UNIWAM: UNIFIED WORLD-ACTION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language-action models benefit from the understanding and reasoning capabilities of pretrained vision-language models, but action-only supervision provides limited grounding in world dynamics. Conversely, world-action models inherit spatiotemporal priors from video generation models, yet remain limited in semantic understanding and reasoning under distribution shifts. We introduce UniWAM, a unified architecture that integrates a physical reasoner, a world generator, and an action predictor to jointly learn semantic understanding of the physical world, visual generation, and action prediction. To adapt the vision-language component to embodied tasks while preserving its inherited language capabilities, we represent low-level robot actions in natural language and introduce a pre-training recipe that assigns complementary supervision from visual question answering (VQA) data, human egocentric data, and robot demonstrations to the appropriate model components. During post-training, future visual noise augmentation reduces reliance on precise future predictions, while history-conditioned flow matching uses encoded action history to initialize action generation. Together, these designs significantly reduce denoising steps while maintaining performance. UniWAM achieves state-of-the-art (SOTA) performance across multiple evaluations, including in-distribution performance, robustness, generalization, instruction following, and long-horizon task execution.

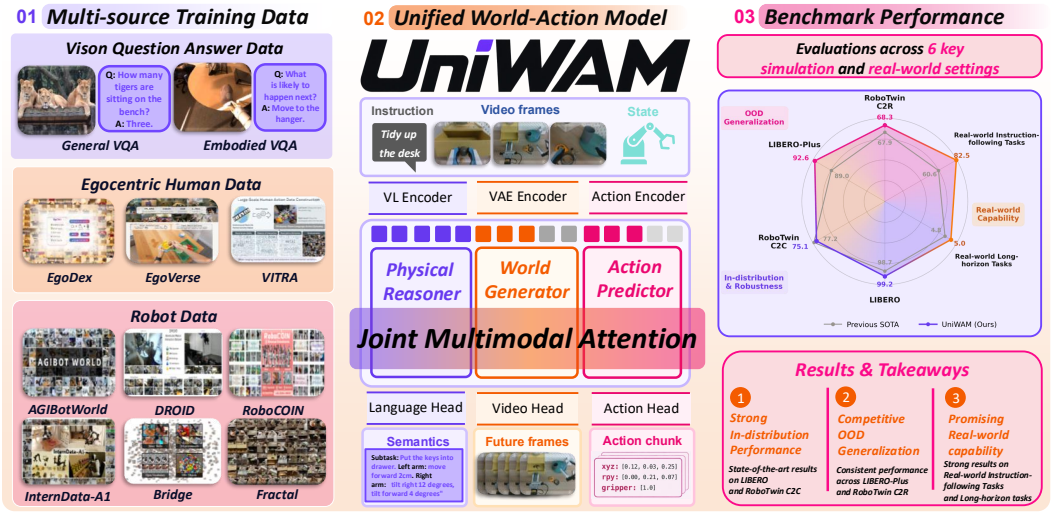


Figure 1: **UniWAM** unifies physical reasoning, visual generation, and action prediction in a unified architecture and pretrains on robot and human egocentric data as well as VQA data. The resulting model exhibits strong robustness, generalization, instruction-following, and reasoning abilities.

1 INTRODUCTION

In recent years, robot foundation models built upon Vision-Language Models (VLMs) (Bai et al., 2025; Beyer et al., 2024; An et al., 2025), namely Vision-Language-Action Models (VLAs) (Black et al., 2025a; Generalist Team, 2025; Bai et al., 2026), have demonstrated strong effectiveness, as they can effectively leverage the semantic understanding capabilities acquired through large-scale

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

VLM pre-training. However, using actions alone as the supervision signal limits the model's ability to understand world dynamics. World-Action Models (WAMs) (Ye et al., 2026; Bi et al., 2026a; Li et al., 2026c; Yang et al., 2026a) inherit strong spatiotemporal priors from Video-Generation Models (VGMs) (Team Wan et al., 2025; Li et al., 2026e) pre-trained on large-scale video data, improving their understanding of dynamics through predicting the next frame, but their understanding in out-of-distribution (OOD) scenarios and reasoning capabilities for complex tasks remain limited. This paper asks: How can we combine the strong reasoning and generalization capabilities of VLMs with the strong dynamics understanding of WAMs, thereby building a truly generalist robot?

To address this question, we introduce UniWAM, integrating a VLM-based physical reasoner, a VGM-based world generator, and an action predictor within a unified architecture, drawing inspiration from the idea of Mixture-of-Transformer (MoT) (Liang et al., 2025; Bi et al., 2026a). The hybrid model jointly supports semantic understanding, visual prediction, and action generation. Figure 1 summarizes the training data, unified architecture, and benchmark performance of UniWAM. **Pre-training** such a hybrid model presents two key challenges: 1) Adapting the VLM to embodied tasks (Wang et al., 2026) requires updating its parameters to acquire embodied knowledge. However, such adaptation must preserve the language capabilities inherited from pretraining, which calls for supervision signals that remain compatible with the VLM's pretraining distribution. To this end, we show that directly representing low-level robot actions using natural language (Zha et al., 2026), thereby aligning the VLM's action supervision with its pretraining input-output distribution, provides an effective way to train the VLM for embodied understanding while preserving its pre-trained capabilities. 2) Another key challenge lies in designing appropriate supervision signals for each expert using heterogeneous data, as differences in data distributions and training objectives may lead to conflicts among the optimization processes of different experts. To this end, we construct a pretraining mixture consisting of Visual Question Answering (VQA), robot data, and human data, and carefully assign their supervision to different components. Specifically, VQA data supervises the VLM to maintain its pretraining knowledge. Human data supervises both the VLM and VGM, thereby fully exploiting the physical knowledge contained in human videos while avoiding supervision from low-precision action labels. Since robot data exhibits the highest physical consistency with downstream post-training and provides the most accurate action annotations, it is used to supervise the VLM, VGM, and action predictor, enabling the learning of precise action outputs.

During post-training, we continue the pre-training principle of maintaining unified capabilities and jointly supervise the three experts to adapt the model's understanding, generation, and prediction capabilities to specific embodied scenarios. We additionally introduce future-visual noise augmentation and history-conditioned flow matching. The former partially perturbs future visual latents while preserving the conditioning frames, encouraging the action expert to extract control-relevant semantics from coarse visual representations rather than rely on precise future predictions, thus strengthening the interaction between visual generation and action prediction. Furthermore, the action expert leverages the dynamics and temporal continuity encoded in historical sequences. It maps the action history into a high-dimensional latent space and uses the resulting representation to initialize the flow-matching process, thereby allowing historical motion information to guide the generation of future action trajectories. Together, these designs substantially reduce the number of denoising steps while maintaining performance, improving inference efficiency. Overall, at the architectural level, we build a MoT that connects language reasoning, video generation, and action prediction through joint attention. Our contributions are summarized in three aspects.

- For pre-training, we construct a three-source pre-training scheme comprising human data, robot teleoperation data, and VQA data, and systematically evaluate the contribution of each data source through source-level ablation studies.
- For post-training, we explicitly formulate how recent action history and perturbed future visual latents are incorporated into the joint flow-matching process.
- Extensive evaluations demonstrate that our UniWAM achieves SOTA performance across multiple dimensions, including in-distribution performance, robustness, generalization, instruction following, and long-horizon task execution.

2 Related Work

Large-scale Pretraining for WAMs. WAMs typically inherit dynamics priors from video generation backbones and pretrain on large-scale robot and egocentric data. DreamZero (Ye et al., 2026)

and Cosmos Policy (Kim et al., 2026) align internet-pretrained video generation models with robot actions. LingBot-VA (Li et al., 2026c) instead pretrains causal video–action prediction on large-scale cross-embodiment data. GigaWorld-Policy (GigaWorld Team et al., 2026) continues pretraining on mixed robot and egocentric video with an action-centered objective, using action-conditioned future-video prediction as supervision. Motus (Bi et al., 2026a), MotuBrain (MotuBrain Team et al., 2026), and Cosmos 3 (NVIDIA et al., 2026) jointly pretrain understanding, generation, and action experts in a MoT. X-WAM (Guo et al., 2026), Hydra-0 (Li et al., 2026b), and flex-π (Yan et al., 2026) pretrain 4D representations beyond RGB video–action modeling. Being-H0.7 (Luo et al., 2026c) first scales this pretraining to 200k hours, and Dyna-2 (Dyna Robotics, 2026) further extends the scale to 1 million hours. ABot-M0.5 (Chen et al., 2026) pretrains an action-unconditioned world model and latent actions on public and synthetic robot videos before specializing to mobile manipulation. However, prior WAMs focus more on video–action modeling and lack language–action modeling. In contrast, we design a hybrid architecture and a unified training recipe, enabling UniWAM to support physical reasoning, visual generation, and action prediction.

Learning from Human Data. Egocentric human videos provide scalable embodied experience beyond costly robot demonstrations. They capture diverse human–object interactions and the resulting physical changes, covering many objects, environments, and long-horizon activities. Recent datasets enrich raw videos with language, human motion, geometry, and action-related signals, making them suitable for large-scale embodied pretraining (Grauman et al., 2022; Damen et al., 2018; Hoque et al., 2026; Liu et al., 2022; Punamiya et al., 2026; Li et al., 2026g;f; Deng & Zhou, 2026). Early work leveraged human data through latent actions, future prediction, and geometric motion cues (Ye et al., 2025; Bjorck et al., 2025; Bu et al., 2025; Luo et al., 2026a; Li et al., 2026d; Zheng et al., 2025). Closer to policy learning, human motions and manipulation trajectories are reconstructed and aligned with robot action spaces for VLA and WAM's pretraining, allowing human data to support both action learning and world modeling (Kareer et al., 2025; Luo et al., 2025; Bi et al., 2026b; Yoshida et al., 2025; Li et al., 2025; Fu et al., 2025; Luo et al., 2026b; Zheng et al., 2026b; Kareer et al., 2026). In this paper, we develop more effective supervision for human data in WAM training and enable co-training with robot data.

