Learning Human-Intention Priors from Large-Scale Human Demonstrations for Robotic Manipulation

Yifan Xie[1,2*] **YuAn Wang**[2*] **Guangyu Chen**[1] **Jinkun Liu**[1]
Yu Sun[2†] **Wenbo Ding**[1‡]

[1]Tsinghua University [2]ByteDance

Abstract

Human videos contain rich manipulation priors, but using them for robot learning remains difficult because raw observations entangle scene understanding, human motion, and embodiment-specific action. We introduce MoT-HRA, a hierarchical vision-language-action framework that learns human-intention priors from large-scale human demonstrations. We first curate HA-2.2M, a 2.2M-episode action-language dataset reconstructed from heterogeneous human videos through hand-centric filtering, spatial reconstruction, temporal segmentation, and language alignment. On top of this dataset, MoT-HRA factorizes manipulation into three coupled experts: a vision-language expert predicts an embodiment-agnostic 3D trajectory, an intention expert models MANO-style hand motion as a latent human-motion prior, and a fine expert maps the intention-aware representation to robot action chunks. A shared-attention trunk and read-only key-value transfer allow downstream control to use human priors while limiting interference with upstream representations. Experiments on hand motion generation, simulated manipulation, and real-world robot tasks show that MoT-HRA improves motion plausibility and robust control under distribution shift.

1 Introduction

Generalist robotic manipulation is increasingly framed as vision-language-action (VLA) modeling, where a single policy conditions on visual observations and language instructions and produces executable actions [1, 2, 3, 4, 5]. This paradigm has benefited from large pretrained vision-language backbones and growing cross-embodiment robot datasets, but its scaling path is still constrained by the cost and sparsity of robot demonstrations. Robot trajectories are expensive to collect, tied to specific hardware, and often cover only a narrow slice of everyday manipulation. In contrast, human videos record abundant object interactions across homes, kitchens, workshops, viewpoints, and task styles [6, 7, 8, 9]. If these videos can be converted into supervision for control, they offer a much broader source of manipulation priors than robot-only data.

However, human videos are not robot demonstrations. Raw clips entangle scene understanding, hand motion, camera motion, task progress, and embodiment-specific constraints. Many segments contain visible hands without purposeful manipulation, while useful interaction clips rarely provide temporally aligned action labels or robot-executable controls. Recent human-centric robot learning methods have begun to exploit this resource by extracting hand motion, aligning visual observations with physical space, or introducing latent bridges between videos and actions [10, 11, 12, 13, 14]. These efforts suggest that human videos contain transferable physical knowledge, but they also expose a core representation problem. To benefit robot learning, a model must separate *where* an interaction should occur, *how* human motion expresses task intent, and *how* a particular robot should execute the task. Collapsing these factors into a single action representation can cause downstream

*: Equal contribution; †: Project lead; ‡: Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

robot loss to overwrite useful human-motion priors, and can make the learned policy brittle when visual appearance, viewpoint, object instance, or embodiment changes.

Our key observation is that human demonstrations are most useful when they are treated not as noisy substitutes for robot actions, but as structured evidence about manipulation intent. A human hand trajectory provides an embodiment-agnostic spatial scaffold, and articulated hand motion captures temporal coordination and contact-oriented motor preference. Robot controls should then be learned as an embodiment-specific realization of this prior. This view suggests a hierarchical representation for human-to-robot transfer: first ground the interaction in 3D space, then model human intention in a structured hand-motion space, and finally adapt the intention-aware representation to robot action chunks. Such a decomposition preserves the reusable parts of human behavior while leaving the final policy free to match the kinematics and action convention of the target robot.

To realize this idea, we present MoT-HRA, a hierarchical VLA framework that learns human-intention priors from large-scale human demonstrations and transfers them to robotic manipulation. We first curate HA-2.2M, a 2.2M-episode action-language dataset reconstructed from heterogeneous human videos through hand-centric filtering, spatial reconstruction, temporal segmentation, and language alignment. Building on this dataset, MoT-HRA factorizes action generation into three coupled experts: a vision-language expert that predicts an embodiment-agnostic 3D trajectory, an intention expert that models a MANO-style human hand-motion prior, and a fine expert that maps the intention-aware representation to robot action chunks. A shared-attention trunk supplies common multimodal context, while read-only key-value transfer implements knowledge insulation [15], allowing downstream robot control to use human priors without overwriting upstream spatial and intention representations. This design turns heterogeneous human videos into an intermediate intention manifold rather than forcing them into robot-specific action labels.

Our main contributions can be summarized as follows:

- We introduce HA-2.2M, a large-scale human-demonstration dataset built from heterogeneous web and egocentric videos. It provides temporally coherent action-language episodes with reconstructed hand-centric spatial supervision, offering a scalable source of manipulation priors beyond robot-only demonstrations.
- We propose MoT-HRA, a hierarchical Mixture-of-Transformer VLA architecture that separates spatial trajectory grounding, latent human-intention modeling, and embodiment-specific robot action generation, with knowledge insulation to reduce destructive interference between human-prior learning and robot policy learning.
- We validate the proposed decomposition on hand motion generation, SimplerEnv, and real-world manipulation tasks, showing improvements in motion plausibility and control robustness. Architectural ablations and a controlled data-scaling study assess the contributions of the hierarchy and human-demonstration scale.

2 Related Work

2.1 Vision-Language-Action Models

Vision-language-action (VLA) models extend pretrained multimodal representations from perception and instruction following to robot control. Representative systems ground visual-language knowledge in robot action datasets and language-conditioned policies [1, 2, 3, 4, 16], while recent work improves action generation through flow-based policies, efficient action tokenization, fine-tuning recipes, spatial representations, latent planning, and world-model interfaces [5, 17, 18, 19, 20, 21, 22]. Other studies focus on preserving pretrained knowledge and strengthening multimodal or world model representations for control [15, 13, 23, 24, 25, 26]. Our work follows this factorized trend by separating visual-language grounding, human-intention modeling, and robot-specific action prediction.

2.2 Human Demonstration Learning

Human demonstration learning studies how passive recordings can supervise robot policies without robot actions. Human and egocentric videos have been used to learn visual representations, rewards, and affordances that expose object-centric goals, contact cues, and task progress [27, 28, 29, 30].

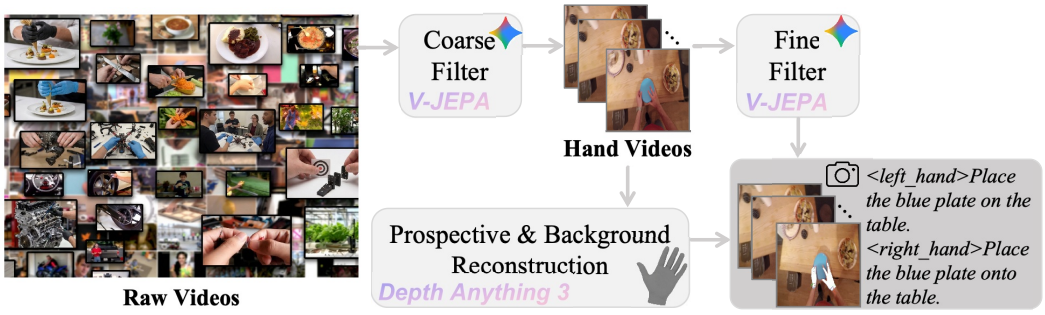


Figure 1: Overview of the HA-2.2M curation pipeline. Large-scale unlabeled human demonstration videos are converted into 2.2M action-language episodes through coarse filtering, perspective and background reconstruction, and fine filtering. These stages retain manipulation-relevant hand videos, recover image-aligned hand motion and scene context, and extract temporally coherent atomic episodes with aligned action descriptions. The pie chart summarizes the source composition of the final dataset.

Recent human-video-to-VLA work frames these signals as representation bridges from unstructured videos to executable actions [31]. Large corpora such as Ego4D, EPIC-KITCHENS, Something-Something-V2, and Ego-Exo4D provide diverse interactions [7, 8, 9, 32], but usually lack aligned control labels, calibrated 3D state, and proprioception.

Existing methods differ in the bridge they recover. Latent-action methods learn representations of inter-frame changes [33, 34, 35, 36, 37], world-model methods transfer future-prediction or generated-video priors to action heads [38, 39, 40], and explicit methods use 2D tracks or affordance prompts [41, 42] or reconstruct 3D wrist, hand, object, and MANO-style motion [43, 44, 45, 10, 11, 14]. Our method follows the explicit 3D line but separates reconstructed human hand motion as an intention prior from the final embodiment-specific action expert, reducing interference when transferring human demonstrations to robot control.

