SWAM: Scaling World Action Models with Cross-Modal Alignment

Yu Zhou[1,2,3,*,‡], Xinyuan Xia[1,2,3,*], Xuankun Yang[1,3,*], Yinghao Yang[4,2,3,*], Wendi Chen[1,2,3]
Bowen Liu[1,2,3,‡], Jianhan Ma[3], Weicheng Duan[5,3], Jun Lv[3], Cewu Lu[1,2,3], Yanwei Li[1,†]

[1]Shanghai Jiao Tong University [2]Shanghai Innovation Institute [3]Noematrix Ltd.
[4]Westlake University [5]Tongji University [*]Equal contribution. [†]Corresponding author.

Dataset Code Webpage

Abstract World action models jointly predict video and actions conditioned on language. Existing studies focus primarily on architectures and training mechanisms, typically treating language–video alignment acquired through video pretraining as an established foundation. However, introducing action learning may impair this alignment. Principled guidance on data and supervision design for maintaining language–video alignment during action learning remains limited. To investigate this, we curate a pool of approximately 47,000 hours of general videos, egocentric videos, UMI demonstrations, and robot demonstrations. Building on this, we systematically study how language–video supervision, language–video–action supervision, and the joint use of these two forms affect WAM performance. In particular, we examine how language annotation quality affects pretrained language–video capabilities, and how action supervision from UMI and robot demonstrations extends these capabilities to language–video–action modeling. To better scale the model, we further conduct detailed studies on joint training recipe and model architecture. Under a fixed training budget, mixing egocentric videos and UMI demonstrations improves control generalization. Model architecture affects how effectively these experiences are used, while scaling pretraining under the selected configuration brings further gains. The resulting recipe, post-trained on RoboTwin Clean demonstrations only, achieves success rates of 88.10% and 61.56% under Easy (Clean-to-Clean) and Hard (Clean-to-Random) conditions, respectively. Our model achieves 40.00% SR across three real-world tasks, compared with 23.33% for the strongest baseline, indicating better generalization under unseen conditions. We release the curated data, training code, and model weights.

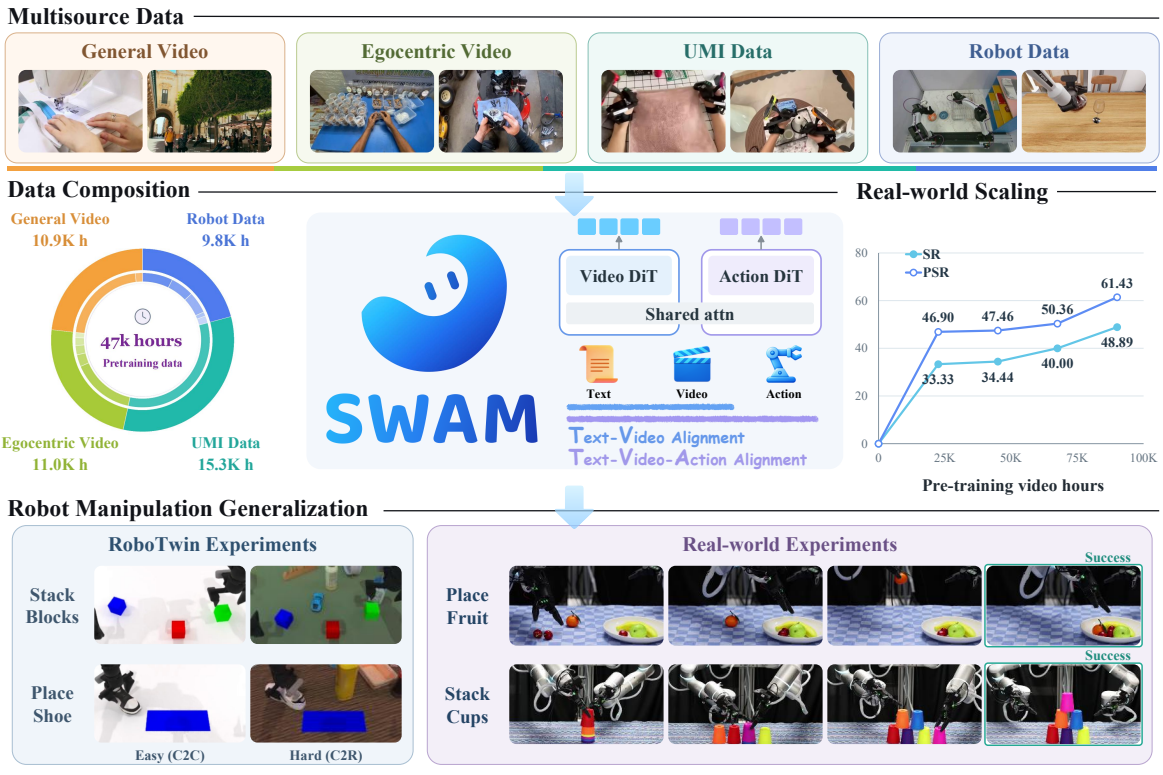


Figure 1. Overview of SWAM. We study how multisource supervision preserves the language–video capabilities acquired during video pretraining and extends them to robot control. We curate a multisource data pool of approximately 47,000 hours to support this study. Experiments show that additional pretraining further improves real-world performance, and that the resulting recipe generalizes to unseen scene conditions in RoboTwin and on real robots.

[‡]Work completed during internships at Shanghai Jiao Tong University.

Contents

1 Introduction **3**

2 Related Work **4**

3 Method **5**

3.1 World Action Modeling 5
3.1.1 Problem Formulation 5
3.1.2 Model Architecture 6
3.1.3 Training Objectives 7
3.2 Language–Video Alignment 8
3.2.1 Video Sources and Curation 8
3.2.2 Language Annotation 9
3.3 Language–Video–Action Alignment 10

4 Experiments **11**

4.1 Experimental Setup 11
4.2 How to establish effective cross-modal alignment? 11
4.2.1 Language–Video Alignment 12
Rich and aligned language supervision improves generalization 12
4.2.2 Language–Video–Action Alignment 13
Human manipulation experience provides action supervision 13
4.3 How to integrate multisource experience? 13
Transitional pretraining on General videos provides a stronger starting point 13
Pretraining exposure to individual data sources improves WAM performance . . . 14
Equal Ego–UMI sampling gives the best generalization 14
Architecture helps preserve pretrained video priors 14
4.4 Does scaling pretraining improve control generalization? 15
Scaling pretraining improves WAM performance 15
4.5 How well does our model perform in unseen environments? 16
Simulation results 16
Real-world results 16
Real-world visualizations 16

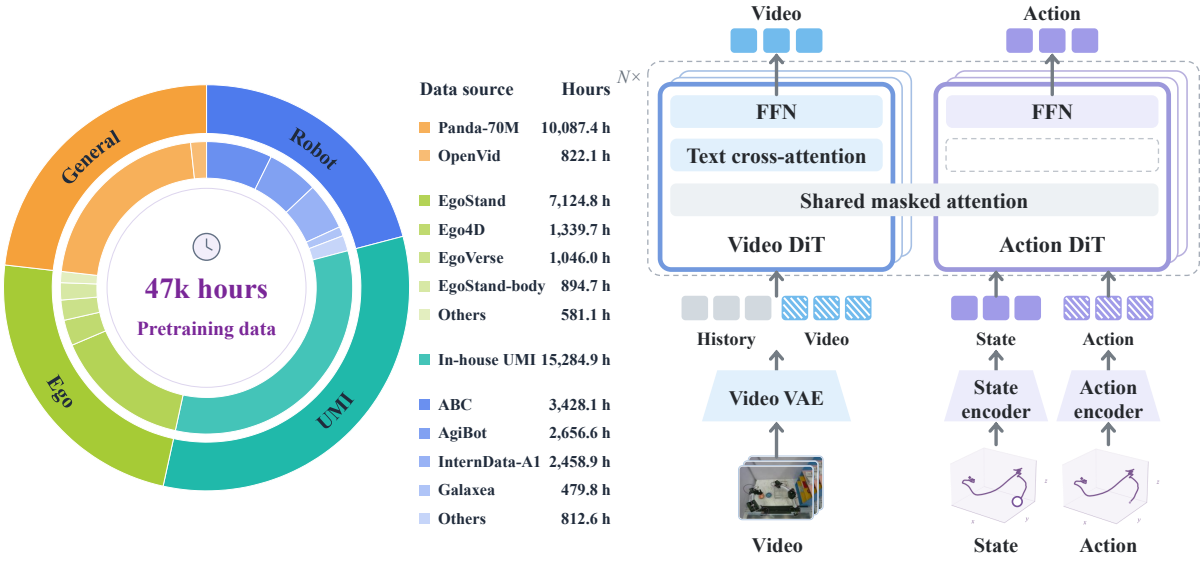
5 Discussion and Conclusion **17**

6 Appendix **23**

1. Introduction

Pretrained video models learn associations between language descriptions, visual content, and visual dynamics, acquiring visual and dynamics priors across scenes, objects, and interactions [1, 2, 3, 4]. These experiences provide a foundation for robots to understand manipulation goals and predict environmental changes [5, 6]. World action models (WAMs) jointly predict video and actions conditioned on language, connecting visual changes to robot actions and providing a route from video pretraining to robot control [7, 8]. This paper studies how to better use these priors to extend language-conditioned video modeling into generalizable action capabilities.

When we introduce action prediction, the language–video capabilities acquired in video pretraining should be preserved. This requires further alignment between language, video, and action. It bridges the gap between language-described targets, visual observation, and executed actions. Existing work on WAMs largely focused on model architectures and training mechanisms [9, 10]. Methods on video pretraining often treat the language–video alignment as an established foundation [7, 8]. However, introducing action learning raises a basic question: how to preserve established language–video prior and further solve language–video–action misalignment? The roles of different experiences and forms of supervision in this process, particularly for control generalization, remain unclear [9, 10, 11, 12].



(a) Multisource Pretraining Data (b) Model Architecture

Figure 2. **Multisource data and model overview.** (a) Composition of the 47k-hour pretraining data pool. (b) Our world action model architecture, with video and action branches connected through shared masked attention.

To study these questions, we first construct a multisource data pool with approximately 47,000 hours, comprising approximately 11,000 hours of general videos, 11,000 hours of egocentric videos, 15,000 hours of in-house UMI demonstrations, and 10,000 hours of robot demonstrations, as shown in Figure 2(a). These sources provide different forms of cross-modal supervision. General and egocentric videos primarily provide language–video supervision, covering general visual dynamics and human interactions, respectively. UMI demonstrations record video and action trajectories of humans manipulating objects with handheld grippers, while robot demonstrations provide observations, states, and actions during robot task execution. Both include language descriptions corresponding to the manipulation process, supporting language–video–action learning in addition to language–video supervision. UMI demonstrations provide scalable human manipulation experience, whereas robot demonstrations provide direct but more limited robot control experience [13, 14, 15]. The sources differ in scene and task coverage and in the completeness of language and action supervision. Multisource pretraining therefore requires combining these experiences appropriately: language–video supervision preserves video pretraining capabilities, while human and robot action supervision establishes video–action correspondences [16, 17, 18].

Building on the constructed multisource experience, we conduct systematic studies at three levels: language–video alignment, language–video–action alignment, and joint learning. Specifically, at the language–video level, we examine how to preserve the visual prior in pretrained video generation with the design of language supervision. At the language–video–action level, we investigate how to utilize action supervision from UMI and robot demonstrations to help action prediction and robot control. In joint training, we study the data recipe and model architecture for better cross-modality modeling. The data recipe determines the coverage of the visual and interaction experience in tasks, scenes, and the proportions of different experiences. The model architecture answers how to conduct cross-modal supervision through modality interactions and language conditioning pathways, as shown in Figure 2(b). We further examine the benefits of scaling training with the resulting data and supervision recipe.

Our study shows that preserving language–video capabilities and extending language–video–action capabilities jointly support control generalization in WAMs. Detailed manipulation descriptions and environmental cues improve language–video supervision, which maintains language–video alignment and thereby improves control generalization. And UMI demonstrations improve robot control in simulation and real scenes, proving the value of human manipulation experience with action supervision for robot learning. In joint learning, the benefits of multisource experience depend both on data and model supervision recipe. Experiments show that increasing training investment under the selected data and model recipe further improves control generalization. These findings provide empirical guidance for data collection and organization, supervision design, and joint training in WAMs. The pretraining recipe derived from these findings achieves success rates of 88.10% and 61.56% under the RoboTwin Easy (Clean-to-Clean) and Hard (Clean-to-Random) settings, respectively. Ours achieves 40.00% SR under the Hard setting across three real-world tasks, exceeding OpenWAM's 23.33%. This result indicates better generalization to unseen objects and backgrounds.