3 METHOD

3.1 MODEL ARCHITECTURE

As shown in Figure 2, UniWAM unifies semantic understanding, language modeling, visual generation, and action prediction in a Mixture-of-Transformers (MoT) architecture (Liang et al., 2025), i.e., physical reasoner, world generator, and action predictor. They exchange information through joint attention. Given the current observation o_t, proprioceptive state s_t, and task instruction I, the model learns the joint distribution of language outputs, future observations, and actions:

$$p_\theta(y, o_{t+1:t+h}, a_{t+1:t+h} \mid o_t, s_t, I)\,, \tag{1}$$

where y is the output language sequence of physical language tokens or answer token. Although UniWAM shares similarities with world action models such as Motus (Bi et al., 2026a) in its decoder-layer structure, it differs in its training recipe, data composition, and overall capabilities.

Modality-specific experts. The physical reasoner projects features from Qwen3-VL-2B-Instruct (Bai et al., 2025), conditioned on o_t and I, into semantic tokens and retains its language head for autoregressive language generation, trained with the language supervision objective in Eq. 4. The world generator uses the Wan2.2-TI2V-5B backbone (Team Wan et al., 2025) to predict future visual flow in a frozen autoencoder's latent space, with the current observation latent kept clean. The action predictor embeds the action chunk with positional information and a current-state token encoding s_t, and predicts its continuous flow field.

Cross-modal interaction. Let X_m^ℓ denote the tokens of modality $m \in \{u, v, a\}$ at layer ℓ, corresponding to understanding, video, and action. After expert-specific normalization and, for video and action, timestep modulation, we project the tokens into a common attention space:

$$Q_m = \widetilde{X}_m^\ell W_m^Q, \qquad K_m = \widetilde{X}_m^\ell W_m^K, \qquad V_m = \widetilde{X}_m^\ell W_m^V. \tag{2}$$

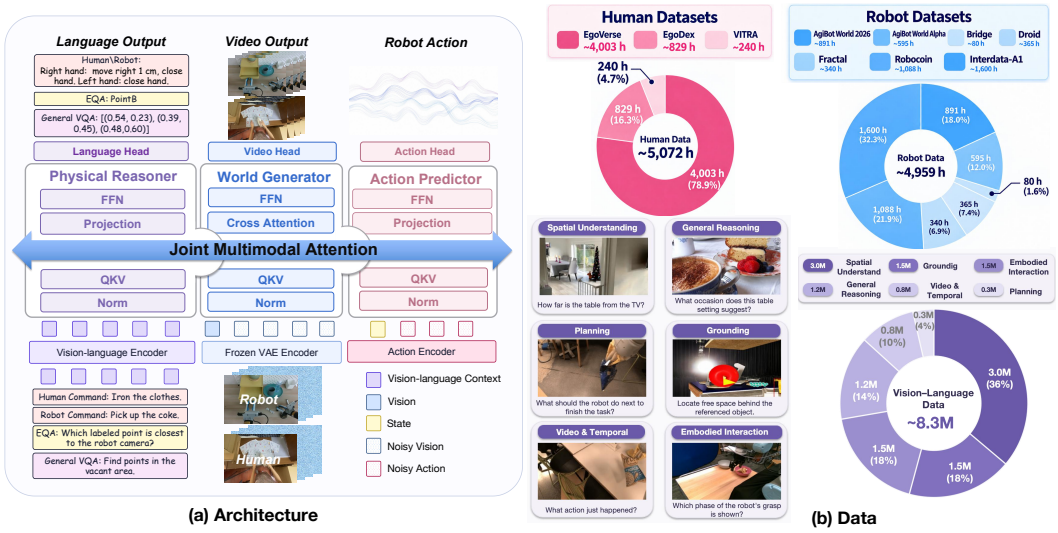


Figure 2: **Overview of our UniWAM. (a)** UniWAM comprises a physical reasoner, a world generator, and an action predictor interconnected through joint multimodal attention. Robot data and human egocentric data are jointly used to train all three experts, with supervision for the physical reasoner provided in the form of physical language. General VQA and Embodied Question Answering (EQA) data are additionally used to train the physical reasoner. **(b)** Overview of the pretraining data: Human and robot data are measured in hours; vision–language data are measured in millions of samples. Percentages are computed within each data category.

For each attention head, the projected sequences are concatenated along the token dimension and jointly attended:

$$O = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right) V, \qquad Q = [Q_v; Q_a; Q_u], \tag{3}$$

where K and V follow the same concatenation order, d is the head dimension, and M is the attention mask. The outputs O are split by modality and returned through expert-specific projections, residual connections, and feed-forward networks. This repeated interaction lets action tokens incorporate both semantic context and evolving visual predictions while preserving modality-specific processing throughout the network.

3.2 Pretraining

3.2.1 Data

Robot Data. Our robot pretraining data combine AgiBotWorld2026 (AgiBot World Team, 2026), AgiBot World Alpha, Bridge (Walke et al., 2023), Droid (Khazatsky et al., 2024), Fractal (Brohan et al., 2023), Robocoin (Wu et al., 2025), and Interndata-a1 (Tian et al., 2025), covering single-arm and dual-arm embodiments. The selected data total 4959 hours, with a per-source breakdown in Figure 2. We curate the training data with a multi-stage filtering pipeline that enforces temporal and kinematic consistency in state–action trajectories while removing corrupted or low-quality visual observations. The detailed procedure is provided in Appendix I.

Human Egocentric Data. Our human data for pretraining are drawn from three large-scale sources: EgoDex (Hoque et al., 2026), EgoVerse (Punamiya et al., 2026), and VITRA (Li et al., 2025). These sources provide 5072 hours of human video spanning a broad range of manipulation tasks, objects, and environments, with rich language and dense motion annotations. We curate subtask-level annotations through a VLM-based automatic annotation pipeline in Appendix E.1 to align language descriptions with the local manipulation behavior in each video segment, rather than only the overall episode goal. Across all three datasets, we temporally align the videos and subtask descriptions with the corresponding head-camera and bilateral wrist trajectories, express wrist motion in the camera coordinate frame at the conditioning time, and filter out samples with missing annotations, invalid temporal alignment, or implausible wrist geometry.

VQA Data. To provide visual–semantic and embodied reasoning supervision for downstream policy learning, we incorporate multi-source visual question answering into pretraining , while helping preserve the physical reasoner's pretrained semantic knowledge (Lin et al., 2026; Gong et al., 2026; Wan et al., 2026). For spatial and scene understanding, we combine SAT (Ray et al., 2024), RoboPoint (Yuan et al., 2024), RefSpatial (Zhou et al., 2026), VST-P (Yang et al., 2025), SenseNova-SI (Cai et al., 2026b), CLEVR (Johnson et al., 2017), and GRiD-3D (Lee et al., 2022), covering grounding, spatial relations, geometric reasoning, and compositional visual understanding. We further include VSI-590K (Yang et al., 2026b), SIMS-VSI (Brown et al., 2025), RoboVQA (Sermanet et al., 2024), and Robo2VLM (Chen et al., 2025a) to provide temporal, task-state, and manipulation-oriented supervision.

3.2.2 Physical Language Supervision

To preserve pretrained VLM knowledge during action supervision, we encode end-effector motions as natural-language descriptions, following LAP (Zha et al., 2026). A fixed coordinate convention and structured templates provide deterministic descriptions of continuous actions at arbitrary precision. We apply this supervision to both human and robot data: task instructions specify the overall goal, while physical-language targets describe the local behavior associated with the current observation. Given observation o_t and instruction I, VLM predicts target tokens autoregressively:

$$\mathcal{L}_{\text{lang}} = -\sum_{j=1}^{T} \log p_\theta(y_j \mid o_t, I, y_{<j}), \tag{4}$$

where loss is computed only on answer tokens. Describing end-effector pose changes provides shared motion semantics across embodiments and complements continuous action supervision.

3.2.3 Training Objectives

Let $\hat{u}^v_\theta$ and $\hat{u}^a_\theta$ denote the visual and action velocity fields predicted by the model with parameters θ at the sampled noise levels τ_v and τ_a, respectively. We supervise these predictions with flow matching losses, using the differences between source and clean target samples as the target velocity fields:

$$\mathcal{L}_v = \mathbb{E}_{\mathcal{D},\tau_a,\tau_v,A_t^1,Z_t^1}\left[\left\|\hat{u}^v_\theta - (Z_t^1 - Z_t^0)\right\|_2^2\right], \qquad \mathcal{L}_{\text{act}} = \mathbb{E}_{\mathcal{D},\tau_a,\tau_v,A_t^1,Z_t^1}\left[\left\|\hat{u}^a_\theta - (A_t^1 - A_t^0)\right\|_2^2\right] \tag{5}$$

where $\mathcal{D}$ is the training data distribution, $A_t^0 = a_{t+1:t+h}$ is the target action chunk, and Z_t^0 contains future visual latents from the frozen video autoencoder. The visual variables and $\hat{u}^v_\theta$ cover only future frames; the conditioning observation o_t does not contribute to the visual loss. We independently sample action and video noise levels τ_a and τ_v from their respective training schedules and form the interpolants $A_t^{\tau_a} = (1-\tau_a)A_t^0 + \tau_a A_t^1$ and $Z_t^{\tau_v} = (1-\tau_v)Z_t^0 + \tau_v Z_t^1$. Both velocity fields are predicted in a joint forward pass. During pretraining, both the action source A_t^1 and the visual source Z_t^1 are Gaussian noise. Generation proceeds from the source ($t = 1$) to the clean target ($t = 0$). The total objective combines action, visual, and language losses:

$$\mathcal{L} = w_a\mathcal{L}_{\text{act}} + w_v\mathcal{L}_v + \mathcal{L}_{\text{lang}}, \tag{6}$$

where w_a and w_v balance the action and visual objectives. The language loss uses the autoregressive objective in Eq. 4, with the answer sequence given by either a physical language description or a visual question answering (VQA) response.