3 Methods

In this section, we describe the full pipeline for learning robot control from large-scale human demonstrations. We first introduce HA-2.2M (Sec. 3.1), a 2.2M-episode dataset with aligned action and language supervision reconstructed from raw human videos. We then present MoT-HRA (Sec. 3.2), a hierarchical Mixture-of-Transformers architecture that converts these demonstrations into a latent intention space connecting multimodal understanding, human motion priors, and embodiment-specific action generation.

3.1 HA-2.2M

Overview HA-2.2M contains 2.2M short human-demonstration episodes from egocentric and third-person views [6, 7, 8, 9]. Rather than reuse heterogeneous source annotations, we reconstruct action-language episodes from raw videos with the pipeline in Fig. 1. This mixture combines egocentric recordings with diverse third-person videos, broadening coverage over objects, scenes, and manipulation styles. At the same time, it introduces noise from passive hand motion, abrupt viewpoint changes, camera motion unrelated to interaction, and mismatched temporal granularity. The final pipeline applies coarse filtering, spatial reconstruction, and fine filtering so that each retained clip corresponds to a short manipulation primitive. These clips serve as pretraining units for learning latent human action priors instead of merely enlarging the visual dataset.

Coarse Filter Raw web videos often show hands without purposeful manipulation, such as idle gestures, transitional motion, or camera motion unrelated to interaction. We therefore use a recall-oriented cascade: Gemini [46] first labels likely hand-centric action segments, and a lightweight V-JEPA-based [47] classifier is then trained on these labels and applied across the full dataset. The first stage removes obviously irrelevant content while preserving diverse manipulation patterns across viewpoints and environments, and the second suppresses visible-hand false positives introduced

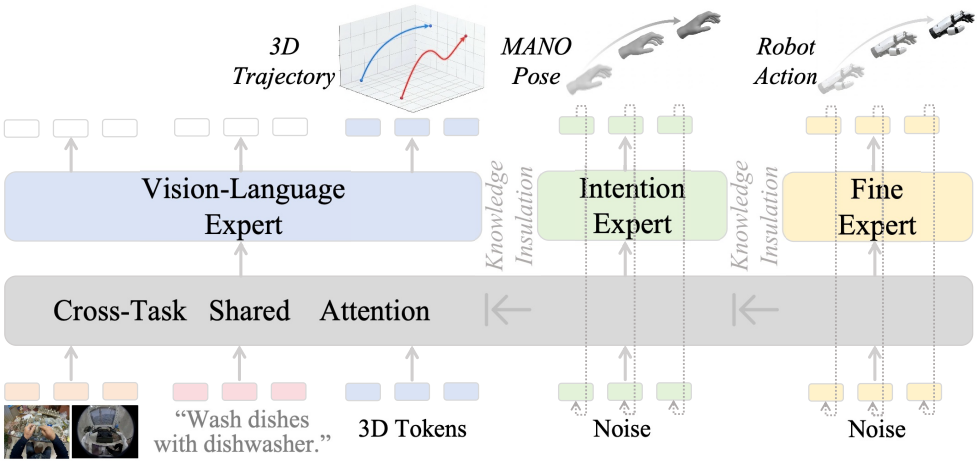


Figure 2: Overview of MoT-HRA. Given an image, a language instruction, and chunk-sized query tokens, the model predicts robot actions with three hierarchical experts: a vision-language expert that autoregressively predicts an embodiment-agnostic discretized 3D trajectory, an intention expert that generates a MANO-style hand-pose sequence by denoising noise conditioned on the spatial plan, and a fine expert that maps the resulting intention-aware representation to embodiment-specific robot actions. A shared-attention trunk supplies common multimodal context, while read-only key-value transfer across experts enforces knowledge insulation and preserves human motion priors during downstream action learning.

by multimodal labeling. This design avoids expensive multimodal calls for every segment while retaining the breadth needed for large-scale pretraining, yielding a cleaner input distribution for the reconstruction stage.

Perspective and Background Reconstruction The retained videos are converted into image-aligned 3D trajectories by reconstructing hand motion and scene context. For each frame, we localize hands with VitPose [48, 49], refine the boxes for robustness under blur and occlusion, and estimate absolute-scale MANO hand pose with HaMeR [50]. Since HaMeR outputs crop-space predictions, we recover crop-to-image translation and scale, map MANO states back to the original frame, and temporally smooth trajectories to reduce jitter. In parallel, Depth Anything 3 [51] predicts relative monocular depth for background layout, which we align to the hand's absolute scale by fitting DA3 values on hand pixels to the MANO reconstruction. This places scene geometry in a hand-consistent 3D frame and gives subsequent segmentation and language alignment access to scaled scene context rather than appearance-only videos.

Fine Filter After reconstruction, we extract atomic episodes and align them with language supervision. A V-JEPA [47] temporal segmentation model, trained on 38K manually annotated segments, predicts action boundaries so that clips correspond to manipulation units rather than arbitrary fixed windows. Because the high-recall boundary predictor can over-segment complex activities, Gemini [46] merges adjacent clips with continuous intent and generates concise action descriptions, including hand-specific tags when the two hands play different roles. The detailed merge rules appear in Appendix A.3. This final stage turns noisy reconstructed trajectories into temporally coherent action-language episodes for HA-2.2M.

3.2 MoT-HRA

Problem Formulation Large-scale human videos provide useful priors for robotic control, but direct observation-to-action learning entangles visual grounding, human motion intent, and embodiment-specific actuation [10, 11, 12]. A single representation must absorb both transferable human manipulation intention and low-level robot actions. MoT-HRA instead factorizes the chunk policy through an embodiment-agnostic spatial plan and a learned human-intention state. Given an

observation image o, a language instruction w, and a chunk horizon H, we approximate

$$p_\theta(a_{1:H} \mid o, w) \approx \int p_{\theta_a}(a_{1:H} \mid o, w, \tau_{1:H}, h^{\text{int}}_{1:H})\, q_{\theta_m}(h^{\text{int}}_{1:H} \mid o, w, \tau_{1:H}) p_{\theta_\tau}(\tau_{1:H} \mid o, w)\, \mathrm{d}\tau\, \mathrm{d}h. \tag{1}$$

Here $\tau_{1:H}$ is an embodiment-agnostic 3D plan, and $h^{\text{int}}_{1:H}$ is the latent intention representation passed to the robot-action factor. The MANO-style pose sequence $m_{1:H}$ supervises the intention branch but is not treated as a robot command. In implementation, the factors are sampled sequentially:

$$\begin{aligned} \tau_{1:H} &\sim p_{\theta_\tau}(\cdot \mid o, w), \\ (m_{1:H}, h^{\text{int}}_{1:H}) &= \Phi_{\theta_m}(\epsilon^m; o, w, \tau_{1:H}), \\ a_{1:H} &= \Phi_{\theta_a}(\epsilon^a; o, w, \tau_{1:H}, h^{\text{int}}_{1:H}), \end{aligned} \tag{2}$$

where $\epsilon^m, \epsilon^a \sim \mathcal{N}(0, I)$, and Φ_{θ_m} and Φ_{θ_a} denote the conditional flow-matching solution maps of the intention and fine experts. The action flow conditions on $h^{\text{int}}_{1:H}$ rather than treating generated hand poses as robot commands. Thus, the 3D trajectory specifies where the interaction should unfold, the intention branch captures human motion regularities, and the final expert adapts this bridge to the target embodiment.

Unified Architecture As shown in Fig. 2, MoT-HRA instantiates this factorization with a Mixture-of-Transformers architecture composed of a cross-task shared-attention trunk and three task-specific experts: a vision-language expert, an intention expert, and a fine expert. The full token sequence is

$$X = \left[x^{\text{img}}_{1:N_v}; x^{\text{txt}}_{1:N_l}; z^{3d}_{1:H}; z^{m}_{1:H}; z^{a}_{1:H}\right], \tag{3}$$

where x^{img} and x^{txt} denote visual and text tokens, and z^{3d}, z^m, and z^a denote trajectory, intention, and action query/noise tokens. The shared trunk builds multimodal context, while each expert updates only its own token span. Later experts read earlier hidden states through read-only key-value caches, yielding *knowledge insulation* [15]: upstream representations remain visible to downstream modules without being overwritten by downstream losses. Concretely, the vision-language/trajectory expert is shielded from MANO and action denoising losses, while the robot-action expert can exploit intention cues without collapsing them into a single entangled latent space. The attention mask in Fig. 3 follows the same hierarchy: image and text tokens attend bidirectionally, 3D tokens use causal attention within their span while attending to image-text context, MANO tokens attend causally within the intention span and read the full prefix, and robot-action tokens attend to all preceding modalities with bidirectional refinement inside the action chunk. Compared with prior MoT and latent-bridge models [13, 52], MoT-HRA grounds the intermediate bridge in reconstructed human hand kinematics.