2. Related Work

World Action Models. World action models aim to connect visual observations with executable robot actions [7]. Early work treated video as a transferable visual plan or dynamic representation. UniPi first generates a language-conditioned visual plan and then recovers actions with an inverse-dynamics module [19]. VPP conditions a policy on predictive representations from a video model [6]. Motus, DreamZero, and LingBot-VA place visual dynamics and action prediction within unified or interactive architectures, exploring discrete tokens, latent actions, joint flow matching, and causal video–action modeling [20, 7, 8]. Fast-WAM and InternW0-Δ further distinguish future supervision during training from action generation at inference, showing that the prediction target and the deployment interface can be separated [21, 17]. VPP2 and LingBot-VA adapt video pretraining to robot manipulation by studying how manipulation data organization, temporal and video representation modeling, and action-branch adaptation can preserve video prediction capabilities while supporting action learning [18, 8]. We build on this foundation and focus on how language–video capabilities from video pretraining can transfer to language–video–action learning and action control.

Multisource Data for Robot Learning. Robot learning increasingly draws on cross-embodiment trajectories, general videos, egocentric videos, and human manipulation demonstrations rather than a single source of robot data. Open X-Embodiment, Octo, and OpenVLA aggregate and filter multi-robot data to study shared representations, observation adaptation, and action-space alignment, showing that cross-embodiment learning requires source normalization and interface alignment [22, 23, 24]. General and egocentric videos expand coverage of scenes, objects, and interaction dynamics, but usually lack action labels directly corresponding to robot execution [25, 26]. EgoMimic and EgoScale study how such human experience can be transferred through view and embodiment alignment, including settings with limited robot data [27, 28]. UMI complements these sources by using portable interfaces to record manipulation trajectories in an end-effector-like interaction format, thereby providing action supervision associated with the manipulation process [13, 14, 15]. EgoWAM, Ego4WAM, and Motus further demonstrate the complementary roles of video, action trajectories, semantic descriptions, and world-prediction targets [29, 10, 20]. The central issue in multisource learning is therefore not simply adding sources, but understanding differences in scene coverage, interaction dynamics, embodiment interfaces, and available supervision. We curate approximately 47,000 hours of multisource data. We use this pool to sys-

tematically examine how experience combinations and supervision design affect WAM pretraining.

3. Method

This section presents the world action model and its multimodal supervision. We first present the problem formulation, model architecture, and training objectives, then describe multisource experience curation and supervision construction. Pretraining uses four types of experience: general videos, egocentric videos, UMI demonstrations, and robot demonstrations. These sources differ in scene and task distributions, data quality, manipulation diversity, and available action supervision. General and egocentric videos provide language–video supervision, while UMI and robot demonstrations additionally provide state and action signals where available and valid. In Table 1 and Figure 3, we summarize the composition, scale, and video-duration distribution of the pretraining pool after data curation and cleaning.

Table 1. **Composition and scale of multisource pretraining data.**

Dataset	Episodes	Frames (M)	Hours
General Videos			**10,909.6**
Panda-70M	2,599,835	1,089.4	10,087.4
OpenVid	256,839	82.5	822.1
Egocentric Videos			**10,986.3**
EgoSuite	294,798	928.9	8,600.6
Ego4D	21,302	144.7	1,339.7
EgoVerse	55,112	113.0	1,046.0
UMI Data			**15,284.9**
In-house UMI	411,754	1,650.8	15,284.9
Robot Data			**9,836.0**
ABC	127,057	370.2	3,428.1
AgiBot	155,650	286.9	2,656.6
InternData-A1	354,192	265.6	2,458.9
Galaxea	19,636	51.8	479.8
RoboCOIN	50,439	45.3	419.1
RoboMind	61,382	42.5	393.5
Total	**4,407,996**	**5,071.5**	**47,016.7**

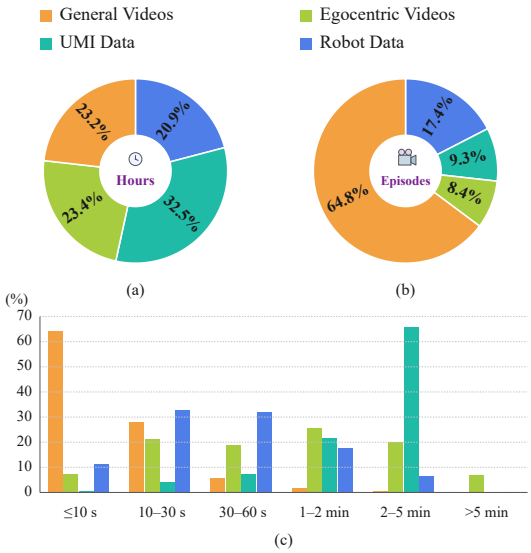


Figure 3. **Pretraining data composition and video duration across sources.** (a) Shares of total video hours. (b) Shares of total episodes. (c) Video-duration distribution within each source.

3.1. World Action Modeling

3.1.1. Problem Formulation

We consider language-conditioned robot manipulation. Given the visual observations acquired so far, a language instruction, and its current proprioceptive state, the robot predicts actions over a future interval. Let n be a prediction start point sampled from a trajectory, $K > 0$ a fixed prediction horizon, $V_{0:n}$ the visual history including the current observation, P the language condition, and S_n the current state. The control task requires learning an action policy π_θ conditioned on visual history, the language instruction, and the current state:

$$A_{n:n+K} \sim \pi_\theta(\cdot \mid V_{0:n}, P, S_n), \tag{1}$$

where $A_{n:n+K}$ denotes the action sequence over the target interval, and θ denotes the model parameters.

Control learning primarily supervised by observation–action pairs must jointly address scene understanding, interaction dynamics, and action generation. We adopt a world action model (WAM) based on a pretrained video model, using the associations between language descriptions and visual dynamics learned during video pretraining and learning correspondences between manipulation processes and executed actions through action supervision. The model jointly predicts video changes and actions over

the same interval. The video sequence $V_{n:n+K}$ describes observable scene evolution during manipulation, while the action sequence $A_{n:n+K}$ describes the corresponding manipulation behavior. The two sequences span the same physical time interval but may use different sampling rates.

We denote the joint predictive distribution of video and actions by p_θ:

$$\begin{aligned} & p_\theta(V_{n:n+K}, A_{n:n+K} \mid V_{0:n}, P, S_n) \\ & = p_\theta(V_{n:n+K} \mid V_{0:n}, P, S_n)\, p_\theta(A_{n:n+K} \mid V_{0:n+K}, P, S_n) \\ & = p_\theta(V_{n:n+K} \mid V_{0:n}, P)\, p_\theta(A_{n:n+K} \mid V_{0:n+K}, P, S_n). \end{aligned} \tag{2}$$

The final equality follows from our modeling choice that the video branch is not conditioned on the proprioceptive state. The first term learns language-conditioned video prediction, corresponding to language–video learning. The second term predicts actions from language, video information, and the current state. Together, the two terms constitute joint language–video–action learning.

Starting from a pretrained video model, we train a unified WAM end to end with a joint prediction objective. The action branch uses visual dynamics and language conditioning information supplied by the video branch. Training retains the video prediction objective while introducing action supervision, aiming to continue using existing language–video modeling capabilities and to learn correspondences between manipulation processes in video and executed actions.

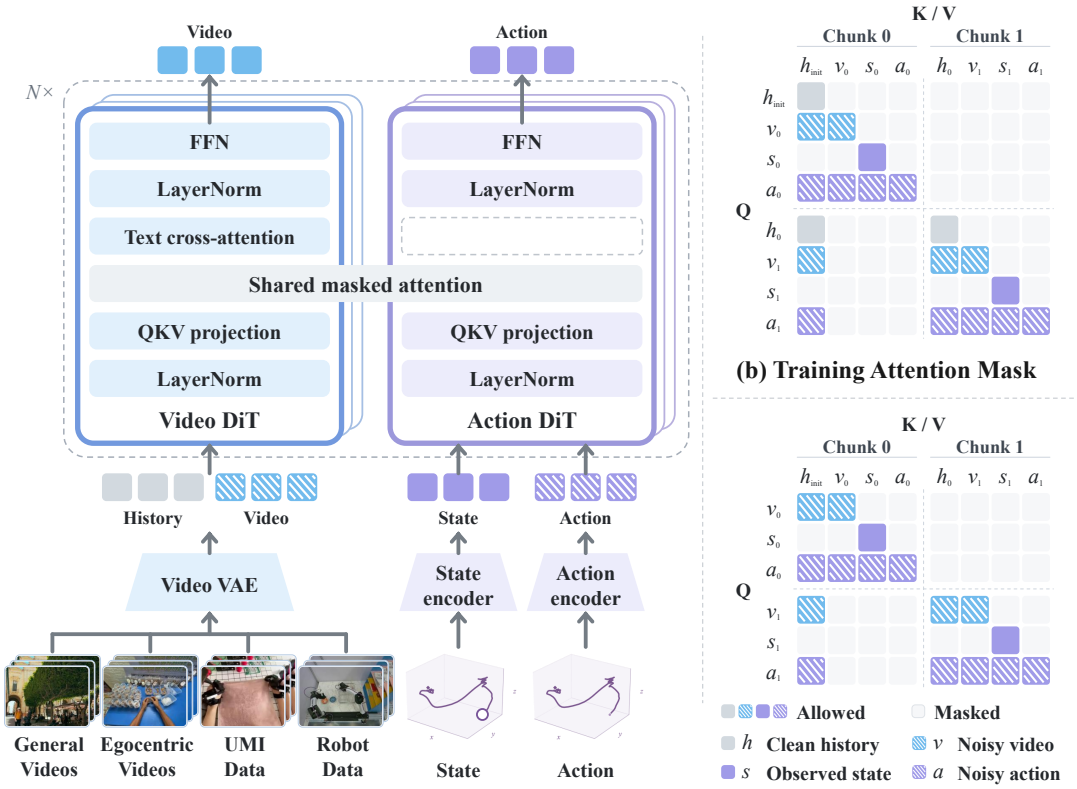
Multisource experience supplies different parts of the supervision for joint prediction. General and egocentric videos provide language–video correspondences for learning language-conditioned visual dynamics. UMI and robot demonstrations additionally provide actions temporally associated with video, enabling action prediction alongside video learning. All samples participate in video learning. Samples with valid action annotations additionally participate in action learning. Missing or invalid action targets contribute no action supervision. Experiences with different supervision coverage can therefore participate in unified WAM pretraining.

3.1.2. Model Architecture

Given visual observations $V = \{v_i\}_{i=1}^{T}$, a language instruction P, and available states S, we adopt a WAM that jointly predicts video and actions. Whereas $n : n + K$ above denotes a trajectory-level prediction interval, we use t below to index temporally aligned video–action chunks. Each chunk contains 2 video latent frames and 48 action steps. For chunk t, let $V_t \in \mathbb{R}^{T_v \times H \times W \times 3}$ denote the corresponding video, $S_t \in \mathbb{R}^{D_s}$ the available state at the chunk start, and $A_t \in \mathbb{R}^{H_a \times D_a}$ the normalized target action sequence. Here, T_v is the number of video frames, H and W are frame height and width, D_s is the state dimension, and H_a and D_a are the number of action steps and the action representation dimension, respectively. A VAE encodes video into latents x_t. Separate encoders map actions and states to tokens. A text encoder maps the language instruction to condition p.

We illustrate the model architecture and the attention masks used during training and inference in Figure 4. The model uses a MoT architecture [30], with the video branch initialized from Wan [2] and the action branch randomly initialized. The action DiT branch comprises action and state encoders, transformer blocks, and an action decoder. Video and action maintain separate residual streams, with joint attention passing information unidirectionally from video to action. The video token stream passes through the video branch and causally attends to current and past video tokens. The action token stream passes through the action branch and attends to video tokens from past and current chunks, as well as state and action tokens within the current chunk. This unidirectional interaction allows samples with video supervision alone to continue training the video branch without directly updating the action branch through the video objective, supporting the use of videos without action supervision.

The video branch receives encoded language conditions through cross-attention. The action branch uses language information only through attention interactions with the video branch and has no direct text cross-attention. This conditioning pathway aims to reuse the language–video correspondences established during video pretraining, allowing action prediction to use semantic information associated with the manipulation process. The video branch predicts flow velocities in video latent space, while the action branch decodes its output tokens into flow velocities in action space.



(a) Model Architecture

(b) Training Attention Mask

(c) Inference Attention Mask

Figure 4. **World action model architecture and attention masks.** (a) The video and action DiT branches are connected through shared masked attention, with text cross-attention in the video branch only. Attention masks for (b) training with clean video history under teacher forcing and (c) generating video and actions chunk by chunk at inference.

