Ropedia W0

Scaling Loco-Manipulation World Models across Embodiments with Human Experience

Tianqi Liu[1,2,*] **Hanfei Zhao**[2,3,*] **Qifeng Liang**[2,4,*] **Hui En Pang**[1,2] **Jiarui Xu**[1,2]
Sayang Mu[1,2] **Penghao Qiu**[1] **YingSheng Zhang**[5] **Jun Wang**[2] **Ziang Cao**[1,2]
Mingze Sun[2,5] **Yukang Cao**[1] **Fangzhou Hong**[1,2] **Zhaoxi Chen**[1,2] **Ziwei Liu**[1,2,†]

[1]Nanyang Technological University [2]Ropedia [3]Duke University
[4]University of California, San Diego [5]Tsinghua University

https://ropedia.github.io/ropedia-w0/

arXiv:submit/8248267 [cs.CV] 18 Oct 2026

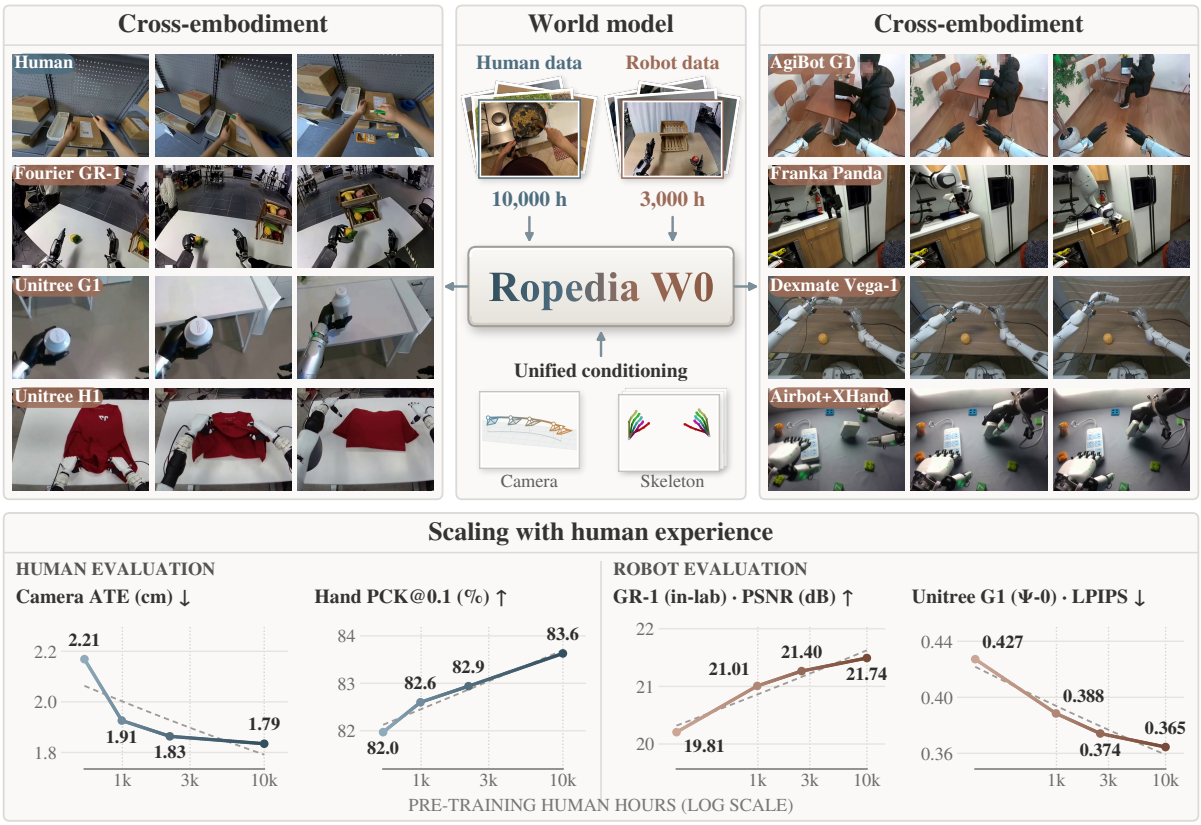


Figure 1 | **Ropedia W0: a unified egocentric world model for cross-embodiment loco-manipulation.** Trained on Xp-Data (14 human and 13 robot public datasets plus in-house recordings; ∼10k h human and ∼3k h robot video), a single shared interface of camera poses and projected kinematic skeletons (*center*) lets one model predict future egocentric video across humans, gripper robots, and dexterous humanoids (*left and right*). Prediction quality on both human and robot evaluations improves as the amount of human-video pre-training is scaled up (*bottom*).

∗ Equal contribution. † Corresponding author.

Abstract

Ropedia

World models enable embodied agents to anticipate action consequences, yet jointly modeling locomotion and manipulation across diverse embodiments remains challenging due to heterogeneous action spaces and limited multimodal supervision. We present **Ropedia W0**, the first egocentric loco-manipulation world foundation model spanning humans and robots with diverse embodiments. We curate **Xp-Data**, comprising 10k hours of egocentric human data and 3k hours of robot data with unified multimodal annotations and systematic quality control. A unified kinematic representation encodes locomotion through camera poses and manipulation through projected skeletons, with robot skeletons combining embodiment-specific forward kinematics with extensive manual keypoint annotation and correction. We also introduce **Xp-Teleop**, a long-horizon loco-manipulation benchmark based on high-quality dexterous humanoid teleoperation data. Evaluations demonstrate state-of-the-art performance across five cross-embodiment benchmarks (DreamDojo, Ψ_0, ω-HOME, Xp-Teleop, and AgiBot) and on held-out human egocentric video, from zero-shot inference to few-shot and full-data post-training. Scaling human pre-training data improves performance on both human and robot evaluations. After distillation, our model supports real-time, minute-long rollouts, enabling live VR interaction. Applications in policy evaluation, model-based planning, and world–action modeling further demonstrate the model's utility for embodied decision-making.

1. Introduction

World models predict how environments evolve in response to actions [42, 14] and support decision-making by imagining future outcomes [43], giving them considerable potential in embodied intelligence. However, most existing models [119, 122, 4] focus on fixed-base, gripper-based manipulation, leaving locomotion-induced viewpoint changes and dexterous hand–object interactions largely underexplored. Moreover, generalizing across embodiments poses a further challenge. DreamDojo [34] replaces latent-action conditioning with robot-specific actions during post-training, and Cosmos3 [1] relies on domain-specific action projections despite a shared backbone; both therefore require a dedicated control interface or additional adaptation for each new embodiment. Therefore, jointly modeling locomotion and manipulation while seamlessly generalizing across diverse embodiments remains an open problem.

Learning a generalizable world model hinges on large-scale, high-quality egocentric video paired with rich multimodal annotations, such as camera trajectories, hand motion, and task descriptions. While robot demonstrations across diverse tasks and embodiments remain expensive to collect, egocentric human video offers an abundant and complementary source of experience on how humans navigate and interact with their surroundings. However, existing public egocentric datasets [39, 40, 46, 113, 27, 2, 78] frequently suffer from incomplete or unreliable annotations. Specifically, many lack metric camera trajectories, provide noisy or missing hand poses, and adopt heterogeneous formats that hinder joint training. Furthermore, the scale of these datasets remains limited, falling short of the data required to train a generalizable world model.

To address these challenges, we present **Ropedia W0**, a unified egocentric world model that scales learning from human and robot videos for cross-embodiment loco-manipulation. We

curate **Xp-Data**, comprising 10k hours of egocentric human data and 3k hours of robot data, with unified geometric, kinematic, and semantic annotations produced through automated processing, manual robot-keypoint annotation, and quality checks. To connect human motion and robot actions, we introduce a shared explicit kinematic representation that combines metric camera poses for locomotion-induced viewpoint changes with projected skeletons for manipulation. The skeletons combine human pose annotations with robot forward kinematics (FK) and manually annotated or corrected robot keypoints, preserving embodiment-specific structure in an image-aligned format. This conditioning interface remains consistent across human pre-training, human–robot mid-training, and optional post-training. We further introduce **Xp-Teleop**, a humanoid teleoperation benchmark for evaluating long-horizon prediction during locomotion and dexterous manipulation.

We evaluate Ropedia W0 on five cross-embodiment benchmarks (DreamDojo, Ψ_0, ω-HOME, Xp-Teleop, and AgiBot), alongside held-out human egocentric video. Results demonstrate strong predictive performance across embodiments and post-training budgets, including zero-shot transfer and data-efficient adaptation. Scaling analyses show that increasing the amount of human pre-training data improves both prediction on human videos and transfer to robots. Ablations establish the complementary roles of camera and skeleton conditioning, while distillation enables real-time, minute-long rollouts and live VR interaction across embodiments. Downstream experiments further demonstrate accurate policy ranking, improved trajectory planning, and more robust action learning in simulation and on real robots.

Our contributions are threefold:

- We introduce **Ropedia W0**, which, to the best of our knowledge, is the first unified egocentric world model for cross-embodiment loco-manipulation. A shared explicit kinematic representation of camera poses and projected skeletons enables scalable learning from human and robot videos.
- We design a human-in-the-loop pipeline that unifies collection, annotation, and quality checks, including extensive manual robot-keypoint annotation and correction of inaccurate FK projections. Consequently, we build **Xp-Data**, comprising over 10k hours of human data and 3k hours of robot data in a shared schema.
- Extensive experiments show that Ropedia W0 achieves state-of-the-art cross-embodiment prediction, including on our long-horizon loco-manipulation **Xp-Teleop** benchmark. We further show that Ropedia W0's performance scales with human data and demonstrate applications in real-time interaction, policy evaluation, model-based planning, and world–action modeling.

2. Related Work

Robotic video world models. Early world models learned compact latent dynamics for control [42, 43]; large video generators yield interactive worlds [14, 88, 89, 75, 24, 74] and ego-motion-controlled driving simulators [47, 98, 35]. Distillation and self-rollout training convert video generators into causal few-step models [130, 50, 26, 72], augmented with long-horizon memory [45, 100]. Robotic world models learn video dynamics from robot trajectories [124, 142, 55, 67, 41, 139, 122, 44], often adapting a video foundation model [3, 108]. Numerical action tokens require embodiment-specific encoding [111, 41]; latent actions inferred from video [14, 128, 36, 38, 34] make unlabeled footage usable, but may entangle actions with other visual changes [133]; spatially aligned renderings provide explicit geometric conditioning [55, 122, 126, 115, 119]. Cross-embodiment models use embodiment-specific action heads,

projections, or adaptation [111, 34, 1], a single rendered interface [115, 119], or factorized human-to-robot recomposition [22]. The last, AnyWorld [22], recomposes human interactions into robot experiences but conditions on source human skeletons rather than target-robot kinematics and is evaluated mainly on manipulation, leaving sustained loco-manipulation untested. Ego-motion is also studied in navigation and driving [8, 47] and in egocentric predictors conditioned on or jointly modeling full-body pose [7, 87, 121]. Ropedia W0 uses an explicit kinematic interface to unify loco-manipulation across humans, gripper robots, and dexterous humanoids.

Human and robot datasets. Large egocentric human collections cover daily and skilled activity [39, 27, 40]; pose-annotated collections add body or hand pose [82, 78, 46, 92], and portable rigs bring hand pose into everyday scenes [109, 76]. Aggregated and in-the-wild demonstrations [86, 58, 107, 30, 118] and large teleoperation sets [15] provide robot actions, calibrated views and task captions in varying combinations, with substantial coverage of tabletop arm and gripper manipulation; humanoid sets include locomotion or dexterous hands [138, 101, 94, 117, 65]. Evaluation data include held-out robot video clips [142, 34, 119], multimodal world-model benchmarks [1], and benchmarks measuring visual quality, physical consistency or embodied utility [132, 84]. Xp-Data unifies annotations for public human and robot data and self-collected egocentric video; Xp-Teleop benchmarks long-horizon dexterous humanoid loco-manipulation with robot-native kinematics.

Policy evaluation, planning, and learning. Policy evaluation extends real-to-sim protocols [62] to video world models through recorded-action rollouts [142, 64, 119] or closed-loop policy interaction [95, 41, 103, 102]. For planning, world models score or refine action proposals using predicted outcomes [93, 28, 53, 34], in pixel or learned latent spaces [140, 5], and support reinforcement fine-tuning and synthetic data generation [143, 56, 61, 120, 54]. Vision-language-action models map observations and language to actions using pretrained vision-language models [13, 59, 12, 52, 11], increasingly trained with human video [128, 16, 15, 123, 80, 57, 117], whereas world–action models couple future video or latent-state prediction with action learning [21, 141, 48, 111, 67, 127, 1, 9, 81, 65, 112, 131]. Ropedia W0 supports these applications in Sec. 6; for world–action modeling, we transfer its pretrained video backbone to initialize such a model rather than adding a dedicated action head.

3. Data

Learning cross-embodiment loco-manipulation requires data that captures both viewpoint changes and manipulator motion under a shared annotation scheme. However, existing datasets [39, 40, 46, 113, 27, 2, 78, 31, 29, 60, 77, 37, 92, 97, 15, 58, 30, 135, 101, 11, 33, 10, 63, 85, 49, 105, 138] fall short of this requirement: most lack essential geometric information such as camera poses, and their hand annotations are often noisy or inaccurate, making them insufficient for training a general-purpose world model.

To address this gap, our **human-in-the-loop data pipeline** unifies collection, multimodal annotation, and quality checks (Fig. 2). We re-annotate public human and robot recordings, with substantial manual robot-keypoint annotation and correction, and collect over 6,000 hours of in-house multi-view video with a four-fisheye-camera rig. This yields Xp-Data, over 10k human and 3k robot training hours with synchronized RGB, camera trajectories, manipulator skeletons, and captions; wrist poses and wide-FOV skeletons are included where available. Our Xp-Teleop benchmark provides about one held-out hour of long-horizon dexterous humanoid loco-manipulation, recorded in the same schema with an off-the-shelf rig [104] (Sec. A.2).

Throughout, an *episode* (robot) or *clip* (human) is a contiguous source recording, a *window* is an

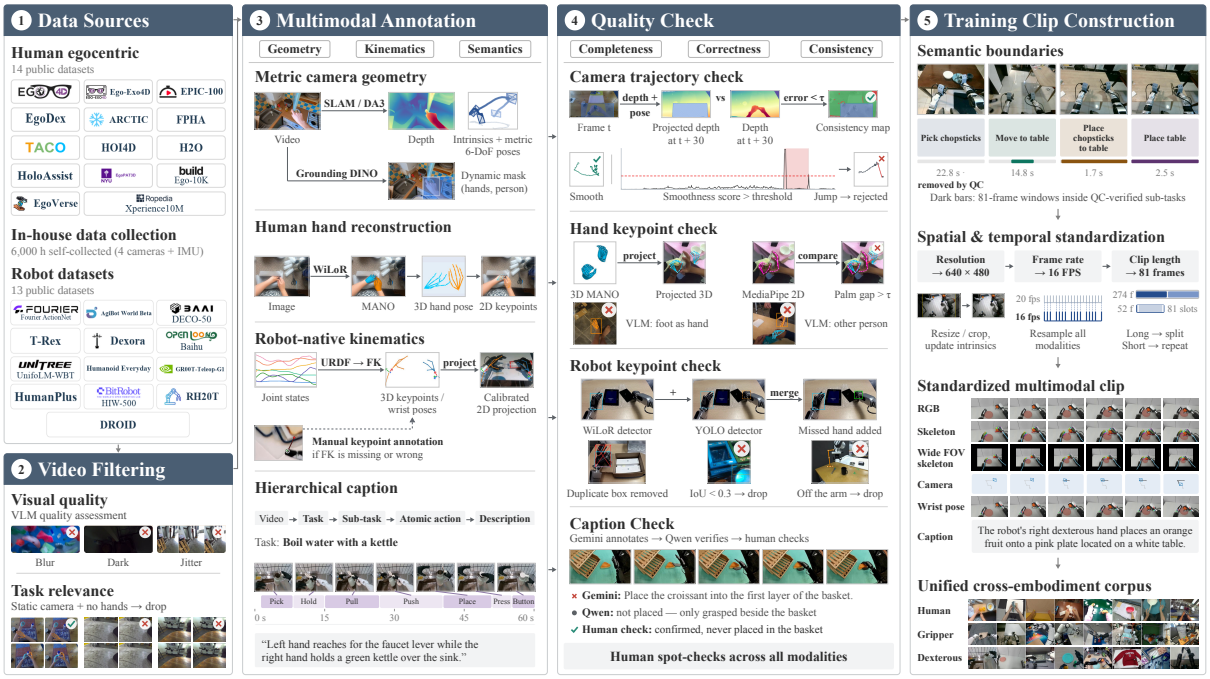


Figure 2 | **Xp-Data curation pipeline.** Heterogeneous human and robot recordings pass through five stages. (1) *Data sources*: 14 public human egocentric datasets, 6,000 h of in-house four-camera recordings, and 13 public robot datasets. (2) *Video filtering* removes low-quality or task-irrelevant clips with a vision-language model. (3) *Multimodal annotation* produces geometry (metric camera poses via SLAM and DA3 [69]), kinematics (human hands via WiLoR [91] and MANO [96], robot keypoints via URDF forward kinematics, manual annotation, and correction), and semantics (hierarchical task, subtask, and atomic-action captions). (4) *Quality check* verifies completeness, correctness, and cross-modal consistency, pairing automated checks with a caption cascade (Gemini [25] annotation, Qwen [6] verification, and human review). (5) *Training-clip construction* cuts quality-verified subtasks into standardized 81-frame, 640 × 480, 16 fps multimodal clips, yielding the unified cross-embodiment corpus.

81-frame training sample extracted from that recording, and a *source* is one (dataset, end-effector) stream. AgiBot World Beta therefore contributes two sources: a parallel-gripper stream and a dexterous-hand head-camera stream.

Data sources. Xp-Data comprises human and robot data. The human collection combines 14 public egocentric datasets [39, 40, 46, 113, 27, 2, 78, 31, 29, 60, 77, 37, 92, 97] with over 6,000 hours of in-house recordings captured using a device equipped with four fisheye cameras, covering diverse daily activities and object interactions. The robot collection draws from 13 public datasets [15, 58, 30, 135, 101, 11, 33, 10, 63, 85, 49, 105, 138] spanning parallel grippers and dexterous hands. Together, these collections provide complementary coverage of activities, camera motion, and manipulator kinematics.

Video filtering. We filter the raw recordings for visual quality and task relevance before annotation. A vision-language model screens each segment for severe blur, poor or uneven illumination, over- and under-exposure, and excessive camera jitter, removing clips unsuitable for reliable geometric and kinematic annotation. We discard segments with neither camera motion nor visible hands.