3.3 Posttraining

In addition to maintaining supervision over all three experts , 1)we further introduce future-frame noise augmentation to promote information transfer between visual and action prediction, and 2) history-conditioned flow matching to improve action prediction efficiency and performance.

3.3.1 Future-frame Noise Augmentation

Joint attention can make action prediction overly dependent on nearly clean future visual latents that are sometimes unreliable at inference. We therefore augment only future-frame latents during

Table 1: **Success rate (%) on both the LIBERO and RoboTwin 2.0 benchmarks.** Gray shading indicates OOD evaluation.

Benchmark	LIBERO					RoboTwin 2.0		
Method	Spatial	Object	Goal	Long	Avg.	C2C	C2R	Avg.
StarVLA (StarVLA Community, 2026)	97.8	98.8	97.4	92.0	96.5	46.50	3.20	24.80
Fast-WAM (Yuan et al., 2026)	98.2	**100.0**	97.0	95.2	97.6	70.20	1.20	37.70
π_0 (Black et al., 2025b)	96.8	98.8	95.8	85.2	94.2	46.42	16.34	31.38
$\pi_{0.5}$ (Black et al., 2025a)	98.8	98.2	98.0	92.4	96.9	70.70	46.00	58.35
X-VLA (Zheng et al., 2026a)	98.2	98.6	97.8	97.6	98.1	68.00	20.90	44.45
Spatial Forcing (Li et al., 2026a)	99.4	99.6	98.8	96.0	98.5	**77.20**	26.74	51.97
Xiaomi-Robotics-0 (Cai et al., 2026a)	98.8	**100.0**	98.8	97.2	98.7	62.90	18.20	40.55
UniWAM (ours)	**99.6**	99.6	**99.2**	**98.4**	**99.2**	75.14	**68.32**	**71.73**

posttraining, while retaining joint attention. With probability 0.5 per sample, we further perturb the interpolated visual input $Z_t^{\tau_v}$ as $\widetilde{Z}_t^{\tau_v} = s_{\text{aug}} Z_t^{\tau_v} + (1 - s_{\text{aug}})\epsilon_v$, where $s_{\text{aug}} \sim \mathcal{U}(0.5, 1)$ and $\epsilon_v \sim \mathcal{N}(0, \mathbf{I})$. Otherwise, the input is unchanged. The current observation latent remains clean, and the original flow targets and timestep embeddings are retained. This input perturbation encourages action prediction to tolerate imperfect future visual representations while preserving the observed task context.

3.3.2 History-conditioned Flow Matching

Recent actions provide a structured prior for predicting subsequent actions. For posttraining, we replace the Gaussian noise source of the action flow with a perturbed chunk of previously executed actions, grounding generation in action history to encourage temporal consistency and support refinement with fewer inference steps. Let $A_t^0 = a_{t+1:t+h}$ denote the target action chunk and $A_{t-1} = a_{t-h+1:t}$ the corresponding history of h actions, sampled at the action rate. We construct the source as $A_t^1 = A_{t-1} + \epsilon_a$, where $\epsilon_a \sim \mathcal{N}(0, \eta^2 \mathbf{I})$, $\eta = 0.02$ controls the perturbation magnitude, and $\mathbf{I}$ denotes the identity matrix.

We write A_t^τ for the action chunk at normalized noise level $\tau \in [0, 1]$, with t indexing control time. The interpolation between the clean target and history source is (Lipman et al., 2023):

$$A_t^\tau = (1 - \tau)A_t^0 + \tau A_t^1, \qquad u_t^a = A_t^1 - A_t^0. \tag{7}$$

The action predictor learns this displacement with a mean-squared flow matching loss conditioned on the current context, while exchanging features with the world generator through joint attention.

4 Experiments

We design our experiments to answer the following questions: **(Q1)** Can our UniWAM successfully perform manipulation tasks under in-domain conditions, complex perturbations, and OOD conditions, respectively? (Section 4.1) **(Q2)** Does the unified design of UniWAM improve instruction following and long-horizon manipulation in the real world? (Section 4.2) **(Q3)** Does our unified model scale effectively with data? (Section 4.3) **(Q4)** How does our physical language shape the learned representations and enable strong generalization? (Section 4.3) **(Q5)** How important are our UniWAM's training designs in policy performance? (Section 4.4)

4.1 Simulated Experiments

We mainly conduct the simulated experiments in LIBERO (Liu et al., 2023) and RoboTwin 2.0 Clean2Clean (Chen et al., 2025b), as well as their OOD variants, LIBERO-Plus (Fei et al., 2025) and RoboTwin 2.0 Clean2rand (Community et al., 2026).

In-distribution Evaluation. LIBERO (Liu et al., 2023) and RoboTwin 2.0 Clean2Clean (C2C) are in-distribution benchmarks, where C2C denotes training on clean demonstrations and evaluating the model on clean settings. As shown in Table 1, UniWAM achieves an average success rate of 99.2% on the standard LIBERO benchmark, outperforming the previous best-performing baseline, Xiaomi-Robotics-0, by 0.5%. This result demonstrates that UniWAM maintains strong in-distribution manipulation performance. Table 1 shows UniWAM also achieves a high success rate of 75.14% in

Table 2: **Success rate (%) under all seven LIBERO-Plus perturbation dimensions (OOD evaluation).** Bold entries denote the highest value in each column among the listed methods. Underlined entries denote the second-highest value in each column.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Total
π_0 (Black et al., 2025b)	79.6	21.1	72.5	84.7	86.2	68.3	69.4	67.4
GR00T-N1.6 (NVIDIA GEAR Team, 2025)	92.6	33.5	80.1	93.6	95.4	93.6	75.0	79.4
OpenVLA-OFT (Kim et al., 2025)	92.8	30.3	85.8	94.9	93.9	89.3	77.6	79.5
Spatial Forcing (Li et al., 2026a)	95.2	47.9	73.5	91.2	95.6	92.2	74.8	80.5
MemoryVLA (Shi et al., 2026)	91.4	48.6	79.4	95.2	95.3	94.0	75.7	81.9
$\pi_{0.5}$ (Black et al., 2025a)	87.6	<u>78.4</u>	80.0	92.6	91.6	91.4	81.8	86.2
ACoT-VLA (Zhong et al., 2026)	**96.6**	70.4	79.7	95.1	<u>97.1</u>	<u>95.9</u>	**85.0**	88.0
Kairos (Kairos Team et al., 2026)	<u>95.5</u>	72.6	<u>86.8</u>	<u>97.7</u>	95.8	**96.8**	81.5	<u>89.0</u>
UniWAM (Ours)	92.0	**89.5**	**92.2**	**97.9**	**97.6**	94.3	<u>84.6</u>	**92.6**

RoboTwin 2.0 C2C. The C2C setting demonstrates the model's bi-manual performance with limited training data.

OOD Evaluation for Robustness and Generalization. LIBERO-Plus (Fei et al., 2025) extends LIBERO with systematic variations along seven dimensions, including background textures, camera viewpoints, language instructions, lighting conditions, object layouts, robot initial states, and sensor noise. As shown in Table 2, UniWAM further achieves the best overall performance on LIBERO-Plus, demonstrating strong robustness under diverse distribution shifts and outperforming all listed VLAs and WAMs. At the dimension level, UniWAM achieves the highest success rates among the listed methods under robot (89.5%), language (92.2%), light (97.9%), and background (97.6%) perturbations. These results are consistent with the intended roles of our training objectives: physical language supervision encourages a shared mapping between semantics and actions, reducing reliance on spurious visual cues; pretraining with VQA data helps preserve language understanding; and large-scale world generation training encourages the model to capture underlying physical dynamics that remain consistent across variations in lighting and background.

In RoboTwin 2.0 Clean2Rand (C2R) (Chen et al., 2025b), models are trained using clean data but tested with domain randomization; UniWAM achieves a success rate of 68.32%, the highest among all compared methods. Its success rate decreases by only 6.82 percentage points from C2C to C2R, compared with 24.70 and 50.46 percentage points for $\pi_{0.5}$ and Spatial Forcing, respectively. These results demonstrate that UniWAM capabilities generalize to different heights, distractors, and backgrounds, which benefit from the physical reasoning and dynamics understanding that are unified in our hybrid model.

4.2 Real-world Experiments

In our real-world experiments, we focus on evaluating the model's instruction-following ability and long-horizon task-solving capability. Experiments are conducted on the AgileX Piper platform, where each arm has 6-DoF joints and a parallel gripper. Our dual-arm system consists of two Piper, each equipped with a wrist-mounted camera, together with an additional third-person camera providing a global view of the workspace. We compare UniWAM against representative WAM and VLA methods, Motus and $\pi_{0.5}$. For all methods, we collect 200 demonstrations per task and jointly train a single generalist policy over all tasks. All models are trained for 80K steps on 8 NVIDIA H100 GPUs under the same training protocol.

Instruction-following Tasks. To evaluate instruction-following capability, we design four complementary tasks that require the policy to ground different linguistic components into appropriate manipulation behaviors; task details are provided in Appendix B. We report two complementary metrics: Instruction-Following Rate (IFR) and Task Success Rate (SR). IFR measures whether the policy correctly grounds the language instruction to the intended object and interaction. In contrast, SR measures whether the instructed task is successfully completed in its entirety.