3.1.3. Training Objectives

For chunk t, x_t and A_t denote the clean video latents and action targets, respectively. Let $\tau \in [0,1]$ denote flow matching time. We sample independent standard Gaussian noise ϵ_v and ϵ_a with the same shapes as their respective targets and construct noisy targets:

$$\begin{aligned} x_t^\tau &= (1-\tau)x_t + \tau\epsilon_v, \\ A_t^\tau &= (1-\tau)A_t + \tau\epsilon_a. \end{aligned} \tag{3}$$

The action encoder maps noisy actions A_t^τ to action tokens. Video and actions within a chunk share τ, sampled independently across chunks. Training uses teacher forcing, with past clean video latents h_t as context. Let u_θ^v and u_θ^a denote the video and action velocity predictors, whose outputs have the same shapes as x_t and A_t, respectively. The action predictor u_θ^a denotes the complete action DiT branch, including action and state encoding, transformer processing, and action decoding. The two flow matching losses [31] are:

$$\begin{aligned} \mathcal{L}_{\text{video}} &= \mathbb{E}\left[\left\|u_\theta^v(x_t^\tau, h_t, p, \tau) - (\epsilon_v - x_t)\right\|_2^2\right], \\ \mathcal{L}_{\text{action}} &= \mathbb{E}\left[\left\|u_\theta^a(A_t^\tau, x_t^\tau, h_t, S_t, p, \tau) - (\epsilon_a - A_t)\right\|_2^2\right]. \end{aligned} \tag{4}$$

Here, θ denotes model parameters, and the expectations are taken over training samples, noise, and flow matching times. The action output depends on the current video x_t^τ and language condition p through the video branch. The two losses are combined with equal weights:

$$\mathcal{L} = \mathcal{L}_{\text{video}} + \mathcal{L}_{\text{action}}. \tag{5}$$

All samples contribute to the video objective. The action objective is computed only for samples with valid action annotations.

3.2. Language–Video Alignment

3.2.1. Video Sources and Curation

In Figure 5, we summarize the curation process used to turn the four experience sources into training-ready video–language examples. We perform quality checks, standardize the video records, segment long videos into subtasks where applicable, and generate multi-level captions before mixing the resulting data for pretraining.

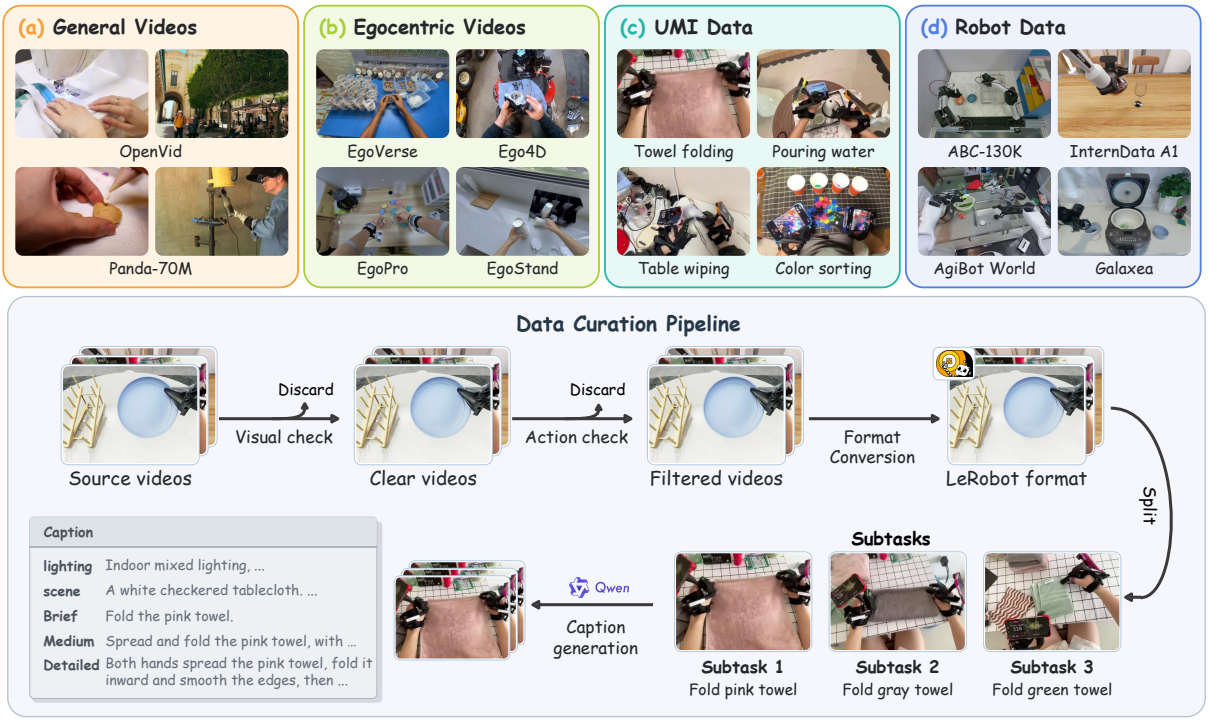


Figure 5. **Multisource training data and curation pipeline.** The upper panel shows representative examples from four data sources. The lower panel illustrates visual checks, action checks, format conversion, subtask segmentation, and caption generation. Captions include action descriptions at three levels of detail, along with scene and lighting information.

General videos. We use OpenVid [32] and Panda70M [33], totaling 10,909.6 hours. These open-source video generation datasets contain videos and text descriptions covering diverse scenes, objects, and visual dynamics. We retain their videos and original descriptions to provide language–video experience beyond manipulation demonstrations.

Egocentric videos. We use Ego4D [25], EgoVerse [26], and EgoSuite [34], totaling 371,212 trajectories and 10,986.3 hours. Ego4D primarily records daily activities across real-life settings. EgoVerse covers more diverse scenes, adding egocentric visual experience across environments. EgoSuite is currently the largest open-source egocentric manipulation video dataset. We retain the videos and language descriptions from these sources.

Data preparation. We package all sources in LeRobot 2.1 format. We retain only samples whose videos, language descriptions, and metadata can be linked. We exclude samples with unassociated video, language, or metadata, videos with encoding or decoding failures, and known low-quality samples.

Video records retain their corresponding language descriptions and metadata. The inventory counts the recorded episodes and videos at their source durations.

Visual content filtering. For general video datasets such as OpenVid and Panda70M, we use original annotations and VLM-based content review to remove animations, static scenes, television programs, and other unsuitable content. For egocentric videos, we use a VLM to assess motion blur, brightness, occlusion, and whether a clip contains meaningful actions, reducing the impact of low-quality imagery and uninformative behavior on training. The UMI and robot demonstrations described below use similar filtering rules.

Video model adaptation on general videos. Our WAM modifies Wan's video modeling design to support robot control, which may disrupt pretrained visual and dynamics priors. To ease this adaptation, we first train the video branch on general videos, whose distribution is close to that of Wan's pretraining data. During this stage, we use the same input frame count, spatial resolution, and causal attention scheme as subsequent WAM training. This stage aims to preserve these priors during adaptation, before introducing embodied data.

3.2.2. Language Annotation

Language annotations describe manipulated objects and executed actions and may additionally include interaction relationships, manipulation processes, and environmental information. For WAM pretraining, what semantic information helps translate video experience into control capabilities? Can these annotations help preserve language–video alignment and improve downstream generalization? We organize language supervision along three dimensions: the level of detail in manipulation descriptions, environmental information, and temporal granularity.

Level of detail. Task descriptions cover executed actions, interaction objects, and their attributes during manipulation. We define three levels of descriptive detail, Brief, Medium, and Detailed, with requirements and examples in Table 2. These levels define controlled language-annotation variants for evaluating how description detail affects the transfer of pretrained video capabilities to robot control.

Table 2. **Language annotation requirements and example descriptions at three levels of detail.** All examples describe the same manipulation segment.

Detail level	Annotation requirements	Example language annotation
Brief	Summarize the main action and core object with a verb–object phrase.	Pick up the snack.
Medium	Summarize the manipulation in one sentence, adding object attributes, spatial relations, or visible outcomes.	Pick up the snack in the light-blue package.
Detailed	Describe the manipulation chronologically, including discernible hand actions, object interactions, and necessary adjustments.	The right hand first touches and grasps the snack in the light-blue package, then lifts it from the table, while the left hand does not participate.

The three descriptions convey the same manipulation. The Brief level specifies the main action and object, the Medium level adds the object's appearance, and the Detailed level further specifies hand involvement and action order.

Scene information. We use Qwen 3.8 Max [35] and Seed 2.1 Pro [36] for subtask segmentation and caption annotation. Scene annotations supplement task descriptions with the working environment, spatial layout, and initial state. Lighting annotations describe brightness, exposure, illumination uniformity, and shadows. For example, the scene annotation for one clip is: "The manipulation takes place on a table covered with a black-and-white striped tablecloth, with various packaged snacks scattered across the surface." Its lighting annotation is: "Indoor lighting is bright and uniform, with no obvious shadows or overexposed regions. All snacks and the manipulation area are clearly visible."

Temporal granularity. Episode descriptions cover an entire trajectory, while Subtask descriptions correspond to local manipulation segments with start and end times. A VLM segments continuous video

into subtasks and outputs each segment's timestamps and description. Descriptive detail and temporal granularity are independent annotation dimensions: the same subtask can have descriptions at different detail levels, and an entire trajectory can also have a brief or detailed description. During training, the Episode setting uses the description of the full trajectory, while the Subtask setting uses the annotation of the corresponding time interval as the language condition.

3.3. Language–Video–Action Alignment

UMI and robot demonstrations follow the video curation and filtering procedures in Section 3.2.1 and use the language annotations described in Section 3.2.2. Building on these language–video correspondences, we align video segments with state and action records to construct language–video–action supervision. We describe the collection and curation of these demonstrations below.

UMI demonstrations. We use 411,754 in-house UMI trajectories totaling 15,284.9 hours. We collect human manipulation demonstrations using RoboPocket [15]. The system integrates a smartphone with a handheld adaptive gripper into a portable collection device. Operators interact with objects by moving, opening, and closing the gripper. Its structure and interaction mechanism resemble robot end effectors, reducing interface differences between human demonstrations and robot execution. Our collection tasks include mouse placement, block sorting, seasoning pouring, towel folding, bimanual snack bagging, and other everyday manipulation tasks. They involve pick-and-place, long-horizon manipulation, large wrist rotations, deformable object manipulation, and bimanual coordination. Collection environments include indoor and outdoor scenes with different lighting and background textures.

During collection, head and bilateral wrist views record the overall manipulation scene and local interaction details, respectively. Smartphone visual–inertial tracking provides action trajectories, while external encoders record gripper opening and closing. A shared coordinate system and time reference establish spatial and temporal correspondences among the videos of both hands, motion trajectories, and gripper states. The collection procedure includes equipment checks and synchronization verification before recording, tracking stability and kinematic feasibility checks during recording, and video and AR trajectory playback afterward. These checks identify anomalies such as localization drift and trajectory jumps and verify task completion.

Robot demonstrations. We curate data from AgiBot World Alpha/Beta/2026 [37], RoboCOIN [38], RoboMIND [39], InternData-A1 [40], Galaxea Open-World [41], and ABC-130K [42]. These sources span different robot embodiments, camera configurations, and manipulation environments. We retain only demonstration subsets collected with bimanual robots that include head and bilateral wrist views and meet our quality requirements, maintaining consistent data interfaces and observation semantics. We exclude subsets with missing views, inconsistent view semantics, insufficient image resolution, overly simple tasks, single-arm robot embodiments, or other issues. After applying the configured episode exclusion lists, we retain 768,356 trajectories totaling 9,836.0 hours.

State–action representation. UMI and robot demonstrations use a shared 16-dimensional bimanual end-effector representation. Each hand has a three-dimensional position, a four-dimensional rotation quaternion in wxyz order, and a one-dimensional gripper value. State and action records are associated with their corresponding video segments. Each source is normalized using its own q1 and q99 statistics.

Trajectory quality. For our in-house UMI demonstrations, we construct action-supervision masks from the state and action validity annotations provided by the collection pipeline. Videos participate in video learning, while valid action labels additionally provide action supervision and invalid labels are masked. For the collected robot demonstrations, we first check action dimensions, missing values, NaNs, non-finite values, and quaternion validity. For subsets with the required joint and end-effector state fields, we further exclude trajectories with abnormal joint angles, excessive changes in joint angles or end-effector poses between adjacent frames, or no movement throughout the trajectory. Trajectories that pass these quality checks are used for action training.

4. Experiments

We aim to use multisource experience to preserve the language–video capabilities acquired during video pretraining and extend them to generalizable robot control. Our experiments examine how language–video alignment, language–video–action alignment, joint training, and pretraining scale affect downstream generalization. We organize our study around the following four questions.

- How to establish effective cross-modal alignment?
- How to integrate multisource experience?
- Does scaling pretraining improve control generalization?
- How well does our model perform in unseen environments?

4.1. Experimental Setup

Pretraining. We pretrain the model using the architecture and multisource corpus introduced above. The video branch is initialized from Wan2.2-TI2V-5B [2], and the action branch is randomly initialized. All model parameters are updated during training. Pretraining and post-training both use the joint video and action objective defined above. We use AdamW with an initial learning rate of 4×10^{-5} and a default global batch size of 256. Training steps and specific settings for each study are given in the corresponding experiment descriptions and table captions.