Vision-Language Expert The vision-language expert is initialized from a pretrained multimodal decoder [53] and converts visual-linguistic context into coarse spatial waypoints. For a chunk of length H, we append H learnable 3D query tokens, each predicting a discretized waypoint $\tau_h = (b^x_h, b^y_h, b^z_h)$ with each coordinate quantized into B bins. The generation process is autoregressive:

$$p_{\theta_\tau}(\tau_{1:H} \mid o, w) = \prod_{h=1}^{H} \prod_{c \in \{x,y,z\}} p_{\theta_\tau}(b^c_h \mid o, w, \tau_{<h}). \tag{4}$$

We use bidirectional attention over image-text tokens and causal attention over the 3D span, so each waypoint depends on the scene, instruction, and previous spatial anchors while remaining independent of future trajectory tokens. The training objective is coordinate-wise cross-entropy:

$$\mathcal{L}_{3d} = -\sum_{h=1}^{H} \sum_{c \in \{x,y,z\}} \log p_{\theta_\tau}(b^c_h \mid o, w, \tau_{<h}). \tag{5}$$

This spatial bottleneck asks the model to localize where the interaction should unfold before predicting hand or robot motion. It improves geometric grounding without tying the representation to a specific embodiment, which is important when transferring priors from human videos to robot control. Compared with regressing embodiment-specific controls directly from pixels, the waypoint representation gives downstream experts a compact spatial scaffold that is easier to supervise from reconstructed human demonstrations.

Intention Expert The intention expert maps the spatial plan to a human-motion prior. Conditioned on the shared multimodal context and predicted 3D trajectory, it denoises Gaussian noise into a MANO-style hand sequence via conditional flow matching [54]. Each hand state is

$$m_h = [w_h, j_h], \qquad w_h \in \mathbb{R}^7, \quad j_h \in \mathbb{R}^{60}, \tag{6}$$

where w_h stores wrist translation and quaternion orientation, and j_h stacks quaternion rotations for 15 finger joints. Quaternion targets are sign-canonicalized, and decoded quaternions are renormalized before kinematic use. Let $\epsilon^m \sim \mathcal{N}(0, I)$, $t \sim \mathcal{U}(0, 1)$, and $x_t^m = (1-t)\epsilon^m + tm$. The intention expert predicts $v_{\theta_m}(x_t^m, t, c^m)$ with condition $c^m = (o, w, \tau_{1:H})$ and is trained with

$$\mathcal{L}_{\text{mano}} = \mathbb{E}_{t,\epsilon^m,m}\left[\frac{1}{7H}\sum_{h=1}^{H}\|v_h^w - (w_h - \epsilon_h^w)\|_2^2 + \frac{1}{60H}\sum_{h=1}^{H}\left\|v_h^j - (j_h - \epsilon_h^j)\right\|_2^2\right] \tag{7}$$

We separate wrist and finger losses so the low-dimensional wrist signal is not dominated by the 60-dimensional joint term. MANO tokens use causal attention within their span, preventing valid early states from attending to future or padded suffix tokens, which otherwise destabilizes convergence. By modeling explicit hand kinematics rather than an unconstrained latent code, the expert turns large-scale human videos into an interpretable intention space aligned with physical interaction [14, 55].

Fine Expert The fine expert maps the intention-aware representation to the final robot action chunk with another conditional flow-matching head. Let $a_{1:H} \in \mathbb{R}^{H \times d_a}$ denote an action chunk for a robot with action dimension d_a. Starting from Gaussian noise $\epsilon^a \sim \mathcal{N}(0, I)$, we define $x_t^a = (1-t)\epsilon^a + ta$ and train $v_{\theta_a}(x_t^a, t, c^a)$ with

$$\mathcal{L}_{\text{act}} = \mathbb{E}_{t,\epsilon^a,a}\left[\|v_{\theta_a}(x_t^a, t, c^a) - (a - \epsilon^a)\|_2^2\right]. \tag{8}$$

The conditioning variable c^a includes image-text features, 3D trajectory states, and the latent intention states $h_{1:H}^{\text{int}}$. Thus, the policy uses the latent states produced while denoising the MANO representation rather than imitating decoded hand poses as targets. These states encode temporal coordination, uncertainty, and human motor preferences while leaving the final controls embodiment-specific. Following Fig. 3, action tokens attend to all preceding modalities and use bidirectional attention within the action chunk, enabling joint refinement of temporally coupled controls. This is well suited to chunked robot control, where commands within the horizon should be optimized together after the intention scaffold is established.

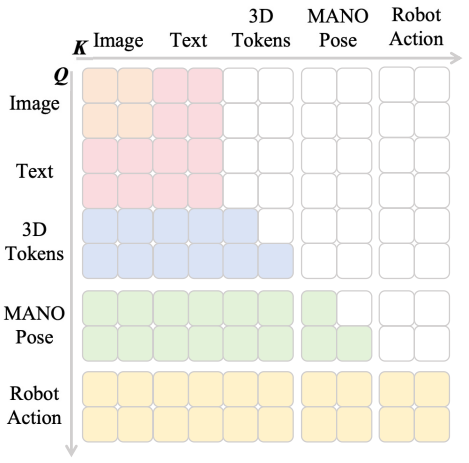


Figure 3: Attention mask of MoT-HRA. Image and text tokens use bidirectional attention to build a shared multimodal context, while 3D trajectory and MANO pose tokens attend to the full prefix but remain causally masked within their own spans. Robot-action tokens attend to all preceding modalities and use bidirectional attention within the action chunk, enabling joint refinement of temporally coupled controls.

Joint Training Strategy We train the three experts jointly with a partially supervised multi-task objective:

$$\mathcal{L} = \lambda_{3d}\mathcal{L}_{3d} + \lambda_m \mathbf{1}_m \mathcal{L}_{\text{mano}} + \lambda_a \mathbf{1}_a \mathcal{L}_{\text{act}}, \tag{9}$$

where $\mathbf{1}_m$ and $\mathbf{1}_a$ indicate whether MANO or robot-action supervision is available. Human-demonstration episodes mainly supervise the trajectory and intention experts, while robot-manipulation episodes supervise the trajectory and fine experts. On robot-only samples, the intention expert still provides latent conditioning through the read-only interface but receives no MANO loss, and downstream action gradients are not propagated into the intention cache. This keeps the human-motion prior stable while the fine expert adapts it to robot control. This training regime lets human videos shape the intention manifold while robot data specializes embodiment-specific actuation, enabling downstream robot inference without MANO annotations and direct generation of executable real-robot action.

Table 1: Hand motion generation results on egocentric [7] and third-person views [59]. ADE and DTW are reported in meters, while Rot and Joint-Rot are reported in degrees. Lower is better for all metrics.

Method	Egocentric View				Third-Person View			
	ADE	DTW	Rot	Joint-Rot	ADE	DTW	Rot	Joint-Rot
Being-H0 [10]	0.185	0.174	38.27	44.03	0.245	0.233	44.91	49.18
VITRA [11]	0.154	0.146	33.26	41.81	0.211	0.201	42.59	41.72
Ours	**0.136**	**0.127**	**28.95**	**34.16**	**0.184**	**0.176**	**38.47**	**40.12**

4 Experiments

4.1 Implementation Details

We train MoT-HRA jointly on HA-2.2M and the real robot dataset [56] with 64 NVIDIA Hopper GPUs under PyTorch FSDP2 [57]. The HA-2.2M dataset supervises the intention expert through MANO-style human-motion targets, the AgiBot-World dataset supervises the fine expert through robot-action targets, and both datasets supervise the vision-language expert through the 3D trajectory prediction objective. Each GPU processes 32 samples, giving a global batch size of 2048, and the model is optimized for 20,000 steps with AdamW and gradient accumulation set to 1. We use a base learning rate of 2.5×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, weight decay 10^{-10}, a 1,000-step warmup, and cosine decay. The chunk horizon is $H = 15$. For MANO generation, we apply classifier-free guidance (CFG) [58] with scale 6.0. Visual inputs are single RGB observations resized to 224×224 with padding and standard data augmentation, and language prompts are tokenized to length 256 with instruction dropout 0.1. Unless otherwise stated, we use EMA, activation checkpointing, and 8 data-loading workers per GPU for memory-efficient large-scale optimization.

4.2 Hand Motion Generation

We compare MoT-HRA with Being-H0 [10] and VITRA [11] on held-out egocentric Ego4D [7] and third-person [59] clips. Although HA-2.2M includes Ego4D videos, the egocentric evaluation split used here is excluded from the training dataset, so it tests held-out egocentric generation rather than memorization of training clips. Third-person view further introduces a larger appearance gap. For each dataset, we evaluate 200 clips, sample five generations per clip with different random seeds, and report average performance. We measure wrist accuracy with average displacement error (ADE, m) and Dynamic Time Warping distance (DTW, m), and pose fidelity with wrist rotation error (Rot, degrees) and finger joint rotation error (Joint-Rot, degrees). Lower is better for all metrics.

