025.

[58] I. Rodin, A. Furnari, D. Mavroeidis, and G. M. Farinella. Predicting the future from first person (egocentric) vision: A survey. *CVIU*, 2021.

A From Egocentric Video to Unified Action Space

A.1 Metric Scale Recovery

The camera trajectory recovered by VGGT is a scale-normalized path $\tilde{T}_{\text{cam}_1}^{\text{cam}_k}$ whose translational component is determined only up to a global scale factor. To recover the metric scale, we align the per-pixel depth map D_k^{norm} from VGGT with the per-pixel metric depth map D_k^{metric} from UniDepth. A per-frame scale is first computed as the median depth ratio over valid pixels,

$$\lambda_k = \text{median}_{(u,v)\in\Omega_k} \frac{D_k^{\text{metric}}(u,v)}{D_k^{\text{norm}}(u,v)}, \tag{7}$$

where Ω_k denotes the set of pixels with valid positive depth values in both D_k^{norm} and D_k^{metric}. The per-frame scales are then aggregated into an episode-level scale $\lambda = \text{median}_{k\in\{1,\dots,K\}} \lambda_k$. The metric camera trajectory is then obtained by scaling only the translational component of the scale-normalized transform,

$$\tilde{T}_{\text{cam}_1}^{\text{cam}_k} = \begin{bmatrix} R_{\text{cam}_1}^{\text{cam}_k} & \tilde{t}_{\text{cam}_1}^{\text{cam}_k} \\ \mathbf{0}^\top & 1 \end{bmatrix}, \qquad T_{\text{cam}_1}^{\text{cam}_k} = \begin{bmatrix} R_{\text{cam}_1}^{\text{cam}_k} & \lambda\, \tilde{t}_{\text{cam}_1}^{\text{cam}_k} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{8}$$

A.2 Video Filtering and Segmentation

As described in Sec. 3.1, we identify hand-object manipulation segments from Ego4D through a two-stage filtering procedure that combines VLM-based temporal segmentation with LLM-based semantic filtering.

VLM-based temporal segmentation. For each Ego4D clip that passes an initial duration filter, a VLM (Qwen3-VL-8B-Instruct [49]) parses the full egocentric video and proposes candidate manipulation segments. The model is prompted to retain only intervals in which the camera wearer purposefully uses their hands to manipulate physical objects, excluding passive observation, walking, waiting, and pure camera motion. For each retained segment, the model outputs a start and end time, an action verb, a list of manipulated objects, and a natural-language task instruction composed from the action and objects. This task instruction serves as the language prompt during pretraining. Adjacent or overlapping segments with the same task description are merged, and a duration filter is applied to remove segments that are too short to contain meaningful manipulation or too long for efficient downstream processing. The full prompt is provided in Fig. 12.

LLM-based semantic filtering. These candidates are then filtered by an LLM (Qwen3-30B-A3B-Instruct [50]) against three semantic criteria: the action must involve hand-object manipulation, the manipulated objects must be artificial physical objects, and the scene must be indoors. Segments involving body parts or other humans as targets, natural or outdoor materials, outdoor activities, or non-manipulation actions are removed. This stage produces the final set of high-confidence indoor manipulation segments that are sampled and extracted at 10 fps for pose estimation. The full prompt is provided in Fig. 13.

Dataset statistics. Fig. 7 visualizes the action verb and object noun distributions of the final pretraining corpus. The verb word cloud reflects the diversity of manipulation actions, while the noun word cloud shows the breadth of manipulated object categories, confirming that the filtering procedure preserves semantic variety suitable for general-purpose pretraining.