Simulation post-training. We use the 50 manipulation tasks in RoboTwin 2.0 [43]. For the main simulation comparisons, we include VLA baselines (StarVLA [44], $\pi_{0.5}$ [45], Xiaomi Robotics-0 [46], and X-VLA [47]) and video-based WAM baselines (Fast-WAM [21], LingBot-VA [8], OpenWAM [9], X-WAM [48], and 4D-WAM [49]). Observations include head and bilateral wrist-camera views and the end-effector state. The model predicts bimanual end-effector actions. Each task provides 50 Clean demonstrations, yielding 2,500 demonstrations in total. Starting from the corresponding pretrained checkpoint, we jointly post-train a single model on all 50 tasks for 20k steps with a global batch size of 128.

Simulation evaluation. We evaluate control generalization on RoboTwin under two settings. Models are post-trained exclusively on Clean demonstrations. **Easy** evaluates the resulting models under the same clean conditions used for post-training (Clean-to-Clean, **C2C**). **Hard** evaluates the same models under randomized conditions, including variations in background textures, clutter, lighting, table height, and camera placement (Clean-to-Random, **C2R**). For each task and setting, we run 25 evaluation episodes for ablation studies and 100 episodes for main results. We report full-task success rate (SR).

Real-world post-training. We use a bimanual Flexiv Rizon 4 platform for cup stacking, towel folding, and gift-in-box. For real-world evaluation, we select StarVLA, $\pi_{0.5}$, Fast-WAM, LingBot-VA, and OpenWAM as representative VLA and WAM baselines. Observations comprise two fisheye wrist-camera views and one fixed head-camera view. All compared models use the same task-specific post-training data. Starting from upstream pretrained weights, we adapt a separate model to each task for 50k steps with a global batch size of 128.

Real-world evaluation. We evaluate under **Easy (C2C)** and **Hard (C2R)** settings. Easy varies the positions of familiar objects, and Hard changes objects, backgrounds, or both. Each task and setting uses ten layouts fixed before evaluation and shared across all models. We run three trials per layout, yielding 30 trials per task and setting. All models use synchronous inference and execute actions at 30 Hz. Success rate (SR) measures the proportion of trials that complete the entire task. Partial success rate (PSR) measures the average fraction of predefined subtasks completed per trial. We compute both metrics per task and average them equally across the three tasks.

4.2. How to establish effective cross-modal alignment?

We examine language–video and language–video–action alignment through the design of language annotations, paired action supervision, and complementary video experience. Downstream control evalu-

ation measures the effects of these supervision choices.

4.2.1. Language–Video Alignment

Detailed and accurate language descriptions improve downstream control generalization. Pretrained video models learn rich information about scenes, objects, and motion from large-scale language–video data. During WAM pretraining, we provide informative language supervision that accurately matches video content. We aim to help the model retain pretrained priors and reduce forgetting of existing language–video capabilities when adapting to robot manipulation data. Across all settings, we pretrain on a mixture of egocentric videos and UMI demonstrations for 20k steps, examining description detail, scene information, and temporal granularity as three aspects of language supervision. For description detail, we compare three levels, Brief, Medium, and Detailed. We also compare settings with and without scene descriptions covering backgrounds, spatial layouts, and lighting. For temporal granularity, we compare Episode descriptions covering complete trajectories with Subtask descriptions of local manipulation segments.

We compare these language supervision choices in Table 3. At each temporal granularity, more detailed descriptions yield higher Hard success rates. At the Subtask level, the Hard success rate increases from 47.20% to 49.60% and 53.52% with Brief, Medium, and Detailed descriptions, respectively. Adding environment information further raises it to 55.04%, achieving the highest Easy and Hard success rates in this group. For temporal granularity, replacing Episode descriptions with Subtask descriptions under the Detailed setting increases the Hard success rate from 51.60% to 53.52%. These results indicate that richer manipulation descriptions, additional environment information, and subtask-level language supervision improve downstream control generalization. We further randomly shuffle the correspondence between Detailed Subtask descriptions and video segments within each trajectory, reducing the Hard success rate from 53.52% to 52.00%. This result further highlights the importance of accurately matched language supervision for language–video alignment.

Table 3. **Caption design and temporal alignment.** We vary description detail, scene information, temporal granularity, and caption–video correspondence. Gray rows indicate the configurations used in our final model.

Granularity	Setting	Easy (C2C)	Hard (C2R)
Caption design			
Episode	Brief	81.36	44.96
Episode	Detailed	82.00	51.60
Subtask	Brief	83.60	47.20
Subtask	Medium	82.08	49.60
Subtask	Detailed	83.68	53.52
Subtask	Detailed + Env.	**83.92**	**55.04**
Temporal alignment			
Subtask	Aligned	**83.68**	**53.52**
Subtask	Shuffled	83.04	52.00

Table 4. **Robot caption variants.** We compare original task instructions, rephrased instructions, and detailed trajectory-specific descriptions.

Caption	Easy (C2C)	Hard (C2R)
Original	84.64	50.80
Rephrased	86.08	54.76
Detailed + Env.	**86.48**	**56.16**

Table 5. **UMI action-label coverage.** All videos are retained as action-label coverage varies.

Action	Easy (C2C)	Hard (C2R)
0%	80.96	46.72
25%	81.44	47.76
50%	83.28	49.92
75%	**84.08**	50.16
100%	83.28	**51.28**

We also find that many trajectories in the open-source robot data we examine share the same simple task instruction. Some datasets use task names or short instructions as video captions, without detailed descriptions of the manipulation process and environment. For this study, we randomly select nine subsets of the robot data, containing 145,430 trajectories but only 592 task-level instruction entries. These shared instructions do not adequately capture differences in manipulation processes and scene conditions across trajectories, limiting fine-grained correspondence between language supervision and trajectory-level visual content. We construct two variants to enrich language supervision. Rephrased provides 100 semantically equivalent paraphrases per task instruction, randomly sampling one during training to increase linguistic diversity. Detailed + Env. provides a detailed description of the specific manipulation process and environment for each trajectory, further enriching the content and diversity of language supervision. Relative to the 50.80% Hard success rate of Original, paraphrasing alone yields a gain of 3.96 percentage points, while Detailed + Env. achieves the highest success rate in this group

at 56.16%, as shown in Table 4. These results further indicate that enriching linguistic expression and descriptive content helps improve language–video alignment and generalization.

4.2.2. Language–Video–Action Alignment

Human manipulation experience provides effective language–video–action supervision. Language–video supervision associates language descriptions with manipulation processes in videos, while paired action labels further provide supervision for language–video–action alignment. Direct post-training from Wan achieves a Hard SR of 34.56%. With robot-data pretraining, it reaches 56.16% in Table 4, demonstrating the value of robot manipulation data. We further examine whether human manipulation experience from UMI can provide effective cross-embodiment language–video–action supervision. To examine the contribution of UMI action supervision, we retain all UMI videos and vary only the proportion of trajectories providing action labels during pretraining. Videos without action labels continue to participate in video learning. At each coverage level, the subset of trajectories providing action supervision remains fixed. All settings use the same pretraining budget, followed by 20k steps of downstream post-training. From Table 5, we observe that as action-label coverage increases from 0% to 100%, the Hard success rate increases across all five coverage levels, rising from 46.72% to 51.28%, a gain of 4.56 percentage points. These results indicate that paired actions from human manipulation experience provide effective supervision for robot control, although UMI data differ visually from robot execution. Given that UMI demonstrations can be collected with handheld devices across diverse scenes, this finding supports human manipulation demonstrations as a source for scaling pretraining experience for robot control.

4.3. How to integrate multisource experience?

After examining the roles of language–video and language–video–action supervision, we study how to integrate multisource experience for pretraining. We first examine the benefits of increasing training on General videos, egocentric videos, and UMI demonstrations. Given differences across sources in data quality, scene and task distributions, and action supervision, we then study how the Ego–UMI sampling ratio affects downstream control generalization under a fixed training budget. Finally, we compare model architectures and conditioning designs under mixed-source training, examining how the pretrained model's existing language–video capabilities can support action learning.

Table 6. **Different data sources strengthen language–video and language–video–action alignment.** We separately increase pretraining exposure to general videos (**General**), egocentric videos (**Ego**), and UMI data (**UMI**).

Data	Steps	Easy (C2C)	Hard (C2R)
General videos			
UMI/Ego 20k	20k	**82.64**	53.84
+ General 25k	45k	**82.64**	**56.48**
Egocentric videos			
UMI only	10k	**83.04**	49.68
+ Ego 10k	20k	82.64	53.84
+ Ego 20k	30k	82.32	**55.60**
UMI data			
Ego only	10k	79.68	42.08
+ UMI 10k	20k	82.64	**53.84**
+ UMI 20k	30k	**83.60**	51.20

Transitional pretraining on General videos provides a stronger starting point. We first examine the role of pretraining on General videos. Their content distribution is close to that of the original video pretraining data, while WAM adjusts input frame count, resolution, and attention computation to support robot control. We therefore first use General videos to help the model adapt to these computational changes under a content distribution close to its pretraining distribution. This transition aims to better preserve existing video priors and provide a foundation for subsequent embodied learning. To this end, we add 25k steps of transitional pretraining using only General videos before 20k steps of mixed

training on egocentric videos and UMI demonstrations, increasing the Hard success rate from 53.84% to 56.48%, as reported in Table 6.

Increasing pretraining exposure to individual data sources improves WAM performance. Egocentric videos broaden the range of egocentric manipulation scenarios, and UMI demonstrations provide manipulation experience paired with actions. We next increase training exposure to egocentric videos or UMI demonstrations while keeping exposure to the other source fixed. We mix egocentric videos and UMI demonstrations during training. Each source's training exposure is calculated from the total training steps and its sampling proportion. As shown in Table 6, increasing training on egocentric videos or UMI demonstrations generally improves downstream control performance.

Equal Ego–UMI sampling yields the best performance in unseen environments. The preceding experiments increase both source exposure and the training budget. We further examine how to allocate a fixed budget across the two sources. Egocentric videos provide diverse visual experience across tasks and scenes, while UMI data provide both video and action supervision. The sampling ratio determines how these forms of supervision are allocated during training. We fix the pretraining budget at 20k steps and compare five Ego-to-UMI sampling ratios of 1:0, 2:1, 1:1, 1:2, and 0:1. In Table 7, we observe that all three mixed settings achieve higher Hard success rates than training on either source alone. Equal sampling achieves the highest Hard success rate of 53.84%, improving over UMI-only training at 51.28% by 2.56 percentage points. These results indicate that an appropriate Ego-to-UMI sampling ratio helps balance language–video and language–video–action supervision and improves control generalization in randomized scenes.

Table 7. **Ego and UMI provide complementary supervision.** We compare five Ego-to-UMI sampling ratios under a fixed 20k-step pretraining budget.

Ratio	Easy (C2C)	Hard (C2R)
1:0	77.04	42.56
2:1	82.72	52.64
1:1	82.64	**53.84**
1:2	83.20	53.28
0:1	**83.28**	51.28

Appropriate architectural choices help preserve pretrained video priors. Under the same 30k-step multisource pretraining setup, we compare Single-stream, in which video and actions share a residual stream, with Dual-stream, which maintains separate residual streams. We also examine the language conditioning pathway of the action branch. We separate video and action feature spaces to accommodate both the reuse of pretrained video representations and the learning of new action representations. The video branch has already learned correspondences between language and visual dynamics during pretraining. The newly introduced text conditioning pathway in the action branch, however, lacks sufficient pretraining. We therefore aim to reuse the semantic information available in the video branch for action prediction, reducing potential interference from the new text conditioning pathway. To this end, we retain language conditioning in the video branch and let the action branch access language information only through video features. As reported in Table 8, Single-stream and standard Dual-stream achieve similar Hard success rates of 52.56% and 52.48%, respectively. The dual-stream model using this language conditioning pathway achieves the highest Hard success rate in this group at 57.36%, a gain of 4.88 percentage points over standard Dual-stream. These results support conditioning the action branch on language through video features.

Table 8. **Model architecture affects control generalization.** We compare video–action architectures (top) and extra image features (bottom).