As shown in Fig. 3(a,b), UniWAM consistently achieves strong performance across tasks with varying levels of linguistic and manipulation complexity, in terms of both task success and instruction following. Averaged over the four tasks, UniWAM achieves a success rate of 67.5% and an instruction-following rate of 82.5%, outperforming $\pi_{0.5}$ by 13.1 and 21.9 percentage points, respectively. The improvement is particularly pronounced in language grounding. On Diverse-Interaction, where the policy must jointly identify the target object and infer the intended interac-

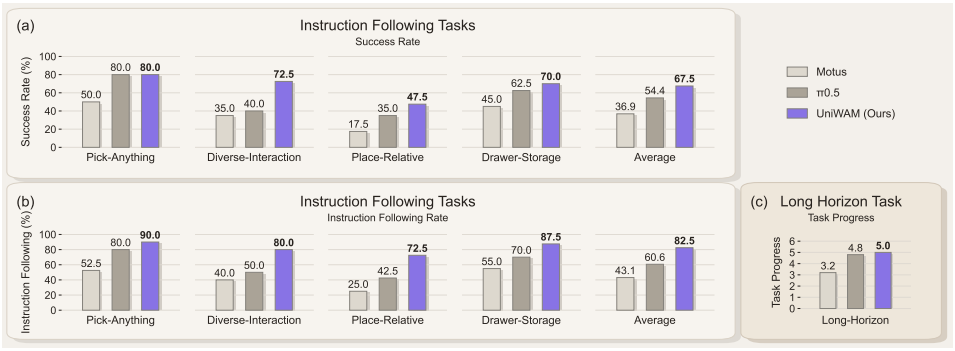
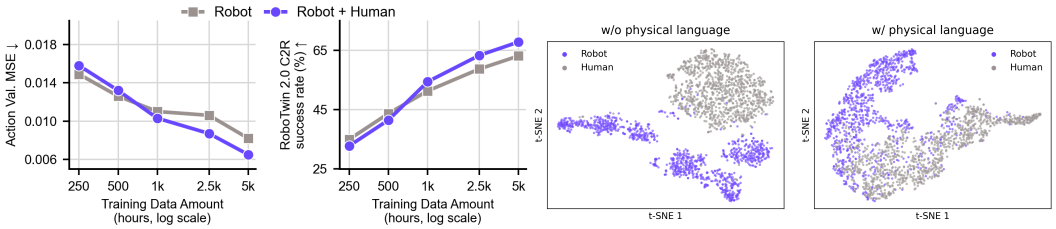


Figure 3: Real-world experiments across four instruction-following tasks and a long-horizon task.



(a) (b)

Figure 4: **Analysis of human–robot co-training.** (a) Data scaling comparison between Robot and Robot + Human. Left: validation MSE; right: RoboTwin 2.0 C2R success rates. The x-axis reports the number of robot-data hours on a logarithmic scale, and Robot + Human augments robot data with a fixed proportion of human data. (b) Visualization of VLM representations from human and robot data. The right plot shows greater overlap between robot (purple) and human (gray) features than the left plot.

tion, UniWAM improves the success rate from 40.0% to 72.5% and the instruction-following rate from 50.0% to 80.0%. These results suggest that our physical language supervision facilitates more reliable grounding of compositional instructions into appropriate manipulation behaviors.

Long-horizon Tasks. We design a multi-stage tabletop organization task to evaluate long-horizon manipulation capability. Since binary task success provides limited information about partial completion in this setting, we use task progress, which ranges from 0 to 6 according to six sequential milestones. A higher score indicates that the policy maintains coherent execution over a longer sequence of dependent manipulation stages. Task details are provided in Appendix B.

As shown in Fig. 3(c), UniWAM achieves an average task progress of 5.0, compared with 4.8 for $\pi_{0.5}$ and 3.2 for Motus. Notably, the policy receives only the high-level instruction at inference time, without an explicit step-by-step description of the required procedure. The model must infer the relevant intermediate subgoals from the scene and its current execution state and progressively complete them over the course of the task. This result indicates that our model exhibits strong long-horizon task understanding, planning, and execution capabilities.

4.3 In-depth Analysis of Human-robot Co-training

Data scaling law emerges. We study how the scale of robot pretraining data affects downstream OOD manipulation performance, and effects brought by added human data. We pretrain the model using 250, 500, 1k, 2.5k, and 5k hours of robot data, with alternative same ratio of human data.

As shown in Figure 4a, increasing the amount of robot data leads to consistent and substantial gains in OOD performance. Average task completion rises monotonically from 35% at 250 hours to 67% at 5k hours, with no signs of saturation in the explored regime. At small data scales, incorporating human data degrades performance. However, once the training data reaches 1k hours, human data begins to yield gains, which become more pronounced as the dataset grows. This pattern suggests

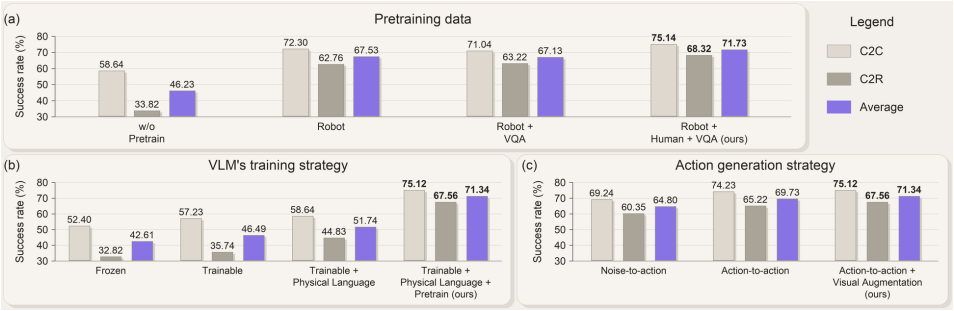


Figure 5: **Ablation studies on RoboTwin 2.0.**

that visual and physical differences between human and robot data at small data scales hinder learning across the two distributions, whereas large-scale data may enable the transfer of shared world dynamics across embodiments, leading to improved generalization. We further observe a remarkably clean log-linear scaling law that validation loss decreases approximately linearly with log data scale, alongside improvements in downstream task success rates.

Human-robot transfer emerges with physical-language pretraining. Figure 4b shows that, with supervision, physical reasoner's representations from robot and human data exhibit greater overlap, particularly in the central and lower regions of the projection. Without this supervision, robot and human samples largely occupy separate regions. These results support the role of physical language supervision in reducing domain gap in the learned representations. This pattern is attributed to shared physical language targets encouraging representations of common manipulation semantics across human and robot embodiments, thereby facilitating the general manipulation capability and improved generalization afforded by mixed-data pretraining.

4.4 Ablation Studies

Multi-source pretraining. Combining robot, human, and VQA data strengthens performance. Figure 5(a) shows the complete three-source configuration achieves the highest success rates among the evaluated variants, reaching 75.14% in C2C, 68.32% in C2R, and 71.73% on average. Compared with training without pretraining, robot-data pretraining improves the average success rate from 46.23% to 67.53%. Adding VQA data primarily improves performance in C2R, likely by enhancing the model's high-level perceptual capabilities. Incorporating human data improves performance in both C2C and C2R, suggesting that human data provides complementary world knowledge and supports fine-grained manipulation across diverse environments.

VLM's training strategy. Training the VLM matters. The cumulative ablation in Figure 5(b) shows that unfreezing the reasoner increases average success from 42.61% to 46.49%. Adding physical language supervision further raises the average to 51.74% and adding pretraining brings the average to 71.34%. This demonstrates that our training recipe is effective and unlocks the benefits of pretraining for VLM.

Action generation strategy. History-based action initialization and future visual augmentation further improve policy performance. As shown in Figure 5(c), action-to-action generation outperforms noise-to-action, increasing average SR from 64.80% to 69.73%. Future visual augmentation further brings the average SR to 71.34%. Further comparisons of A2A and N2A generation in terms of performance and inference speed are provided in Appendix C.

5 Conclusion

We presented UniWAM, a unified framework that learns physical semantic reasoning, world modeling, and action prediction together. Through physical language supervision and component-specific co-training on VQA, human egocentric, and robot data, UniWAM acquires embodied knowledge while preserving pretrained semantic capabilities. This work suggests that pretraining a unified world-action architecture on heterogeneous data offers a promising path toward generalist robot models.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript, organize its structure, and assist with code writing. They were not used to generate experimental data, execute the reported evaluations, or generate references. We reviewed all AI-assisted work and take responsibility for the final content of this paper, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies simulated and planned real-world robotic manipulation. The current experiments do not involve human participants or personal data. A deployed system could nevertheless cause physical damage when instructions are misinterpreted or control fails under distribution shift. Any real-world evaluation should therefore use workspace constraints, emergency-stop procedures, and human supervision appropriate to the platform.

REPRODUCIBILITY STATEMENT

The manuscript records the model formulation and the configuration values that can be verified from the current implementation. Appendix H collects these values, and Appendix G reports all available RoboTwin task-level results. The code and instructions for use are also included in the supplementary-materials archive.

REFERENCES

AgiBot World Team. AgiBot World 2026. `https://huggingface.co/datasets/agibot-world/AgiBotWorld2026`, 2026.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Shuanghao Bai, Wenxuan Song, Jiayi Chen, Yuheng Ji, Zhide Zhong, Jin Yang, Han Zhao, Wanqi Zhou, Zhe Li, Pengxiang Ding, et al. Embodied robot manipulation in the era of foundation models: Planning and learning perspectives. *IEEE Transactions on Robotics*, 2026.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 35101–35113, June 2026a.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18135–18143, 2026b.

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

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$:

a Vision-Language-Action Model with Open-World Generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40. PMLR, 2025a. URL https://proceedings.mlr.press/v305/black25a.html.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025b. doi: 10.15607/RSS.2025.XXI.010.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alexander Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael S Ryoo, Grecia Salazar, Pannag R Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan H Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. RT-1: Robotics Transformer for Real-World Control at Scale. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023. doi: 10.15607/RSS.2023.XIX.025.

Ellis Brown, Arijit Ray, Ranjay Krishna, Ross Girshick, Rob Fergus, and Saining Xie. Sims-v: Simulated instruction-tuning for spatial video understanding. *arXiv preprint arXiv:2511.04668*, 2025.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Learning to Act Anywhere with Task-centric Latent Actions. In *Proceedings of Robotics: Science and Systems*, June 2025. doi: 10.15607/RSS.2025.XXI.014. URL https://www.roboticsproceedings.org/rss21/p014.html.

Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, Feng Qiu, Heng Qu, Yifei Su, Qiao Sun, Dong Wang, Donghao Wang, Yunhong Wang, Rujie Wu, Diyun Xiang, Yu Yang, Hangjun Ye, Yuan Zhang, and Quanyun Zhou. Xiaomi-Robotics-0: An Open-Sourced Vision-Language-Action Model with Real-Time Execution. *arXiv preprint arXiv:2602.12684*, 2026a. URL https://arxiv.org/abs/2602.12684.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Tongxi Zhou, et al. Scaling spatial intelligence with multimodal foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7879–7890, 2026b.

Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025a.

Ronghan Chen, Yandan Yang, Zuojin Tang, Dongjie Huo, Tong Lin, Haoning Wu, Haoyun Liu, Yuzhi Chen, Lulu Zheng, Botai Yuan, Tianlun Li, Mingxin Wang, Dekang Qi, Bin Hu, Wei Mei, Yuze Xuan, Haolong Yang, Yanqing Zhu, Mu Xu, Zhiheng Ma, and Xinyuan Chang. ABot-M0.5: Unified Mobility-and-Manipulation World Action Model. *arXiv preprint arXiv:2607.00678*, 2026. URL https://arxiv.org/abs/2607.00678.

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

XPolicyLab Community, Tianxing Chen, Yue Chen, Tian Nian, Zijian Cai, Guangyu Chen, Wenwei Lin, Qiwei Liang, Peicheng Xiang, Kailun Su, et al. Xpolicylab: A unified standard and open ecosystem for robot policy evaluation and deployment. *arXiv preprint arXiv:2608.09892*, 2026.

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

Yufan Deng and Daquan Zhou. Humannet: Scaling human-centric video learning to one million hours. *arXiv preprint arXiv:2605.06747*, 2026.

Dyna Robotics. Dyna-2: A 1-million-hour scaling law for world-action models. https://dyna.co/dyna-2, August 2026. URL https://dyna.co/dyna-2.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Yankai Fu, Ning Chen, Junkai Zhao, Shaozhe Shan, Guocai Yao, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Metis: Multi-source egocentric training for integrated dexterous vision-language-action model. *arXiv preprint arXiv:2511.17366*, 2025.

Generalist Team. GEN-0: Embodied foundation models that scale with physical interaction. Generalist AI Blog, November 2025. URL https://generalistai.com/blog/gen-0.

GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, et al. GigaWorld-Policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026. URL https://arxiv.org/abs/2603.17240.

Xicheng Gong, Qiwei Li, Peiran Xu, and Yadong Mu. Extending embodied question answering from perception to decision. *arXiv preprint arXiv:2605.25813*, 2026.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Jun Guo, Qiwei Li, Peiyan Li, Zilong Chen, Nan Sun, Yifei Su, Heyun Wang, Yuan Zhang, Xinghang Li, and Huaping Liu. Unified 4D World Action Modeling from Video Priors with Asynchronous Denoising. *arXiv preprint arXiv:2604.26694*, 2026. URL https://arxiv.org/abs/2604.26694.

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Kairos Team, Fei Wang, Shan You, Qiming Zhang, Tao Huang, Zuoyi Fu, Zhisheng Zheng, Yunlong Xi, Feng Lv, Xiaoming Wu, Zeyu Liu, Cong Wan, Pu Li, Ruiqing Yang, Xiaoou Li, Wei Wang, Kangkang Zhu, Yuwei Zhang, Shi Fu, Zheng Zhang, Xiaoning Wu, Xuzeng Fan, Dacheng Tao, and Xiaogang Wang. Kairos: A regret-aware native world-action model stack for physical ai. *arXiv preprint arXiv:2606.16533*, 2026. URL https://arxiv.org/abs/2606.16533.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of Human to Robot Transfer in Vision-Language-Action Models. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. URL https://arxiv.org/abs/2512.22414.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-Tuning Vision-Language-Action Models: Optimizing Speed and Success. In *Proceedings of Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.017. URL https://www.roboticsproceedings.org/rss21/p017.html.

Moo Jin Kim, Yihuai Gao, Tsung-Yi Lin, Yen-Chen Lin, Yunhao Ge, Grace Lam, Percy Liang, Shuran Song, Ming-Yu Liu, Chelsea Finn, and Jinwei Gu. Cosmos Policy: Fine-tuning video models for visuomotor control and planning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/748becc400a57c0e31cfe6a2e7951467-Abstract-Conference.html.

Jae Hee Lee, Matthias Kerzel, Kyra Ahrens, Cornelius Weber, and Stefan Wermter. What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 1039–1045. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/145. URL https://doi.org/10.24963/ijcai.2022/145. Main Track.

Fuhao Li, Wenxuan Song, Han Zhao, Jingbo Wang, Pengxiang Ding, Donglin Wang, Long Zeng, and Haoang Li. Spatial Forcing: Implicit Spatial Representation Alignment for Vision-language-action Model. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=euMVC1DO4k.

Hongyu Li, Bowen Wen, Xinghao Zhu, Yixuan Wang, Yilun Du, Yunzhu Li, George Konidaris, Stan Birchfield, Soha Pouya, Chenran Li, and Yan Chang. Hydra-0: Action flow for generalist world modeling and control. *arXiv preprint arXiv:2608.18077*, 2026b. URL https://arxiv.org/abs/2608.18077.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Luyao Zhang, Mingrui Yu, Zelin Gao, Nan Xue, Boyu Zhou, Xing Zhu, Mingyu Ding, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026c. doi: 10.15607/RSS.2026.XXII.016.

Qiwei Li, Xicheng Gong, Xinghang Li, Peiyan Li, Quanyun Zhou, Hangjun Ye, Jiahuan Zhou, and Yadong Mu. Rotvla: Rotational latent action for vision-language-action model. *arXiv preprint arXiv:2605.13403*, 2026d.

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.

Xinghang Li, Jun Guo, Qiwei Li, Long Qian, Hang Lai, Yueze Wang, Hongyu Yan, Jiahang Cao, Xi Chen, Jingen Qu, et al. Xiaomi-robotics-u0: Unified embodied synthesis with world foundation model. *arXiv preprint arXiv:2607.11643*, 2026e.

Yihang Li, Xuelong Wei, Jingzhou Luo, Yingjing Xiao, Yibo Bai, Guangyuan Zhou, Teng Zou, Chenguang Gui, Jiajun Wen, He Zhang, et al. Egolive: A large-scale egocentric dataset from real-world human tasks. *arXiv preprint arXiv:2604.23570*, 2026f.

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, et al. Open-aoe: An open egocentric manipulation dataset and toolchain for embodied learning. *arXiv preprint arXiv:2607.14183*, 2026g.

Weixin Liang, Lili Yu, Liang Luo, Srinivasan Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2411.04996.

Fanqi Lin, Kushal Arora, Jean Mercat, Haruki Nishimura, Paarth Shah, Chen Xu, Mengchao Zhang, Mark Zolotas, Maya Angeles, Owen Pfannenstiehl, et al. A systematic study of data modalities and strategies for co-training large behavior models for robot manipulation. *arXiv preprint arXiv:2602.01067*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, 2023.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022.

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.

Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable vla pretraining in the wild. *arXiv preprint arXiv:2602.21736*, 2026a.

Hao Luo, Ye Wang, Wanpeng Zhang, Sipeng Zheng, Ziheng Xi, Chaoyi Xu, Haiweng Xu, Haoqi Yuan, Chi Zhang, Yiqing Wang, Yicheng Feng, and Zongqing Lu. Being-H0.5: Scaling Human-Centric Robot Learning for Cross-Embodiment Generalization. In *Conference on Robot Learning*, 2026b. URL https://arxiv.org/abs/2601.12993.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-H0.7: A Latent World-Action Model from Egocentric Videos. *arXiv preprint arXiv:2605.00078*, 2026c. URL https://arxiv.org/abs/2605.00078.

MotuBrain Team, Chendong Xiang, Fan Bao, Haitian Liu, Hengkai Tan, Hongzhe Bi, et al. MotuBrain: An advanced world action model for robot control. *arXiv preprint arXiv:2604.27792*, 2026. URL https://arxiv.org/abs/2604.27792.

NVIDIA et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

NVIDIA GEAR Team. GR00T N1.6: An improved open foundation model for generalist humanoid robots. NVIDIA Research technical report, December 2025. URL https://research.nvidia.com/labs/gear/gr00t-n1_6/. Released December 15, 2025.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Joshua Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Ryan Co, Hye-Young Chung, Renee Zbizika, Jinyun Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Wenkai Xuan, Chenyu Yang, Xi Wang, James Fort, Richard Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Robert K. Katzschmann, Marc Pollefeys, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. EgoVerse: An Egocentric Human Dataset for Robot Learning from Around the World. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.092.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

Hao Shi, Bin Xie, Yingfei Liu, Lin Sun, Fengrong Liu, Tiancai Wang, Erjin Zhou, Haoqiang Fan, Xiangyu Zhang, and Gao Huang. MemoryVLA: Perceptual-cognitive memory in vision-language-action models for robotic manipulation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=54U3XHf7qq.

StarVLA Community. StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing. *arXiv preprint arXiv:2604.05014*, 2026. URL https://arxiv.org/abs/2604.05014.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. InternData-A1: Pioneering High-Fidelity Synthetic Data for Pre-training Generalist Policy. *arXiv preprint arXiv:2511.16651*, 2025. URL https://arxiv.org/abs/2511.16651.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Hanwen Wan, Dafeng Chi, Linbo Zhai, Tianao Shen, Yuzheng Zhuang, Tianle Zhang, Peidong Liu, Liang Lin, and Xiaoqiang Ji. Cometvla: Co-training on an embodied data pyramid towards physical understanding. *arXiv preprint arXiv:2608.30289*, 2026.