As shown in Tab. 1, MoT-HRA performs best across both datasets, improving wrist alignment and articulated pose quality. The advantage on Ego4D indicates that the learned intention prior improves both global motion and local hand configuration, while the gains on third-person view evaluation set suggest that this prior is not limited to the visual style of the training data. The comparatively smaller improvement on fine-grained joint rotation also indicates that articulated finger motion remains the most difficult case under strong domain transfer. Qualitative comparisons in Fig. 4 show the same trend: baselines often produce wrist jitter, inconsistent trajectories, or implausible finger poses, whereas our method yields smoother and more anatomically plausible motion. These results support the 3D-to-MANO decomposition, which anchors the interaction spatially before refining it with a structured human-motion prior.

4.3 Robotic Manipulation

SimplerEnv Benchmark We evaluate MoT-HRA on SimplerEnv [60] to test visually grounded control under distribution shifts and assess its potential for transfer to real-world manipulation. We focus on WidowX tasks, which vary illumination, background texture, and camera viewpoint, requiring the policy to preserve geometric grounding under distracting observations. This makes the benchmark a stringent test of whether the learned intention representation improves control rather than only fitting human-motion metrics. We compare with a diverse set of strong base-

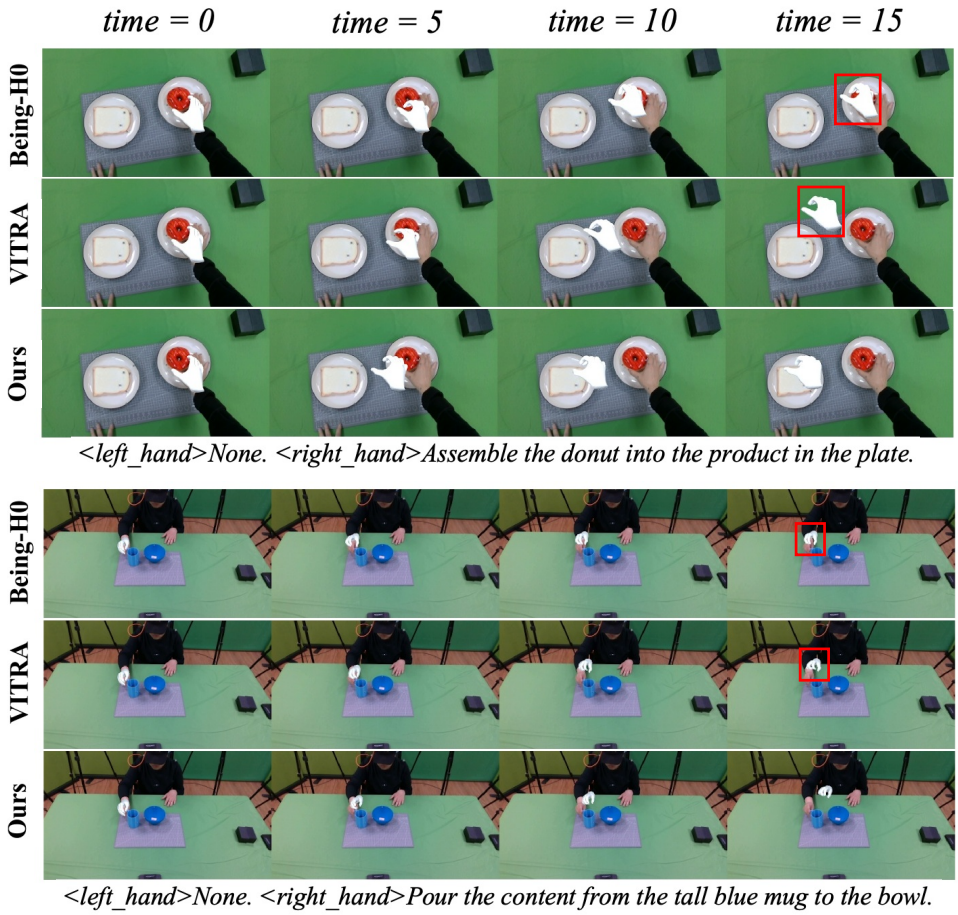


Figure 4: Qualitative comparison in egocentric (top) and third-person (bottom) views. Our method generates more stable and plausible hand motions. Red boxes highlight typical failure cases of the baselines.

lines, including RoboVLMs [19], OpenVLA-OFT [18], π_0 [5], π_0-FAST [17], SpatialVLA [20], and ThinkACT [21].

As shown in Tab. 2, our method achieves the best overall performance and the most consistent behavior across tasks. The gains are most visible on *Spoon* and *Carrot*, which require accurate spatial grounding and stable placement under distracting visual changes. This supports the role of the hierarchical intention representation in preserving fine-grained control under distribution shift. Although SpatialVLA is strongest on *Eggplant*, MoT-HRA remains competitive there while avoiding over-specialization to a single scenario.

Table 2: Success rates (%) on SimplerEnv-WidowX [60] tasks. Spoon denotes *Put Spoon on Towel*, Carrot denotes *Put Carrot on Plate*, Stack denotes *Stack Green Block on Yellow Block*, and Eggplant denotes *Put Eggplant in Yellow Basket*.

Method	Spoon	Carrot	Stack	Eggplant	Average
RoboVLMs [19]	45.8	20.8	4.2	79.2	37.5
OpenVLA-OFT [18]	34.2	30.0	30.0	72.5	41.7
π_0 [5]	29.1	0.0	16.6	62.5	27.1
π_0-FAST [17]	29.1	21.9	10.8	66.6	32.1
SpatialVLA [20]	16.7	25.0	29.2	**100.0**	42.7
ThinkACT [21]	58.3	37.5	8.7	70.8	43.8
Ours	**78.1**	**62.5**	**40.6**	83.3	**66.1**

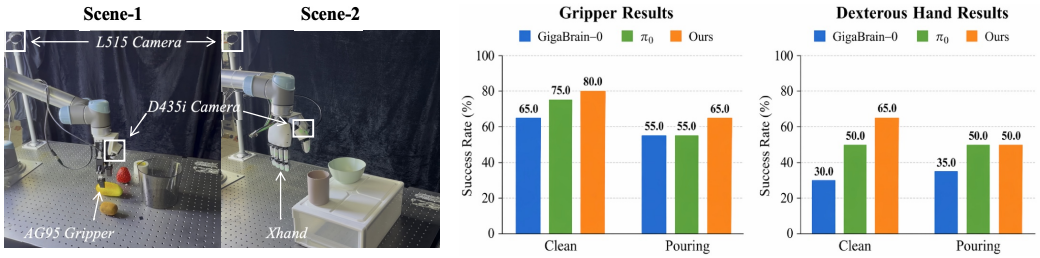


Figure 5: Real-world evaluation on long-horizon manipulation tasks. Both gripper and dexterous-hand settings are tested on *Clean* and *Pouring* under out-of-distribution changes in object position, category, and appearance.

Table 3: Ablation study of key architectural components. ADE, DTW, Rot, and Joint-Rot are evaluated on Ego4D hand motion generation, where lower is better. SimplerEnv Average reports the average success rate (%) on SimplerEnv-WidowX, where higher is better.

3D Trajectory	Intention Expert	Knowledge Insulation	ADE	DTW	Rot	Joint-Rot	SimplerEnv Average
			0.205	0.196	40.16	44.90	48.4
✓			0.182	0.173	36.64	42.53	52.1
✓	✓		0.140	0.133	30.45	35.97	62.7
✓	✓	✓	**0.136**	**0.127**	**28.95**	**34.16**	**66.1**

Real-World Experiments We further evaluate MoT-HRA on real robots with both a parallel gripper and a dexterous hand. Each embodiment is tested on long-horizon *Clean* and *Pouring* tasks after post-training with 150 task-specific trajectories, and evaluation includes out-of-distribution changes in object position, category, and color. Each method is evaluated over 20 trials and compared with π_0 [5] and GigaBrain-0 [22]. As shown in Fig. 5, MoT-HRA achieves more reliable task completion across both embodiments, suggesting that the human-intention priors remain effective under realistic visual and object variation. Additional implementation details and qualitative rollouts appear in Appendix A.4.