B Training Details

The model comprises a 3B visual-language prefix and a 0.6B action expert. As described in Sec. 3.2, training follows a two-stage recipe. The egocentric human video pretraining stage is further divided into a warm-up phase and a full training phase. During warm-up, the visual-language prefix is frozen and only the action expert is trained, allowing the randomly initialized action expert to reach

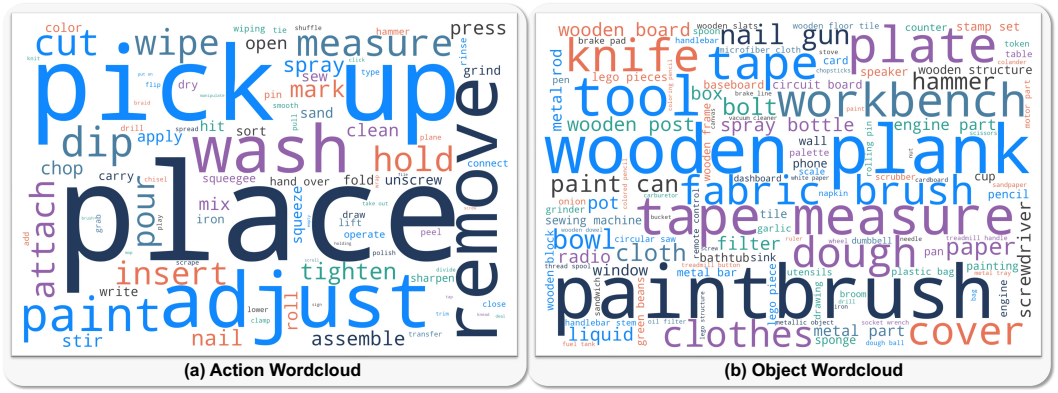

Figure 7: **Pretraining corpus statistics.** Word cloud of **(a)** action verbs and **(b)** manipulated objects in the final pretraining corpus after filtering, showing broad coverage of manipulation actions and object categories.

a reasonable operating point before joint optimization. The full training phase then unfreezes all parameters and trains the entire model end-to-end. The robot-specific training stage initializes from the pretrained checkpoint and fine-tunes on task-specific robot data. Table 1 summarizes the hyperparameters for each phase. Both ActiveMimic and ActiveMimic$_{\text{wrist-only}}$ share the same training configuration; the only difference is that ActiveMimic$_{\text{wrist-only}}$ is supervised with the 18D wrist action instead of the full 27D action.

	Stage 1: Warm-up	**Stage 1: Train**	**Stage 2**
Frozen modules	VLM	None	None
Batch size	64	64	64
Optimizer	AdamW	AdamW	AdamW
LR schedule	Cosine decay	Cosine decay	Cosine decay
Peak LR	1×10^{-4}	1×10^{-4}	1×10^{-4}
Final LR	1×10^{-5}	1×10^{-5}	1×10^{-5}
Warmup steps	200	1,000	100
Total steps	2,000	500,000	$\sim$5 epochs
Action horizon	50	50	50

Table 1: **Training hyperparameters for each phase.** Stage 2 trains for approximately 5 epochs on each task.

C Experimental Details

C.1 Task Setup

Table 2 consolidates the detailed specifications for each evaluation task. All four tasks are executed on the same robot platform described in Sec. 4.1.

C.2 Robustness Evaluation

Restocking under flashing lighting. We evaluate all models on the Restocking task under alternating red/green/blue flashing light, using the same checkpoint as the main results, with 81 trials per condition. As shown in Fig. 9(a), ActiveMimic achieves the highest success rate under the flashing condition (79.0%) and shows the smallest absolute drop among all models (-11.1% from 90.1%). ActiveMimic$_{\text{wrist-only}}$ drops by 24.7% to 58.0%, suggesting that camera motion supervision contributes to the robustness. ActiveMimic$_{\text{sft-only}}$ and MotoVLA both collapse to 0%, indicating that pretraining without active-perception-aware supervision provides no protection against visual perturbations. π_0 shows comparable robustness (75.3%), dropping 11.1%.