Yihao Wang, Pengxiang Ding, Lingxiao Li, Can Cui, Zirui Ge, Xinyang Tong, Wenxuan Song, Han Zhao, Wei Zhao, Pengxu Hou, Siteng Huang, Yifan Tang, Wenhui Wang, Ru Zhang, Jianyi Liu, and Donglin Wang. VLA-Adapter: An effective paradigm for tiny-scale vision-language-action model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18638–18646, 2026. doi: 10.1609/aaai.v40i22.38931. URL https://ojs.aaai.org/index.php/AAAI/article/view/38931.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, Zhaoye Long, Runtian Xu, Yue Wang, Chong Liu, Dihan Wang, Ziqiang Ni, Xiang Yang, You Liu, Ruoxuan Feng, Lei Zhang, Denghang Huang, Chenghao Jin, Anlan Yin, Xinlong Wang, Zhenguo Sun, Junkai Zhao, Mengfei Du, Mingyu Cao, Xiansheng

Chen, Hongyang Cheng, Xiaojie Zhang, Yankai Fu, Ning Chen, Cheng Chi, Sixiang Chen, Huaihai Lyu, Xiaoshuai Hao, Yequan Wang, Bo Lei, Dong Liu, Xi Yang, Yance Jiao, Tengfei Pan, Yunyan Zhang, Songjing Wang, Ziqian Zhang, Xu Liu, Ji Zhang, Caowei Meng, Zhizheng Zhang, Jiyang Gao, Song Wang, Xiaokun Leng, Zhiqiang Xie, Zhenzhen Zhou, Peng Huang, Wu Yang, Yandong Guo, Yichao Zhu, Suibing Zheng, Hao Cheng, Xinmin Ding, Yang Yue, Huanqian Wang, Chi Chen, Jingrui Pang, YuXi Qian, Haoran Geng, Lianli Gao, Haiyuan Li, Bin Fang, Gao Huang, Yaodong Yang, Hao Dong, He Wang, Hang Zhao, Yadong Mu, Di Hu, Hao Zhao, Tiejun Huang, Shanghang Zhang, Yonghua Lin, Zhongyuan Wang, and Guocai Yao. RoboCOIN: An Open-Sourced Bimanual Robotic Data Collection for Integrated Manipulation. *arXiv preprint arXiv:2511.17441*, 2025. URL https://arxiv.org/abs/2511.17441.

Ge Yan, Jinghao Liu, Yuzhi Fan, Lei Cai, Minwen Liao, Jesse Zhang, and Dieter Fox. Flex-π: A multi-stream world-action model with compute flexibility. *arXiv preprint arXiv:2608.10860*, 2026.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, Qiao Sun, Pengwei Wang, Lingqiao Liu, Yan Wang, Yuxiang Gao, Feras Dayoub, and Haoang Li. 4D-WAM: Infusing Spatiotemporal Awareness into World Action Models through Trajectory Fields. *arXiv preprint arXiv:2608.08023*, 2026a. URL https://arxiv.org/abs/2608.08023.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. In *International Conference on Learning Representations*, volume 2026, pp. 78185–78225, 2026b.

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

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. URL https://arxiv.org/abs/2602.15922.

Tomoya Yoshida, Shuhei Kurita, Taichi Nishimura, and Shinsuke Mori. Developing vision-language-action model from egocentric videos. *arXiv preprint arXiv:2509.21986*, 2025.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do World Action Models Need Test-time Future Imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

Lihan Zha, Asher James Hancock, Mingtong Zhang, Tenny Yin, Yixuan Huang, Dhruv Shah, Allen Z. Ren, and Anirudha Majumdar. LAP: Language-Action Pre-training Enables Zero-Shot Cross-Embodiment Transfer. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.203.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, Tai Wang, Ya-Qin Zhang, Jingjing Liu, and Xianyuan Zhan. X-VLA: Soft-Prompted Transformer as Scalable Cross-Embodiment Vision-Language-Action Model. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=kt51kZH4aG.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, et al. Flare: Robot learning with implicit world modeling. *arXiv preprint arXiv:2505.15659*, 2025.

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

Linqing Zhong, Yi Liu, Yifei Wei, Ziyu Xiong, Si Liu, and Guanghui Ren. Acot-vla: Action chain-of-thought for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8152–8162, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhong_ACoT-VLA_Action_Chain-of-Thought_for_Vision-Language-Action_Models_CVPR_2026_paper.html.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, et al. Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. *Advances in Neural Information Processing Systems*, 38:28404–28481, 2026.

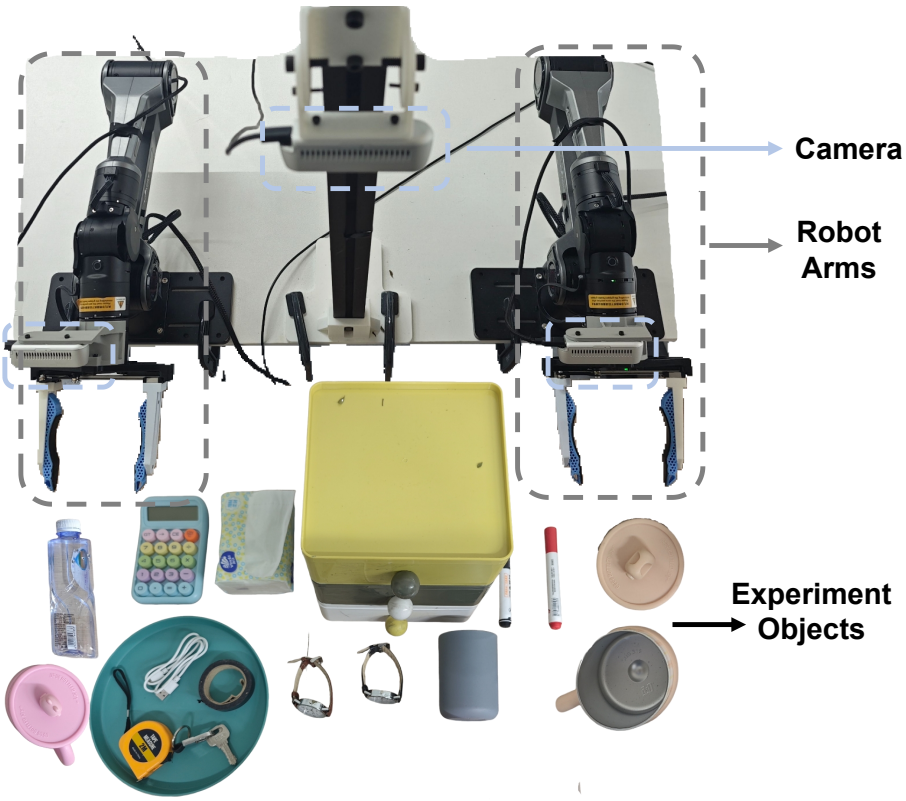


Figure 6: **Real-world experimental setup.** Our setup consists of two robot arms, three cameras for observation, and a diverse set of everyday objects used in the manipulation experiments.

A ABSTRACT

In this supplementary material, we provide the following information:

- Section B describes the real-world tasks and experimental setup.
- Section C compares action-generation strategies across inference-step counts.
- Section D describes the physical action template shared across robot and human data.
- Sections E and E.1 describe the human egocentric datasets and our automatic annotation pipeline.
- Section F details the RoboTwin 2.0 evaluation protocol.
- Section G reports task-level RoboTwin 2.0 results.
- Section H summarizes implementation details.
- Section I describes robot-data processing and quality control.

B REAL-WORLD TASKS

Objects and Robot Setup. Our real-world experiments involve a diverse set of household objects, as illustrated in Fig. 6, including a calculator, tissue box, drawer, markers of different colors, plate, tape measure, keys, data cable, tape, orange cup, pink cup, cup lid, brown watch, black watch, plastic cup, and water bottle. All experiments are conducted using the dual-arm setup described in the main text. For object-centric manipulation, the arm spatially closer to the target object is responsible for grasping it, while both arms can participate when bimanual coordination is required. Each task is evaluated over 40 trials.