4.4 Ablation Study

Architectural Components As shown in Tab. 3, the first row can be regarded as a π_0-style direct VLA architecture, which maps visual-language observations to actions without explicit 3D trajectory grounding, intention experts, or knowledge insulation. Adding the 3D trajectory branch improves both hand-motion metrics and the SimplerEnv average, indicating that spatial grounding provides a useful scaffold for downstream control. Introducing the intention expert brings a larger gain, especially for ADE, DTW, and articulated pose accuracy, showing that explicit human-intention modeling is crucial for coherent wrist trajectories and plausible hand poses.

Knowledge insulation further improves the full model, raising the SimplerEnv average while also reducing motion-generation errors. This suggests that separating intention abstraction from policy learning helps avoid negative interference in large-scale heterogeneous training. Overall, the improvement from the π_0-style baseline to the full MoT-HRA model supports the contribution of the hierarchical structure, although these ablations do not isolate architecture from changes in model capacity.

Human-Demonstration Scale To assess the effect of human-data scale, we construct nested, source-stratified HA-2.2M subsets of 0.10M, 0.50M, 1.00M, and 2.20M unique episodes. We jointly train each subset with the same AgiBot-World data [56], fixing the architecture, optimization schedule (Sec. 4.1), and human-to-robot sampling ratio. Only the number of unique human-video episodes changes.

The robot-only reference retains the full architecture without HA-2.2M data, distinguishing it from the architectural baseline in Sec. 4.4. We use the same hand-motion and SimplerEnv evaluation protocols as above.

Table 4: Effect of HA-2.2M scale under a fixed joint-training protocol. Hours denote approximate human-video duration. ADE is reported in meters on the held-out Ego4D and OakInk evaluations; SimplerEnv Avg. is the average success rate (%) on the four WidowX tasks in Tab. 2. The zero-data row is the robot-only reference; its hand-motion results are not reported. Best reported results are bold.

Human episodes	Approx. hours	Ego4D ADE ↓	OakInk ADE ↓	SimplerEnv Avg. ↑
0	0	–	–	58.7
0.10M	89	0.179	0.232	54.0
0.50M	443	0.162	0.217	59.7
1.00M	886	0.148	0.205	60.4
2.20M	1,950	**0.136**	**0.184**	**66.1**

Table 4 shows consistent improvements across the four nonzero human-data scales. Ego4D ADE decreases from 0.179 to 0.136 m and OakInk ADE from 0.232 to 0.184 m, while SimplerEnv success rises from 54.0% to 66.1%. Transfer is not uniformly beneficial at small scales: 0.10M human episodes yield a lower success rate than the robot-only reference (58.7%), whereas the observed success rates exceed this reference from 0.50M onward. Limited human-demonstration diversity may contribute to the initial degradation, but this experiment does not isolate its cause. The full dataset gives the best reported results on all three evaluations and improves SimplerEnv success by 7.4 percentage points over robot-only training. These results support the value of broader human-demonstration coverage within the evaluated scale range.

5 Conclusion

We presented HA-2.2M, a 2.2M-episode action-language dataset reconstructed from human videos, and MoT-HRA, a hierarchical VLA framework for learning human-intention priors. MoT-HRA separates spatial grounding, human-motion modeling, and embodiment-specific action generation, enabling human videos to supervise robot policies without robot actions in every clip. Experiments on hand motion generation, SimplerEnv, and real-world manipulation show better motion plausibility and control robustness, with ablations supporting each component. The data-scaling study further shows improving motion accuracy and robot success across the tested nonzero human-data scales, with the full dataset outperforming robot-only training by 7.4 percentage points on SimplerEnv. These results suggest that structured human-intention representations can help scale robot policies beyond robot-only data.

Limitations MoT-HRA is still limited by the quality and coverage of automatically curated human demonstrations. Noise, ambiguous hand-object contacts, and imperfect action-language alignment may weaken the learned intention prior. Our evaluations cover representative hand-motion and manipulation tasks, but not highly dynamic interactions, multi-object long-horizon planning, or very different embodiments. Future work should improve data verification, expand embodiment coverage, and add failure detection for more reliable open-world deployment.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant No. 2024YFB3816000 and the Shenzhen Science and Technology Program under Grant Nos. QNXMA20250701093320029 and KJZD20240903100905008.

References

[1] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

[2] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[3] Oier Mees, Dibya Ghosh, Karl Pertsch, Kevin Black, Homer Rich Walke, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, et al. Octo: An open-source generalist robot policy. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

[4] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pages 2679–2713. PMLR, 2025.

[5] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. pi_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

[6] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019.

[7] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017.

[10] 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.

[11] Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, et al. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025.

[12] Yicheng Feng, Wanpeng Zhang, Ye Wang, Hao Luo, Haoqi Yuan, Sipeng Zheng, and Zongqing Lu. Spatial-aware vla pretraining through visual-physical alignment from human videos. *arXiv preprint arXiv:2512.13080*, 2025.

[13] BeingBeyond Team. Being-h0.7: A latent world-action model from egocentric videos. `https://research.beingbeyond.com/being-h07`, 2026.

[14] Mingfei Chen, Yifan Wang, Zhengqin Li, Homanga Bharadhwaj, Yujin Chen, Chuan Qin, Ziyi Kou, Yuan Tian, Eric Whitmire, Rajinder Sodhi, et al. Flowing from reasoning to motion: Learning 3d hand trajectory prediction from egocentric human interaction videos. *arXiv preprint arXiv:2512.16907*, 2025.

[15] Danny Driess, Jost Tobias Springenberg, Brian Ichter, Lili Yu, Adrian Li-Bell, Karl Pertsch, Allen Z Ren, Homer Walke, Quan Vuong, Lucy Xiaoyang Shi, et al. Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705*, 2025.

[16] Xiaomi Robotics Team, Jun Guo, Piaopiao Jin, Jason Li, Peiyan Li, Yingyan Li, Futeng Liu, Wanli Peng, Optimus Qin, Yifei Su, et al. Xiaomi-Robotics-1: Scaling vision-language-action models with over 100K hours of real-world trajectories. *arXiv preprint arXiv:2607.15330*, 2026.

[17] Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. Fast: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

[18] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025.

[19] Xinghang Li, Peiyan Li, Long Qian, Minghuan Liu, Dong Wang, Jirong Liu, Bingyi Kang, Xiao Ma, Xinlong Wang, Di Guo, et al. What matters in building vision–language–action models for generalist robots. *Nature Machine Intelligence*, pages 1–15, 2026.

[20] Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

[21] Chi-Pin Huang, Yueh-Hua Wu, Min-Hung Chen, Yu-Chiang Frank Wang, and Fu-En Yang. Thinkact: Vision-language-action reasoning via reinforced visual latent planning. *arXiv preprint arXiv:2507.16815*, 2025.

[22] GigaBrain Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jie Li, Jiagang Zhu, Lv Feng, et al. Gigabrain-0: A world model-powered vision-language-action model. *arXiv preprint arXiv:2510.19430*, 2025.

[23] Yifan Xie, Mingyang Li, Shoujie Li, Xingting Li, Guangyu Chen, Fei Ma, Fei Richard Yu, and Wenbo Ding. Universal visuo-tactile video understanding for embodied interaction. *arXiv preprint arXiv:2505.22566*, 2025.

[24] Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

[25] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[26] Yushan Liu, Peibo Sun, Shoujie Li, Yifan Xie, Lingfeng Zhang, Xintao Chao, Shiyuan Dong, Fang Chen, Xiao-Ping Zhang, and Wenbo Ding. OA-WAM: Object-addressable world action model for robust robot manipulation. *arXiv preprint arXiv:2605.06481*, 2026.

[27] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, 2022.

[28] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations*, 2023.

[29] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[30] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *IEEE International Conference on Robotics and Automation*, 2023.

[31] 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. *TechRxiv preprint*, 2026.

[32] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[33] Seonghyeon Ye, Joel Jang, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, 2025.

[34] Xiaoyu Chen, Junliang Guo, et al. Igor: Image-goal representations are the atomic control units for foundation models in embodied ai. *arXiv preprint arXiv:2411.00785*, 2024.

[35] Yi Chen, Yuying Ge, et al. Moto: Latent motion token as the bridging language for learning robot manipulation from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[36] Qingwen Bu, Yanting Yang, et al. Univla: Learning to act anywhere with task-centric latent actions. In *Robotics: Science and Systems*, 2025.

[37] Yifan Xie, Hekun Tian, Jinkun Liu, YuAn Wang, Qiao Sun, and Wenbo Ding. GeoLAM: Learning geometry-grounded latent actions from unlabeled human videos. *arXiv preprint arXiv:2609.17099*, 2026.

[38] Hongtao Wu, Ya Jing, et al. Unleashing large-scale video generative pre-training for visual robot manipulation. In *International Conference on Learning Representations*, 2024.

[39] Chi-Lam Cheang, Guangzeng Chen, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[40] Homanga Bharadhwaj, Debidatta Dwibedi, et al. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Conference on Robot Learning*, 2025.