Multimodal annotation. We annotate geometry, kinematics, and semantics in a unified format. For geometry, we obtain camera intrinsics and metric trajectories from available calibration or estimate them using our in-house SLAM pipeline or DA3 [69]. Grounding DINO [73] identifies

dynamic regions, such as hands and people, for exclusion from camera-trajectory consistency checks. For human kinematics, we project WiLoR-reconstructed [91] MANO [96] keypoints into images. Captions describe hands, objects, and interactions at task, subtask, and atomic-action levels.

Robot keypoint annotation and correction. Direct FK projections often misalign with the hands and grippers in the public robot recordings we process, and many releases lack reliable camera calibration. We therefore invest substantial effort in manually annotating image-space keypoints and correcting inaccurate projections. FK provides initial estimates where joint states and calibration are available; manual annotations supply or refine the image-aligned robot skeletons used for training. Wrist poses are derived through FK where available.

Quality checks. We verify annotation completeness, correctness, and cross-modal consistency. Camera trajectories are checked for depth-reprojection consistency and temporal smoothness, following [110]. Human hand annotations are checked both geometrically and semantically: projected 3D keypoints are compared against independent 2D detections from MediaPipe [79] and a fine-tuned MMPose [83] model, while a vision-language model flags false detections and hands belonging to other people. Robot keypoints are cross-checked by two detectors fine-tuned on our manually annotated robot-hand data, which recover missed hands and remove duplicate or misaligned detections. Qwen [6] verifies Gemini [25] captions, with human review resolving discrepancies. Human spot checks across modalities identify systematic annotation errors.

Training clip construction. From subtasks that pass quality checks, we build 81-frame clips at 640×480 (with adjusted intrinsics), each uniformly resampled from a contiguous source interval. A language model [25] merges the keyframe captions within a window into one window-level caption. Each clip provides synchronized RGB, projected skeletons, camera parameters, and a caption (a default caption is used for the small fraction of robot windows that lack a generated one), plus wide-FOV skeletons and wrist poses where available, in a format shared across human and robot data.

The overall curation funnel is summarized in Fig. S1 (Appendix A.1). Fig. 3 summarizes the composition of Xp-Data: the most frequent caption terms, the distribution of environments and tasks for the robot and human portions, and representative unified skeleton annotations across human hands, dexterous robot hands, and parallel grippers. Together they show that the corpus spans diverse scenes and activities while sharing a single annotation format across embodiments.

4. Model

Ropedia W0 combines a unified kinematic control interface with a video diffusion model for cross-embodiment egocentric loco-manipulation prediction (Fig. 4). The interface jointly captures viewpoint changes and manipulation through camera poses and projected skeletons, allowing human and robot motion to share a common visual format while preserving embodiment-specific geometry. We first formulate the prediction task and introduce a shared representation of camera and manipulator motion (Sec. 4.1). We then describe the conditioning architecture (Sec. 4.2), training recipe (Sec. 4.3), and distillation (Sec. 4.4).

4.1. Representation: Unified Kinematic Control

Egocentric loco-manipulation requires jointly modeling locomotion-induced viewpoint changes and manipulator motion. We encode these complementary factors through camera

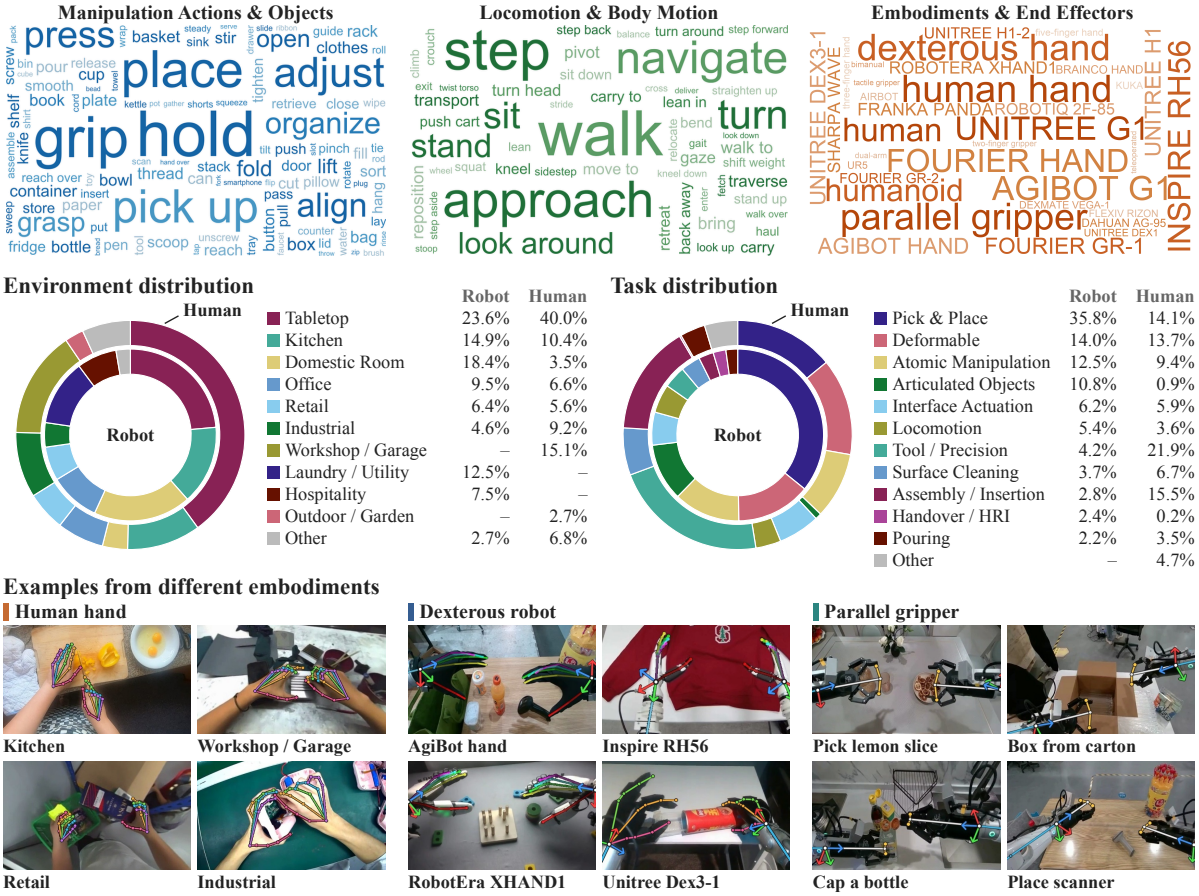


Figure 3 | **Overview of Xp-Data.** Top: caption terms. Middle: environment and task distributions for robot (inner ring) and human (outer ring) data. Bottom: unified skeleton annotations across human hands, dexterous robot hands, and parallel grippers.

poses and projected kinematic skeletons, providing explicit viewpoint control and spatially aligned manipulation conditioning. Across embodiments, joint-space actions differ in dimensionality and semantics, whereas native skeletons can be rendered in a common image-space format. Together, these conditions provide a unified interface for humans, grippers, and dexterous humanoids while preserving embodiment-specific geometry.

Given an initial image I_0, a text condition c_{txt}, and a per-frame kinematic condition sequence $c_{1:T}$, Ropedia W0 predicts future egocentric frames:

$$\hat{I}_{1:T} \sim p_\theta(I_{1:T} \mid I_0, c_{\text{txt}}, c_{1:T}), \qquad c_t = \left(c_t^{\text{cam}}, c_t^{\text{skel}}\right), \tag{1}$$

where c_{txt} denotes the encoded text prompt, and c_t^{cam} and c_t^{skel} denote the camera-ray map and projected skeleton rendering, respectively.

Camera motion for locomotion. We express camera poses relative to the initial frame and encode them as dense Plücker-ray maps. Let $\mathbf{R}_t$ map directions from the current camera frame to the initial camera frame, and let $\mathbf{o}_t$ denote the camera center in this reference frame. Given camera intrinsics $\mathbf{K}_t$ and a homogeneous pixel coordinate $\tilde{\mathbf{u}}$, we compute

$$\mathbf{d}_t(\mathbf{u}) = \frac{\mathbf{R}_t \mathbf{K}_t^{-1} \tilde{\mathbf{u}}}{\left\| \mathbf{R}_t \mathbf{K}_t^{-1} \tilde{\mathbf{u}} \right\|_2}, \qquad c_t^{\text{cam}}(\mathbf{u}) = \left[\mathbf{o}_t \times \mathbf{d}_t(\mathbf{u}) \,;\, \mathbf{d}_t(\mathbf{u})\right]. \tag{2}$$

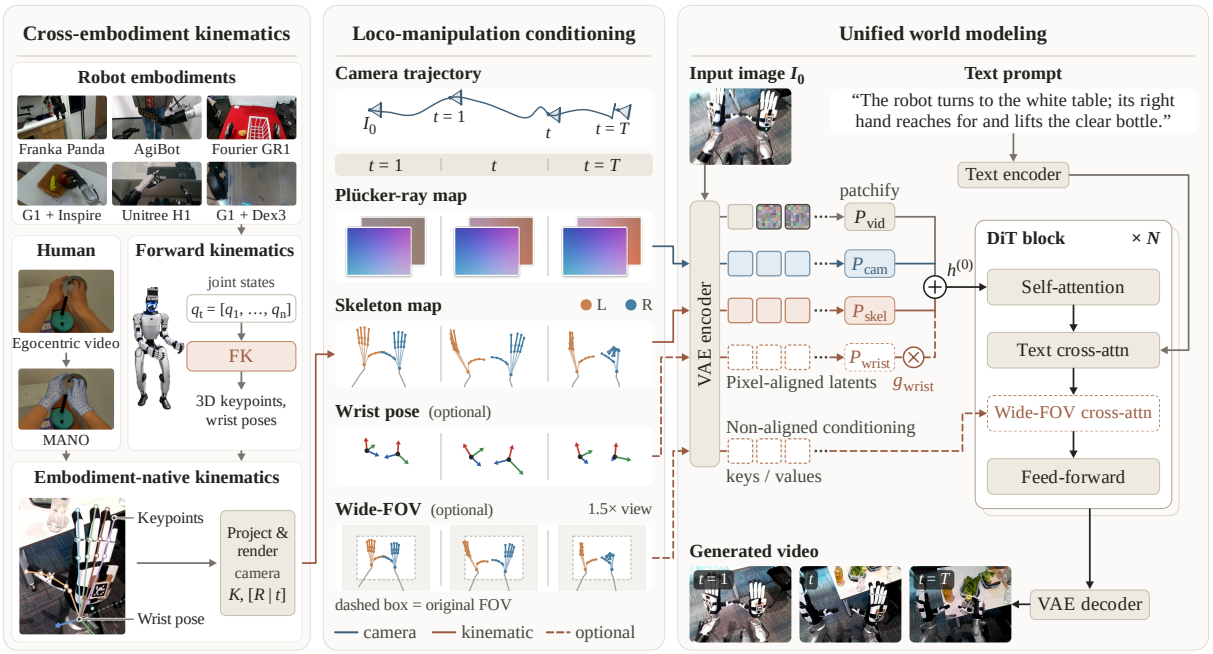


Figure 4 | **Overview of Ropedia W0. Left:** Human pose annotations and robot forward kinematics provide kinematic structure and wrist poses; manual robot-keypoint annotation and correction align robot skeletons with the observed images. **Middle:** Camera trajectories are encoded as Plücker-ray maps, while manipulation is represented by skeleton maps, optionally augmented with wrist-pose and wide-FOV renderings. **Right:** A video diffusion transformer predicts future frames from an initial image, a text prompt, and these control signals. Pixel-aligned conditions are fused through additive projections, while wide-FOV features are injected via cross-attention.

This six-channel encoding associates camera geometry with each image location, providing spatially aligned conditioning for video generation. Camera translations are expressed in metric units, maintaining a consistent physical scale across embodiments.

Kinematic skeletons for manipulation. Human hand keypoints are obtained from pose annotations. For curated robot videos, we combine embodiment-native URDF forward kinematics with manual image-space keypoint annotation and correction (Sec. 3). Where joint states and calibration are available, the initial FK projection of keypoint j on robot embodiment e is

$$\tilde{\mathbf{u}}_{t,j}^{(e)} = \pi\left[\mathbf{K}_t\left(\mathbf{R}_t^{\mathrm{c}\leftarrow\mathrm{b}}\mathcal{F}_j^{(e)}(\mathbf{q}_t) + \mathbf{t}_t^{\mathrm{c}\leftarrow\mathrm{b}}\right)\right], \tag{3}$$

where $\mathcal{F}_j^{(e)}$ maps joint states $\mathbf{q}_t$ to a keypoint position in the robot base frame, $(\mathbf{R}_t^{\mathrm{c}\leftarrow\mathrm{b}}, \mathbf{t}_t^{\mathrm{c}\leftarrow\mathrm{b}})$ transforms it into the current camera frame, and π denotes perspective division. We render the resulting image-aligned keypoints with their native connectivity into c_t^{skel}, preserving gripper opening geometry and dexterous finger configurations. Projection into the camera view at each timestep captures how these configurations appear under simultaneous camera and manipulator motion.

When available, wrist-pose renderings c_t^{wrist} and wide-FOV skeletons c_t^{wide} supplement the core representation with explicit wrist orientation and kinematic context outside the target field of view.

4.2. Architecture: Kinematics-Conditioned World Model

Ropedia W0 adopts a latent video diffusion transformer conditioned on the initial observation, text, and kinematic controls. We instantiate it at two scales: a dense 5B model and an A14B mixture-of-experts (MoE) model based on Wan2.2-I2V-A14B [108] (approximately 27B total and 14B active parameters). Spatially aligned conditions are added to the video tokens, while additional context is incorporated through cross-attention.

Spatially aligned conditioning. We encode camera-ray maps and skeleton renderings into condition latents z^{cam} and z^{skel}, aligned with the video latents in space and time. Separate projections map these latents to the transformer's token dimension:

$$h^{(0)} = P_{\text{vid}}(z_\tau) + P_{\text{cam}}(z^{\text{cam}}) + P_{\text{skel}}(z^{\text{skel}}), \tag{4}$$

where z_τ denotes the video latents at noise level τ, and P_{vid} is the pretrained video projection. The camera and skeleton projections are initialized to zero, preserving the backbone's behavior at initialization. Text embeddings c_{txt} enter through the backbone's cross-attention layers.

Additional kinematic context. Wrist-pose renderings follow the same aligned injection scheme, adding $g_{\text{wrist}} P_{\text{wrist}}(z^{\text{wrist}})$ to $h^{(0)}$, where $g_{\text{wrist}} \in \{0, 1\}$ indicates annotation availability. Wide-FOV skeletons extend beyond the target view and lack direct pixelwise correspondence with the video. Their encoded features therefore enter through an additional cross-attention branch. We initialize the wrist projection and the wide-FOV branch's output projection to zero, and disable each optional branch when its annotations are unavailable.

4.3. Training: From Human Experience to Cross-Embodiment

We train Ropedia W0 in two stages: human pre-training and human–robot mid-training. Both stages share the camera-and-skeleton conditioning interface, enabling the model to transfer motion-conditioned prediction learned from human videos to diverse robot embodiments.

Human pre-training. We first train on large-scale egocentric human videos conditioned on text, camera trajectories, and projected hand skeletons. This stage learns the visual consequences of camera and manipulator motion across diverse environments and interactions, establishing a foundation for cross-embodiment transfer.

Human–robot mid-training. We continue training on a balanced mixture of human and robot data, sampling each with equal probability. Robot motion is expressed through the same conditioning interface using embodiment-native kinematics. This stage introduces robot-specific appearance, manipulator geometry, and interaction patterns while maintaining exposure to diverse human experience. We also incorporate wrist-pose and wide-FOV skeleton conditions during this stage, using availability gates to accommodate samples with different annotation coverage.

4.4. Distillation: Real-Time Interactive Rollouts

We first obtain a block-causal, 4-step autoregressive generator using Causal Forcing [144, 137], then extend it to long-horizon streaming generation using LongLive [125], retaining image, camera, skeleton, and text conditioning. Causal initialization produces the few-step generator; subsequent asymmetric distribution-matching distillation (DMD) contrasts the clean-latent predictions of a frozen bidirectional real denoiser μ_ψ (with classifier-free guidance) and an online

Table 1 | **Comparison of humanoid teleoperation datasets** across capabilities (locomotion and dexterity), task horizon (duration, locomotion range, and number of subtasks), and available recording modalities (RGB frame rate, depth, IMU, whole-body joints, tactile, calibration, and keypoints). A check vs. cross marks whether each property is provided; datasets are grouped by whether the lower body is used during collection (tabletop vs. whole-body). Xp-Teleop is the only one offering synchronized long-horizon loco-manipulation with the full modality set.

Dataset	Robot / hand	Capabilities		Task horizon			Modalities						
		Loco.	Dex.	Duration (s)	Loco. range (m)	Subtasks	RGB@fps	Dpt.	IMU	WB joints	Tac.	Cal.	Kpts.
Tabletop teleoperation													
RoboMIND (Tien Kung)	Tien Kung / Inspire	×	✓	21.7	0	2.2	640×480@30	✓	×	×	×	×	×
GR00T-Teleop-GR1	GR-1 / Fourier	×	✓	14.3	0.21	1.0	1280×800@20	×	×	✓	×	×	×
Unitree G1 Dex3	G1 / Dex3	×	✓	27.4	0	1.0	640×480@30	×	×	×	×	×	×
Unitree G1 Dex1	G1 / Dex1 gripper	×	×	28.7	0	1.1	640×480@30	×	×	✓	×	×	×
PH2D (robot)	H1 / Inspire	×	✓	12.6	0	2.6	1280×720@30	×	×	×	×	×	×
Loco-manipulation teleoperation (legs commanded)													
Humanoid Everyday	G1 / Dex3; H1 / Inspire	✓	✓	20.4	0.84	2.0	640×480@30	✓	✓	✓	✓	×	×
Ψ_0 (real)	G1 / Dex3	✓	✓	40.1	0.40	2.6	640×480@30	×	×	×	×	×	×
ω-HOME (ω_0)	G1 / Inspire	✓	✓	30.4	0.33	1.1	640×360@30	×	×	✓	×	×	×
UnifoLM-WBT (Inspire)	G1 / Inspire	✓	✓	39.8	0.45	1.7	640×480@30	×	×	✓	×	×	×
UnifoLM-WBT (Dex1)	G1 / Dex1 gripper	✓	×	39.8	0.45	1.7	640×480@30	×	×	✓	×	×	×
Xp-Teleop (ours)	G1 / Inspire	✓	✓	118.6	1.48	3.6	1280×720@30	✓	✓	✓	✓	✓	✓

bidirectional fake denoiser μ_ϕ trained on the generator's samples:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = \mathbb{E}_{\tau,\epsilon}\left[\frac{\mu_\phi(z_\tau,\tau) - \mu_\psi(z_\tau,\tau)}{\|z_0 - \mu_\psi(z_\tau,\tau)\|_1}\frac{\partial z_0}{\partial \theta}\right]. \tag{5}$$

Here z_0 is the student output and z_τ its noised version; the update is normalized per sample by the real denoiser's correction magnitude [129], and we apply several critic updates per generator update. LongLive subsequently adapts the few-step generator for long-video streaming with a bounded KV cache (a local attention window plus an attention sink). The student streams fixed-size latent blocks at 21.3 fps at 640 × 480 for minute-long interactive rollouts.