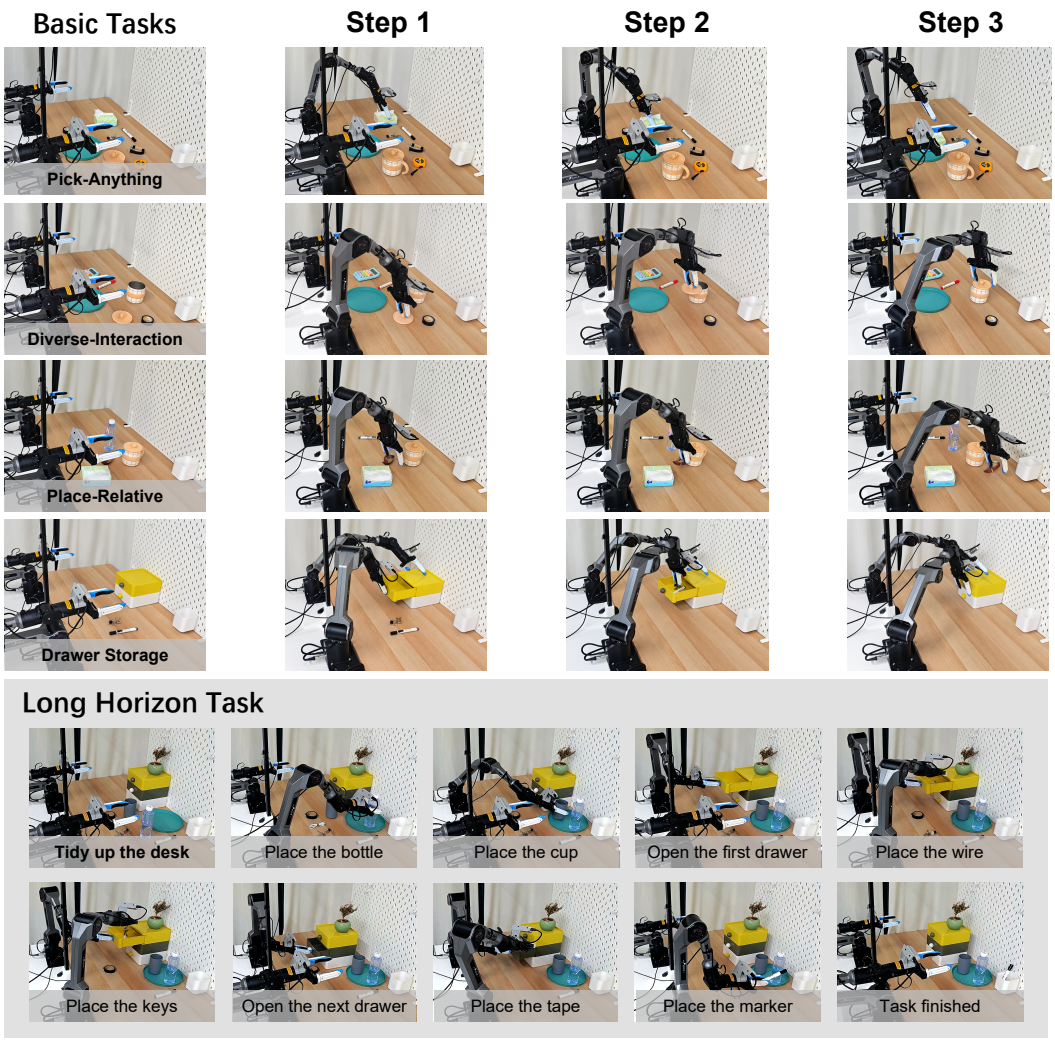


Figure 7: **Overview of our real-world evaluation tasks.** The top panel shows representative execution sequences for four instruction-following tasks: Pick-Anything, Diverse-Interaction, Place-Relative, and Drawer Storage. The bottom panel presents a long-horizon task, where the robot sequentially completes multiple manipulation stages.

Instruction-following tasks. As illustrated in Fig. 7, the four tasks for evaluating instruction-following capabilities are shown as follows: **(1) Pick-Anything** requires the robot to identify and grasp a target object from a multi-object scene, primarily evaluating noun grounding and object discrimination. **(2) Diverse-Interaction** requires the robot to perform different interactions on multiple objects according to the instruction, jointly testing noun and verb grounding. **(3) Place-Relative** requires the robot to pick a specified object and place it at a designated spatial relation to another object, evaluating noun grounding and spatial-relation understanding, particularly the interpretation of prepositions. **(4) Drawer-Storage** requires the dual-arm system to manipulate a drawer and place the target object inside, testing bimanual coordination, spatial reasoning.

Pick-Anything. Each scene contains eight objects randomly selected from the object set and scattered across the tabletop, together with a plate placed at a fixed location. The instruction follows the form "pick up [target object] and place it on the plate." This task primarily evaluates object-level language grounding and the ability to distinguish the instructed target from multiple distractors.

Diverse-Interaction. Each scene contains a cup, a cup lid, a calculator, two additional graspable objects, and a plate. The instructions require the robot to perform different interactions according to the specified object and action, including placing a target object on the plate, placing the lid onto

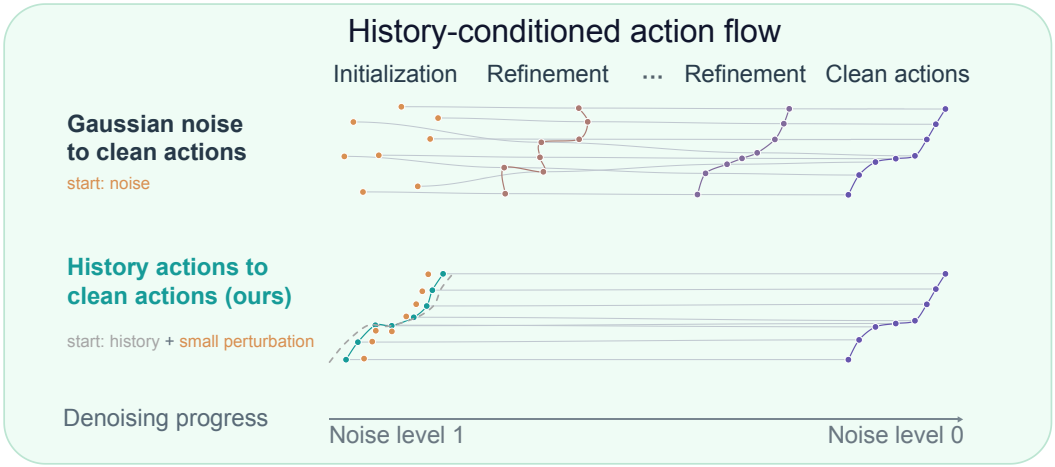


Figure 8: **The illustration for our action denoising process.**

Table 3: Action generation performance and inference time with different numbers of inference steps. N2A denotes noise-to-action and A2A denotes action-to-action generation. C2C and C2R denote clean2clean and clean2rand success rates (%). Average is the arithmetic mean of C2C and C2R. Inference time is reported in milliseconds. Values marked † are illustrative estimates, not measured results.

Inference steps	N2A				A2A			
	C2C ↑	C2R ↑	Average ↑	Time (ms) ↓	C2C ↑	C2R ↑	Average ↑	Time (ms) ↓
1	45.00†	39.00†	42.00†	157	68.00†	61.50†	64.75†	157
2	51.00†	45.00†	48.00†	315	72.30	66.48	69.39	315
4	56.32	50.30	53.31	629	75.14	68.32	71.73	629
10	70.17	63.29	66.73	1571	75.18	68.40	71.79	1571

the cup, or activating the calculator. The same object can be associated with different manipulation behaviors across instructions. The task therefore requires the policy to jointly ground both object semantics and action semantics, rather than relying solely on object recognition.

Place-Relative. Each scene contains five objects placed at varying locations on the tabletop. The robot is instructed to pick up one specified object and place it either to the left or to the right of another specified object, e.g., "place the pink cup to the left of the calculator." Successful execution requires simultaneously identifying both the source and reference objects and grounding the specified spatial relation. This task evaluates compositional object grounding and spatial-relation understanding.

Drawer Storage. The scene contains a two-level drawer and up to three candidate objects. The instruction specifies both a target object and a target drawer level, requiring the robot to place the designated object into either the upper or lower drawer. Completing the task involves object identification, drawer-level grounding, sequential manipulation, and bimanual coordination, as one arm interacts with the drawer while the other handles the target object.

Long-horizon Tasks. The long-horizon task is composed of 6 subtasks: (1) placing the first cup onto a plate, (2) placing the second cup onto the plate, (3) opening the drawer, (4) placing the designated daily-use object or tool inside its corresponding drawer, (5) closing the drawer, and (6) placing a marker into its designated holder.

C Action Generation with Different Inference Steps

As shown in Table 3, A2A achieves strong performance with only 2–4 inference steps, even outperforming N2A with ten steps. This demonstrates that A2A can deliver both high action-generation quality and fast inference. We therefore use four-step A2A as the default inference setting.

D Physical Action Templates

We use a shared physical-language template to describe end-effector action changes in robot and human data. At each timestep, the template summarizes the net changes over a sliding action window; absolute pose inputs are first converted to per-step changes. Translation magnitudes are expressed in centimeters and rounded to the nearest integer, while rotation magnitudes are expressed in degrees and rounded to the nearest 10 degrees. Positive and negative changes map to "move forward" and "move back" along x, "move up" and "move down" along z, and "move left" and "move right" along y. Rotations are described as tilting left/right (roll), tilting back/forward (pitch), or rotating counterclockwise/clockwise (yaw). Components with no change are omitted. When gripper values are available, the final value in the window determines whether the text says "open gripper" (value at least 0.5) or "close gripper" (value below 0.5). For bimanual actions, the summaries are prefixed with "Left arm:" and "Right arm:". This shared template provides a consistent language interface for action supervision across the two data sources.

The following are illustrative template outputs, not examples taken from recorded demonstrations:

> *Single arm:* "move forward 3 cm, move left 2 cm, tilt left 10 degrees, rotate counterclockwise 20 degrees, open gripper."
> *Bimanual:* "Left arm: move up 4 cm, close gripper. Right arm: move back 2 cm, open gripper."

E Details of Human Egocentric Data

EgoDex (Hoque et al., 2026) captures dexterous tabletop manipulation through first-person human videos paired with dense 3D hand and finger tracking. EgoVerse (Punamiya et al., 2026) comprises human demonstrations collected across diverse real-world settings and participants, accompanied by language and human-motion annotations. VITRA (Li et al., 2025) transforms in-the-wild human activity videos into atomic manipulation segments with language descriptions and reconstructed hand and camera trajectories; we use its processed subsets derived from Ego4D (Grauman et al., 2022), Ego-Exo4D (Grauman et al., 2024), and EPIC-KITCHENS (Damen et al., 2018).

E.1 Automatic Annotation of Egocentric Manipulation Videos

EgoVerse and VITRA already provide fine-grained subtask annotations. Since EgoDex does not provide fine-grained subtask annotations, we develop **EgoANT**, a general annotation pipeline that converts long, untrimmed egocentric manipulation videos into temporally aligned subtask supervision. EgoANT requires only raw egocentric video as external input, making it applicable to other in-the-wild egocentric video datasets without requiring dataset-specific action annotations. Given an episode V, the pipeline produces a sequence $\{(t_i^s, t_i^e, y_i)\}_{i=1}^N$, where $[t_i^s, t_i^e]$ denotes the temporal extent of an atomic manipulation event and y_i is a concise language description of the completed action. EgoANT decomposes the annotation process into two stages: *coarse-to-fine temporal segmentation*, which determines when individual manipulation events occur, and *semantic annotation*, which determines what action is performed within each resulting segment.