	Restocking	Reaching	Finding	Pouring
Object specifications				
Target object	Water bottle	Cubic block	Yogurt	Cup
Object height (cm)	17	5	16	9
Object diameter/side (cm)	5.5	5	5	7
Environment specifications				
Shelf tier heights (cm)	70 / 100 / 130	–	–	–
Shelf tier size (cm)	92.5 × 35	–	–	–
Obstacle height (cm)	–	24	–	–
Init. region (cm)	–	20 × 20	15 × 25 (×2)	20 × 20 (×2)
Training and evaluation				
Demonstrations	270	30	60	90
Evaluation trials	81	18	36	45
Metric	SR + Pts	SR	SR	SR

Table 2: **Detailed task specifications.** Init. region denotes the randomized object initialization area; (×2) indicates two separate regions. SR = success rate; Pts = average points per trial (1 for pickup + 1 for placement).

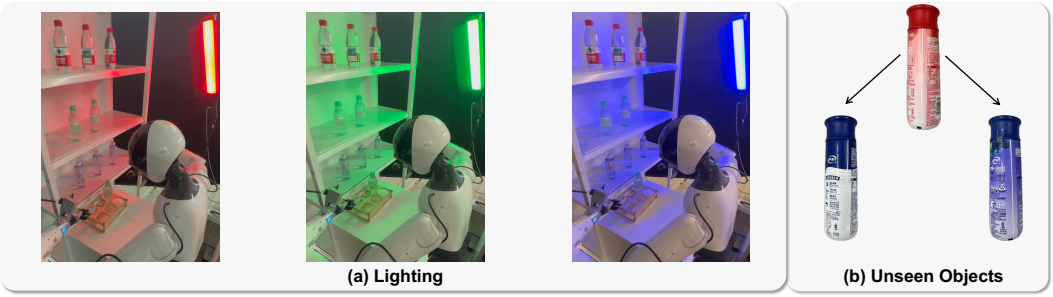

Figure 8: **Robustness evaluation setup.** **(a)** Restocking under alternating red, green, and blue flashing light. **(b)** Finding with two unseen yogurt variants (different packaging, identical shape and size) not present in training demonstrations. The training yogurt is shown at the top; the two unseen variants are shown below.

Finding with unseen objects. We replace the training yogurt with two unseen yogurt variants (different packaging, identical shape and size) and evaluate on the Finding task with 36 trials per condition. As shown in Fig. 9(b), ActiveMimic maintains the highest success rate among all models (72.2%) and the smallest drop (-19.5% from 91.7%). π_0 drops 22.2% to 63.9%. ActiveMimic$_{\text{wrist-only}}$ drops 33.4% to 47.2%, while ActiveMimic$_{\text{sft-only}}$ and MotoVLA drop to 27.8% and 11.1%, respectively. The larger gaps relative to the lighting experiment reflect that visual appearance is particularly load-bearing for object localization, yet ActiveMimic's active-perception pretraining provides the strongest generalization.

C.3 Failure Case Analysis

Fig. 10 presents representative failure cases of the *w/o head* condition on Restocking. All three failures occur at the placement point. In the first case, the arm reaches the correct shelf tier and lateral position but the placement motion is imprecise and knocks over the shelf. In the second case, the arm places the bottle on the correct tier but at the wrong lateral position. In the third case, the arm targets the wrong tier entirely. All three failures stem from removing the head camera, which severs the visual loop that the pretrained model relies on to coordinate head and hand movements during active perception. Without this feedback, the head-hand coordination acquired during egocentric human video pretraining breaks down, producing increasingly coarse placement errors.

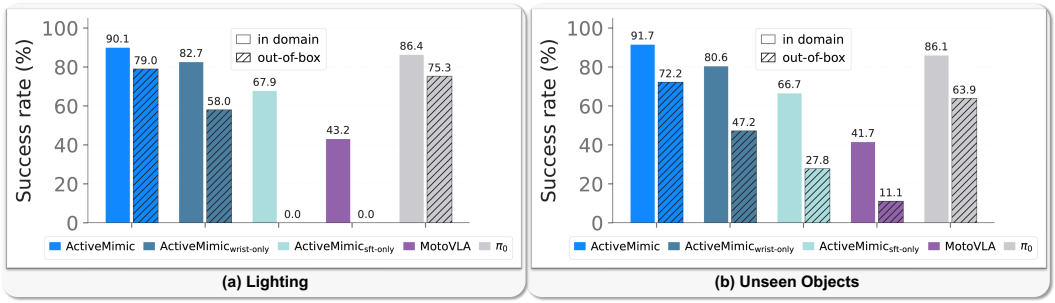

Figure 9: **Robustness evaluation.** **(a)** Restocking under alternating red/green/blue flashing light. **(b)** Finding with unseen yogurt objects (different packaging, identical shape and size). Solid bars denote in-domain (normal) conditions; hatched bars denote out-of-domain conditions. ActiveMimic achieves the highest success rate under both perturbations and exhibits the smallest absolute drop among all models.