[41] Chuan Wen, Xingyu Lin, et al. Any-point trajectory modeling for policy learning. In *Robotics: Science and Systems*, 2024.

[42] Jianwei Yang, Reuben Tan, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[43] Yuzhe Qin, Yueh-Hua Wu, et al. Dexmv: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision*, 2022.

[44] Hanzhi Chen, Boyang Sun, et al. Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[45] Ruihan Yang, Qinxi Yu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[47] 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.

[48] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*, 35:38571–38584, 2022.

[49] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose++: Vision transformer for generic body pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):1212–1230, 2023.

[50] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[51] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[52] Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, et al. Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. *arXiv preprint arXiv:2601.12993*, 2026.

[53] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*, 2024.

[54] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

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

[56] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[57] Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, et al. Pytorch fsdp: experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*, 2023.

[58] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[59] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022.

[60] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. In *RSS 2024 Workshop: Data Generation for Robotics*, 2024.

[61] Xingting Li, Yifan Xie, Han Liu, Wei Hou, Guangyu Chen, Shoujie Li, and Wenbo Ding. Master micro residual correction with adaptive tactile fusion and force-mixed control for contact-rich manipulation. *arXiv preprint arXiv:2603.15152*, 2026.

A Supplemental Material

A.1 Broader Impacts

This work aims to reduce the dependence of robot learning on expensive robot-specific demonstrations by extracting structured manipulation priors from large-scale human videos. If developed responsibly, such priors could make robot policies easier to adapt across embodiments and environments, lowering the cost of assistive, household, and industrial manipulation systems. At the same time, the use of web-scale and egocentric videos raises important concerns about privacy, consent, representation bias, and dataset provenance. Models trained on human demonstrations may inherit demographic, geographic, cultural, or task-distribution biases from the source videos, and imperfectly grounded policies can still fail in ways that create physical safety risks. We therefore view MoT-HRA as a research step toward more data-efficient manipulation rather than a complete deployment solution. Practical use should include careful dataset governance, removal of sensitive content, transparent documentation of data sources and filtering criteria, human oversight during deployment, and safety validation under the target robot embodiment and environment. The learned intention prior should also not be used to infer private human attributes beyond the manipulation behavior required for the task.

A.2 Future Works

Future work can extend MoT-HRA along both data and modeling axes. On the data side, stronger verification of hand-object contact, temporal boundaries, and language alignment would reduce noise in HA-2.2M, while broader coverage of bimanual, tool-use, deformable-object, and force-sensitive interactions would make the learned intention prior more useful for complex manipulation. On the modeling side, the current hierarchy captures spatial plans and MANO-style hand motion but does not explicitly model contact forces, object dynamics, uncertainty calibration, or long-horizon task structure. Incorporating contact-aware scene representations, physics-informed objectives, and memory over multiple action chunks could improve reliability on multi-stage tasks. Another promising direction is closed-loop human-to-robot data refinement: human videos provide scalable priors, robot trials expose embodiment-specific failures, and active data collection can target cases where the intention prior and robot action expert disagree. Finally, evaluating the method across more robot morphologies and safety-critical out-of-distribution settings will be necessary before drawing stronger conclusions about general-purpose deployment.

A.3 HA-2.2M

Episode Representation Each HA-2.2M episode is designed to be a supervision unit for the hierarchy in Sec. 3.2, rather than a generic video-caption pair. An episode contains an RGB clip, source and timestamp metadata, a normalized action phrase, hand-side validity masks, an image-aligned 3D hand trajectory $\tau_{1:H}$, and a MANO-style hand sequence $m_{1:H}$. When both hands are visible, we keep separate left/right trajectories and annotate the dominant hand when the action has an asymmetric role, e.g., one hand stabilizes a container while the other inserts an object. The clip is resampled to the model horizon H during training, while the original timestamps are retained so that temporal boundaries can be audited without losing the source granularity.

The dataset is intentionally heterogeneous. HowTo100M contributes broad web-scale task diversity, Ego4D and EPIC-KITCHENS contribute egocentric hand-object interactions in daily environments, and Something-Something-V2 contributes short object-centric state changes. Table 5 summarizes the final source composition. We remove any clips that overlap with held-out evaluation splits before training, and all splits are constructed at the source-video level to avoid near-duplicate leakage across train and evaluation episodes.

Coarse Filtering Protocol The coarse filter is optimized for recall, since missed useful demonstrations cannot be recovered later. Gemini [46] is first used on a diverse seed pool to label whether a clip contains visible hand-object manipulation, whether the motion causes an observable state change, and whether the clip is likely to be useful for learning manipulation priors. We then train the V-JEPA-based classifier [47] on these labels and apply it to the full candidate pool. The classifier suppresses common false positives, including idle hand gestures, camera-carried ego-motion

Table 5: Source composition of HA-2.2M after the full curation pipeline.

Source	Episodes	Primary contribution
HowTo100M [6]	1.40M	Web-scale tasks, tools, and scenes
Ego4D [7]	0.63M	First-person manipulation and daily activities
EPIC-KITCHENS [8]	0.12M	Kitchen actions with fine hand-object context
Something-Something-V2 [9]	0.05M	Short object-state transformations
Total	2.20M	Action-language episodes with hand-centric supervision

without interaction, social scenes in which hands are visible but not manipulating objects, pure navigation, extreme motion blur, and clips where the relevant hand is too small or too occluded to reconstruct. We keep ambiguous candidates only when the hand remains visible for a temporally contiguous interval and the surrounding frames suggest contact or imminent contact with an object. This asymmetric criterion favors coverage early in the pipeline, while later reconstruction and fine filtering remove clips that fail geometric or semantic checks.

Spatial Reconstruction and Quality Control The reconstruction stage converts a retained clip into supervision compatible with both the 3D trajectory expert and the MANO intention expert. For every frame, VitPose [48, 49] provides hand keypoints and boxes. HaMeR [50] estimates MANO parameters in crop coordinates, and we map them back to the image frame using the crop scale and translation before temporal smoothing. Depth Anything 3 [51] supplies relative depth cues for the surrounding scene, which are used to stabilize image-aligned 3D waypoints and to reject frames whose hand motion is inconsistent with the local depth layout.

We apply three quality gates before an episode enters the final dataset. First, a visibility gate requires enough valid hand detections over the candidate interval and rejects clips with frequent hand identity swaps. Second, a kinematic gate removes physically implausible wrist jumps, unstable quaternion estimates, and hand sequences whose articulated pose collapses under occlusion. Third, a scene-consistency gate checks that the recovered trajectory is compatible with the image evidence, e.g., the hand should not drift far from the visible hand mask or move independently of a contacted object throughout the whole interval. These gates are deliberately simple and conservative: they do not certify perfect contact physics, but they remove the dominant failure modes that would otherwise teach the intention expert jittery or non-causal hand motion.

Temporal Segmentation The fine filter converts long reconstructed videos into atomic episodes. A V-JEPA temporal segmentation model, trained with 38K manually annotated boundaries, proposes intervals whose internal motion is coherent and whose endpoints align with visible changes in hand-object state. We prefer action boundaries at contact onset, object release, insertion completion, container opening, tool contact, or the completion of a transfer, rather than at arbitrary fixed time windows. Gemini then reviews neighboring intervals with the reconstructed motion summaries and action candidates, merging over-segmented intervals only when they form one short causal operation. This step is important because a robot policy benefits more from compact state-changing primitives than from captions that merely describe every motor phase.

Action Phrase Normalization and Merge Rules Action phrases are normalized to objective, externally observable descriptions. We avoid subjective intent, hidden mental states, and overly broad task names such as "prepare food" when the clip only shows a local operation. The preferred templates are: (i) `[verb] [object]`, e.g., *open the cabinet door*; (ii) `use [tool] to [verb] [target object]`, e.g., *use a towel to wipe the table*; and (iii) `put/place/insert [object] [preposition] [destination]`, e.g., *place the cup on the table*. When source captions are multilingual or noisy, we translate and normalize them into these templates while preserving object and destination names whenever they are visually grounded.

The merge rule follows the atomic state-change principle. Adjacent clips are merged only if they involve the same primary object, or a direct by-product of that object, and no unrelated manipulation occurs between them. For transfer actions, the final spatial relation is the label: *pick up the cup* followed by *place the cup on the table* becomes *place the cup on the table*; *pick up the bowl* followed by *put the bowl into the sink* becomes *put the bowl into the sink*. We do not emit phrases such as

"pick up the cup and then place it" because they overemphasize the preparatory grasp rather than the resulting state. For short causal chains with an immediate by-product, we keep a compact combined phrase, e.g., *peel the banana and discard the peel*, *take the item out of the package*, *pour liquid from the bottle*, *wipe the surface and discard the tissue*, or *put the chopped ingredient into the bowl*. Repeated strokes on the same surface may be merged into one wiping, brushing, or cutting episode, but repeated interactions with different objects remain separate episodes.