5. Xp-Teleop

Xp-Teleop is a humanoid teleoperation benchmark for evaluating long-horizon prediction during locomotion and dexterous manipulation. It comprises 27 long-horizon takes (about one hour of egocentric video at 30 fps) recorded on a Unitree G1 humanoid with two six-DoF dexterous hands through a whole-body teleoperation rig assembled from off-the-shelf hardware and Unitree's open-source framework [104]. Two operators drive locomotion and manipulation concurrently, so walking and dexterous interaction occur within the same take. Every take is recorded with synchronized RGB-D, camera poses, projected skeletons, robot states, language, tactile, and head-IMU signals in the same schema as Xp-Data, and is divided into 543 human-annotated atomic segments spanning 30 skill classes (median 18 segments per take). We keep failed grasps and recoveries so that the benchmark also probes whether a world model can simulate off-nominal outcomes, not only successful executions.

Fig. 5 overviews the activities, representative episodes, synchronized modalities, and action labels, and Table 1 compares Xp-Teleop with other humanoid teleoperation datasets in task coverage and available modalities. Full rig, control, recording, and comparison details are provided in Sec. A.2.

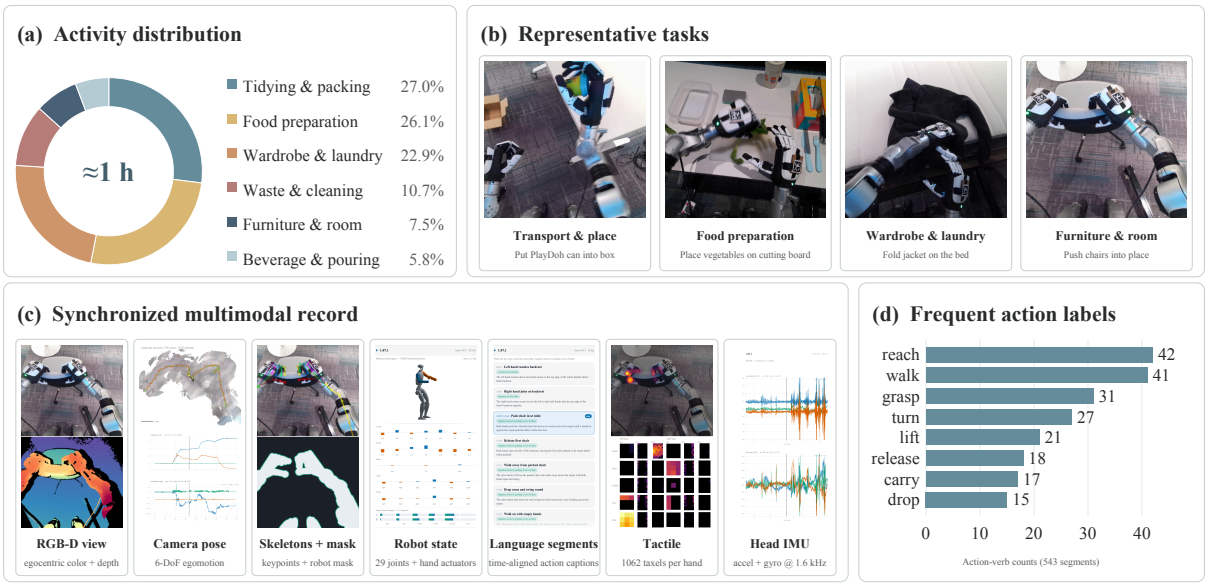


Figure 5 | **Overview of Xp-Teleop.** About one hour of dexterous-humanoid teleoperation spanning six everyday household activity groups, from tidying and food preparation to laundry and cleaning (a). Each take is a long-horizon episode that chains locomotion with manipulation, such as putting objects into a box or folding a jacket on the bed (b). Takes are recorded as a single synchronized multimodal stream (egocentric RGB-D, 6-DoF camera pose, projected hand/robot skeletons and masks, 29-joint robot state, time-aligned language segments, per-hand tactile, and head IMU; c), and their 543 annotated segments are dominated by loco-manipulation verbs such as *reach*, *walk*, and *grasp* (d).

6. Experiments

6.1. Setup

Benchmarks. We evaluate action-conditioned video prediction on five cross-embodiment benchmarks spanning distinct embodiments and data sources: (i) the DreamDojo benchmarks [34], three Fourier GR-1 robot evaluation sets from DreamDojo: In-lab (100 clips), EgoDex (65 clips), and DreamDojo-HV (68 clips); (ii) Ψ_0 [117], 100 evaluation windows of dexterous humanoid manipulation recorded with the Ψ_0 vision-language-action (VLA) system (50 seen-task, 50 unseen-task); (iii) ω-HOME [65], 100 evaluation windows of household loco-manipulation on a Unitree G1 humanoid with dexterous five-finger hands across 24 tasks (50 seen, 50 from five held-out tasks); (iv) Xp-Teleop (Sec. 5), 100 evaluation windows of long-horizon dexterous humanoid teleoperation, split by scene (50 windows from held-out takes of training scenes, 50 from unseen scenes); and (v) AgiBot [15], 50 held-out windows of gripper-based manipulation, where all compared models include AgiBot data in training. In addition, we report zero-shot prediction on 200 held-out human egocentric clips (EgoSuite and OpenAoE; Sec. 6.2).

Metrics. We report visual quality with PSNR, SSIM [116], and LPIPS [134]. For benchmarks with camera motion we additionally evaluate camera-trajectory accuracy on poses recovered from the generated video, namely absolute trajectory error (ATE), relative pose error (RPE), and relative rotation error (RRE), as well as hand-following accuracy via keypoint PCK@0.1 (PCKm on matched hands, with detection rate Det). Where noted, we also report video-quality metrics (VBench [51] subject/background consistency, dynamic degree, and aesthetic quality) and FVD [106]. Full metric definitions and the per-benchmark evaluation protocol are given in Sec. C.1.

Post-training protocol. For each benchmark, we evaluate three adaptation settings: *no post-training* (direct transfer), *few-shot* post-training using a small number of episodes per training task, and *full-data* post-training using the entire training split. Throughout, *zero-shot* (equivalently *no post-training* / direct transfer) means the model is evaluated on a target dataset it was never trained on. The target *embodiment* may still appear in the multi-embodiment mid-training mixture through *other* datasets, but the target evaluation dataset itself is never used for training. "Seen" vs. "unseen" splits within a benchmark refer to seen vs. held-out *tasks/scenes* of that benchmark, not to the embodiment. We evaluate both training stages of Ropedia W0: *pretrain* rows are initialized from the human-only pre-training stage, and *midtrain* rows from the subsequent mid-training stage that mixes human and multi-embodiment robot data. Checkpoints are always selected on a held-out validation split, never on the benchmark itself.

Baselines. We compare against generalist robot world models with different action interfaces: DreamDojo (2B and 14B), which conditions on continuous latent actions pretrained from large-scale human video, and Cosmos3 (Nano, 16B; Super, 64B), which treats action as a first-class modality with per-domain action encoders. On the egocentric human-video and AgiBot benchmarks we additionally report two recent action-conditioned world models where available, OSCAR [119] and RynnWorld-Teleop [136]. For data-efficiency comparisons, we additionally post-train the same architecture as ours initialized from the public video-generation weights ("official base"), isolating the contribution of our pre-training. All baselines are post-trained with their official recipes.

6.2. Benchmark Results

Human egocentric video. We evaluate on 200 clips from EgoSuite [68] and OpenAoE [66], two datasets unseen during training by both Ropedia W0 and all compared methods. Ropedia W0 achieves the lowest camera-trajectory errors, the highest hand-following accuracy, and the best visual quality (PSNR, SSIM, LPIPS) (Table 2). These results demonstrate strong generalization to unseen human egocentric datasets, in both motion following and visual quality.

DreamDojo benchmark. Following DreamDojo's protocol, we post-train on its GR-1 teleoperation data and evaluate on three test sets: In-lab, EgoDex, and DreamDojo-HV. Ropedia W0 outperforms DreamDojo across all evaluated post-training budgets (1%–100%), with the largest gains in the low-data regime, demonstrating data-efficient transfer to the target embodiment (Table 3).

Ψ_0 and ω-HOME. On these dexterous-humanoid benchmarks (Table 4; breakdowns in Table S3), Ropedia W0 leads across all adaptation settings. On both, an A14B midtrain variant already matches or exceeds every baseline's full-data result with little or no post-training. Without post-training, the A14B midtrain variant also yields the smallest seen–unseen gap on both benchmarks (0.06/0.69 dB on Ψ_0/ω-HOME); this narrowing does not persist after few-shot or full-data post-training (Table S3). Qualitative comparisons on ω-HOME (Figs. S13 and S14) illustrate the generated hand motion and object interactions against ground-truth frames.

Xp-Teleop. On long-horizon loco-manipulation split into seen and unseen scenes, A14B midtrain leads all three settings over the best baseline, and its zero-shot result exceeds every baseline's full-data result (Table 4). Post-training lifts unseen scenes in step with seen ones (Table S3), consistent with adaptation rather than pure memorization. Ropedia W0 also leads on long-horizon autoregressive rollouts ($\approx$32 s on ω-HOME and $\approx$30 s on Xp-Teleop), with full tables in the appendix (Tables S4 and S5).

AgiBot. All compared checkpoints include AgiBot World data; on held-out gripper-

Table 2 | **Results on a 200-clip EgoSuite [68] and OpenAoE [66] evaluation set** (100 clips each). Zero-shot egocentric human-video prediction, reported as (a) camera- and hand-action following (trajectory errors ATE/RPE/RRE, hand position error HandL2, and PCK-based hand-keypoint accuracy with detection rate Det) and (b) video prediction and generation quality (PSNR/SSIM/LPIPS, VBench [51] consistency/dynamics/aesthetics, and FVD [106]). ATE/RPE are in cm, RRE in degrees, HandL2 in pixels, and PSNR in dB; PCK/PCKm/Det and Cons./Dyn/Aes are reported as percentages. Best bold, second-best underlined; full protocol in Sec. C.1.

Method	(a) Camera and hand action following						
	ATE↓	RPE↓	RRE↓	HandL2↓	PCK@.1↑	PCKm@.1↑	Det↑
Cosmos3-Nano (16B)	1.931	0.170	1.584	46.9	78.94	84.03	92.08
Cosmos3-Super (64B)	2.031	0.178	2.045	48.3	77.89	83.32	91.83
OSCAR (2B)	3.480	0.273	6.258	42.1	77.60	84.46	90.55
RynnWorld-Teleop (5B)	3.543	0.274	6.109	49.1	74.67	82.37	88.78
DreamDojo (2B)	2.652	0.224	4.716	69.8	58.84	70.14	81.44
DreamDojo (14B)	2.652	0.227	3.730	65.9	60.25	72.61	80.00
Ropedia W0 (5B), pretrain	1.793	0.165	1.078	**32.7**	83.63	**88.66**	93.15
Ropedia W0 (5B), midtrain	1.884	0.170	1.131	33.2	**83.85**	88.56	**93.78**
Ropedia W0 (A14B), pretrain	**1.542**	**0.152**	0.980	36.0	81.27	86.85	92.20
Ropedia W0 (A14B), midtrain	1.750	0.164	**0.970**	36.6	81.38	86.67	92.62

Method	(b) Video prediction and generation quality							
	PSNR↑	SSIM↑	LPIPS↓	Cons.↑	Dyn↑	Aes↑	FVD_{16}↓	FVD_{full}↓
Cosmos3-Nano (16B)	18.66	0.6244	0.2753	93.66	74.00	40.68	71.9	178.9
Cosmos3-Super (64B)	18.36	0.6119	0.2884	93.77	72.50	40.89	68.5	175.3
OSCAR (2B)	16.21	0.5473	0.3818	93.15	68.50	**41.32**	121.3	286.4
RynnWorld-Teleop (5B)	16.94	0.5765	0.3649	**94.26**	68.00	40.05	114.0	277.5
DreamDojo (2B)	16.75	0.5711	0.3813	92.01	67.00	38.85	128.1	322.5
DreamDojo (14B)	17.76	0.6014	0.3435	91.99	62.00	39.11	121.4	317.8
Ropedia W0 (5B), pretrain	18.90	0.6409	0.2444	93.64	75.00	40.73	73.0	189.9
Ropedia W0 (5B), midtrain	18.91	0.6402	0.2453	93.71	72.00	41.15	70.1	190.1
Ropedia W0 (A14B), pretrain	**19.77**	**0.6749**	**0.2381**	93.55	76.50	40.20	**64.0**	**155.9**
Ropedia W0 (A14B), midtrain	19.34	0.6691	0.2462	93.45	**79.00**	40.23	72.4	168.0

manipulation windows, A14B midtrain leads all baselines and 5B is on par with the strongest (OSCAR [119]), so the advantage is not limited to transfer settings (Table 5).

Single-checkpoint cross-embodiment prediction. A single midtrain checkpoint predicts all eight embodiments on the cross-embodiment set with no post-training on these evaluation datasets (*zero-shot*; the embodiments appear in mid-training only through other datasets), reaching 19.83 PSNR at 5B and 21.98 at A14B (Table S6). For a matched comparison, we fine-tune Cosmos3 on the same robot mixture; even so it reaches only 19.02 PSNR, below our 5B and ≈3 dB below A14B. Mid-training drives this transfer, improving prediction most before any target-specific adaptation (pretrain/midtrain contrast, Table 4).

6.3. Ablation Studies and Analysis

Effectiveness of control conditions. We train separate variants without skeleton or camera conditioning and evaluate them on the cross-embodiment set (Table S6). Removing skeleton conditioning reduces PSNR by 3.8/1.8 dB for A14B/5B, respectively, highlighting its importance for visual fidelity. Removing camera conditioning yields smaller average drops of 0.3/0.1 dB, as six of the eight embodiments exhibit little camera motion. On the embodiments with moving viewpoints, however, removing camera conditioning reduces PSNR by 1.3 dB and leads to less generated motion. These results demonstrate the complementary roles of the two control

Table 3 | **Comparison on the DreamDojo [34] benchmarks under different post-training settings.** Despite having only **5B** parameters, our model consistently outperforms the larger DreamDojo (14B) baseline across all metrics. Best results are in **bold**, and second best are underlined.

Data	Model	In-lab			EgoDex			DreamDojo-HV		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1%	DreamDojo (2B)	18.840	0.731	0.254	18.584	0.749	0.243	16.996	0.724	0.256
	DreamDojo (14B)	18.983	0.743	0.244	18.070	0.747	0.246	16.643	0.716	0.262
	Ropedia W0 (5B)	20.854	0.764	0.213	20.569	0.778	0.208	19.045	0.760	0.217
	Ropedia W0 (A14B)	**23.028**	**0.826**	**0.190**	**22.354**	**0.830**	**0.188**	**20.788**	**0.810**	**0.198**
10%	DreamDojo (2B)	19.238	0.744	0.244	19.091	0.768	0.230	17.210	0.729	0.250
	DreamDojo (14B)	19.348	0.751	0.235	19.057	0.770	0.225	17.511	0.739	0.241
	Ropedia W0 (5B)	21.021	0.786	0.206	20.776	0.801	0.199	19.290	0.787	0.207
	Ropedia W0 (A14B)	**23.151**	**0.827**	**0.185**	**22.485**	**0.835**	**0.181**	**20.929**	**0.816**	**0.191**
100%	Cosmos-Predict2.5 [3]	20.576	0.774	0.222	19.952	0.787	0.219	18.274	0.754	0.236
	DreamDojo (2B)	21.114	0.774	0.222	20.411	0.775	0.226	18.813	0.747	0.238
	DreamDojo (14B)	21.413	0.788	0.208	20.525	0.787	0.213	18.924	0.751	0.228
	Ropedia W0 (5B)	21.738	0.797	0.202	21.230	0.810	0.196	19.814	0.795	0.204
	Ropedia W0 (A14B)	**23.217**	**0.828**	**0.186**	**22.731**	**0.845**	**0.177**	**21.086**	**0.826**	**0.187**

Table 4 | **World-model evaluation on Ψ_0 [117], ω-HOME [65], and Xp-Teleop** under three post-training settings: no post-training (zero-shot transfer), few-shot, and full-data, reporting PSNR, SSIM, and LPIPS. For Ropedia W0 we report both the pretrain (human-only) and midtrain (human + robot) stages at the 5B and A14B scales. Best bold, second-best underlined.

Benchmark	Model / initialization	No post-training			Few-shot			Full-data		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ψ_0	DreamDojo (14B)	14.45	0.580	0.432	14.43	0.579	0.438	13.78	0.569	0.449
	Cosmos3-Nano (16B)	13.74	0.563	0.465	14.71	0.584	0.409	16.55	0.636	0.320
	Cosmos3-Super (64B)	13.78	0.562	0.461	14.94	0.591	0.400	15.73	0.613	0.352
	Ropedia W0 (5B), pretrain	15.17	0.587	0.394	15.59	0.606	0.364	16.26	0.625	0.330
	Ropedia W0 (5B), midtrain	15.22	0.594	0.385	15.59	0.603	0.369	16.41	0.628	0.325
	Ropedia W0 (A14B), pretrain	15.87	0.614	0.368	17.83	0.664	0.272	18.94	0.692	0.241
	Ropedia W0 (A14B), midtrain	**17.36**	**0.656**	**0.291**	**18.28**	**0.674**	**0.260**	**19.07**	**0.698**	**0.232**
ω-HOME	DreamDojo (14B)	14.70	0.642	0.436	14.44	0.636	0.448	16.16	0.676	0.344
	Cosmos3-Nano (16B)	16.24	0.686	0.336	17.24	0.712	0.293	18.84	0.744	0.235
	Cosmos3-Super (64B)	15.90	0.675	0.353	16.84	0.700	0.309	18.46	0.735	0.247
	Ropedia W0 (5B), pretrain	17.35	0.698	0.295	18.35	0.728	0.250	19.02	0.738	0.233
	Ropedia W0 (5B), midtrain	17.82	0.714	0.277	18.30	0.721	0.259	19.25	0.742	0.228
	Ropedia W0 (A14B), pretrain	18.40	0.723	0.265	20.43	0.768	0.197	21.49	0.785	0.175
	Ropedia W0 (A14B), midtrain	**19.55**	**0.755**	**0.217**	**20.71**	**0.775**	**0.190**	**21.77**	**0.788**	**0.168**
Xp-Teleop	DreamDojo (14B)	13.49	0.445	0.392	13.00	0.423	0.448	12.91	0.433	0.406
	Cosmos3-Nano (16B)	13.69	0.461	0.375	14.51	0.489	0.320	14.98	0.508	0.293
	Cosmos3-Super (64B)	13.85	0.462	0.364	14.40	0.482	0.324	14.81	0.496	0.300
	Ropedia W0 (5B), pretrain	14.52	0.465	0.336	15.50	0.518	0.296	15.90	0.520	0.277
	Ropedia W0 (5B), midtrain	14.98	0.489	0.320	15.52	0.509	0.293	15.89	0.517	0.279
	Ropedia W0 (A14B), pretrain	14.95	0.496	0.343	17.98	0.600	0.215	18.26	0.601	0.209
	Ropedia W0 (A14B), midtrain	**16.42**	**0.537**	**0.259**	**18.25**	**0.601**	**0.205**	**18.74**	**0.618**	**0.196**

conditions, with camera conditioning particularly important for loco-manipulation.