Figure 10: **Representative failure cases of ActiveMimic without the head camera on Restocking.** All three failures occur at the placement point. From left to right: (1) correct shelf tier and lateral position, but the placement motion is imprecise and knocks over the shelf; (2) correct tier, wrong lateral position; (3) wrong tier entirely. All three stem from severing the visual loop that the pretrained model relies on to coordinate head and hand movements during active perception.

C.4 Representational Transfer: K Sensitivity Analysis

Sec. 4.5 reports representational transfer results at $K = 10$. Fig. 11 extends this analysis to $K \in \{5, 10, 15, 20\}$. Across all values of K, ActiveMimic maintains consistently higher top-K% activation overlap than ActiveMimic$_{\text{wrist-only}}$, confirming that the conclusion is robust to the choice of K.

D Limitations and Future Directions

Data scale. The current pretraining corpus comprises approximately 10 hours of filtered egocentric manipulation video from Ego4D. While this scale already yields significant gains over training from scratch, substantially larger egocentric corpora (*e.g.*, the full Ego4D [21] or Ego-Exo4D [51]) are readily available and can be incorporated with the same automated procedure, which we expect to further strengthen the pretrained representations.

Embodiment diversity. All real-world experiments use a single humanoid platform. Because the pretraining stage is embodiment-agnostic, operating on egocentric video without any robot-specific input, extending to other embodiments only requires the robot-specific training stage with corresponding demonstrations. Validating this across a broader range of platforms is a natural next step.

Label fidelity. Action labels are obtained from vision-based pose estimation rather than hardware-recorded trajectories, which inevitably introduces estimation noise. Nonetheless, the pretrained models achieve strong downstream performance, suggesting that the learning objective is robust to

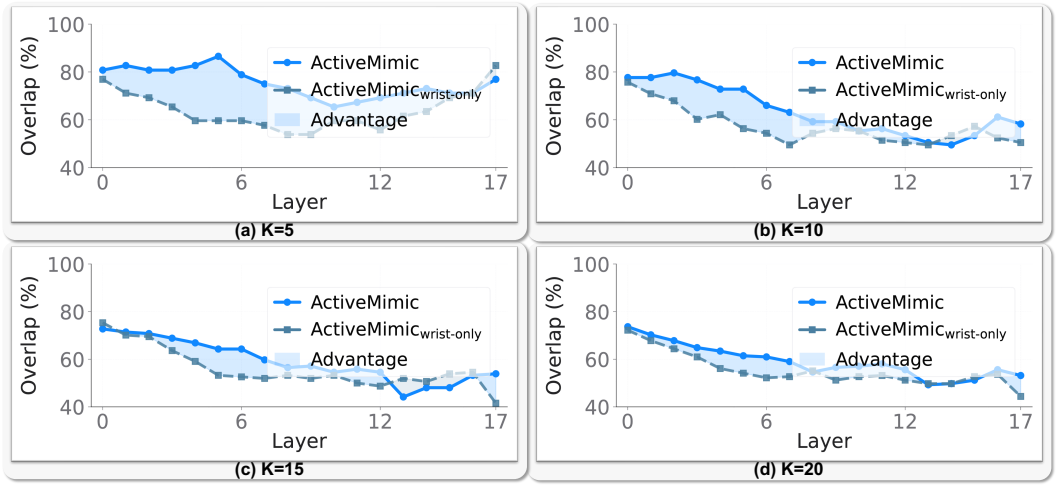

Figure 11: **K sensitivity analysis for representational transfer.** Top-$K\%$ activation overlap between full-view and head-view inference conditions for ActiveMimic and ActiveMimic$_{\text{wrist-only}}$ across all action-expert layers, evaluated at $K = 5, 10, 15, 20$. The shaded area indicates the advantage of ActiveMimic over ActiveMimic$_{\text{wrist-only}}$. ActiveMimic maintains consistently higher overlap across all K values, confirming that the conclusion in Sec. 4.5 is robust to the choice of K.