	Easy (C2C)	Hard (C2R)
Architecture		
Single-stream	**86.48**	52.56
Dual-stream	84.96	52.48
Ours	84.80	**57.36**
Extra features		
Native	**83.44**	**49.52**
+ CLIP [50]	82.40	42.48
+ DINOv2 [51]	82.00	39.76
+ SigLIP2 [52]	82.40	42.40
+ Qwen3-VL	81.60	42.96

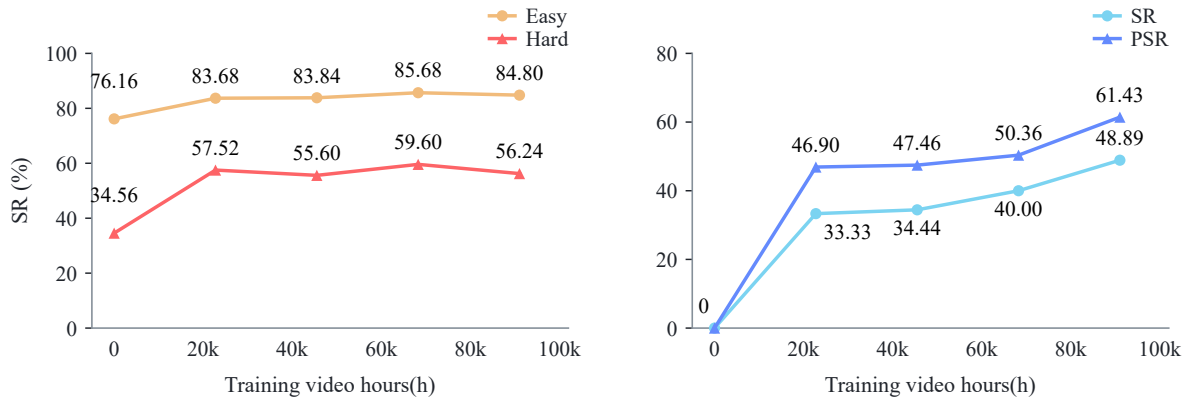
For visual conditioning, we examine whether image-encoder features can provide benefits. We pretrain for 20k steps with equal sampling of egocentric videos and UMI data. Using Native as the baseline, we separately add image features extracted by CLIP, DINOv2, SigLIP2, and Qwen3-VL 2B [35] through cross-attention to provide additional visual conditioning. Native achieves the highest Easy and Hard success rates in this group, at 83.44% and 49.52%, respectively. Adding features from the four image encoders lowers the Hard success rate to 39.76%–42.96%, as summarized in Table 8. These results support retaining the pretrained video model's native visual inputs. The fusion pathway for the additional

image features has not undergone the same large-scale language–video pretraining as the video branch. This may lead the model to rely more on appearance cues in the current training distribution, weakening generalization to unseen scenes.

4.4. Does scaling pretraining improve control generalization?

Building on the preceding supervision design and multisource training recipe, we evaluate different pretrained checkpoints in simulation and on real robots. Within each evaluation domain, we use the same downstream post-training settings to examine how additional pretraining affects control generalization.

Increasing pretraining investment in multisource experience provides a stronger WAM. We study how model capabilities scale with pretraining, measured in training video hours. Here, 50k pretraining steps correspond to approximately 45k video hours. We use a mixture of egocentric videos and UMI demonstrations with equal sampling weights. For this study, we combine the configurations marked by gray rows in Tables 3, 7, and 8, such as using Detailed Subtask descriptions with scene and lighting information as language conditions. The 0-hour setting skips multisource pretraining and proceeds directly to downstream post-training from the video-model initialization.



(a) RoboTwin (b) Real World

Figure 6. **Scaling pretraining improves control generalization.** We vary training video hours on a fixed corpus of multisource experience, using the same downstream post-training protocol within each domain. (a) RoboTwin SR under the Easy (C2C) and Hard (C2R) settings. (b) Real-world SR and PSR under the Hard setting, averaged across three tasks. The 0-hour setting uses the video-model initialization without additional pretraining and achieves no success in real-world evaluation.

As shown in Figure 6(a), multisource pretraining improves both Easy and Hard success rates over the 0-hour baseline. The highest Hard success rate reaches 59.60%, compared with 34.56% without pretraining. In Figure 6(b), we show that real-world Hard SR and PSR, averaged equally across three tasks, increase consistently as training video hours grow from 22.5k to 90k, rising from 33.33% and 46.90% to 48.89% and 61.43%, respectively. The 0-hour setting fails to complete any task. These results show continued gains within the evaluated range, supporting longer pretraining on both sources for control under real-world scene changes.

Table 9. **Scaling pretraining on human experience and robot data improves generalization across unseen environmental conditions.** We first pretrain on egocentric videos and UMI demonstrations (**Stage 1**), then continue training on robot data (**Stage 2**). The middle four columns evaluate single-factor environmental changes.

Stage 1	Stage 2	Easy (C2C)	Background	Lighting	Clutter	Height	Hard (C2R)
0k	0k	76.16	59.12	76.24	51.28	57.68	34.56
20k	0k	82.64	72.16	84.08	68.56	70.32	53.84
75k	0k	85.68	77.92	85.60	78.48	75.44	59.60
75k	25k	**88.08**	**82.16**	**88.08**	**80.72**	**76.96**	**61.52**

After 75k steps of pretraining on egocentric videos and UMI demonstrations, we continue pretraining on robot data for another 25k steps and apply the same RoboTwin post-training settings. In Table 9, we observe that success rates under individual changes in background, lighting, clutter, and table height improve as pretraining investment increases, with the Hard success rate ultimately reaching 61.52%.

4.5. How well does our model perform in unseen environments?

Building on the preceding findings and training recipe, we compare our models pretrained on our full multisource dataset with existing policies in RoboTwin simulation and real-world manipulation.

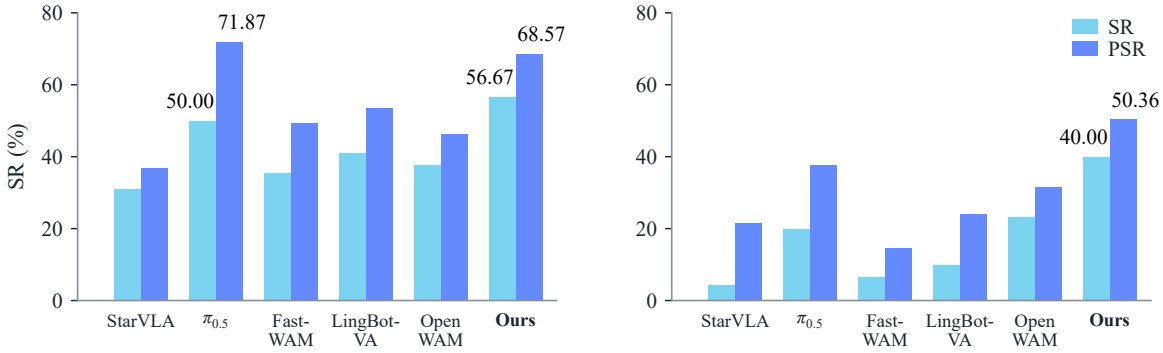
Simulation results. We compare VLA and WAM methods on the 50 manipulation tasks in RoboTwin 2.0. Each model is post-trained on Clean demonstrations and evaluated under Easy (C2C) and Hard (C2R) settings. As reported in Table 10, our model achieves a Hard success rate of 61.56%, outperforming all compared methods and exceeding the strongest baseline, OpenWAM, by 12.86 percentage points. Its Easy success rate is 88.10%, slightly below OpenWAM's 89.30%, while its average success rate across both settings reaches 74.83%, the highest among the compared methods. Our model and OpenWAM perform similarly under Clean conditions, while our model achieves a higher overall success rate under the Hard setting, indicating better generalization across scene conditions.

Table 10. **Our model outperforms VLA and WAM baselines in simulation and real-world evaluation.** We compare Easy (C2C) and Hard (C2R) settings, reporting SR in simulation and SR/PSR on real-world tasks.

Model	RoboTwin			Real-world			
	Easy	Hard	Avg.	Easy SR	Easy PSR	Hard SR	Hard PSR
VLA							
StarVLA [44]	60.03	14.06	37.05	31.11	36.87	4.44	21.55
$\pi_{0.5}$ [45]	70.66	46.00	58.33	50.00	**71.87**	20.00	37.78
Xiaomi Robotics-0 [46]	62.86	18.22	40.54	–	–	–	–
X-VLA [47]	67.96	20.95	44.46	–	–	–	–
WAM							
Fast-WAM [21]	77.76	1.92	39.84	35.56	49.33	6.67	14.52
LingBot-VA [8]	77.04	36.16	56.60	41.11	53.61	10.00	24.01
OpenWAM [9]	**89.30**	48.70	69.00	37.78	46.35	23.33	31.67
X-WAM [48]	69.98	25.84	47.91	–	–	–	–
4D-WAM [49]	81.50	41.80	61.65	–	–	–	–
Ours	88.10	**61.56**	**74.83**	**56.67**	68.57	**40.00**	**50.36**

Real-world results. We evaluate cup stacking, towel folding, and gift-in-box on a bimanual Flexiv platform. All three tasks start from the same pretrained checkpoint and undergo 50k steps of task-specific post-training. Each task-specific model is evaluated under both Easy (C2C) and Hard (C2R) settings. All compared methods use the same task-specific post-training data. Under the Hard setting, our model achieves the highest SR and PSR among the evaluated methods, as shown in Table 10 and Figure 7. Averaged equally across the three tasks, SR reaches 40.00%, exceeding OpenWAM's 23.33% by 16.67 percentage points. PSR reaches 50.36%, exceeding $\pi_{0.5}$'s 37.78% by 12.58 percentage points. These results show improvements in both full-task success and subtask completion under unseen object and background conditions, supporting the generalization benefits of the training recipe in real-world robot manipulation.

Real-world visualizations. Figure 8 shows how our model and the baselines perform the three tasks, with light red dashed boxes highlighting the baselines' main execution errors. In cup stacking, Fast-WAM misplaces the first cup of the second layer, causing it to fall onto the cups below, and cannot complete the stack afterward. For towel folding, our model lifts and spreads both sides before completing both folds, whereas OpenWAM lifts only one side and fails to fold the towel properly. In gift-in-box, $\pi_{0.5}$ grasps the object but gradually loosens its gripper, causing the object to fall, and remains unsuccessful after repeated attempts. Our model retains its grasp and places the object into the box. To provide



(a) Real-world (Easy) (b) Real-world (Hard)

Figure 7. **Our method achieves the best performance under the Hard setting in real-world evaluation.** Results are averaged equally across three real-world tasks.

a fuller view of our model's execution, Figure 9 presents head and external camera recordings. The gift-in-box model generalizes to fruit placement with a different scene, manipulated objects, and target container. Trained with same-color cups on a gray tablecloth, the cup-stacking model also completes rainbow stack with multicolored cups on a patterned tablecloth. These examples illustrate how our model transfers learned manipulation skills to unfamiliar objects and scenes.

5. Discussion and Conclusion

In this work, we studied how to extend pretrained language–video capabilities to generalizable robot control with multisource experience. Using approximately 47,000 hours of experience, we systematically investigated language–video alignment, language–video–action alignment, and their joint learning. We found detailed and accurate language descriptions preserve pretrained language–video capabilities. Meanwhile, action supervision from UMI demonstrations further supports robot control, demonstrating the value of human manipulation experience for robot learning. Our findings prove the importance of experience organization and supervision usage. The resulting training recipe demonstrates generalization to unseen conditions in simulation and real-world evaluations, with further gains from continued pretraining. We release the data, code, and model weights to support future research.

Upon the pretrained video model, the proposed method offers limited support for complex reasoning and explicit multi-stage planning, despite outperforming VLA models in our control generalization evaluations. The limited historical context may also hinder the execution of long-range tasks. In the future, we plan to explore memory mechanisms and hierarchical planning to support tasks that require reasoning over past observations and planning subsequent actions.

Acknowledgments

We thank Zijun Xu for his contributions to the early exploration of this work. We also thank Yutong Dong for her efforts in real-world deployment.

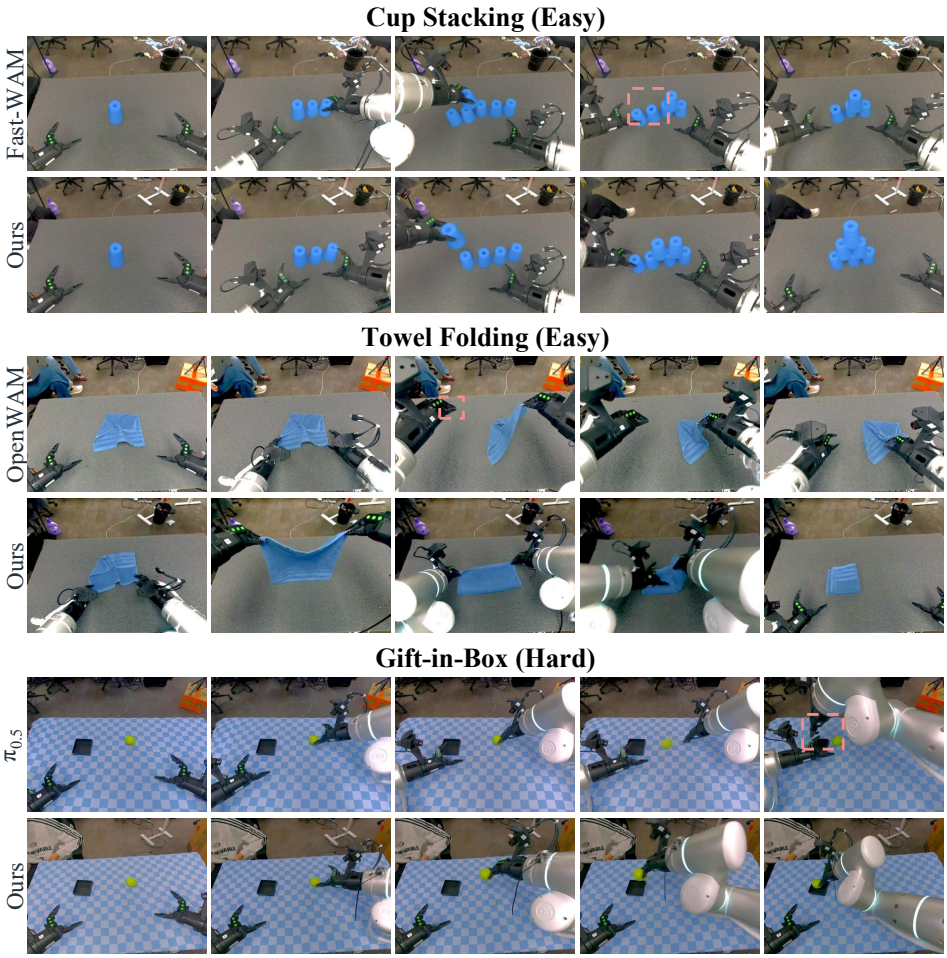


Figure 8. **Real-world execution comparisons.** Our model completes cup stacking and towel folding under the Easy setting and gift-in-box under the Hard setting, while the corresponding baseline executions fail. Light red dashed boxes highlight the baselines' main execution errors.