Coarse-to-fine temporal segmentation. We first temporally subsample the full episode and organize consecutive sampled frames into timestamped contact sheets. This representation compresses a long video into a compact set of visual observations while preserving long-range temporal context and explicit timestamps. A VLM processes the contact sheets from the full episode and predicts a coarse decomposition into completed manipulation events. The segmentation focuses on interactions that produce meaningful state changes, such as picking or placing an object, opening or closing a container, pouring, or using a tool, while avoiding transient approach, adjustment, and withdrawal

Table 4: **Final EgoANT configuration used for human-data annotation.** Each row summarizes the visual or contextual input and VLM used in one stage of the annotation pipeline.

Stage	Input / Configuration	Model
Global segmentation	Timestamped contact sheets; 0.5 s sampling and 20 frames per sheet	Qwen3.6-27B
Local refinement	Local timestamped visual context, coarse proposal, and episode-level context	Qwen3.6-27B
Segment annotation	Fixed temporal segment with complementary visual and contextual annotation passes	Qwen3.5-397B
Candidate selection	Segment context and candidate subtask descriptions	Qwen3.5-397B

motions as independent subtasks. The full-episode prediction serves as a coarse proposal rather than the final segmentation. For each coarse proposal, we construct a local temporal view around the corresponding portion of the video and again represent it using timestamped visual observations. The VLM re-examines this local context together with the coarse hypothesis and episode-level context, and re-segments the interval into completed atomic manipulation events. This local refinement corrects inaccurate boundaries, removes spurious motion fragments, and can decompose an under-segmented coarse proposal into multiple subtasks. Applying this refinement across the episode produces the final temporal segmentation.

Segment-level semantic annotation. Once the temporal boundaries are fixed, semantic labeling is performed independently for each segment. We sample visual observations within the corresponding interval and use a VLM to describe the completed manipulation with a concise action phrase, including the manipulated object and its destination or resulting state when observable. Rather than relying on a single annotation pass, EgoANT generates multiple candidate descriptions using complementary visual sampling and contextual conditioning strategies, including independent raw-frame annotations and relabeling conditioned on intermediate descriptions produced by earlier stages. A final VLM-based selection stage compares these candidate descriptions with the segment context and selects the annotation that best matches the visible completed manipulation, favoring concrete object–action–destination/state descriptions while avoiding unsupported objects or actions. The selected language annotation is paired with the refined temporal interval to form the final timestamped subtask annotation. We apply EgoANT to EgoDex and align the resulting subtask intervals with the corresponding egocentric observations and human wrist trajectories for pretraining.

The final configuration used in our data processing is summarized in Table 4.

F Details of Evaluation on RoboTwin 2.0

RoboTwin 2.0 is a benchmark consisting of 50 bimanual manipulation tasks (Chen et al., 2025b). We train the model on all 50 tasks with 50 clean demonstrations per task and evaluate it on clean (C2C) and randomized (C2R) settings individually for 100 episodes, where the latter evaluates the generalization abilities.

G RoboTwin 2.0 Task-Level Results

Table 5 reproduces every task-level value supplied in the experiment record. The table is the authoritative location for these individual percentages; the main text reports only the values needed to expose aggregate performance and important failure boundaries.

H Implementation Details

Model architecture. UniWAM combines Qwen3-VL-2B-Instruct as the physical reasoner and Wan2.2-TI2V-5B as the world generator with an action predictor in a Mixture-of-Transformers architecture. The three experts comprise 30 layers each. The world generator has a hidden dimension of 3,072 and 24 attention heads; the physical reasoner and action predictor have hidden dimensions

Table 5: Task-level UniWAM success rates (%) on RoboTwin 2.0. C2R denotes clean2rand and C2C denotes clean2clean. C2C and C2R values are reproduced from the supplied experiment record. Average is their arithmetic mean.

Task	C2R	C2C	Average	Task	C2R	C2C	Average
adjust_bottle	72	72	72.00	place_can_basket	35	36	35.50
beat_block_hammer	56	65	60.50	place_cans_plasticbox	86	97	91.50
blocks_ranking_rgb	67	64	65.50	place_container_plate	93	97	95.00
blocks_ranking_size	38	36	37.00	place_dual_shoes	77	87	82.00
click_alarmclock	92	100	96.00	place_empty_cup	94	98	96.00
click_bell	99	100	99.50	place_fan	69	76	72.50
dump_bin_bigbin	94	96	95.00	place_mouse_pad	55	59	57.00
grab_roller	100	100	100.00	place_object_basket	50	54	52.00
handover_block	18	24	21.00	place_object_scale	64	69	66.50
handover_mic	81	89	85.00	place_object_stand	78	81	79.50
hanging_mug	16	16	16.00	place_phone_stand	64	75	69.50
lift_pot	92	100	96.00	place_shoe	79	79	79.00
move_can_pot	46	46	46.00	press_stapler	96	98	97.00
move_pillbottle_pad	71	78	74.50	put_bottles_dustbin	38	44	41.00
move_playingcard_away	93	90	91.50	put_object_cabinet	28	40	34.00
move_stapler_pad	53	58	55.50	rotate_qrcode	41	46	43.50
open_laptop	75	90	82.50	scan_object	33	42	37.50
open_microwave	44	76	60.00	shake_bottle	71	89	80.00
pick_diverse_bottles	78	76	77.00	shake_bottle_horizontally	70	92	81.00
pick_dual_bottles	85	92	88.50	stack_blocks_three	52	74	63.00
place_a2b_left	64	65	64.50	stack_blocks_two	91	91	91.00
place_a2b_right	80	85	82.50	stack_bowls_three	82	88	85.00
place_bread_basket	81	84	82.50	stack_bowls_two	91	94	92.50
place_bread_skillet	68	81	74.50	stamp_seal	70	87	78.50
place_burger_fries	76	95	85.50	turn_switch	70	86	78.00
Mean	68.32	75.14	71.73				

of 512 and 1,024, respectively. The model uses 14-dimensional robot states and actions, predicts eight future frames at 384×320 resolution, and generates action chunks of length 16.

Pretraining. Pretraining runs for seven days on 32 NVIDIA H800 GPUs. The joint objective weights the video, action, and language losses in the ratio $1 : 1 : 0.1$:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{action}} + 0.1\,\mathcal{L}_{\text{language}}. \tag{8}$$

Post-training. For RoboTwin post-training, we use a batch size of 5 for 200,000 optimization steps. We optimize with AdamW, a learning rate of 5×10^{-5}, and weight decay of 0.01. The learning-rate schedule is linear with 200 warmup steps; gradient norm is clipped at 0.5. The Wan and VLM backbones use bfloat16 precision. Future-frame noise augmentation is applied with probability 0.5, with the visual latent retention scale sampled uniformly from $[0.5, 1.0]$.

Inference. The action flow is initialized from a recent action-history chunk of length 16 with Gaussian perturbation of standard deviation 0.02, while the future video latents are initialized from Gaussian noise. We use 2–4 flow integration steps at inference.

I Robot Data Processing

Robot datasets collected from different platforms and acquisition pipelines often exhibit heterogeneous failure modes in both trajectory signals and visual observations. We therefore apply a unified quality-control procedure before training. Our pipeline evaluates each trajectory from three complementary perspectives: temporal and statistical reliability, geometric consistency, and visual validity.

I.1 Temporal and Statistical Trajectory Screening

Abrupt-transition screening. We first identify abrupt transients in state and action signals. For each scalar trajectory, a smooth reference signal is obtained using cascaded median filtering followed by Savitzky–Golay smoothing. Instead of relying on a single pointwise deviation criterion,

we jointly consider the discrepancy from the smooth reference and higher-order temporal variations. We flag a time step as anomalous when the residual exceeds its threshold and either the acceleration or jerk exceeds its respective threshold. This criterion suppresses isolated non-physical spikes while being less sensitive to gradual changes in the underlying motion. Because the numerical characteristics differ across robot embodiments and data sources, the corresponding thresholds are configured on a per-dataset basis.

State–action temporal consistency. We next verify whether the action stream is temporally compatible with the observed robot-state evolution. For dimensions shared by the state and action representations, we first smooth both trajectories and estimate their relative temporal offset using cross-correlation. After compensating for the estimated lag, we compute directional agreement from their first-order differences. Episodes with directional agreement below a dataset-specific threshold are considered inconsistent and excluded. For datasets whose actions are represented as increments, the action sequence is first integrated to recover absolute values before this comparison.

Distribution-based outlier removal. We further suppress rare numerical outliers using robust statistics computed independently for each relevant signal dimension. Let q_{01} and q_{99} denote the 1st and 99th percentiles, respectively, and retain values within $[q_{01} - \alpha(q_{99} - q_{01}),\ q_{99} + \alpha(q_{99} - q_{01})]$, where α controls the tolerance beyond the central percentile range. This step prevents a small number of extreme measurements from distorting subsequent normalization. Gripper dimensions are excluded from this filtering due to their bimodal distributions.

I.2 Geometry-Aware Consistency Checks

Joint–EEF consistency. For datasets with reliable kinematic metadata, we examine the consistency between recorded joint states and logged end-effector poses through forward kinematics. The comparison is used to identify systematic discrepancies arising from joint-angle conventions, end-effector or TCP definitions, rotation representations, and base-frame assumptions. This check is applied only when the joint measurements, robot model, and end-effector semantics are sufficiently specified. It is otherwise skipped for the corresponding dataset.

Coordinate-frame and orientation alignment. Different datasets may adopt different base-frame and end-effector orientation conventions. For datasets with well-defined coordinate semantics, we apply dataset-specific rotation corrections to align them to a common world-frame convention, with the positive x-axis corresponding to the robot's forward direction. Such corrections are applied only when the underlying frame definitions are explicitly available. Otherwise, the original orientation representation is retained.

I.3 Visual Observation Quality

Video-quality screening. We filter visually invalid observations, including black frames, corrupted frames, blurred frames, and prolonged static segments. Static periods are identified jointly from visual, state, and action signals to distinguish redundant observations from task-relevant transitions. Key interaction frames, such as gripper-closure events, are preserved even when the corresponding visual changes are small.