moderate label noise. Label quality can be further improved as better off-the-shelf pose estimation methods become available.

Loco-manipulation. The current evaluation focuses on stationary tabletop and shelf manipulation. Extending ActiveMimic to loco-manipulation on humanoid robots [19, 52, 53, 54], where the robot must coordinate locomotion and manipulation simultaneously, is a promising direction, as egocentric video datasets already contain abundant walking-while-manipulating footage that could serve as pretraining data.

Human-robot interaction. Egocentric video covers diverse daily activities, many of which naturally involve interactions with other people [55, 56, 57, 58]. Extending ActiveMimic to human-robot interaction scenarios, where the robot must perceive and respond to human actions in shared workspaces, is a compelling direction that could leverage this inherent property of egocentric data.

```
You are an expert in understanding egocentric videos involving hand-object interactions.

Please watch the entire egocentric video carefully and identify all time segments where
the camera wearer is performing a specific, goal-directed task that involves direct
interaction between their hands and a specific object.

A valid task must satisfy the following conditions:
- The person's hands are actively manipulating or interacting with a physical object
- The action has a clear purpose, such as "washing a dish", "opening a bottle", or
  "tightening a screw"
- Segments where the person is not using their hands to manipulate any object | such as
  walking, turning their head, looking around, standing still, observing, or waiting |
  should be excluded

The total duration of the video is [VIDEO_DURATION] seconds.

For each detected task segment, provide:
1. The start time (in seconds, integer only)
2. The end time (in seconds, integer only)
3. A concise description of the specific task being performed

Each description must include:
- The main manipulation action (a verb like "pick up", "place", "insert", "open", etc.)
- A list of one or more objects that are being manipulated
- A short natural language instruction generated from the action and objects

The segments may overlap in time if multiple tasks are performed in close succession
or simultaneously.

Return the results strictly in the following JSON format:
[
  {"start": 4, "end": 9, "action": "open", "objects": ["bag"], "task": "Open the bag"},
  {"start": 9, "end": 15, "action": "place", "objects": ["apple", "plate"],
   "task": "Place the apple on the plate"}
]
```

Figure 12: **Prompt used for VLM-based temporal segmentation.** The model identifies manipulation segments from egocentric video and outputs structured annotations including a natural-language task instruction that serves as the language prompt during pretraining.

```
You are a task filter for egocentric video clips.

You will be given a JSON object that represents a single video clip. Each clip contains
a list of task segments. Your job is to extract only segments that are suitable for
training a Vision-Language-Action (VLA) model focused on indoor hand-object manipulation.

Each segment includes:
- start, end: time in seconds
- action: verb describing the action
- objects: list of physical objects being interacted with
- task: natural language description

Filtering criteria | keep only segments that satisfy all three:
1. The action involves hand-object manipulation (e.g., pick up, cut, fold, assemble,
   insert, tighten, wipe, pour, etc.)
2. The object(s) must be artificial, physical items (tools, containers, utensils,
   electronics, furniture, fabric, household goods). Exclude: body parts (leg, hand,
   arm), people (man, woman, person), natural materials (plant, soil, mud, grass, tree).
3. The scene is likely indoors. Exclude: gardening, farming, outdoor repair, digging,
   planting, handling mud/branches/natural terrain.

Return a JSON object:
{"clip_uid": "...", "status": "success", "filtered_segments": [...]}
```

Figure 13: **Prompt used for LLM-based semantic filtering.** The model retains only segments involving indoor hand-object manipulation of artificial objects.