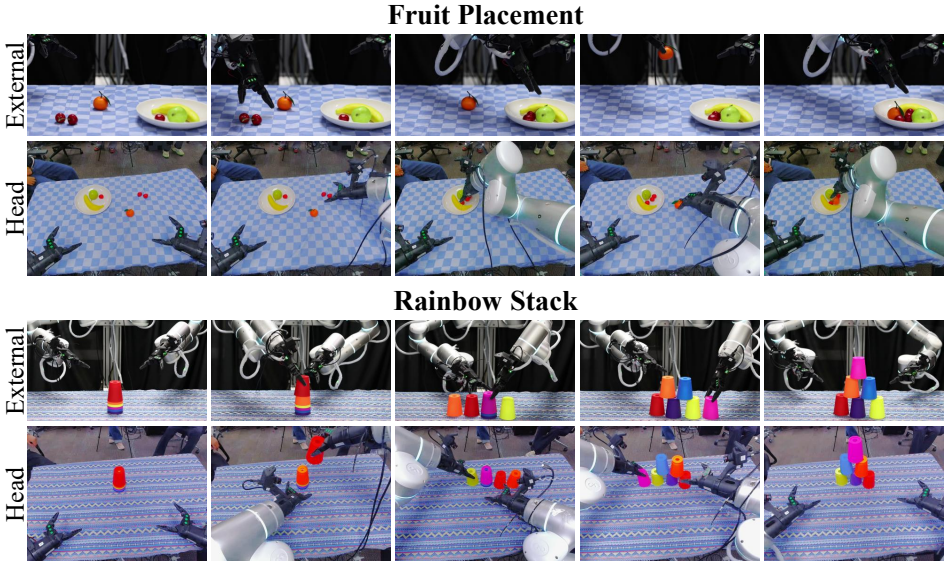


Figure 9. **Generalization examples from head and external camera views.** Fruit placement and rainbow stack demonstrate generalization to new objects, containers, and scene appearances.

References

[1] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. Hunyuanvideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[2] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[3] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *ICLR*, 2025.

[4] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

[5] Hongtao Wu, Ya Jing, Chilam Cheang, Guangzeng Chen, Jiafeng Xu, Xinghang Li, Minghuan Liu, Hang Li, and Tao Kong. Unleashing large-scale video generative pre-training for visual robot manipulation. In *ICLR*, 2024.

[6] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video Prediction Policy: A Generalist Robot Policy with Predictive Visual Representations. In *ICML*, volume 267, 2025.

[7] 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.

[8] 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.

[9] Heng Yu, David D Yuan, Juze Zhang, Changan Chen, Yao Feng, Michelle Baldonado, Steve Cousins, Li Fei-Fei, Jiajun Wu, and Ehsan Adeli. OpenWAM: An Open Framework for Composable World-Action Models. *arXiv preprint arXiv:2610.07922*, 2026.

[10] Zhihao Sun, Liu Liu, Xinjiang Wang, Haoyi Jiang, Wei Feng, Huiqiang Zhang, Xiaosong Jia, Zhizhong Su, and Zuxuan Wu. Ego4WAM: What Matters When Scaling Egocentric Human Data for Robot Learning? *arXiv preprint arXiv:2609.40341*, 2026.

[11] Jewon Yeom, Hanseul Kim, Jeongjae Park, Sungmok Jung, Jaejin Lee, and Taesup Kim. What Makes Video World Model Latents Action-Relevant: Prediction over Reconstruction. *arXiv preprint arXiv:2606.07687*, 2026.

[12] Haoyi Jiang, Liu Liu, Xinjiang Wang, Zhihao Sun, Zequn Chen, Sen Wang, Xinjie Wang, Xia Chen, Jingfeng Yao, Weiheng Zhao, et al. Rethinking Representations for World-Action Modeling. *arXiv preprint arXiv:2609.38163*, 2026.

[13] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal Manipulation Interface: In-The-Wild Robot Teaching Without In-The-Wild Robots. In *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024.

[14] Zhaxizhuom Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, et al. Fastumi: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning*, pages 3069–3093. PMLR, 2025.

[15] Junjie Fang, Wendi Chen, Han Xue, Fangyuan Zhou, Tian Le, Yi Wang, Yuting Zhang, Jun Lv, Chuan Wen, and Cewu Lu. Robopocket: Improve robot policies instantly with your phone. *arXiv preprint arXiv:2603.05504*, 2026.

[16] Jiacheng Wei, Jerry Bai, Xiaoyu Yue, Zidong Wang, Xiaoyang Guo, Cheng Chen, Fanqi Pu, Fan Wu, Zhixu Yue, Yizhuo Li, et al. XPACE: Joint World and Action Modeling from Heterogeneous Experience. *arXiv preprint arXiv:2609.17372*, 2026.

[17] Xingyu Miao, Zizun Li, Baole Fang, Kaiwen Song, Tenghui Wang, Hanxue Zhang, Yating Wang, Xudong Li, Yuping He, Xueyuan Wei, et al. InternW0-Δ: A World Action Model Bridging Predictive Dynamics and Actions with 20K+ Hours of Open Data. *arXiv e-prints*, 2026.

[18] Yanjiang Guo, Haodong Yan, Zhide Zhong, Zhongru Zhang, Qingyuan Yang, Qingzhou Lu, Xiaoyu Chen, Yen-Jen Wang, Shuying Deng, Chenghan Yang, et al. Video Prediction Policy 2: Predict Better, Act Better. *arXiv preprint arXiv:2610.10270*, 2026.

[19] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning Universal Policies via Text-Guided Video Generation. In *NeurIPS*, 2023.

[20] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *CVPR*, pages 35101–35113, 2026.

[21] 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.

[22] OX-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. *arXiv preprint arXiv:2310.08864*, 2023.

[23] Dibya Ghosh, Homer Rich Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Quan Vuong, Ted Xiao, Pannag R. Sanketi, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An Open-Source Generalist Robot Policy. In *Robotics: Science and Systems XX, Delft, The Netherlands, July 15-19, 2024*, 2024.

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

[25] 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 *CVPR*, pages 18995–19012, 2022.

[26] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, et al. Egoverse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*, 2026.

[27] 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 *ICRA*, pages 13226–13233. IEEE, 2025.

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

[29] Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World Action Models Beyond Pixels with In-the-Wild Egocentric Human Data. In *Robot World Models*, 2026.

[30] Weixin Liang, Lili Yu, Liang Luo, Srini 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*.

[31] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *ICLR*, 2023.

[32] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A Large-Scale High-Quality Dataset for Text-to-video Generation. In *ICLR*, 2025.

[33] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. In *CVPR*, pages 13320–13331, 2024.

[34] Lightwheel. EgoSuite-Open100K: A Fully Annotated Open Egocentric Human Dataset. https://huggingface.co/collections/LightwheelAI/egosuite-open100k, August 2026. Dataset.

[35] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[36] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

[37] 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.

[38] Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025.

[39] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

[40] Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. In *CVPR*, pages 976–985, 2026.

[41] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxy Open-World Dataset and G0 Dual-System VLA Model. In *2026 IEEE International Conference on Robotics and Automation (ICRA)*, pages 20749–20756. IEEE, 2026.

[42] Arthur Allshire, Himanshu Gaurav Singh, Ritvik Singh, Adam Rashid, Hongsuk Choi, David McAllister, Justin Yu, Yiyuan Chen, Huang Huang, Pieter Abbeel, et al. Scalable Behavior Cloning with Open Data, Training, and Evaluation. *arXiv preprint arXiv:2606.27375*, 2026.

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

[44] StarVLA Community. StarVLA: A Lego-like Codebase for Vision-Language-Action Model Developing. *arXiv preprint arXiv:2604.05014*, 2026.

[45] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025.

[46] Rui Cai, Jun Guo, Xinze He, Piaopiao Jin, Jie Li, Bingxuan Lin, Futeng Liu, Wei Liu, Fei Ma, Kun Ma, et al. Xiaomi-robotics-0: An open-sourced vision-language-action model with real-time execution. *arXiv preprint arXiv:2602.12684*, 2026.

[47] Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *ICLR*, 2026.

[48] 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.

[49] Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, et al. 4d-wam: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026.

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

[52] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

A. Data Annotation Details

General videos retain their original captions. For egocentric videos, UMI demonstrations, and robotic demonstrations, we construct manipulation annotations using Qwen 3.8 Max primarily for subtask segmentation and Seed 2.1 Pro for caption annotation. The procedure consists of the following steps.

Subtask Segmentation. We use Qwen 3.8 Max to divide egocentric videos and UMI recordings into temporally ordered, non-overlapping subtasks according to action intent, recording each subtask's start and end time. Robotic demonstrations are short, so each complete episode is treated as one subtask without further segmentation.

Caption Annotation. We use Seed 2.1 Pro to annotate each egocentric or UMI subtask and each robotic episode. The model identifies whether the segment contains manipulation and whether the visual evidence is sufficiently clear, and describes the scene, lighting, and actions at multiple levels of detail. Segments that contain no manipulation or whose visual evidence is insufficient are discarded entirely.

Action Quality Filtering. For UMI and robotic data, we further screen the associated state and action trajectories. Samples that fail one or more of the following checks are removed from action training:

- Missing or non-finite values, including NaNs, invalid dimensions, or invalid quaternions.
- Excessive absolute joint values, with any $|q| > 20$ rad.
- Adjacent-frame joint jumps, with any $|q_{t+1} - q_t| > 1$ rad.
- End-effector position jumps, with the Euclidean distance between adjacent 3D positions exceeding 0.2 m.
- End-effector orientation jumps, with a rotation-angle difference exceeding 90° for either arm between adjacent frames.
- Persistent abnormal jerk in the joint or end-effector trajectories.

Condensed versions of the segmentation and caption annotation prompts are provided below.

Split Prompt (Condensed)

Role. You are a professional robot manipulation video analysis expert.
Goal. Split a video into continuous, non-overlapping atomic subtasks. The input video is sampled at 2 fps, with 0.5 seconds between consecutive frames. The overall task is: `task_name`.
Workflow.

1. Watch the complete video and identify every meaningful atomic action in chronological order.
2. Split by action intent.
3. Ensure that subtasks are sequential, with no overlap or gap.
4. Name objects accurately.

Output. Output standardized JSON containing the description, start time, and end time of every subtask.

```
[
  {
    "description": "subtask description",
    "start_time": 0.500,
    "end_time": 2.500
  }
]
```

Caption Prompt (Condensed)

Role. You are a professional first-person manipulation video analysis expert and dense action annotation specialist. Describe scenes and continuous manipulation actions accurately, objectively, and concisely.
Task. Watch the input video and use only visually observable evidence. The reference task, `task_annotation: {task_name}`, is only a weak hint; all descriptions must be supported by the visible video content.
Segment-type classification. The input is a temporal segment cut from a complete video and may contain no actual manipulation, for example waiting, posture adjustment, hovering, or observation. First determine:

- `segment_type = manipulation` when the segment contains an actual operation on a task object.
- `segment_type = no_action` when the segment contains no operation on a task object.

Visual-clear classification. The field `visual_clear` only describes the quality of the visual evidence:

- Set `visual_clear = true` when the main objects and actions are sufficiently visible to judge, even with ordinary motion blur or local occlusion.
- Set `visual_clear = false` when severe blur, darkness, occlusion, corruption, or missing views make the main objects or actions unreliable to judge.