Human pre-training and scaling. To isolate the effect of our pre-training, we post-train the same architecture initialized from either public video-generation weights ("official base") or our pretrained weights (Table S7). With 1% of the post-training data, Ropedia W0 5B improves PSNR by 4.8/7.1/6.6 dB over the official base and already outperforms the official base post-trained on 100% of the data. It retains an approximately 2.3 dB advantage under full-data post-training. The gains are largest in the low-data regime and remain substantial under

Table 5 | **Comparison on AgiBot** [15] (50 held-out gripper windows). VBench [51] consistency, aesthetics, and dynamic degree. Best bold, second-best underlined.

Model / initialization	Fidelity			Video quality			
	PSNR↑	SSIM↑	LPIPS↓	S-Cons.↑	B-Cons.↑	Aes.↑	Dyn.
DreamDojo (2B), AgiBot post-trained	18.26	0.750	0.218	0.9383	0.9520	0.4318	0.54
DreamDojo (14B), AgiBot post-trained	18.30	<u>0.755</u>	<u>0.210</u>	<u>0.9416</u>	<u>0.9526</u>	0.4374	0.48
Cosmos3-Nano (16B)	17.70	0.742	0.230	0.9159	0.9262	0.4373	0.72
Cosmos3-Super (64B)	17.28	0.732	0.241	0.9181	0.9274	0.4396	0.68
OSCAR (2B)	18.56	0.746	0.217	0.9281	0.9458	**0.4477**	0.66
Ropedia W0 (5B), midtrain	<u>18.60</u>	0.750	0.214	**0.9437**	**0.9573**	0.4378	0.60
Ropedia W0 (A14B), midtrain	**20.88**	**0.790**	**0.154**	0.9368	0.9525	<u>0.4407</u>	0.64

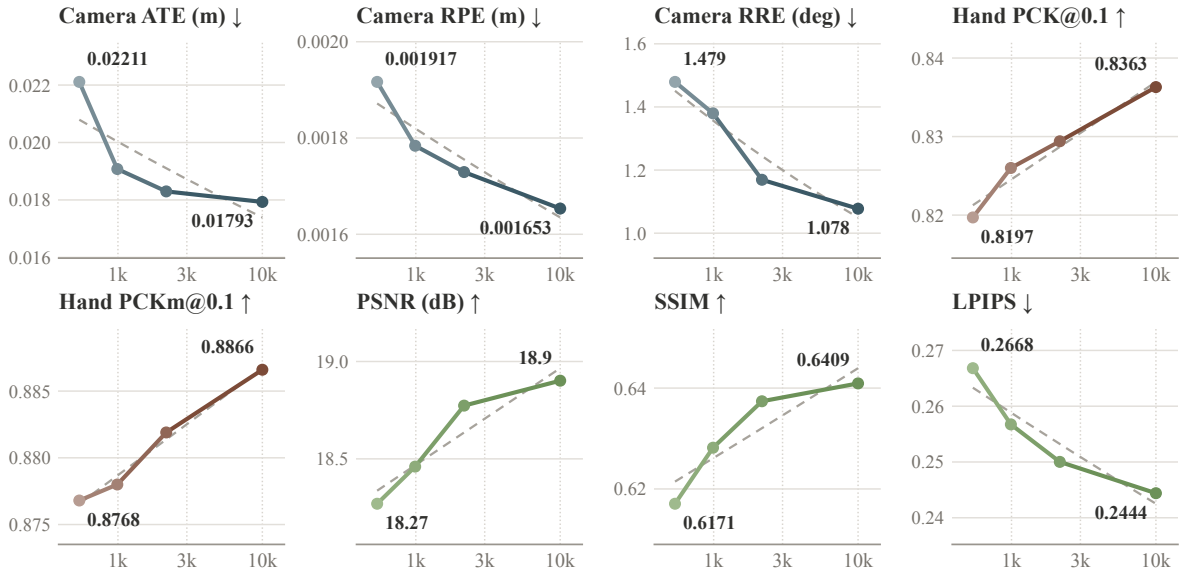


Figure 6 | **Scaling on held-out human video.** Zero-shot evaluation of the 5B model on 200 held-out egocentric clips versus human-video pre-training hours (log scale). The eight panels report camera ATE, RPE, and RRE; hand PCK@0.1 and PCKm@0.1; and PSNR, SSIM, and LPIPS. PCK counts missed hand detections as failures, whereas PCKm scores matched hands only; see Sec. C.1 for the evaluation protocol. Dashed lines indicate fitted trends.

full-data post-training. Prediction quality also improves as we scale the amount of human-video pre-training data, showing a consistent, monotonic improvement on held-out human video (Fig. 6), in human-to-robot transfer to Ψ_0, Xp-Teleop, and ω-HOME (Fig. 7), and on the DreamDojo GR-1 cross-evaluation sets (Fig. 8); see Sec. C.2 for the scaling protocol.

Model scaling. Beyond data, increasing the model from 5B to A14B improves prediction consistently across every regime: +2.2 dB PSNR on the cross-embodiment set (19.83 → 21.98, Table S6), +2.3 dB in-distribution on AgiBot (18.60 → 20.88, Table 5), and, without any post-training, +2.1/1.7/1.4 dB on Ψ_0/ω-HOME/Xp-Teleop (Table 4). Model-size and data scaling are thus complementary, jointly driving the gains reported above.

Distillation. The distilled 4-step 5B student reaches 21.3 fps on a single GPU (7× the 50-step teacher, past the 20 fps real-time line), while matching the teacher's fidelity from short clips (6.3 s) out to minute-long rollouts (Table S1; Sec. B.2).

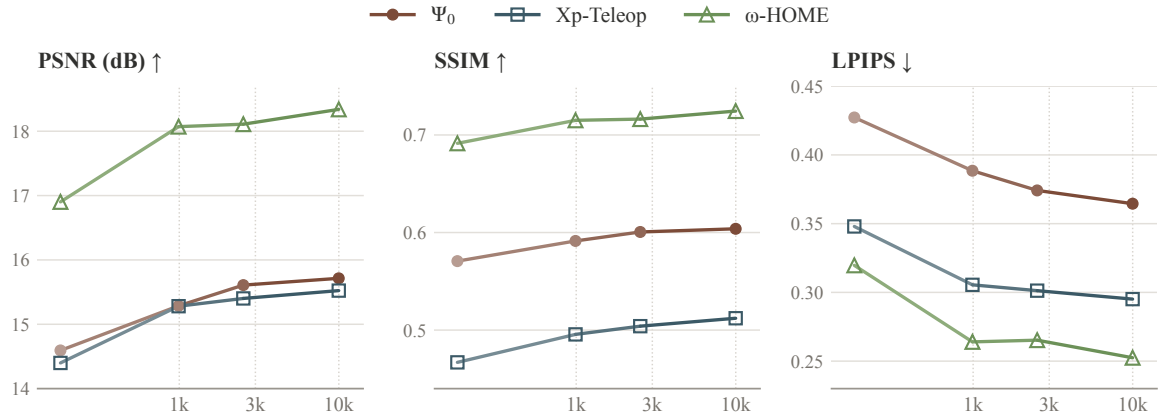


Figure 7 | **Human-to-robot transfer scaling.** PSNR, SSIM, and LPIPS on Ψ_0 [117], Xp-Teleop, and ω-HOME [65] versus human-video pre-training hours (log scale); each point is a pre-training checkpoint post-trained on the target benchmark before evaluation. Colors and marker shapes distinguish benchmarks; lines connect measured checkpoint results.

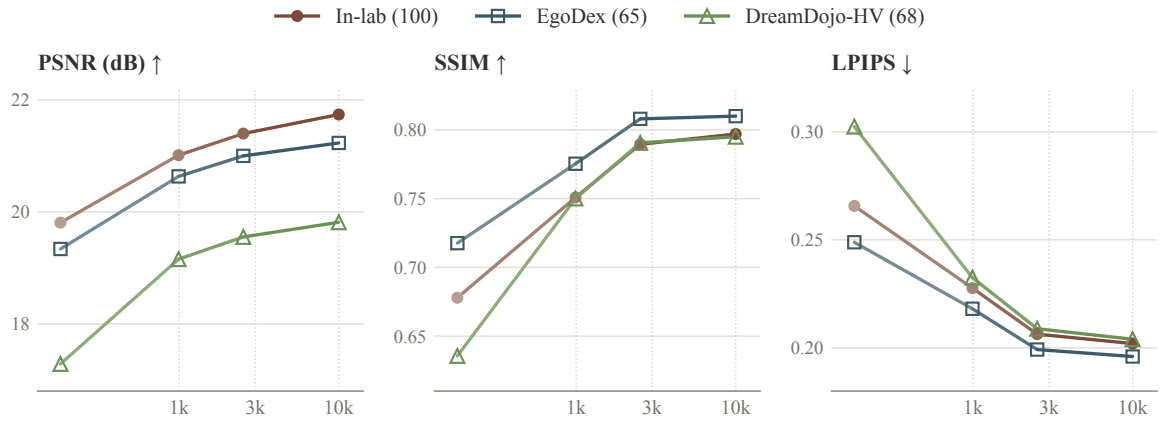


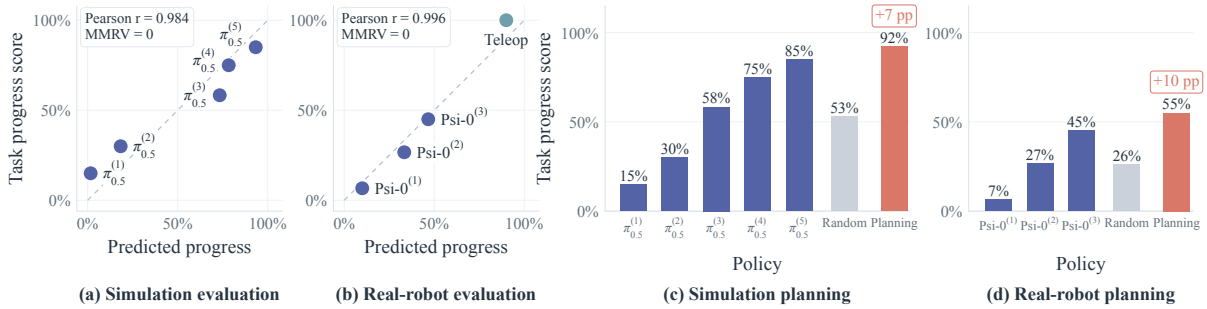
Figure 8 | **Cross-evaluation scaling on the DreamDojo [34] benchmark.** PSNR, SSIM, and LPIPS on the In-lab, EgoDex [46], and DreamDojo-HV evaluation sets versus human-video pre-training hours (log scale), under an identical GR-1 post-training budget; the largest-scale point uses Ropedia W0 (5B). Colors and marker shapes distinguish evaluation sets; lines connect measured checkpoint results.

6.4. Downstream Applications

We study four downstream uses of Ropedia W0, spanning simulation and real robots: policy evaluation, model-based planning, world–action modeling, and live teleoperation.

Policy evaluation. Used as a proxy environment, Ropedia W0 evaluates policy rollouts using a task-progress rubric following WorldEval [64] and SIMPLER [62]. Its scores closely match ground-truth evaluations, achieving Pearson correlations of $r = 0.984$ across five $\pi_{0.5}$ [52] checkpoints in RoboTwin 2.0 [23] and $r = 0.996$ across three Ψ_0 checkpoints and an expert-teleoperation reference on a real Unitree G1 loco-manipulation task. Both evaluations yield MMRV $= 0$ and preserve the ground-truth checkpoint rankings (Fig. 9(a,b)), supporting the use of Ropedia W0 for policy comparison and checkpoint selection.

Model-based planning. Following DreamDojo [34], Ropedia W0 is used for closed-loop planning: it rolls out multiple candidate action sequences, an external value model scores the



(a) Simulation evaluation (b) Real-robot evaluation (c) Simulation planning (d) Real-robot planning

Figure 9 | **Policy evaluation and world-model-based planning.** (a,b) Predicted vs. ground-truth task-progress scores in simulation and on the real robot (dashed: perfect agreement). (c,d) Task-progress scores for individual policies and world-model-based candidate selection; *Random* selects one candidate rollout uniformly at random, so its expected score equals the mean over candidates.

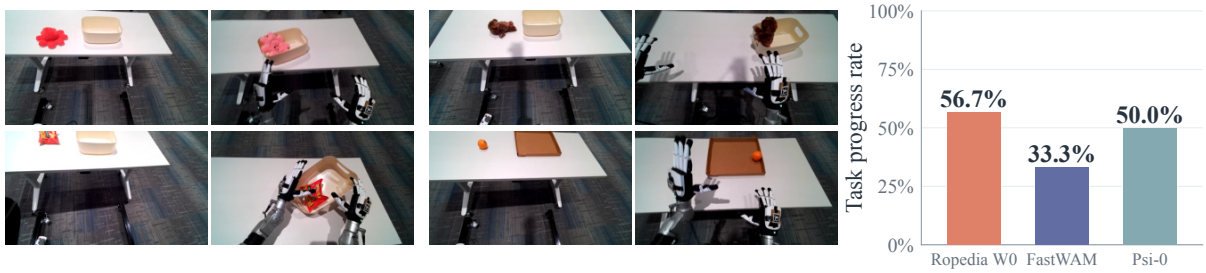
Table 6 | **World–action modeling on LIBERO [71] and LIBERO-Plus [32].** Success rates (%) with FastWAM [131] or Ropedia W0 init (LIBERO-Plus held out); best bold. Improvement is the absolute gain, relative gain over FastWAM in gray.

Demos/task	Initialization	LIBERO	LIBERO-Plus							
			Camera	Robot	Language	Lighting	Background	Noise	Layout	**Overall**
full (≈43)	FastWAM	96.95	17.32	44.32	67.47	79.96	52.04	38.66	60.92	50.06
	Ropedia W0	**97.00**	**20.14**	**51.68**	**73.52**	**88.05**	**61.43**	**48.09**	**66.30**	**56.82**
	Improvement	+0.05 (0.1%)	+2.82 (16.3%)	+7.36 (16.6%)	+6.05 (9.0%)	+8.09 (10.1%)	+9.39 (18.0%)	+9.43 (24.4%)	+5.38 (8.8%)	**+6.76 (13.5%)**
10	FastWAM	87.30	**18.32**	23.10	43.59	55.27	29.46	**28.36**	42.36	33.58
	Ropedia W0	**92.20**	10.94	**24.26**	**50.94**	**70.30**	**40.43**	21.99	**49.84**	**36.71**
	Improvement	+4.90 (5.6%)	-7.38 (-40.3%)	+1.16 (5.0%)	+7.35 (16.9%)	+15.03 (27.2%)	+10.97 (37.2%)	-6.37 (-22.5%)	+7.48 (17.7%)	**+3.13 (9.3%)**
3	FastWAM	60.00	7.57	12.58	25.31	**51.32**	**19.05**	15.12	27.15	21.44
	Ropedia W0	**62.65**	**9.76**	**13.87**	**34.74**	46.31	17.38	**20.17**	**30.23**	**23.97**
	Improvement	+2.65 (4.4%)	+2.19 (28.9%)	+1.29 (10.3%)	+9.43 (37.3%)	-5.01 (-9.8%)	-1.67 (-8.8%)	+5.05 (33.4%)	+3.08 (11.3%)	**+2.53 (11.8%)**

predicted rollouts, and the selected plan is executed in the environment. We evaluate planning on fixed-view manipulation tasks in simulation and on the real robot. The executed plan reaches 92% task progress vs. 53% (random selection) / 85% (best fixed checkpoint) in RoboTwin 2.0, and 55% vs. 26% / 45% on the real G1 (Fig. 9(c,d)).

World–action modeling. A world–action model jointly predicts future frames and the actions that produce them, so a stronger world model should be a better initialization. We test this by initializing the video backbone of FastWAM [131] from our pretrained 5B world model and leaving its action head and training recipe unchanged. This improves action learning: trained on LIBERO [71] and evaluated directly on both LIBERO and the held-out LIBERO-Plus [32], it raises LIBERO-Plus success from 50.06% to 56.82% and improves every few-shot setting (Table 6), and on a real G1 loco-manipulation task it reaches 56.7% task progress vs. 33.3% for FastWAM, competitive with the Ψ_0 VLA policy itself (50.0%; Fig. 10).

Live teleoperation. A Quest 3 headset maps the user's head and hand motion to camera poses and skeletons, and the distilled 5B student generates egocentric video in real time, letting users interactively drive viewpoint and manipulation across embodiments (Fig. 11; videos in the supplement).



(a) Real-robot executions **(b) Policy comparison**

Figure 10 | **World–action modeling.** (a) Real-robot executions of the policy initialized from Ropedia W0, shown as four chronological pairs of recorded keyframes with different objects, containers, and placements. (b) Aggregate real-robot task progress for Ropedia W0 initialization, FastWAM [131], and the Ψ_0 [117] policy.