We explicitly do not merge clips merely because they belong to the same high-level activity. For example, opening a drawer, picking up a spoon, and stirring a cup are kept separate unless the video shows them as one uninterrupted object-transfer primitive with a single final state. This distinction reduces language-action ambiguity and makes the supervision better aligned with chunked robot control, where each training unit should correspond to a short horizon with a coherent spatial goal.

Auditing and Dataset Governance Because HA-2.2M is automatically curated, we use stratified audits over source dataset, viewpoint, action category, and reconstruction confidence to inspect failure patterns and update filtering prompts and thresholds. The audit focuses on four error types: non-manipulation false positives, incorrect temporal boundaries, wrong object or destination in the language phrase, and unstable hand reconstruction. We also retain provenance metadata and filtering decisions for each episode, which allows problematic sources or action categories to be removed without rebuilding the full dataset. Following the concerns discussed in Broader Impacts, the dataset construction avoids using action labels to infer private attributes of people in the videos, and sensitive or personally identifying content is filtered when detected. These measures do not remove all bias or noise, but they make HA-2.2M a more transparent and controllable source of human-intention supervision for MoT-HRA.

A.4 More Experimental Details

A.4.1 Training Details

We provide additional implementation details for reproducibility. MoT-HRA is initialized from a pretrained PaliGemma2-3B-224 [53] vision-language backbone, while the trajectory, intention, and fine-action heads are trained jointly. Unless otherwise specified, the model is trained on the mixture of HA-2.2M and AgiBot-World [56]. Human-video episodes provide supervision for the 3D trajectory and MANO-style intention branches, robot episodes provide supervision for the 3D trajectory and fine-action branches, and missing targets are removed from the loss by branch-specific masks. We pad all continuous state/action targets to a maximum dimension of 134, corresponding to two 67-D hand states, and supervise only valid dimensions. This padding-and-mask formulation lets single-hand, bimanual, and robot-action samples share the same transformer interface without forcing unavailable hand sides or embodiment-specific joints to contribute loss.

Following the main implementation details, the chunk horizon is set to $H = 15$. We resize each RGB observation to 224×224 with padding and apply training-time image augmentation only on the training split. Language instructions are tokenized to a maximum length of 256, and instruction dropout is set to 0.1. Continuous states and actions are normalized with fixed per-embodiment statistics before denoising and are unnormalized only for evaluation or execution. For the flow-matching objectives, noisy action/intention chunks are sampled from Gaussian noise and interpolated with the target chunk; the model predicts the corresponding velocity field with masked MSE. Wrist and articulated-finger dimensions are aggregated by hand components so that the low-dimensional wrist signal is not dominated by the larger joint-rotation block. We use unit weights in the joint objective unless an ablation states otherwise, and gradients from robot-action losses are not propagated into the cached intention representation.

All reported main experiments use PyTorch FSDP2 [57] on 64 NVIDIA Hopper GPUs. Each GPU processes 32 samples, giving a global batch size of 2048 with one gradient-accumulation step. We optimize for 20,000 steps using AdamW with learning rate 2.5×10^{-5}, $(\beta_1, \beta_2) = (0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 10^{-10}. The learning rate is linearly warmed up for 1,000 steps and then follows cosine decay. We maintain an exponential moving average (EMA) of model parameters, enable activation checkpointing for SigLIP encoder layers and Gemma decoder/expert layers, and use 8 dataloader workers per GPU. Checkpoints are saved every 5,000 optimization steps; unless noted, evaluation uses the EMA checkpoint at the end of training.

A.4.2 Inference Details

All inference results are obtained with a single NVIDIA A100 80G GPU. We run the EMA checkpoint in evaluation mode with gradients disabled. At test time, the same image and language preprocessing as training is used except that data augmentation and instruction dropout are disabled. The input RGB observation is resized to 224×224 with padding, the instruction is tokenized, and continuous states are normalized with the fixed training statistics. After prediction, action chunks are unnormalized, invalid padded dimensions are discarded, and delta-controlled dimensions are converted back to absolute commands when required by the benchmark or robot interface.

MoT-HRA generates outputs in the same hierarchical order as the factorization in Sec. 3.2. The vision-language expert first produces the 3D trajectory tokens. Conditioned on this spatial plan, the intention expert samples the MANO-style hand-motion sequence with classifier-free guidance [58]; we use guidance scale 6.0 for all MANO-generation experiments. The fine expert then predicts the embodiment-specific action chunk conditioned on the image-language context, predicted trajectory states, and read-only intention states. For both continuous denoising branches, we start from standard Gaussian noise and integrate the learned flow with 10 uniformly spaced Euler steps from noise to data. We keep the full chunk of $H = 15$ actions for offline metrics. For closed-loop robot rollout, the policy is queried repeatedly and only the earliest actions in the chunk are executed before re-planning, which reduces compounding error while preserving temporally coordinated chunk prediction.

For stochastic hand-motion generation, we follow the main hand-motion evaluation protocol and sample five trajectories per clip using different random seeds, then average the metrics across samples. For deterministic robot-control benchmarks, we fix the random seed and use a single sampled action chunk per observation to avoid test-time selection bias. All baselines are evaluated with their released or reported inference settings whenever available; otherwise, we use the same image resolution, action horizon, and normalization protocol as MoT-HRA for a controlled comparison.

A.4.3 Evaluation Metrics

Average Displacement Error (ADE) Following hand-trajectory forecasting practice [14], ADE measures the mean Euclidean error of predicted wrist positions over the full prediction horizon. Let $\hat{\mathbf{p}}_t^{s,k}, \mathbf{p}_t^s \in \mathbb{R}^3$ denote the predicted and ground-truth wrist positions for hand side $s \in \{L, R\}$, timestep $t \in \{1, \dots, H\}$, and stochastic sample $k \in \{1, \dots, K\}$, and let $\mu_t^s \in \{0, 1\}$ indicate whether that hand state is valid. We average over all valid hand-timestep pairs and over the $K = 5$ generated samples used in the main experiments, rather than selecting the best sample, so the metric reflects expected generation quality.

$$\mathrm{ADE} = \frac{1}{K \sum_s \sum_{t=1}^{H} \mu_t^s} \sum_{k=1}^{K} \sum_{s \in \{L,R\}} \sum_{t=1}^{H} \mu_t^s \left\| \hat{\mathbf{p}}_t^{s,k} - \mathbf{p}_t^s \right\|_2 . \tag{10}$$

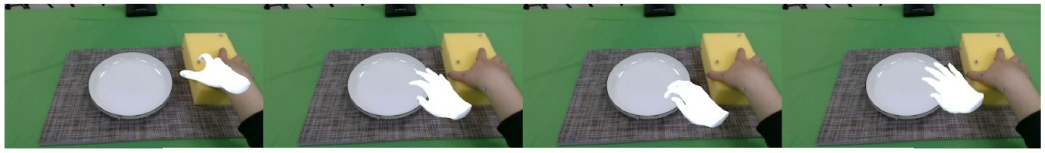
Dynamic Time Warping Distance (DTW) DTW evaluates whether the predicted wrist trajectory follows the same spatial path as the ground truth while allowing small temporal shifts, such as an early contact or a delayed release. For each valid hand side, we remove padded timesteps and compute the minimum-cost monotone alignment path between the predicted and ground-truth wrist-position sequences, then average the normalized alignment costs across samples and hands. This complements ADE by penalizing trajectory-shape mismatch without over-penalizing phase offsets that preserve the intended manipulation path. Here $T_s = \sum_{t=1}^{H} \mu_t^s$ is the valid trajectory length for side s, $\mathcal{H}_v = \{s : T_s > 0\}$ is the set of visible hand sides, and $\Pi(T_s, T_s)$ denotes all monotone DTW paths from $(1, 1)$ to (T_s, T_s).

$$\begin{aligned} \pi_\star^{s,k} &= \arg \min_{\pi \in \Pi(T_s, T_s)} \frac{1}{|\pi|} \sum_{(i,j) \in \pi} \left\| \hat{\mathbf{p}}_i^{s,k} - \mathbf{p}_j^s \right\|_2 , \\ \mathrm{DTW} &= \frac{1}{K |\mathcal{H}_v|} \sum_{k=1}^{K} \sum_{s \in \mathcal{H}_v} \frac{1}{|\pi_\star^{s,k}|} \sum_{(i,j) \in \pi_\star^{s,k}} \left\| \hat{\mathbf{p}}_i^{s,k} - \mathbf{p}_j^s \right\|_2 . \end{aligned} \tag{11}$$

Rotation Error (Rot) Rot measures wrist-orientation fidelity in degrees using the geodesic distance on $SO(3)$, which is the standard angular distance induced by the relative rotation matrix.