Scene. Describe the initial operating environment in one sentence. Mention only the visible workspace, relevant background objects, approximate positions, and initial state.
Lighting. Describe brightness, exposure, uniformity, shadows, overexposure, underexposure, blur, or occlusion in one sentence.
action_detail. Describe the manipulation as one chronological paragraph, including manipulated objects, discernible hand actions, object trajectories, interactions, state changes, and necessary adjustments.
action_medium. Describe the manipulation in one sentence, adding object attributes, spatial relations, or visible outcomes.
action_short. Summarize the main action and core object with a 2–8 word verb-object phrase.
Output. Return only one valid JSON object with the following fields. Use exactly `manipulation` or `no_action` for `segment_type`, and a JSON boolean for `visual_clear`. Write descriptions in English. For `no_action`, set the three action fields to `none`; otherwise describe only observed actions.

```
{
  "segment_type": "manipulation",
  "visual_clear": true,
  "lighting": "",
  "scene": "",
  "action_detail": "",
  "action_medium": "",
  "action_short": ""
}
```

B. Training Details

B.1. Common Settings

Optimization. Pretraining and post-training share the optimization settings in Table 11. Batch sizes, training steps, and learning-rate schedules vary across training stages, as detailed in Section B.2.

Table 11. **Shared optimization settings.** Values shared across pretraining and post-training.

Configuration	Value
Optimizer	AdamW
Learning rate	4×10^{-5}
Adam (β_1, β_2)	(0.95, 0.999)
Adam ϵ	10^{-8}
Weight decay	10^{-5}
Gradient clip	1.0
Warmup ratio	5%
Random seed	42

Model inputs and trainable components. We use three camera views. Each view is resized to 128×160 and arranged into a 2×2 grid, yielding a final 256×320 composite frame. The head-camera view is duplicated across the top row, while the left- and right-wrist views occupy the bottom row. During training, each view is randomly cropped to 95% of its original extent before resizing, followed by color jitter with brightness, contrast, saturation, and hue parameters of 0.3, 0.4, 0.5, and 0.08, respectively.

We represent bimanual end-effector states and actions as 16-dimensional vectors and predict actions in chunks of 48 steps.

B.2. Training Configurations

Initialization. In pretraining, we initialize the video DiT from Wan2.2-TI2V-5B [2] and load its pretrained UMT5-XXL text encoder and Wan2.2 video VAE. The action branch, including the action and state encoders, transformer blocks, and action decoder, is randomly initialized. For each experiment, post-training starts from the corresponding pretrained weights.

Pretraining. Training uses BF16 mixed precision. The default global batch size for ablation studies is 256. We increase it to 512 for pretraining in the scaling study and subsequent robot-data pretraining described in Section 4.4, as well as for pretraining the final models evaluated in Section 4.5. For the scaling study, we sample egocentric videos and UMI demonstrations with equal probability and use detailed subtask descriptions with scene and lighting information. We evaluate checkpoints at 25k, 50k, 75k, and 100k steps, as shown in Figure 6. Table 12 compares pretraining in the scaling study with simulation and real-world post-training.

Simulation post-training. We jointly train a single model on all 50 RoboTwin 2.0 tasks using 50 Clean demonstrations per task, following the same post-training protocol across simulation experiments. Each sample contains left-wrist, right-wrist, and head-camera views and the current bimanual end-effector state. The same three-view preprocessing and augmentation settings described above are used during simulation post-training. We use absolute end-effector states and actions and normalize both with their q99 statistics.

Real-world post-training. For each real-world task, we train an independent model using task-level instructions. We apply this procedure to cup stacking, towel folding, and gift-in-box.

Table 12. **Comparison of training configurations.**

Configuration	**Pretraining**	**Post-training**	
	Scaling study	Simulation	Real world
Global batch size	512	128	128
Training steps	100,000	20,000	50,000
Learning-rate schedule	Constant	Cosine	Cosine

C. Real-World Evaluation Protocol

C.1. Post-Training Data Preparation

Post-training data. We adapt models separately to cup stacking, towel folding, and gift-in-box using task-specific data for post-training following general-purpose pretraining on egocentric videos and UMI demonstrations. Table 13 summarizes the post-training data for each task.

Table 13. **Task-specific post-training data for real-world evaluation.**

Task	Episodes	Hours
Cup stacking	2,748	25.788
Towel folding	2,852	15.112
Gift-in-box	4,399	21.715

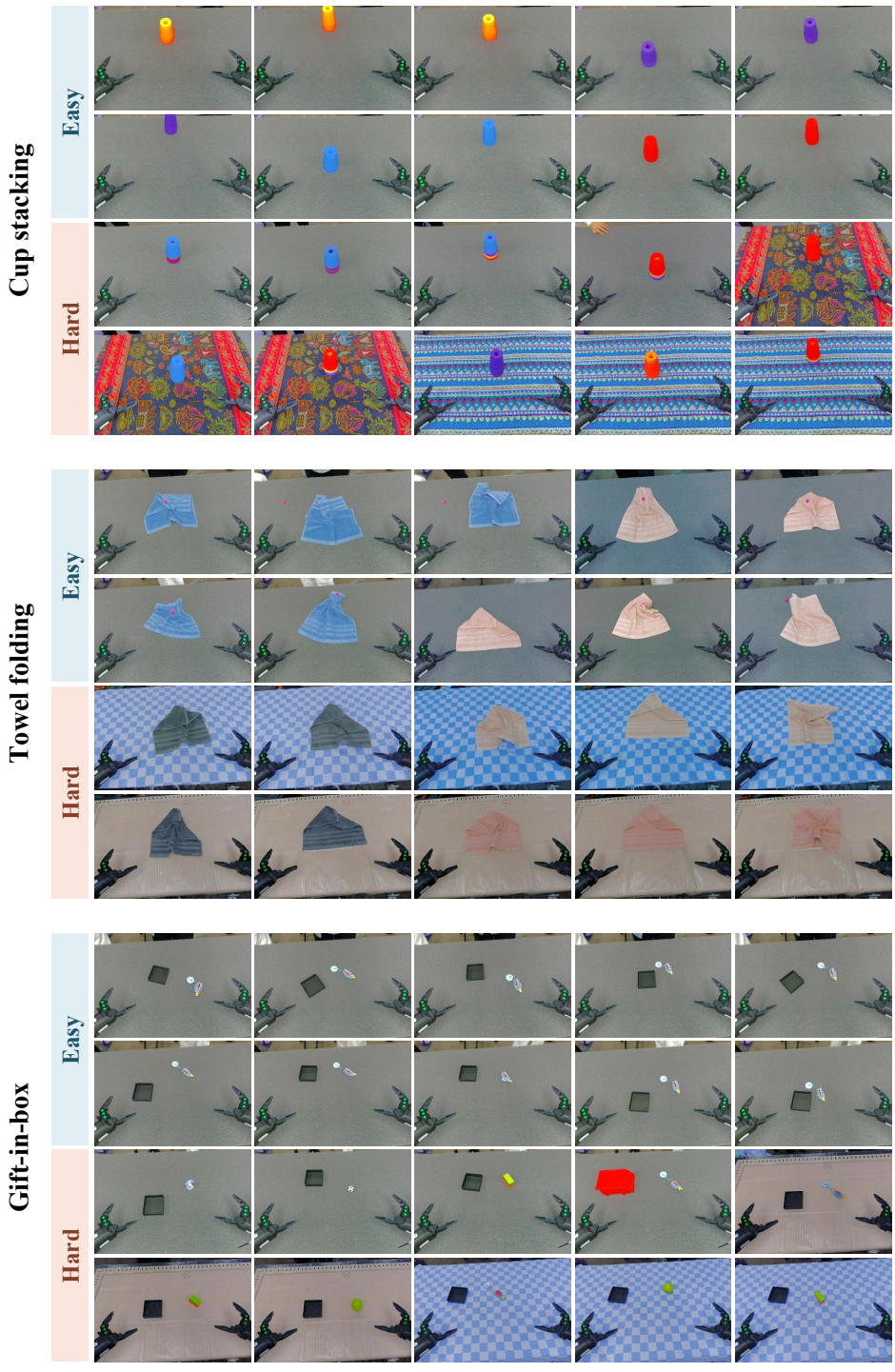


Figure 10. **Real-world evaluation layouts under Easy and Hard settings.** All models use the same ten layouts for each task and setting, with three trials per layout.

Observations and actions. We use a bimanual Flexiv Rizon 4 platform. Observations comprise two fisheye wrist-camera views and one fixed head-camera view. Actions encode the end-effector position, orientation, and continuous gripper width for each arm. For OpenWAM, we follow its official default configuration and use a 6D rotation representation. All other models represent orientation with a wxyz quaternion, yielding a 16-dimensional action vector across both arms.

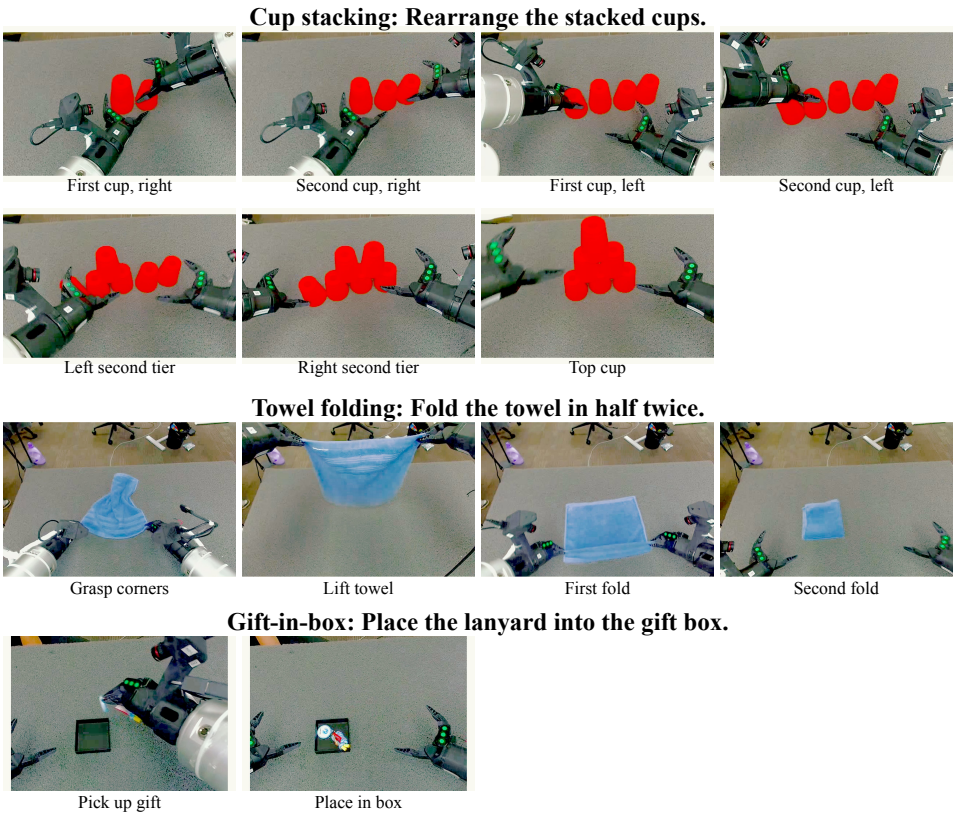


Figure 11. **Subtask illustrations for the three real-world tasks.**

C.2. Evaluation Protocols and Metrics

Evaluation protocols. We evaluate under **Easy (C2C)** and **Hard (C2R)** settings. Easy settings retain the scene conditions and physical objects used in post-training, including their shapes and colors, while varying their initial configurations across layouts. Hard settings vary across scene and object attributes, including background, tabletop material, object instance, shape, color, size, and texture. The evaluation layouts are illustrated in Figure 10. All models use synchronous inference and execute actions at 30 Hz.

Layouts and trials. A layout specifies the initial object configurations, target positions, and scene appearance. For each task and setting, ten layouts are fixed before evaluation and shared across all models. We run three trials per layout to account for variability in object placement and model sampling, yielding 30 trials per task and setting. To reproduce a layout, we align the fixed live head-camera view with a reference photograph using an image overlay.

Scoring and termination. Each task is divided into predefined, equally weighted subtasks, scored by reviewing execution videos. The protocol permits up to five additional autonomous replanning attempts after an initial failed attempt, retaining the current environment state and execution history without human assistance. All models follow the same retry and termination rules. A trial is marked as a failure if the cup tower collapses or an object leaves the workspace.

Metrics. We report success rate (SR) and partial success rate (PSR). For a task with $N = 30$ trials and K subtasks, let $s_{ik} \in \{0, 1\}$ indicate whether subtask k succeeds in trial i. Complete-task success requires all

subtasks to succeed, so its indicator is $z_i = \prod_{k=1}^{K} s_{ik}$. The two metrics are

$$\text{SR} = \frac{1}{N}\sum_{i=1}^{N} z_i, \qquad \text{PSR} = \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K} s_{ik}. \tag{6}$$

SR measures complete-task success. PSR measures progress at the subtask level. The three-task score is the equally weighted mean of the task-level rates, computed separately for Easy and Hard and expressed as a percentage.