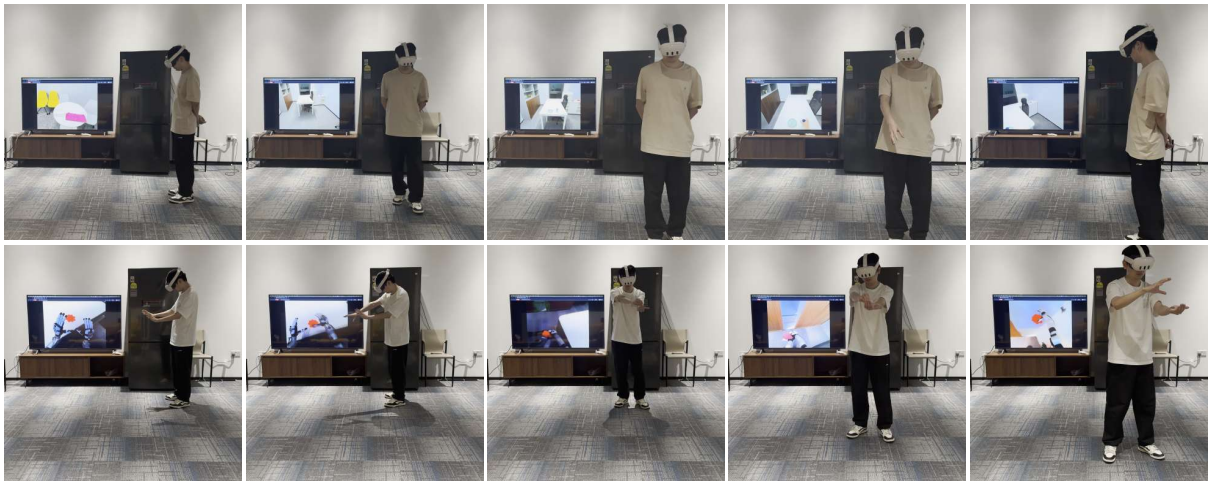


Figure 11 | **Live teleoperation with Ropedia W0.** On-site recordings show the headset user and the monitor displaying model-generated feedback. Five chronological keyframes show interactions with human (top) and robot (bottom) embodiments.

7. Conclusion

We presented Ropedia W0, a unified egocentric world model for loco-manipulation across humans and robots with grippers or dexterous hands. Its explicit kinematic representation combines camera poses for viewpoint motion with projected skeletons for manipulation, preserving embodiment-native geometry within a shared visual format. This interface enables learning from human and robot data and remains consistent from pre-training to post-training. Trained on Xp-Data with 10k hours of human video and 3k hours of robot data, Ropedia W0 demonstrates strong zero-shot generalization, data-efficient adaptation, and competitive prediction across diverse benchmarks, including our new Xp-Teleop. Our scaling analysis further shows that increasing human-video pre-training improves prediction on held-out human video and transfer to robot benchmarks, highlighting human experience as a scalable source of supervision for cross-embodiment world modeling. Beyond video prediction, the model supports policy evaluation, model-based planning, and world–action modeling, while distillation enables real-time interactive simulation.

References

[1] Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, and 1 others. 2026. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*.

[2] Build AI. 2025. Egocentric-10K.

[3] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, and 1 others. 2025. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*.

[4] Hadi Alzayer, Wenlong Huang, Haonan Chen, Christopher Luey, Lvmin Zhang, Maneesh Agrawala, Gordon Wetzstein, Li Fei-Fei, Yilun Du, Jiajun Wu, and 1 others. 2026. Masked visual actions for unified world modeling. *arXiv preprint arXiv:2607.19343*.

[5] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, and 1 others. 2025. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*.

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

[7] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. 2026. Whole-body conditioned egocentric video prediction. *Advances in Neural Information Processing Systems*, 38:164375–164418.

[8] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. 2025. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15791–15801. IEEE.

[9] Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, and 1 others. 2026. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 35101–35113.

[10] BitRobot, Unitree, and Hugging Face. 2026. HIW-500: Humanoids In-the-Wild Dataset for Robot Learning. https://bitrobot-foundation.github.io/humanoids-in-the-wild-500-hours/.

[11] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, and 1 others. 2025. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*.

[12] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, and 1 others. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*.

[13] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, and 1 others. 2023. RT-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

[14] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, and 1 others. 2024. Genie: Generative interactive environments. In *Proceedings of the International Conference on Machine Learning*.

[15] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, and 1 others. 2025. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*.

[16] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. 2025. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. *arXiv preprint arXiv:2505.06111*.

[17] BWM Team. 2026. BWM: A low-cost high-fidelity world simulator for robot learning. *arXiv preprint arXiv:2607.29302*.

[18] Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. 2026. PhysX-3D: Physical-grounded 3D asset generation. *Advances in Neural Information Processing Systems*, 38:93771–93784.

[19] Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. 2026. PhysX-Anything: Simulation-ready physical 3D assets from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5839–5848.

[20] Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. 2026. PhysX-Omni: Unified Simulation-Ready Physical 3D Generation for Rigid, Deformable, and Articulated Objects. *arXiv preprint arXiv:2605.21572*.

[21] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, and 1 others. 2024. GR-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*.

[22] Cheng Chen, Jerry Bai, Jiacheng Wei, Boyu Chen, Xiaoji Zheng, Fan Wu, Minghao Yang, Tianrun Chen, Ruibo Li, Xiaoyu Yue, and 1 others. 2026. AnyWorld: Factorized Egocentric World Models for Cross-Embodiment Generalization. *arXiv preprint arXiv:2608.29242*.

[23] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, and 1 others. 2025. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*.

[24] Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 2025. 4DNeX: Feed-Forward 4D Generative Modeling Made Easy. *arXiv preprint arXiv:2508.13154*.

[25] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

[26] Jiaxing Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. 2026. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, volume 2026, pages 85802–85822.

[27] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and 1 others. 2022. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130(1):33–55.

[28] Maximilian Du and Shuran Song. 2026. DynaGuide: Steering diffusion polices with active dynamic guidance. *Advances in Neural Information Processing Systems*, 38:44192–44221.

[29] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 2023. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954. IEEE.

[30] Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. 2024. RH20T: A comprehensive robotic dataset for learning diverse skills in one-shot. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 653–660. IEEE.

[31] Irving Fang, Yuzhong Chen, Yifan Wang, Jianghan Zhang, Qiushi Zhang, Jiali Xu, Xibo He, Weibo Gao, Hao Su, Yiming Li, and 1 others. 2024. EgoPAT3Dv2: Predicting 3D action target from 2D egocentric vision for human-robot interaction. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3036–3043. IEEE.

[32] Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, and 1 others. 2025. LIBERO-Plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*.

[33] Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. 2024. HumanPlus: Humanoid shadowing and imitation from humans. *arXiv preprint arXiv:2406.10454*.

[34] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, and 1 others. 2026. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*.

[35] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. 2024. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 37:91560–91596.

[36] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. 2025. AdaWorld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*.

[37] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. 2018. First-person hand action benchmark with RGB-D videos and 3D hand pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419. IEEE.

[38] Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. 2026. Learning latent action world models in the wild. In *Proceedings of the International Conference on Machine Learning*.

[39] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, and 1 others. 2022. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

[40] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, and 1 others. 2024. 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*, pages 19383–19400.

[41] Yanjiang Guo, Lucy Shi, Jianyu Chen, and Chelsea Finn. 2026. Ctrl-World: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, volume 2026, pages 6121–6138.

[42] David Ha and Jürgen Schmidhuber. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440.

[43] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2025. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653.

[44] D HO, J MONAS, JT REN, and C YU. 2025. 1X World Model: evaluating bits, not atoms. https://www.1x.tech/1x-world-model.pdf.

[45] Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, and 1 others. 2025. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*.

[46] Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, and 1 others. 2026. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pages 4218–4237.

[47] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. 2023. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*.

[48] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. 2024. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*.

[49] Qianxi Hua, Xinyue Li, Zheng Yan, Yang Li, Chi Zhang, Yongyao Li, and Yufei Liu. 2026. VTouch++: A Multimodal Dataset with Vision-Based Tactile Enhancement for Bimanual Manipulation. *arXiv preprint arXiv:2604.20444*.

[50] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. 2025. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*.

[51] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, and 1 others. 2024. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818. IEEE.

[52] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, and 1 others. 2025. $\pi_{0.5}$: a Vision-Language-Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*.

[53] Arnav Kumar Jain, Vibhakar Mohta, Subin Kim, Atiksh Bhardwaj, Juntao Ren, Yunhai Feng, Sanjiban Choudhury, and Gokul Swamy. 2026. A smooth sea never made a skilled sailor: Robust imitation via learning to search. *Advances in Neural Information Processing Systems*, 38:62676–62707.

[54] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, and 1 others. 2025. DreamGen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*.

[55] Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, and 1 others. 2025. EnerVerse-AC: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*.

[56] Zhennan Jiang, Kai Liu, Yuxing Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Zhengtao Zhang, Chao Yu, Haoran Li, and Dongbin Zhao. 2026. World4RL: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. *IEEE Robotics and Automation Letters*.

[57] Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. 2025. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*.

[58] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, and 1 others. 2024. DROID: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*.

[59] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. OpenVLA: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

[60] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. 2021. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10118–10128. IEEE.

[61] Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, and 1 others. 2025. VLA-RFT: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*.

[62] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, and 1 others. 2024. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*.

[63] Xukun Li, Yu Sun, Lei Zhang, Bosheng Huang, Yibo Peng, Yuan Meng, Haojun Jiang, Shaoxuan Xie, Guocai Yao, Alois Knoll, and 1 others. 2026. DECO: Decoupled Multimodal Diffusion Transformer for Bimanual Dexterous Manipulation with a Plugin Tactile Adapter. *arXiv preprint arXiv:2602.05513*.

[64] Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. 2025. WorldEval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*.

[65] Zhe Li, Zhenzhe Zhang, Yangyang Wei, Wenjie Zhang, Xichen Yuan, Peiyuan Zhi, Gen Li, Xinying Guo, Fengjie Gao, Jianfei Yang, and 1 others. 2026. ω-0: A Latent Predictive World Action Model for Concurrent Humanoid Loco-Manipulation. *arXiv preprint arXiv:2608.06375*.

[66] Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, and 1 others. 2026. Open-AoE: An Open Egocentric Manipulation Dataset and Toolchain for Embodied Learning. *arXiv preprint arXiv:2607.14183*.

[67] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, and 1 others. 2026. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pages 88446–88463.

[68] Lightwheel. 2026. EgoSuite-Open100K. https://huggingface.co/collections/LightwheelAI/egosuite-open100k.

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

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

[71] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. 2023. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36:44776–44791.

[72] Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. 2025. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*.

[73] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, and 1 others. 2024. Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pages 38–55. Springer.

[74] Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, and 1 others. 2025. Light-X: Generative 4D Video Rendering with Camera and Illumination Control. *arXiv preprint arXiv:2512.05115*.

[75] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. 2025. Free4D: Tuning-free 4D Scene Generation with Spatial-Temporal Consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 25571–25582.

[76] Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. 2025. EgoZero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*.

[77] Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. 2024. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21740–21751. IEEE.

[78] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2022. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20981–20990. IEEE.

[79] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and 1 others. 2019. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*.

[80] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. 2025. Being-H0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*.

[81] Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, and 1 others. 2026. LDA-1B: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv preprint arXiv:2602.12215*.

[82] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, and 1 others. 2024. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *Proceedings of the European Conference on Computer Vision*, pages 445–465. Springer.

[83] MMPose Contributors. 2020. OpenMMLab Pose Estimation Toolbox and Benchmark. https://github.com/open-mmlab/mmpose.

[84] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. 2026. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–958. IEEE.

[85] Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, and 1 others. 2026. T-Rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*.

[86] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, and 1 others. 2024. Open X-Embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 6892–6903. IEEE.

[87] Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. 2026. EgoControl: Controllable egocentric video generation via 3D full-body poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4269–4279.

[88] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, and 1 others. 2024. Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model.

[89] Jack Parker-Holder, Shlomi Fruchter, and 1 others. 2025. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4.

[90] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4172–4182. IEEE.

[91] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12242–12254. IEEE.

[92] Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y Zhu, and 1 others. 2026. EgoVerse: An egocentric human dataset for robot learning from around the world. *arXiv preprint arXiv:2604.07607*.

[93] Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. 2026. Inference-time enhancement of generative robot policies via predictive world modeling. *IEEE Robotics and Automation Letters*.

[94] Ri-Zhao Qiu, Shiqi Yang, Xuxin Cheng, Chaitanya Chawla, Jialong Li, Tairan He, Ge Yan, David J Yoon, Ryan Hoque, Lars Paulsen, and 1 others. 2025. Humanoid policy˜ human policy. *arXiv preprint arXiv:2503.13441*.

[95] Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. 2026. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, volume 2026, pages 78932–78957.

[96] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2017. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):1–17.

[97] Ropedia. 2026. Xperience-10M: A Large-Scale Egocentric Multimodal Dataset with Structured 3D/4D Annotations. Dataset.

[98] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. 2025. Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523*.

[99] Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, and 1 others. 2026. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226.

[100] Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. 2025. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*.

[101] Fourier ActionNet Team and Yao Mu. 2025. ActionNet: A dataset for dexterous bimanual manipulation. *Fourier ActionNet dataset release*.

[102] Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, and 1 others. 2025. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*.

[103] Wei-Cheng Tseng, Jinwei Gu, Qinsheng Zhang, Hanzi Mao, Ming-Yu Liu, Florian Shkurti, and Lin Yen-Chen. 2025. Scalable policy evaluation with video world models. *arXiv preprint arXiv:2511.11520*.

[104] Unitree Robotics. 2024. XR-Teleoperate: An Open-Source Teleoperation Framework and Data Collection Toolkit for Embodied Intelligence. `https://github.com/unitreerobotics/xr_teleoperate`. Accessed: 2026-02.

[105] Unitree Robotics. 2026. UnifoLM-WBT-Dataset. `https://huggingface.co/collections/unitreerobotics/unifolm-wbt-dataset`.

[106] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new metric for video generation. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net.

[107] 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, and 1 others. 2023. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of the Conference on Robot Learning*, pages 1723–1736. PMLR.

[108] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, and 1 others. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

[109] Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C Karen Liu. 2024. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. *arXiv preprint arXiv:2403.07788*.

[110] Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. 2026. VGGT-Ω. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[111] Lirui Wang, Kevin Zhao, Chaoqi Liu, and Xinlei Chen. 2025. Learning real-world action-video dynamics with heterogeneous masked autoregression. *arXiv preprint arXiv:2502.04296*.

[112] Siyin Wang, Junhao Shi, Zhaoyang Fu, Xinzhe He, Feihong Liu, Chenchen Yang, Yikang Zhou, Zhaoye Fei, Jingjing Gong, Jinlan Fu, and 1 others. 2026. World action models: The next frontier in embodied ai. *arXiv preprint arXiv:2605.12090*.

[113] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, and 1 others. 2023. HoloAssist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20213–20224. IEEE.

[114] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1905–1914. IEEE.

[115] Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. 2025. Precise action-to-video generation through visual action prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12713–12724. IEEE.

[116] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

[117] Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, and 1 others. 2026. Ψ_0: An Open Foundation Model Towards Universal Humanoid Loco-Manipulation. *arXiv preprint arXiv:2603.12263*.

[118] Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, and 1 others. 2024. RoboMIND: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*.

[119] Zhuoyuan Wu and Jun Gao. 2026. OSCAR: Omni-Embodiment Action-Conditioned World Model for Robotics. *arXiv preprint arXiv:2606.04463*.

[120] Junjin Xiao, Yandan Yang, Xinyuan Chang, Ronghan Chen, Feng Xiong, Mu Xu, Wei-Shi Zheng, and Qing Zhang. 2025. World-Env: Leveraging world model as a virtual environment for VLA post-training. *arXiv preprint arXiv:2509.24948*.

[121] Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. 2026. EgoTwin: Dreaming body and view in first person. In *International Conference on Learning Representations*, volume 2026, pages 28642–28659.

[122] Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. 2026. Kinema4D: Kinematic 4D world modeling for spatiotemporal embodied simulation. *arXiv preprint arXiv:2603.16669*.

[123] Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, and 1 others. 2025. EgoVLA: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*.

[124] Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. 2023. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*.

[125] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. 2025. LongLive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*.

[126] Xiuyu Yang, Bohan Li, Shaocong Xu, Nan Wang, Chongjie Ye, Zhaoxi Chen, Minghan Qin, Yikang Ding, Zheng Zhu, Xin Jin, and 1 others. 2025. ORV: 4D occupancy-centric robot video generation. *arXiv preprint arXiv:2506.03079*.

[127] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, and 1 others. 2026. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*.

[128] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, and 1 others. 2025. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pages 28213–28239.

[129] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. 2024. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*.

[130] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. 2025. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22963–22974. IEEE.

[131] Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. 2026. FastWAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*.

[132] Hu Yue, Siyuan Huang, Yue Liao, Shengcong Chen, Pengfei Zhou, Liliang Chen, Maoqing Yao, and Guanghui Ren. 2025. EWMBench: Evaluating scene, motion, and semantic quality in embodied world models. *arXiv preprint arXiv:2505.09694*.

[133] Chuheng Zhang, Tim Pearce, Pushi Zhang, Kaixin Wang, Xiaoyu Chen, Wei Shen, Li Zhao, and Jiang Bian. 2026. What do latent action models actually learn? *Advances in Neural Information Processing Systems*, 38:146676–146697.

[134] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595. IEEE.

[135] Zongzheng Zhang, Jingrui Pang, Zhuo Yang, Kun Li, Minwen Liao, Saining Zhang, Guoxuan Chi, Jinbang Guo, Huan-ang Gao, Modi Shi, and 1 others. 2026. Dexora: Open-source VLA for high-DoF bimanual dexterity. *arXiv preprint arXiv:2605.18722*.

[136] Haoyu Zhao, Xingyue Zhao, Hangyu Li, Biao Gong, Kehan Li, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. 2026. Rynnworld-teleop: An action-conditioned world model for digital teleoperation. *arXiv preprint arXiv:2607.06558*.

[137] Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. 2026. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*.

[138] Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharov, Vitor Guizilini, and Yue Wang. 2025. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*.

[139] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. 2025. Tesseract: learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*.

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

[141] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. 2025. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. *arXiv preprint arXiv:2504.02792*.

[142] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. 2025. IRASim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9834–9844. IEEE.

[143] Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. 2026. WMPO: World model-based policy optimization for vision-language-action models. In *International Conference on Learning Representations*, volume 2026, pages 62486–62502.

[144] Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. 2026. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*.