<left_hand>None. <right_hand>Press the button on the scale.



<left_hand>None. <right_hand>Wipe the plate with the sponge.



<left_hand>Scrape the content from the gray-green porcelain bottle to the plate. <right_hand>None.

Figure 6: Additional hand-motion generation examples.

Predicted and ground-truth wrist rotations are converted from normalized quaternions to rotation matrices $\hat{\mathbf{R}}^{s,k}_{t,0}, \mathbf{R}^{s}_{t,0} \in SO(3)$ before evaluation, with index 0 denoting the wrist. This metric is important because two trajectories can have similar wrist positions while implying different grasp approach directions or palm orientations.

$$
\begin{aligned}
d_{\text{geo}}(\hat{\mathbf{R}}, \mathbf{R}) &= \frac{180}{\pi} \cos^{-1}\left(\text{clip}\left(\frac{\text{tr}(\hat{\mathbf{R}}^{\top}\mathbf{R}) - 1}{2}, -1, 1\right)\right), \\
\text{Rot} &= \frac{1}{K \sum_s \sum_{t=1}^{H} \mu_t^s} \sum_{k=1}^{K} \sum_{s \in \{L,R\}} \sum_{t=1}^{H} \mu_t^s d_{\text{geo}}\left(\hat{\mathbf{R}}^{s,k}_{t,0}, \mathbf{R}^{s}_{t,0}\right).
\end{aligned}
\tag{12}
$$

Joint Rotation Error (Joint-Rot) Joint-Rot extends the wrist rotation metric to the articulated MANO finger joints, directly testing whether the intention expert generates plausible local hand configuration rather than only accurate global wrist motion. For each valid hand state, we evaluate the geodesic angular error for the $J = 15$ MANO finger joints after quaternion sign canonicalization and normalization, and then average over joints, timesteps, hand sides, and stochastic samples. This metric is stricter than Rot because it penalizes implausible or contact-inconsistent finger articulation even when the wrist pose is close to the ground truth.

$$
\text{Joint-Rot} = \frac{1}{KJ \sum_s \sum_{t=1}^{H} \mu_t^s} \sum_{k=1}^{K} \sum_{s \in \{L,R\}} \sum_{t=1}^{H} \sum_{j=1}^{J} \mu_t^s d_{\text{geo}}\left(\hat{\mathbf{R}}^{s,k}_{t,j}, \mathbf{R}^{s}_{t,j}\right), \quad J = 15. \tag{13}
$$

A.4.4 More Examples

Figure 6 provides additional qualitative hand-motion generation results beyond Fig. 4. Across these examples, MoT-HRA preserves the two properties that are most important for using human motion as an intermediate intention representation. First, the predicted wrist trajectories remain spatially anchored to the manipulated object, rather than drifting with background clutter or camera-view variation. Second, the articulated hand poses evolve smoothly over the action horizon, producing grasp- and release-like configurations that are consistent with the visible object interaction.

Figure 7 shows detailed real-world rollouts for the two long-horizon tasks summarized in Fig. 5. For *Clean*, the language instruction is "Clean everything on the table into the bucket." This task requires

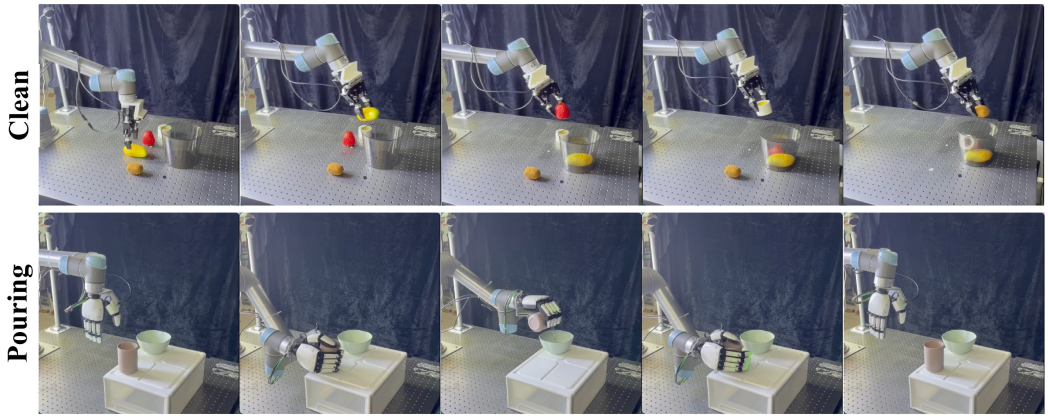


Figure 7: Additional real-world rollouts. The *Clean* task uses the instruction "Clean everything on the table into the bucket." The *Pouring* task uses the instruction "Pour the water from the cup into the bowl." The examples show that MoT-HRA can maintain spatial grounding across multi-stage manipulation under object and appearance variation.

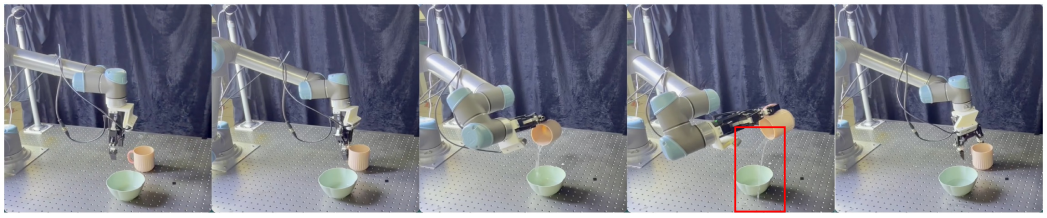


Figure 8: Representative failure case on the real-world *Pouring* task. The robot executes the intended source-to-target sequence but spills water outside the bowl, highlighting the need for stronger contact-, force-, and dynamics-aware modeling.

repeated object localization, grasp selection, transport, and release into a common container while the visual scene changes after each object is removed. The rollout illustrates that the policy does not merely move toward the nearest salient object; it repeatedly re-grounds the next manipulation target and executes a sequence of spatially consistent pick-and-place primitives. For *Pouring*, the instruction is "Pour the water from the cup into the bowl." This task stresses a different aspect of the hierarchy: the policy must approach the cup, establish a stable grasp, move above the bowl, and rotate the end effector so that the liquid is transferred into the target container. The successful sequence suggests that the human-intention prior helps organize the task into interpretable stages, while the fine expert adapts this prior to the robot's embodiment-specific control interface.

We also include a representative failure case in Fig. 8. During a *Pouring* rollout, the robot reaches the cup and initiates the pouring motion, but the water spills outside the bowl. This failure is informative because the high-level intention is largely correct: the policy identifies the source cup, moves it toward the target bowl, and attempts the appropriate rotation. The error occurs in the final contact- and dynamics-sensitive phase, where small deviations in cup pose, bowl alignment, tilt rate, or liquid state can change the outcome. This supports the limitation discussed: MoT-HRA captures spatial plans and MANO-style human motion priors, but it does not explicitly model fluid dynamics, contact forces, or uncertainty-aware recovery. Future extensions with contact-aware state estimation, force/torque feedback [61], and physics-informed objectives may reduce such failures, especially for tasks whose success depends on precise material flow rather than only geometric reachability.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the dataset and hierarchical-model contributions. Section 4.4 reports the observed scaling trend and distinguishes the proposed explanation for the small-data degradation from an established cause.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations paragraph in Section 5 discusses reconstruction noise and evaluation scope; Section 4.4 also notes that small amounts of human data can reduce robot success.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the model, training objectives, optimization settings, and evaluation protocols in the methods, experiments, and Appendix A.4. Section 4.4 specifies the nested, source-stratified subsets and the settings held fixed across scales.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: A public release link for the code and curated dataset is not provided in this version. The paper describes the curation pipeline, architecture, and experimental protocols, but these descriptions do not constitute open access to the artifacts.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix A.4 provide optimization, preprocessing, inference, and evaluation details. Section 4.4 describes the controlled data-scaling protocol.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report aggregate metrics without error bars, confidence intervals, or statistical significance tests. The scaling results describe the observed trend across the tested dataset sizes, and the explanation for the small-data degradation remains a hypothesis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section 4.1 and Appendix A.4 report the GPU types and counts, batch size, optimization steps, and inference hardware. Per-run training time, training-memory usage, and total project compute are not reported.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

Justification: Appendix A.1 discusses potential benefits for robot learning and risks related to privacy, dataset bias, provenance, and physical safety.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper poses no such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly credit all used assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: This paper introduces the new dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This paper describes the usage of LLMs.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.