Task details. Figure 11 shows the three real-world tasks and their corresponding ordered subtask decompositions. Cup stacking requires arranging cups into a stable pyramid, towel folding requires folding a towel in half twice, and gift-in-box requires placing a gift inside an open gift box. We divide these tasks into seven, four, and two subtasks, respectively, corresponding to observable milestones in task execution. These decompositions define the task-specific subtask counts K used in the PSR calculation above.

D. Detailed Results

D.1. RoboTwin Task Success Rates

Table 14. **Per-task success rates on RoboTwin 2.0 under Easy and Hard settings.** Each task is evaluated over 100 episodes per setting. All values are percentages.

Task	Easy	Hard	Task	Easy	Hard
Adjust Bottle	97	81	Place Can Basket	84	36
Beat Block Hammer	97	55	Place Cans Plasticbox	99	74
Blocks Ranking RGB	98	80	Place Container Plate	96	87
Blocks Ranking Size	87	30	Place Dual Shoes	56	33
Click Alarmclock	100	75	Place Empty Cup	100	77
Click Bell	100	85	Place Fan	87	48
Dump Bin Bigbin	91	66	Place Mouse Pad	73	29
Grab Roller	100	80	Place Object Basket	84	56
Handover Block	75	31	Place Object Scale	94	67
Handover Mic	78	40	Place Object Stand	98	83
Hanging Mug	3	1	Place Phone Stand	87	28
Lift Pot	100	56	Place Shoe	98	82
Move Can Pot	90	72	Press Stapler	81	78
Move Pillbottle Pad	95	58	Put Bottles Dustbin	90	82
Move Playingcard Away	98	56	Put Object Cabinet	78	57
Move Stapler Pad	58	22	Rotate QRcode	90	50
Open Laptop	93	81	Scan Object	95	57
Open Microwave	86	4	Shake Bottle	100	98
Pick Diverse Bottles	94	74	Shake Bottle Horizontally	99	98
Pick Dual Bottles	99	85	Stack Blocks Three	93	82
Place A2B Left	92	52	Stack Blocks Two	100	99
Place A2B Right	95	59	Stack Bowls Three	84	80
Place Bread Basket	87	67	Stack Bowls Two	99	90
Place Bread Skillet	92	57	Stamp Seal	94	29
Place Burger Fries	97	60	Turn Switch	44	51

D.2. Real-World Task Success Rates

D.2.1. Baseline Comparison

Figure 12 reports SR and PSR for each real-world task and their average under the Easy and Hard settings.

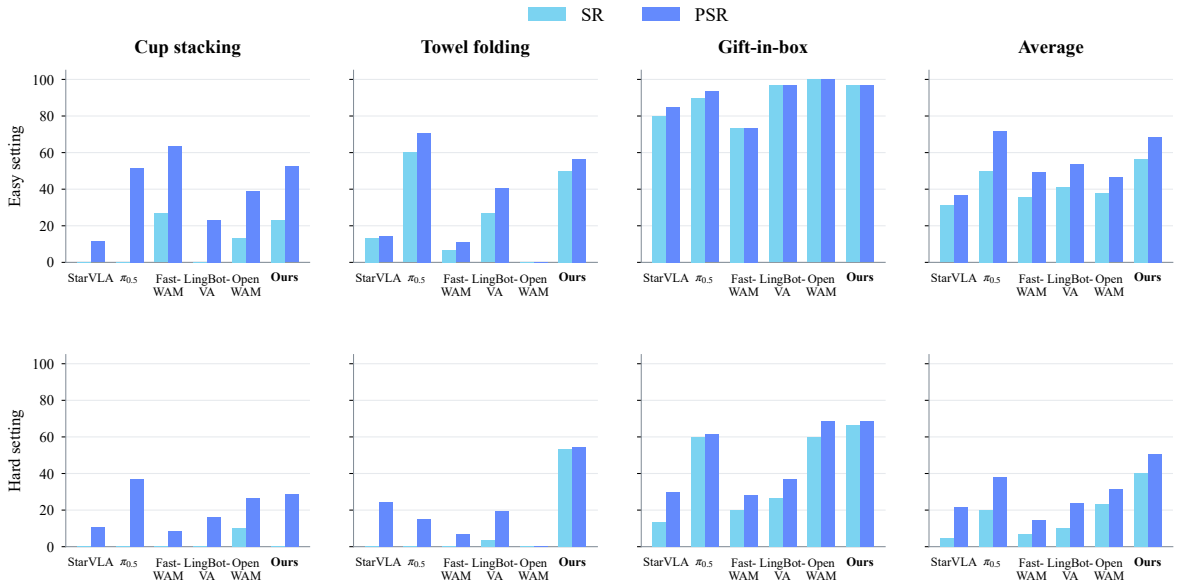


Figure 12. **Real-world comparison of our model and baseline methods.**

D.2.2. Pretraining Scaling

Figure 13 compares the 0-hour setting with models pretrained for 25k, 50k, 75k, and 100k steps, corresponding to approximately 22.5k, 45k, 67.5k, and 90k training video hours, respectively. At 0 hours, SR and PSR are both 0% for every task under both settings.

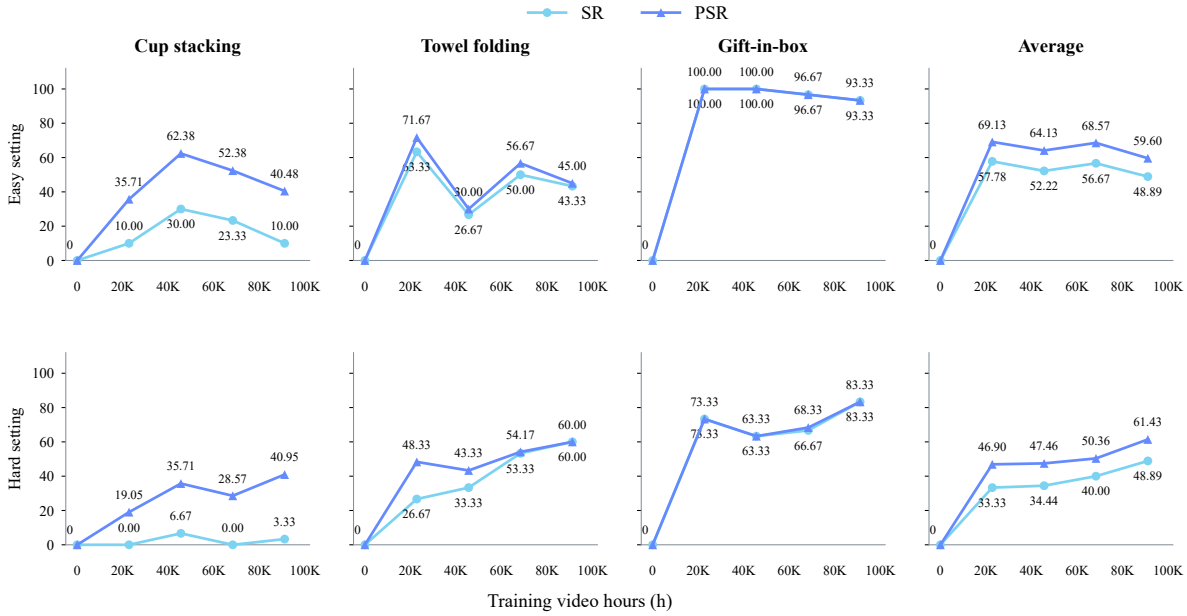


Figure 13. **Real-world performance with increasing pretraining video hours.**

E. Real-World Execution Examples

Figures 14–16 show execution examples under Hard settings. Head-camera frames are sampled at uniform time intervals. Numbers indicate the order of the sampled frames.

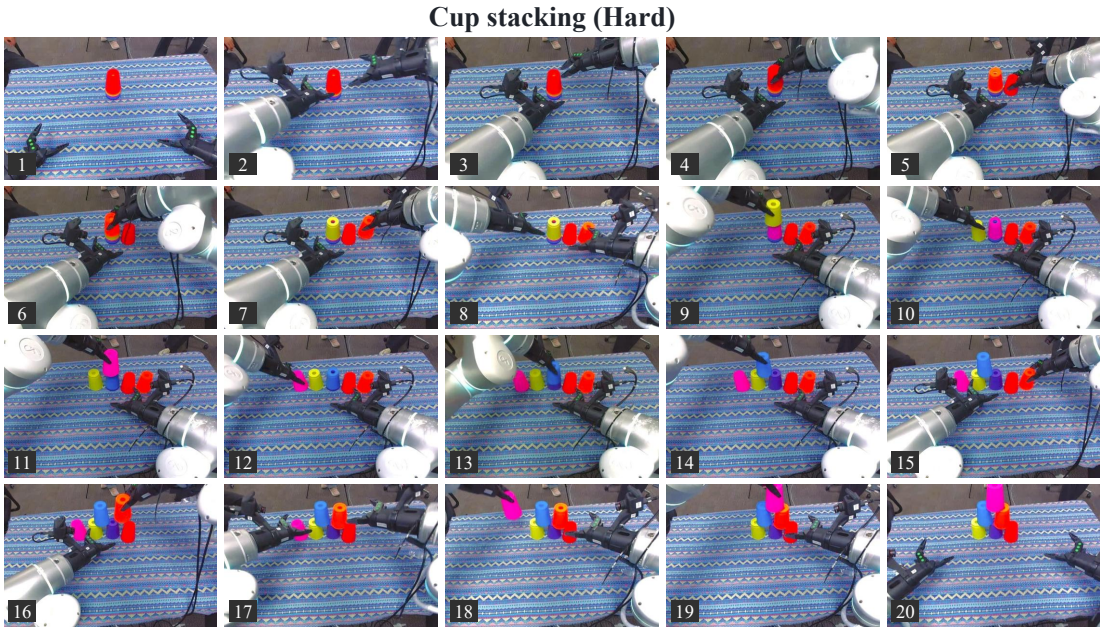


Figure 14. **Cup stacking.** The robot builds a three-tier pyramid from nested cups.

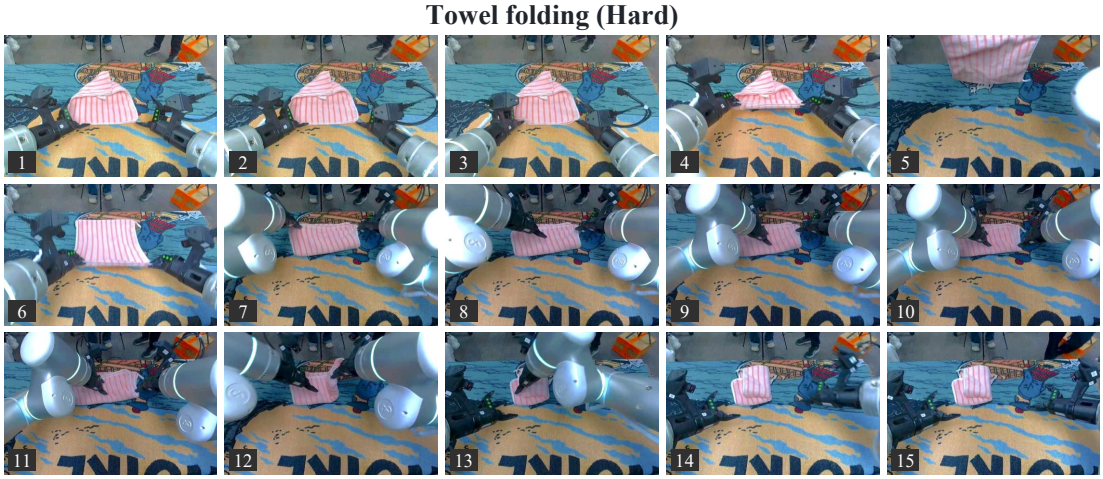


Figure 15. **Towel folding.** The robot spreads the towel and folds it twice.

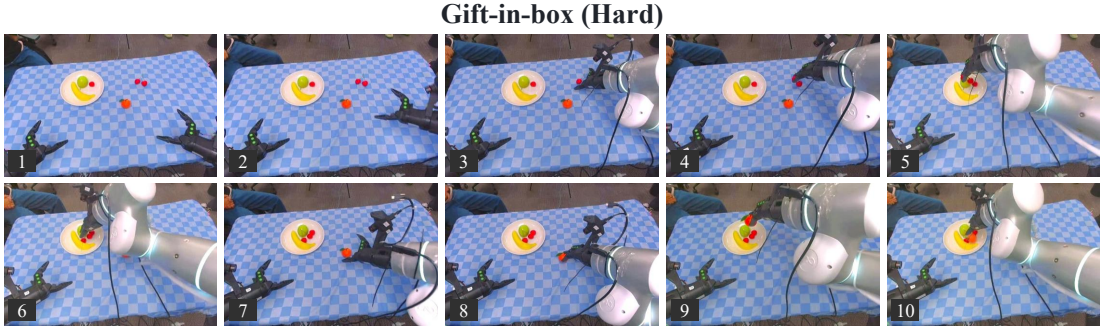


Figure 16. **Gift-in-box.** The robot places the gift in the box and releases it.