Appendix Contents

A Data and Benchmark Details **28**

A.1 Data Curation and Annotation Examples 28

A.2 Xp-Teleop Collection and Benchmark Details 32

B Model and Implementation Details **34**

B.1 Backbone and Conditioning 34

B.2 Distillation and Interactive Simulation 35

B.3 Wide-FOV Skeleton Conditioning 37

C Evaluation Protocols **39**

C.1 Human Egocentric Video Evaluation 39

C.2 Scaling Protocol 39

D Additional Quantitative Results **40**

D.1 Human Evaluation by Dataset 40

D.2 Seen and Unseen Robot Evaluation 40

D.3 Long-Horizon Rollout Evaluation 41

D.4 Conditioning and Pretraining Ablations 42

D.5 WorldArena Comparison against BWM 43

E Qualitative Comparisons **44**

E.1 Cross-Embodiment Robot Prediction 44

E.2 Zero-Shot Loco-Manipulation 48

E.3 Post-Training and Data Efficiency 50

E.4 Additional Long-Horizon Rollouts 53

F Limitations and Future Work **54**

A. Data and Benchmark Details

We first document the curated training corpus and its unified annotations, then describe the held-out Xp-Teleop benchmark, including its collection setup and dataset-comparison criteria.

A.1. Data Curation and Annotation Examples

Figure S1 summarizes the curation pipeline. The examples below illustrate the resulting scene coverage and unified motion annotations.

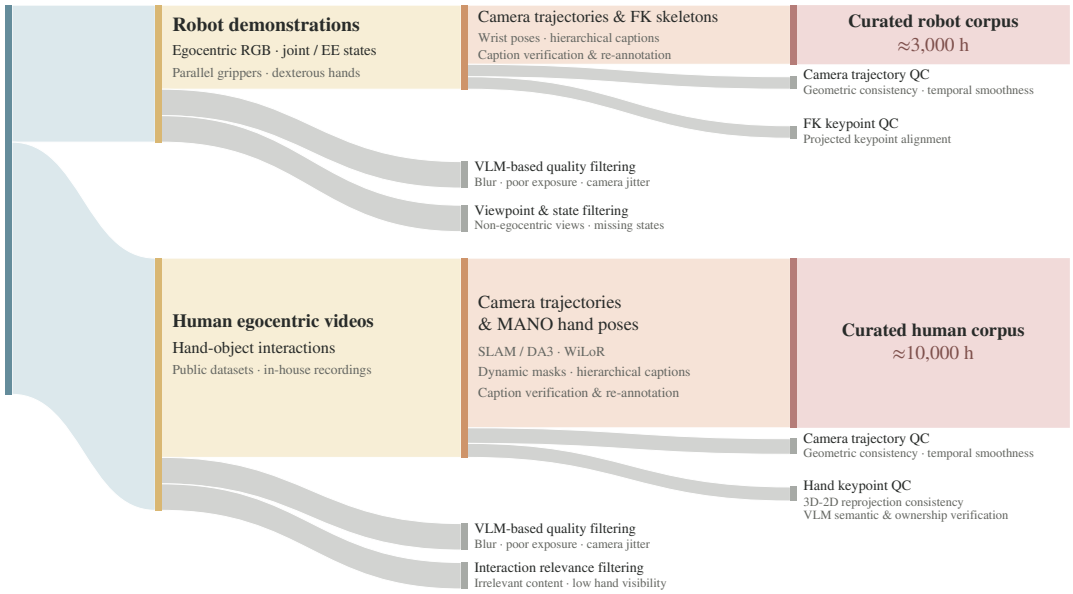


Figure S1 | **Data curation funnel for Xp-Data.** Robot and human egocentric recordings pass through source screening, motion annotation, and quality control; gray bands mark data removed at each step, and the surviving streams form the unified corpus of ∼3,000 robot hours and ∼10,000 human hours.

Figures S2–S4 visualize the corpus across human and robot recordings. Figure S3 shows example robot frames spanning different environments, tasks, objects, and embodiments. Figures S2 and S4 overlay the unified motion annotations on the same kinds of frames: projected hand or manipulator skeletons in the image, with the per-window camera trajectory shown in the inset.

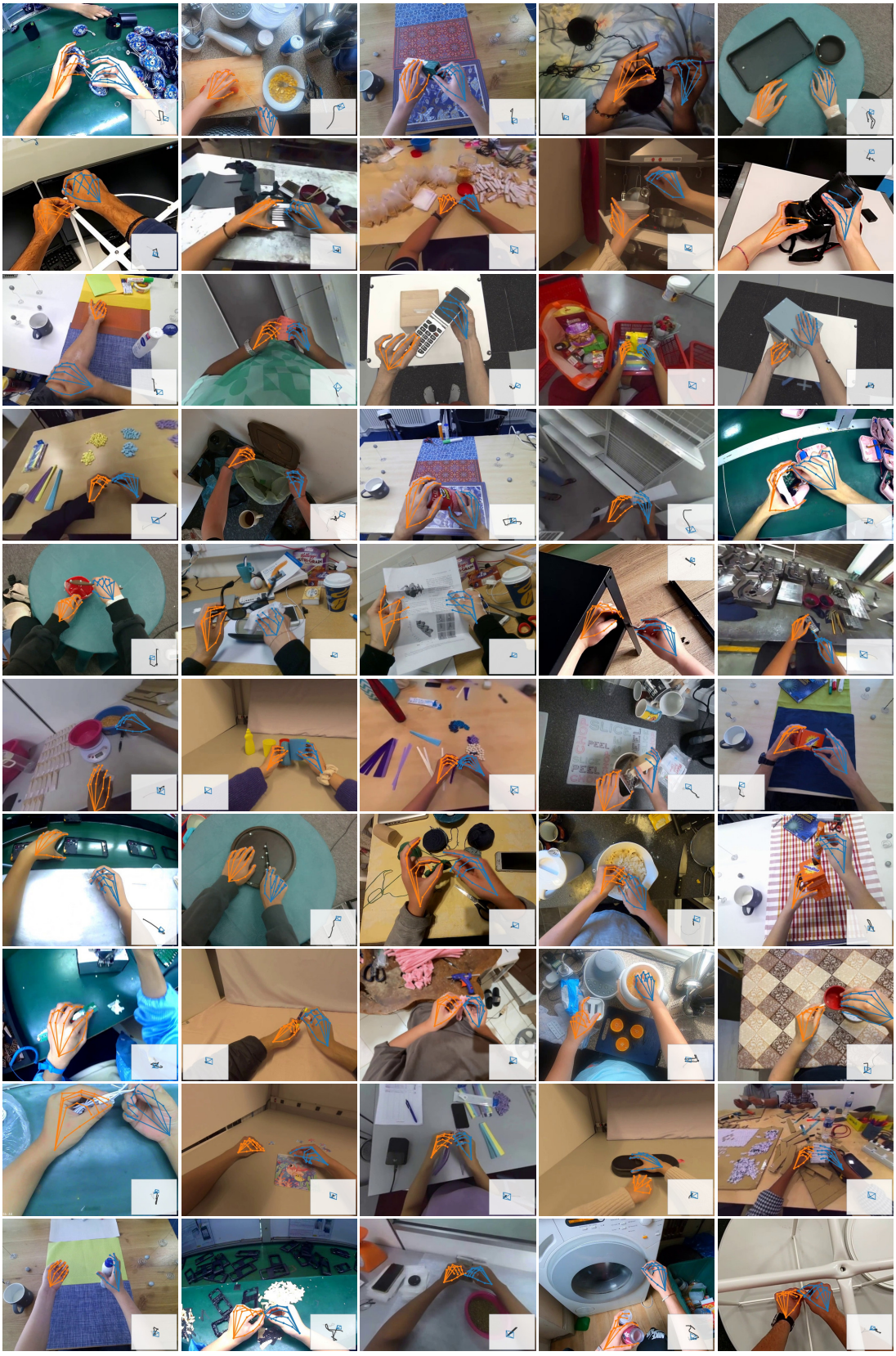


Figure S2 | **Motion annotations on the egocentric human data.** Projected 21-point hand skeletons are overlaid on each frame, with the two hands drawn in different colors; the inset shows the camera trajectory of the surrounding window.

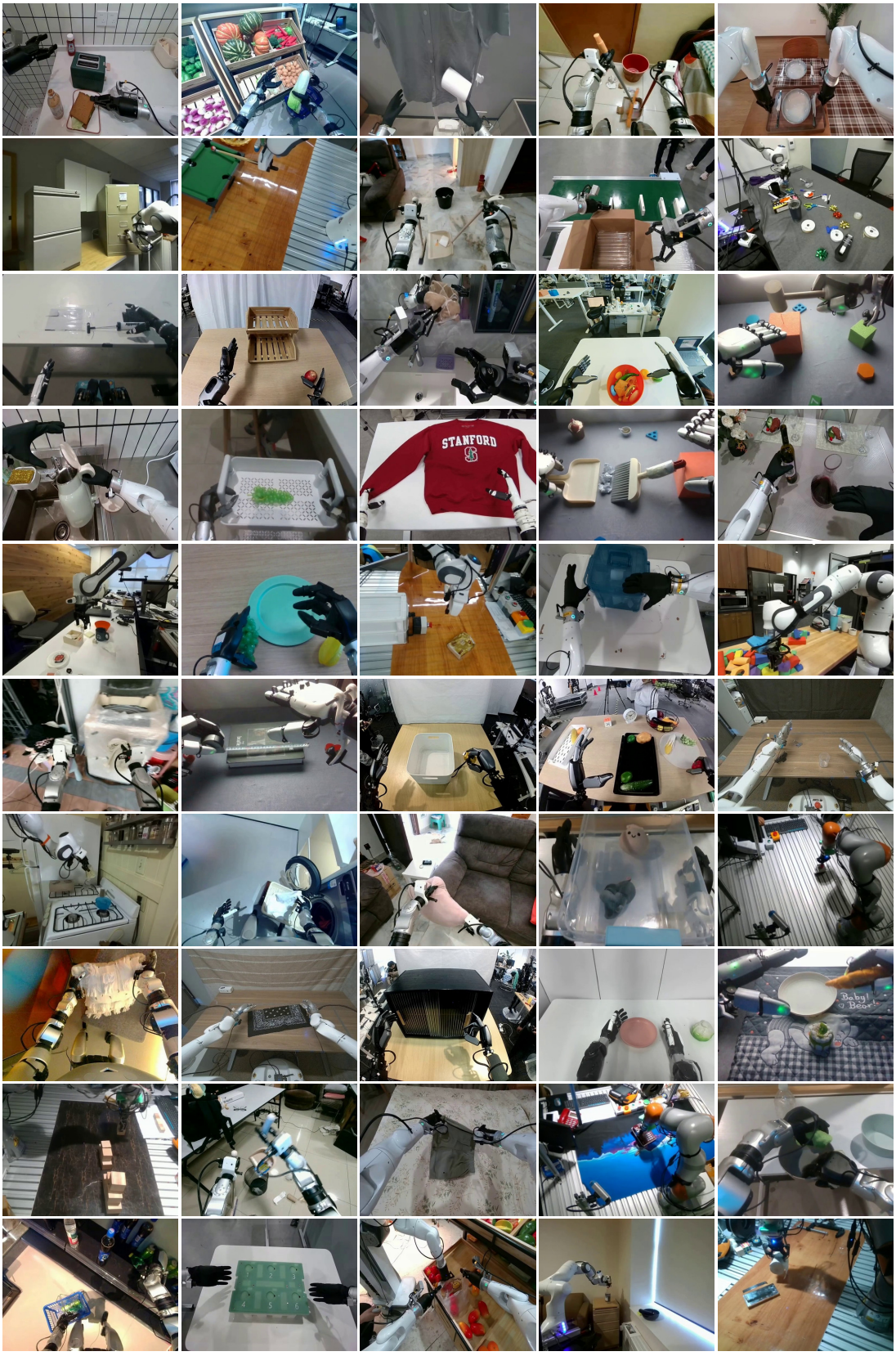


Figure S3 | **Example frames from the robot data in Xp-Data.** The grid spans dexterous five-finger hands, three-finger hands, and parallel grippers on single-arm, dual-arm, and humanoid platforms, across household, retail, industrial, and tabletop scenes.



Figure S4 | **Motion annotations on the robot data.** Image-aligned skeletons combine embodiment-native FK projections with manual keypoint annotation and correction (Sec. 3), covering dexterous hands and grippers; the inset shows the camera trajectory of the surrounding window.

A.2. Xp-Teleop Collection and Benchmark Details

The benchmark overview (Fig. 5) and the dataset comparison (Table 1) appear in the main text (Sec. 5); here we provide the full rig, control, recording, and comparison details.

Rig. A Unitree G1 humanoid with 29 actuated degrees of freedom is equipped with two Inspire RH56DFX dexterous hands, each with six actuated degrees of freedom. An Azure Kinect DK mounted on a custom 3D-printed helmet rigidly attached to the robot's head records the egocentric RGB-D view. Because the G1 has no actuated neck joints, the camera moves only with the waist and mobile base, matching the motion represented by our camera-pose control channel (Sec. 4).

Control. Two operators control the robot concurrently, allowing locomotion and manipulation to occur within the same take. The manipulation operator wears a Meta Quest 3, whose controller-free hand tracking provides targets to a Pinocchio-based inverse-kinematics solver for the two 7-DoF arms and to a dexterous-hand retargeting module. The locomotion operator independently commands the robot's base velocity using Unitree's R3 remote.

Recording. Each take records the robot and hand joint states, commanded targets, R3 commands, headset tracking, and Kinect stream on a common clock. From these signals, we derive the two action channels described in Sec. 4: camera poses and hand skeletons. The hand skeletons are computed through forward kinematics and projected into the image using the calibrated helmet-to-camera extrinsics. We retain every take initiated by the operators, including failed grasps and subsequent recoveries, because Xp-Teleop is intended to evaluate whether a world model can simulate off-nominal outcomes as well as successful executions.

Dataset composition and annotation granularity. Xp-Teleop contains 27 long-horizon teleoperation takes, totaling approximately one hour of egocentric video recorded at 30 fps. The takes are divided into 543 human-annotated atomic segments, with each take containing between 8 and 42 segments (median 18). This annotation is substantially finer-grained than the common convention of assigning a single instruction to an entire episode. The resulting segments cover 30 skill classes. The segment captions also identify the manipulated objects and describe the temporal progression of each task, providing language supervision aligned with the recorded observations and actions.

Scope of the dataset comparison. Table 1 compares publicly described dexterous-humanoid teleoperation datasets in terms of robot capabilities, task horizon, and available recording modalities. We divide the datasets into two groups according to how the lower body is used during data collection. In the tabletop-teleoperation group, the robot's legs remain locked even if the release provides base commands or whole-body state. In the loco-manipulation group, locomotion is actively commanded and can occur within the same episode as manipulation.

Task-horizon metrics. The three task-horizon columns report episode-level means over the analyzed portion of each dataset. *Duration* is the wall-clock length of one episode, computed as the number of recorded frames divided by the dataset's frame rate (30 Hz for all datasets except GR00T-Teleop, 20 Hz). Frame counts are taken from each dataset's own metadata. For our dataset we report the released task windows, not the raw recordings. RoboMIND does not document its frame rate; we assume 30 Hz.

Loco. Range measures how much space an episode covers. For each episode we take the 3-D

position trajectory of the robot base (or, when only a camera trajectory is available, the head camera), compute its axis-aligned bounding box, and report the box diagonal in metres. We use the bounding-box diagonal rather than integrated path length, because path length at 30 Hz is dominated by body sway and pose-estimation jitter (on our data it inflates the value by roughly $7\times$).

Pose sources. Ours: visual-inertial SLAM (metric). Humanoid-Everyday: on-board odometry. UnifoLM-WBT (new schema): the robot state estimator's base pose. All other datasets provide no metric base pose; we run DA3-Streaming with its metric configuration on the egocentric video (10 fps, 504 px input) to obtain camera poses. Because monocular metric SLAM has a systematic scale error, we calibrate it once against ground truth: on the 879 Humanoid-Everyday locomotion episodes the DA3 box diagonal exceeds the odometry value by a median factor of 1.39 (IQR 1.28–1.55), and on our 27 episodes the median factor against SLAM is 1.03; we divide all DA3-derived values by the pooled median 1.39. After calibration the estimator reproduces ground truth within 10% on both datasets (ours 1.48 vs 1.61 m; Humanoid-Everyday 0.84 vs 0.82 m). For AgiBot-Beta we estimate on a length-stratified sample of 30 episodes. Unitree G1 Dex1/Dex3, PH2D, and RoboMIND Tien Kung are reported as 0.

Subtasks counts how many sequential sub-goals an episode's task instruction specifies. We apply one rule to every dataset: the instruction is lower-cased and split into clauses at commas, semicolons, periods and the connectives "and", "then", "until", "upon", "after"; a clause counts as one subtask if its first or second content word is an action verb from a fixed list (pick, place, put, take, grasp, carry, walk, go, enter, turn, open, close, push, pull, pour, wipe, fold, locate, find, hang, press, insert, ...). Locomotion clauses count the same as manipulation clauses. Every instruction counts at least 1. Instruction sources: the per-episode task string for LeRobot datasets, the HDF5 description attribute for PH2D and RoboMIND, the task name for ω-HOME, and the human-written task instruction of each take for our dataset (one 19-step instruction excluded as an outlier). AgiBot-Beta ships no instruction text. Because subtask boundaries follow the wording and granularity of each release's instructions, this count is an annotation-level measure of task horizon rather than a platform-independent action count.

Modality criteria. The modality columns describe information available in the analyzed dataset release. For *Kpts.*, we also count image-space keypoints that can be reconstructed from released joint states and camera calibration using a publicly available URDF. *RGB@fps* reports the resolution and frame rate of the principal egocentric RGB stream. *Dpt.* requires a usable released depth stream; placeholder or zero-filled channels do not qualify. *IMU* requires explicit inertial-sensor measurements. Estimated root poses, torso poses, or camera trajectories alone are not counted as IMU data. *Tac.* denotes released tactile measurements.

WB joints requires recorded joint states covering both the upper and lower body, including the legs. Arm-only or hand-only states, estimated root or torso poses, and commanded base velocities do not satisfy this criterion. This column describes the coverage of the recorded state and does not require the leg joints to exhibit substantial motion during the released episodes. Consequently, a dataset may provide whole-body joint states while still belonging to the tabletop group.

Cal. requires sufficient camera calibration to relate robot kinematics to the image observations, including both camera intrinsics and the relevant extrinsics. Camera trajectories or extrinsics provided without intrinsics do not satisfy this criterion.

Kpts. denotes temporally aligned image-space keypoints, either provided directly or obtainable by projecting the relevant robot kinematics using released joint states, sufficient camera

calibration, and a publicly available URDF.

Dataset-specific qualifications. For Unitree Dex1, whole-body joint states are available only in the v3 release; v2 provides arm and gripper states but not the corresponding lower-body joint states. Although v3 supports base commands, the observed variation of its leg joints is below 0.03 rad in the analyzed recordings. We therefore treat its measured locomotion range as zero and place it in the tabletop group.

Ψ_0 supports locomotion, but the analyzed release does not provide leg-joint states. Its released depth channel is zero-filled and is therefore not counted as usable depth. RoboMIND entries refer specifically to its Xsens dexterous-hand subset. This subset includes depth observations but does not provide hand-joint states. Humanoid Everyday task-horizon statistics are computed from its loco-manipulation subset, and its tactile measurements are available only for the G1/Dex3 embodiment. The analyzed ω-HOME files do not contain a depth stream.

B. Model and Implementation Details

This section describes the video backbones, distillation and interactive inference, and the optional wide-FOV skeleton-conditioning branch. Figures and measurements are grouped with the component they illustrate.

B.1. Backbone and Conditioning

Ropedia W0 predicts future egocentric video from an initial visual observation and factorized camera and manipulator controls (Sec. 4.1). The following details describe the pretrained backbones used by this shared conditioning interface.

Latent video diffusion backbones. We instantiate Ropedia W0 at two scales using the Wan2.2 family of pretrained video diffusion models [108]. Wan2.2-TI2V-5B uses a dense diffusion transformer with approximately 5B parameters. Wan2.2-I2V-A14B uses two experts specialized for different noise regimes: a high-noise expert and a low-noise expert. Only one expert is active at each denoising step, yielding approximately 14B active parameters. In our experiments, the two experts are adapted separately and used in their respective noise regimes during inference.

Both backbones operate on spatiotemporal latents produced by a video VAE and follow the diffusion transformer architecture [90]. Text embeddings are incorporated through cross-attention, while sinusoidal timestep embeddings provide adaptive normalization and residual modulation within each transformer block. We train them with the latent flow-matching objective [70]: the video VAE encodes each clip into latents, the first temporal latent is retained as clean visual context, and noise and the loss are applied only to future latents. The camera-ray and skeleton conditioning branches introduced in Sec. 4.2 augment these backbones through zero-initialized projections, preserving the pretrained model's behavior at initialization.

B.2. Distillation and Interactive Simulation

Distillation recipe. We first obtain a 4-step autoregressive generator using Causal Forcing [144, 137], then extend it to long-video streaming using LongLive [125], retaining camera–skeleton conditioning.

Few-step distillation. An autoregressive diffusion model provides causal initialization. Asymmetric DMD then trains the causal generator using a frozen bidirectional real-score teacher (with classifier-free guidance) and an online bidirectional fake-score model. The generator starts from the causal initialization, while both score networks initialize from the bidirectional camera–skeleton model. The fake-score model learns a denoising objective on generated samples.

Long-video extension. LongLive adapts the few-step generator for streaming with within-block bidirectional and cross-block causal attention. A bounded KV cache retains a local window and an attention sink.

Figure S5 illustrates distillation; Fig. S6 shows subsequent long-video streaming.

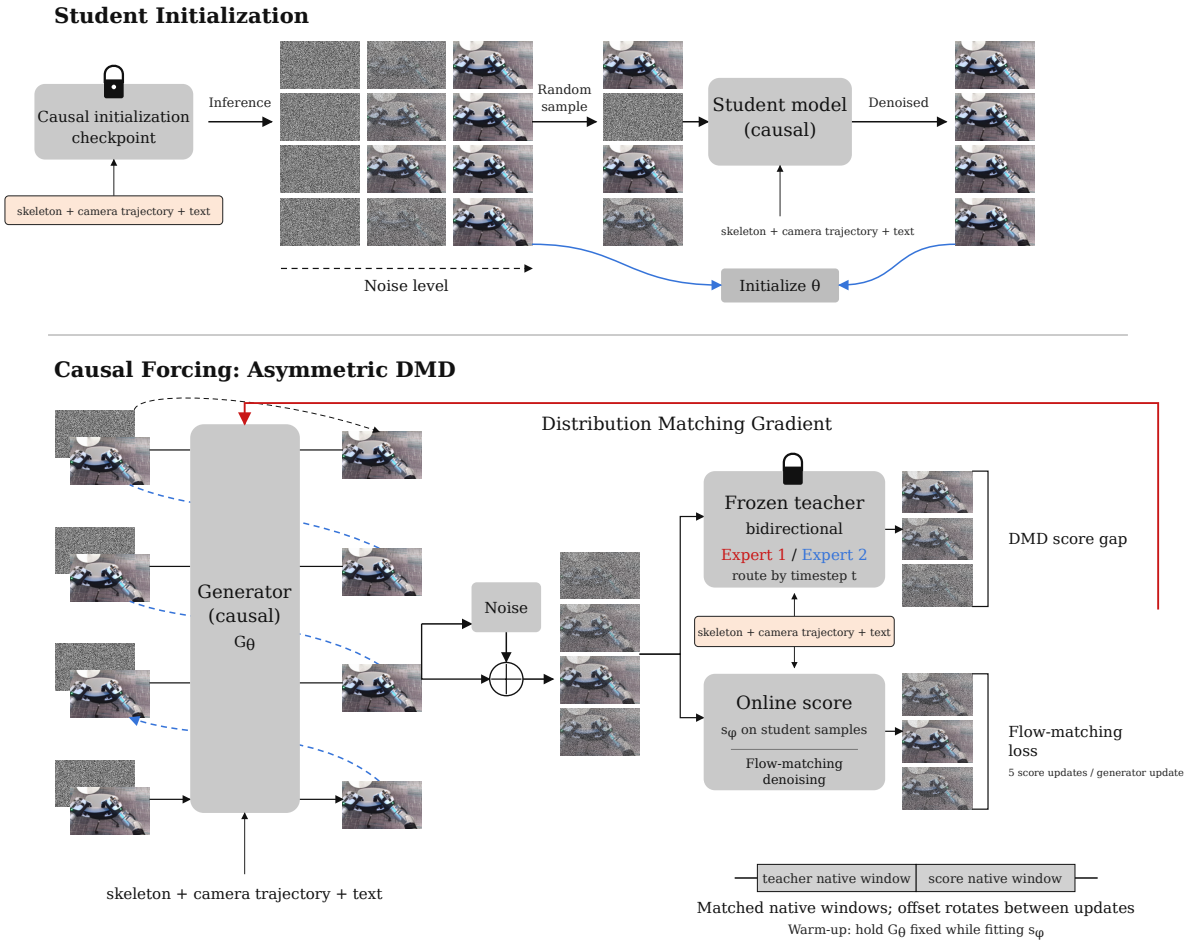


Figure S5 | **Distillation of Ropedia W0 into a few-step causal student.** Top: causal distillation initializes the few-step autoregressive generator. Bottom: asymmetric DMD updates the causal generator G_θ using a frozen bidirectional real-score teacher and an online-trained bidirectional fake-score model. The diagram illustrates the A14B MoE backbone; the 5B model follows an analogous procedure with a dense backbone. LongLive subsequently extends the few-step model for long-video streaming (Fig. S6).

Interactive inference and throughput. On a single GPU at 640 × 480, the 4-step student generates a 125-frame clip in about 2 s of autoregressive generation and about 6 s including VAE decoding (≈21 fps), compared with about 41 s for the 50-step teacher (a 7× end-to-end speed-up), and streams its first block within roughly 1.4 s.

Figure S6 illustrates the streaming rollout mechanism; Table S1 reports quality and throughput across rollout lengths.

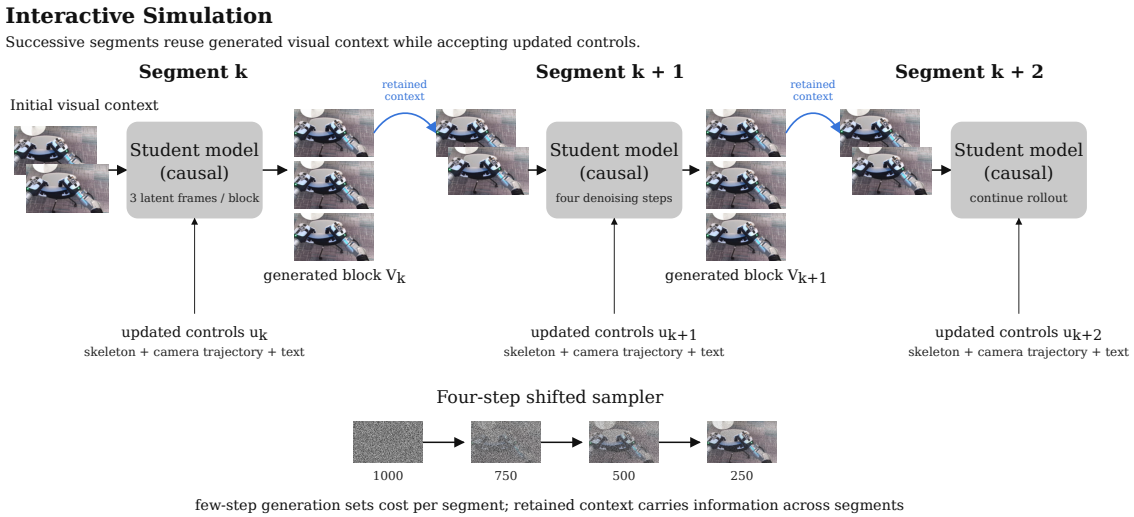


Figure S6 | **LongLive-based streaming with the distilled student.** Successive segments reuse the previously generated visual context while accepting updated controls (skeleton, camera trajectory, and text). Each block is produced by the few-step causal student, so the per-segment cost is bounded while retained context carries information across segments, enabling real-time long-horizon rollout.

Table S1 | **Distillation quality–speed comparison** (GR1, 5B). The 4-step student matches or exceeds the 50-step teacher at every rollout length (bold) while running 7× faster.

Model	Rollout	FPS↑	PSNR↑	SSIM↑	LPIPS↓
Teacher (50-step)	6.3 s	3.0	15.95	0.528	0.248
	19 s		14.34	0.454	0.321
	32 s		14.13	0.438	0.348
	61 s		13.64	0.414	0.385
Student (4-step)	6.3 s	**21.3**	**16.12**	**0.538**	**0.247**
	19 s		**14.70**	**0.468**	**0.312**
	32 s		**14.44**	**0.451**	**0.338**
	61 s		**13.98**	**0.428**	**0.373**

B.3. Wide-FOV Skeleton Conditioning

Head cameras often crop the manipulator. On the WorldArena [99] (RoboTwin 2.0) test set, the narrow skeleton touches the image border in 99.4% of frames; a 1.5× field-of-view render raises in-frame manipulator coverage from 0.37 to 0.56. The optional wide-FOV branch (Sec. 4.2) feeds this render through per-block cross-attention, providing arm configuration that is absent from the narrow view.

Figures S7 and S8 compare predictions without and with this branch against ground truth when the arm works near or beyond the frame edge. Columns identify the variants; the yellow box marks the head-camera field of view within the 1.5× canvas. Both examples report per-episode border-band keypoint error.

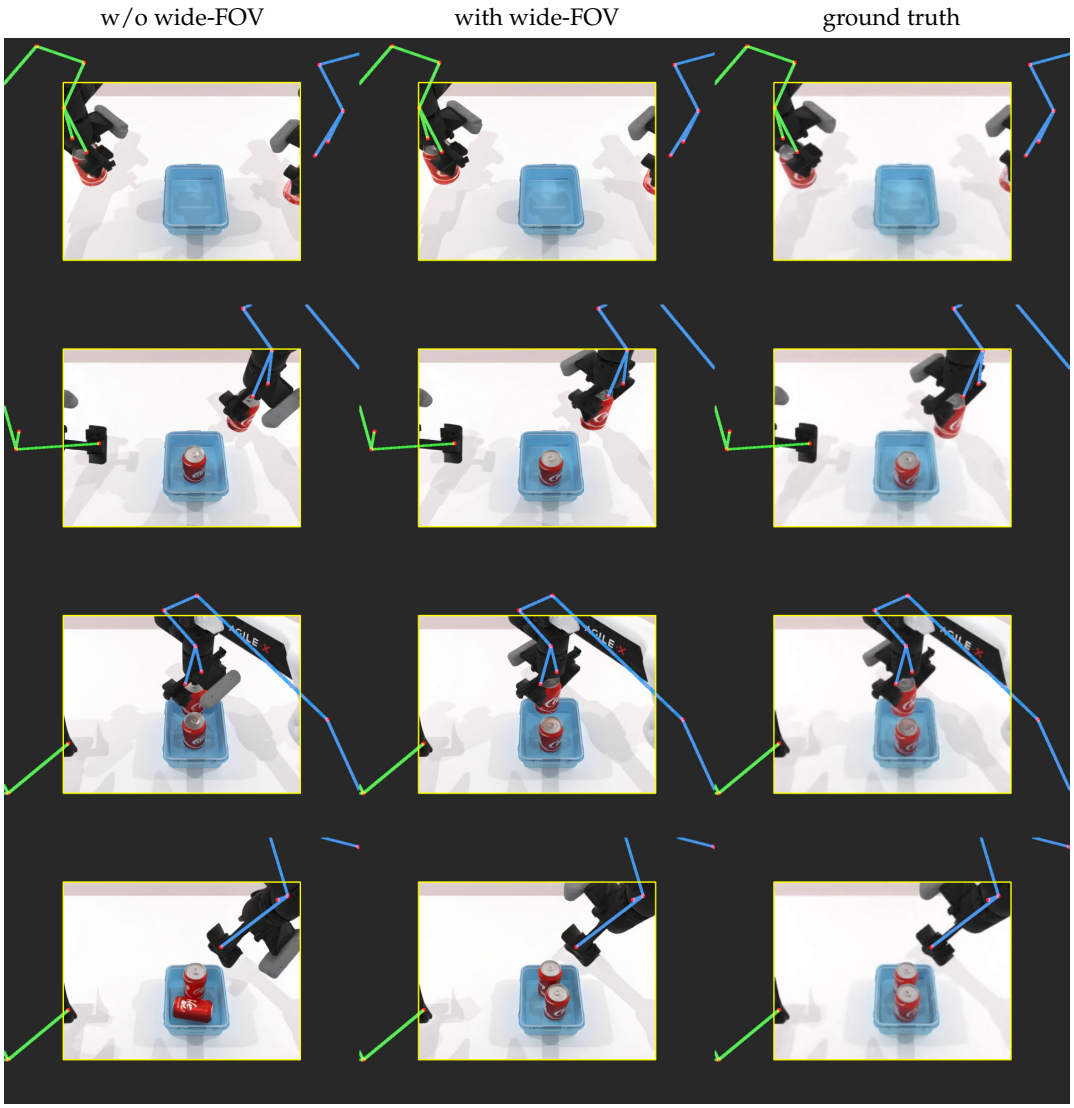


Figure S7 | **Wide-FOV skeleton branch, example 1** (four moments in time order; columns labeled above). Wide-FOV skeleton conditioning improves the alignment of the visible arm near the image boundary. Here the outer border-band keypoint error drops from 8.9 to 6.9 px, and the wide branch is closer to the ground truth on 91% of frames.

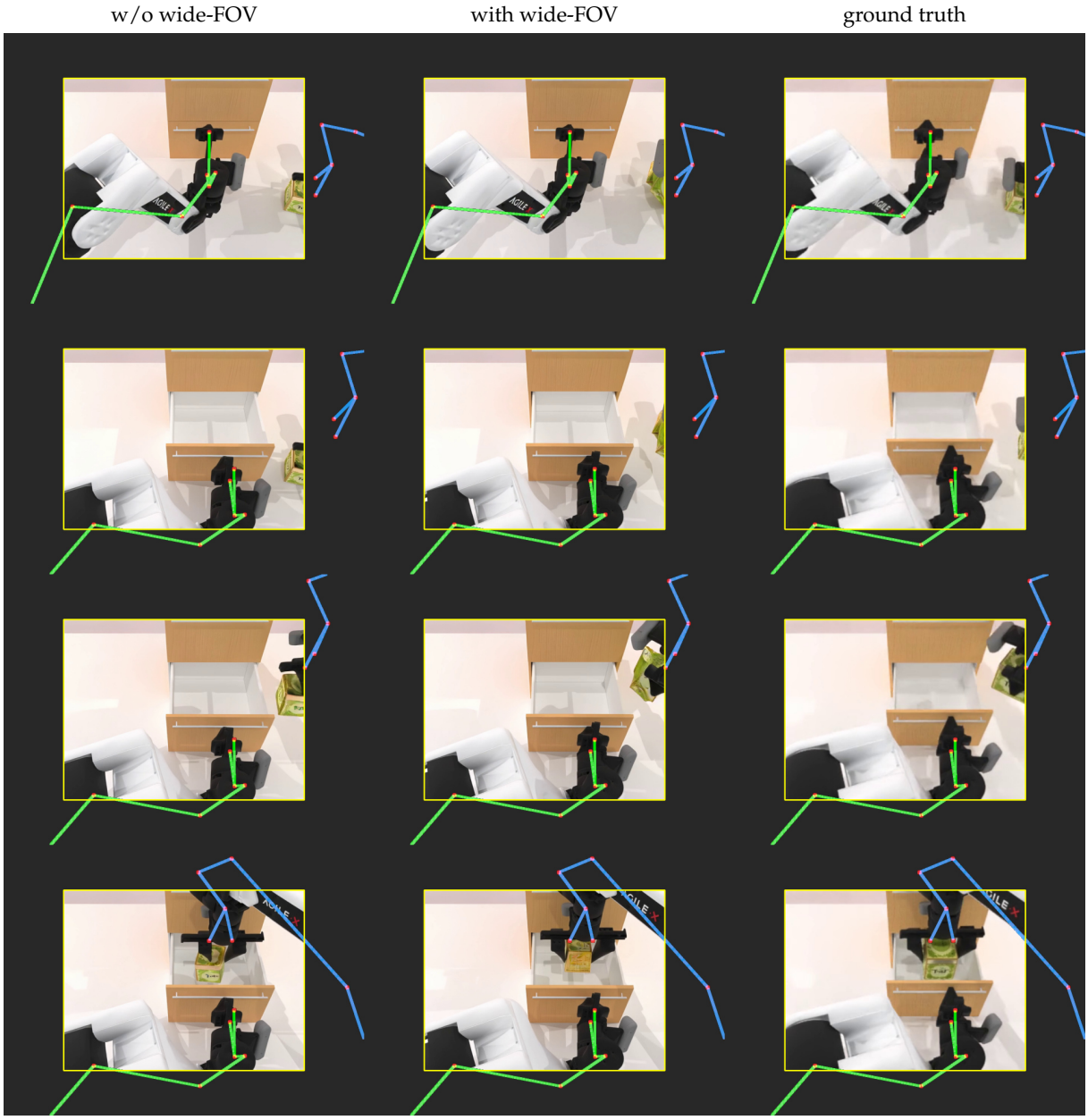


Figure S8 | **Wide-FOV skeleton branch, example 2** (four moments in time order; columns labeled above). When the arm extends beyond the camera view, wide-FOV skeleton conditioning improves the prediction of its visible portion. Here the border-band keypoint error drops from 14.0 to 11.7 px, with the wide branch closer to the ground truth on 93% of frames.

C. Evaluation Protocols

We specify the human-video evaluation procedure and the protocol used for the scaling curves. Supplementary benchmark results are collected in Sec. D.

C.1. Human Egocentric Video Evaluation

The pooled 200-clip results (100 per dataset from EgoSuite [68] and OpenAoE [66]) are reported in the main text (Table 2). Here we specify the evaluation-set composition, protocol, and metric definitions; the per-dataset breakdown is reported in Sec. D.1.

Evaluation set. We evaluate on 100 clips per dataset (200 total), sampled across high, medium, and low camera-motion strata.

Evaluation protocol. Clips contain 81 frames at 640×480 and 16 fps. Methods receive the same initial frame and, when supported, the same caption, with seed 0. Camera and hand conditions derive from the same camera annotations and projected hand keypoints, converted to each method's native interface. Each baseline is run under its native conditioning interface: DreamDojo with latent actions, Cosmos3 with joint camera-and-hand conditioning.

Metrics. ATE is absolute trajectory error after rigid SE(3) alignment without scale fitting; RPE measures stepwise translation error, and RRE measures rotation error relative to the first frame. Camera poses are estimated from generated videos with DA3 [69]. Hand metrics use MediaPipe [79] detections at confidence 0.3, matched to projected annotations. PCK@.1 uses a threshold of 10% of the image diagonal and penalizes missed detections; PCKm@.1 evaluates matched hands only, and Det is the detection rate. Cons., Dyn, and Aes denote VBench subject consistency, dynamic degree, and aesthetic quality. FVD_{16} uses non-overlapping 16-frame windows, while FVD_{full} uses the full sequences. FVD is recomputed for each evaluation set rather than averaged across datasets.

In the human evaluation tables, ATE and RPE are reported in centimeters, RRE in degrees, HandL2 in pixels, and PSNR in dB. PCK, PCKm, Det, Cons., Dyn, and Aes are reported as percentages; SSIM, LPIPS, and FVD have no physical units.

C.2. Scaling Protocol

The scaling figures appear in the main text: held-out human video at hand-detection confidence 0.3 (Fig. 6), human-to-robot transfer to Ψ_0/Xp-Teleop/ω-HOME (Fig. 7), and the DreamDojo cross-evaluation sets (Fig. 8). We detail the shared protocol here. Each point is a checkpoint taken at the indicated amount of human-video pre-training data (hours of video processed), with the data mixture held fixed. Because training stays within a single epoch, the hours processed at each point equal the amount of *unique* human data seen, so these curves measure data scaling rather than repeated exposure. For held-out human video the checkpoint is evaluated zero-shot, whereas for the DreamDojo and transfer benchmarks it is post-trained on the target with the same recipe before evaluation.

For transfer scaling, post-training uses one clip for 50 steps on Ψ_0 and 144 clips for five epochs on Xp-Teleop.

D. Additional Quantitative Results

The main benchmark summaries appear in Tables 4, 5, and 2. Here we provide per-dataset human results, seen/unseen robot breakdowns, long-horizon comparisons, ablations, and the WorldArena comparison against BWM. Distillation quality and throughput are reported alongside the implementation in Sec. B.2.

D.1. Human Evaluation by Dataset

Table S2 separates the pooled human-video results in Table 2 into EgoSuite and OpenAoE. The evaluation protocol and metric units are specified in Sec. C.1.

Table S2 | **Per-dataset results on EgoSuite [68] and OpenAoE [66].** Each set contains 100 clips from Table 2. ATE is in cm and PSNR in dB. Best and second-best bold and underlined.

Method	EgoSuite (100)			OpenAoE (100)		
	ATE↓	PSNR↑	FVD_{16}↓	ATE↓	PSNR↑	FVD_{16}↓
Cosmos3-Nano (16B)	1.122	19.75	111.5	2.740	17.56	139.9
Cosmos3-Super (64B)	1.322	19.55	110.4	2.740	17.16	<u>132.7</u>
OSCAR (2B)	2.586	16.87	173.7	4.374	15.54	221.8
RynnWorld-Teleop (5B)	2.642	17.98	182.4	4.444	15.90	200.3
DreamDojo (2B)	1.892	17.94	173.8	3.412	15.57	241.1
DreamDojo (14B)	1.890	18.95	178.8	3.414	16.56	207.9
Ropedia W0 (5B), pretrain	1.012	20.45	109.8	2.574	17.35	140.0
Ropedia W0 (5B), midtrain	1.038	20.38	<u>103.0</u>	2.729	17.43	139.4
Ropedia W0 (A14B), pretrain	**0.872**	**21.48**	**94.8**	**2.212**	<u>18.07</u>	**125.8**
Ropedia W0 (A14B), midtrain	<u>0.988</u>	<u>20.51</u>	114.4	<u>2.511</u>	**18.17**	135.8

D.2. Seen and Unseen Robot Evaluation

Table S3 breaks down the main dexterous-humanoid results (Table 4) into seen and unseen subsets for Ψ_0, ω-HOME, and Xp-Teleop, under no post-training, few-shot, and full-data adaptation.

Table S3 | **Seen/unseen breakdown on Ψ_0 [117], ω-HOME [65], and Xp-Teleop (PSNR↑).** Δ is the seen−unseen gap. Best bold, second-best underlined.

Model / initialization	No post-training			Few-shot			Full-data		
	Seen	Unseen	Δ	Seen	Unseen	Δ	Seen	Unseen	Δ
Ψ_0									
DreamDojo (14B)	14.90	13.99	0.91	14.85	14.02	0.83	14.13	13.43	0.70
Cosmos3-Nano (16B)	13.83	13.65	0.18	15.04	14.37	0.67	17.10	16.00	1.10
Cosmos3-Super (64B)	13.82	13.73	0.09	15.27	14.60	0.67	16.23	15.22	1.01
Ropedia W0 (5B), pretrain	15.34	15.01	0.33	15.78	15.39	0.39	16.67	15.85	0.82
Ropedia W0 (5B), midtrain	15.45	14.99	0.46	15.75	15.44	0.31	16.68	16.13	0.55
Ropedia W0 (A14B), pretrain	15.96	15.78	0.18	17.90	17.77	0.13	19.24	18.64	0.60
Ropedia W0 (A14B), midtrain	**17.39**	**17.33**	0.06	**18.58**	**17.98**	0.60	**19.25**	**18.90**	0.35
ω-HOME									
DreamDojo (14B)	15.41	13.98	1.43	15.14	13.74	1.40	17.16	15.17	1.99
Cosmos3-Nano (16B)	16.79	15.69	1.10	17.88	16.60	1.28	19.79	17.89	1.90
Cosmos3-Super (64B)	16.45	15.35	1.10	17.46	16.22	1.24	19.25	17.67	1.58
Ropedia W0 (5B), pretrain	18.18	16.52	1.66	18.87	17.83	1.04	19.79	18.26	1.53
Ropedia W0 (5B), midtrain	18.52	17.13	1.39	18.89	17.71	1.18	20.00	18.50	1.50
Ropedia W0 (A14B), pretrain	18.87	17.93	0.94	21.06	19.81	1.25	22.97	20.01	2.96
Ropedia W0 (A14B), midtrain	**19.90**	**19.21**	0.69	**21.50**	**19.92**	1.58	**23.29**	**20.25**	3.04
Xp-Teleop									
DreamDojo (14B)	14.60	12.39	2.21	13.98	12.01	1.97	13.93	11.90	2.03
Cosmos3-Nano (16B)	14.72	12.66	2.06	15.14	13.89	1.25	15.74	14.22	1.52
Cosmos3-Super (64B)	14.90	12.80	2.10	15.15	13.64	1.51	15.63	13.99	1.64
Ropedia W0 (5B), pretrain	15.40	13.64	1.76	16.37	14.63	1.74	16.76	15.05	1.71
Ropedia W0 (5B), midtrain	15.84	14.11	1.73	16.37	14.68	1.69	16.75	15.03	1.72
Ropedia W0 (A14B), pretrain	16.09	13.80	2.29	18.85	17.11	1.74	19.18	17.34	1.84
Ropedia W0 (A14B), midtrain	**17.25**	**15.59**	1.66	**19.15**	**17.35**	1.80	**19.61**	**17.87**	1.74

D.3. Long-Horizon Rollout Evaluation

Tables S4 and S5 compare autoregressive rollouts against the baselines: approximately 32 s zero-shot on ω-HOME and 30 s on Xp-Teleop. The tables report aggregate quality and how prediction quality changes over the rollout.

Table S4 | **Long-horizon rollout on ω-HOME [65]** (≈32 s, zero-shot; 12 chained windows). Overall metrics are frame-averaged; per-segment PSNR at ≈1/14/31 s shows degradation over the rollout. Best **bold**, second-best underlined.

Model / initialization	Overall			Per-segment PSNR↑		
	PSNR↑	SSIM↑	LPIPS↓	seg. 1	seg. 6	seg. 12
DreamDojo (14B)	14.17	0.641	0.579	22.32	13.36	13.02
Cosmos3-Nano (16B)	13.91	0.573	0.576	25.57	12.75	11.11
Cosmos3-Super (64B)	13.74	0.579	0.570	25.13	12.58	11.14
Ropedia W0 (5B), pretrain	13.64	0.651	0.516	23.92	12.57	10.35
Ropedia W0 (5B), midtrain	13.93	0.664	0.506	26.33	12.74	10.43
Ropedia W0 (A14B), pretrain	15.20	0.696	0.478	25.03	14.03	12.40
Ropedia W0 (A14B), midtrain	**16.48**	**0.738**	**0.376**	29.33	15.20	12.35

Table S5 | **Long-horizon rollout on Xp-Teleop** (≈30 s; 11 autoregressively chained windows, each restarting from the model's own last frame). Best **bold**, second-best underlined; Dyn. is best when closest to the ground-truth value (0.70).

Model / initialization	Hand following			Fidelity			Video quality		
	Err. (px)↓	Miss↓	Err.@120↓	PSNR↑	SSIM↑	LPIPS↓	B-Cons.	Aes.↑	Dyn.
DreamDojo (14B)	62.6	77.1%	106.9	10.00	0.301	0.620	0.891	0.449	0.55
Cosmos3-Nano (16B)	55.8	50.9%	88.5	11.10	0.338	0.579	0.892	0.500	0.55
Cosmos3-Super (64B)	52.5	61.5%	94.0	10.76	0.302	0.587	0.884	0.478	0.50
Ropedia W0 (5B), midtrain	19.8	45.3%	65.2	10.83	0.336	0.573	0.892	0.469	0.70
Ropedia W0 (A14B), midtrain	**12.0**	**6.9%**	**19.5**	**13.47**	**0.433**	**0.444**	**0.906**	**0.510**	0.70

D.4. Conditioning and Pretraining Ablations

Table S6 evaluates the camera and skeleton conditioning channels on the cross-embodiment set. Table S7 compares our pretrained initialization with public video-generation weights at matched post-training data budgets. Qualitative examples of the initialization comparison are grouped in Sec. E.3.

Table S6 | **Cross-embodiment set** (400 evaluation windows over eight embodiments). Cosmos3 [1] is fine-tuned on the same robot mixture used for Ropedia W0's mid-training; neither model is post-trained on the evaluation windows. Each ablation variant is trained without one conditioning channel. Best bold, second-best underlined.

Model	Variant	PSNR↑	SSIM↑	LPIPS↓
Cosmos3	Nano (16B)	19.02	0.739	0.197
	Super (64B)	18.77	0.736	0.202
Ours 5B	Full	19.83	0.760	0.167
	w/o Plücker	19.71	0.755	0.178
	w/o skeletons	18.03	0.736	0.214
Ours A14B	Full	**21.98**	**0.802**	**0.128**
	w/o Plücker	21.65	0.791	0.145
	w/o skeletons	18.17	0.744	0.208

Table S7 | **Effect of pretraining.** Our pretrained base vs. the same architecture from public video-generation weights ("official base"), post-trained on 1%/10%/100% of the GR-1 data; best bold.

Data	Base	In-lab			EgoDex			DreamDojo-HV		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
1%	Official base	16.04	0.495	0.446	13.42	0.394	0.548	12.43	0.348	0.548
	Ropedia W0	**20.85**	**0.764**	**0.213**	**20.57**	**0.778**	**0.208**	**19.05**	**0.760**	**0.217**
10%	Official base	18.89	0.630	0.287	18.64	0.687	0.267	16.95	0.623	0.307
	Ropedia W0	**21.02**	**0.786**	**0.206**	**20.78**	**0.801**	**0.199**	**19.29**	**0.787**	**0.207**
100%	Official base	19.41	0.650	0.275	18.87	0.695	0.265	17.48	0.645	0.293
	Ropedia W0	**21.74**	**0.797**	**0.202**	**21.23**	**0.810**	**0.196**	**19.81**	**0.795**	**0.204**

D.5. WorldArena Comparison against BWM

We compare Ropedia W0 (A14B) with the Boundless World Model (BWM) [17] on WorldArena [99] Track 1 (RoboTwin 2.0 [23], `fixed_scene_task` test set, 1000 episodes). Both models use their official code and public checkpoints, the same official evaluation pipeline at 640×480, and the same Real-ESRGAN [114] super-resolved first frame ("BWM + SR"). We report our original-length run.

Ropedia W0 leads on 12 of the 15 metrics and by +5.9 EWMScore over BWM + SR (Table S8). The margin is dominated by **Trajectory Accuracy** (+29.1: the arm being where the commanded action says it should be), followed by background/subject consistency and instruction following. Fig. S9 shows this qualitatively.

Table S8 | **WorldArena [99] Track 1 comparison against BWM [17]** (RoboTwin 2.0 [23], 1000 episodes; identical official pipeline and inputs for both). EWMScore is the mean of the 15 metrics; best in bold.

Metric	BWM + SR	Ropedia W0 (A14B)
EWMScore	61.01	**66.92**
Image Quality	51.54	**53.57**
Aesthetic Quality	**43.09**	41.04
JEPA Similarity	90.93	**96.00**
Dynamic Degree	17.49	**25.48**
Flow Score	8.49	**15.75**
Motion Smoothness	66.13	**67.55**
Subject Consistency	68.08	**79.51**
Background Consistency	71.78	**84.17**
Photometric Consistency	**35.95**	27.98
Interaction Quality	70.22	**77.42**
Trajectory Accuracy	34.64	**63.78**
Depth Accuracy	94.71	**98.56**
Perspectivity	96.04	**97.72**
Instruction Following	76.22	**85.50**
Semantic Alignment	**89.79**	89.73

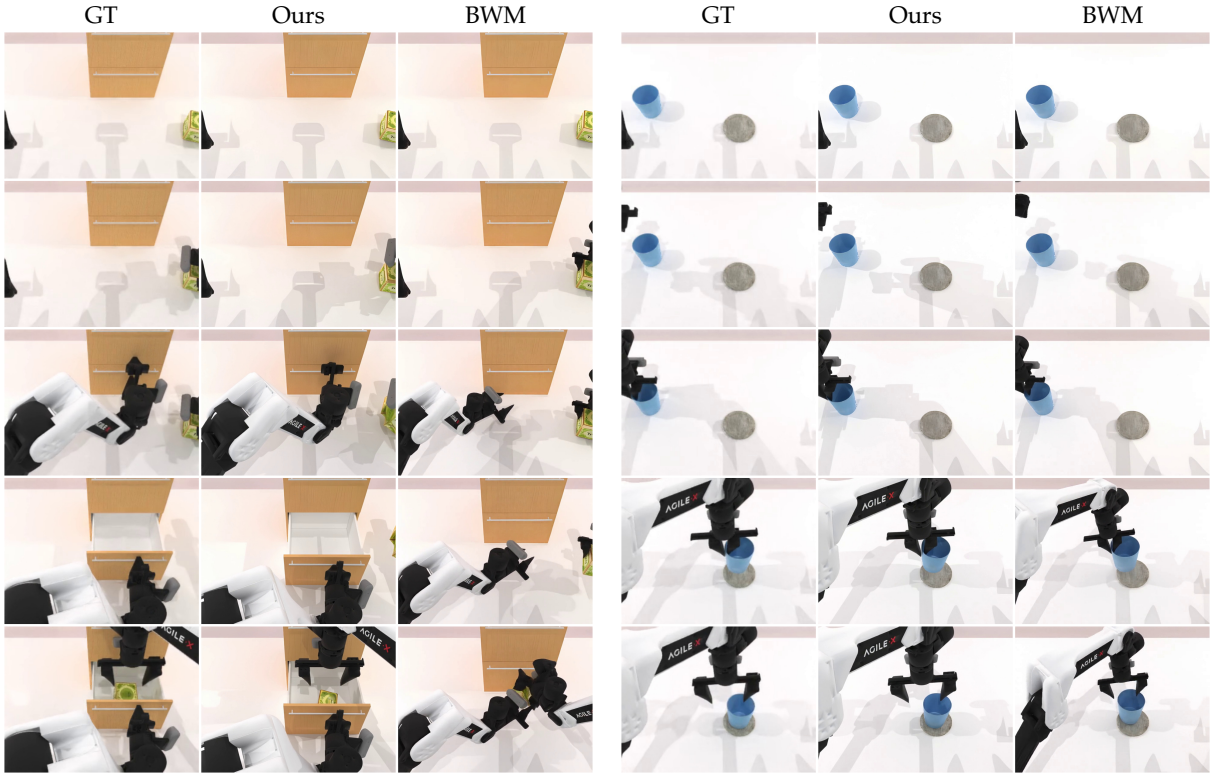


Figure S9 | **Qualitative comparison on WorldArena** (left: drawer + box; right: cup onto coaster). Each panel: five chronological frames (rows); columns are *GT* | *Ropedia W0 (A14B)* | *BWM* (Table S8). Ropedia W0 follows the commanded arm trajectory and completes the task like the ground truth (opening the drawer and placing the box, precisely seating the cup), whereas BWM's arm drifts and under-moves (it never opens the drawer).

E. Qualitative Comparisons

We group qualitative comparisons by evaluation setting. Compared methods receive the same initial frame and conditioning through their native interfaces. Ground-truth and predicted frames are aligned in time; figures from the project-page videos use the same displayed frame indices across methods.

E.1. Cross-Embodiment Robot Prediction

Figures S10–S12 compare predictions on held-out clips from Fourier, AgiBot, Baihu, Dexora, RH20T, and DROID. These examples cover different manipulators, objects, and interaction types; the captions specify the adaptation setting.

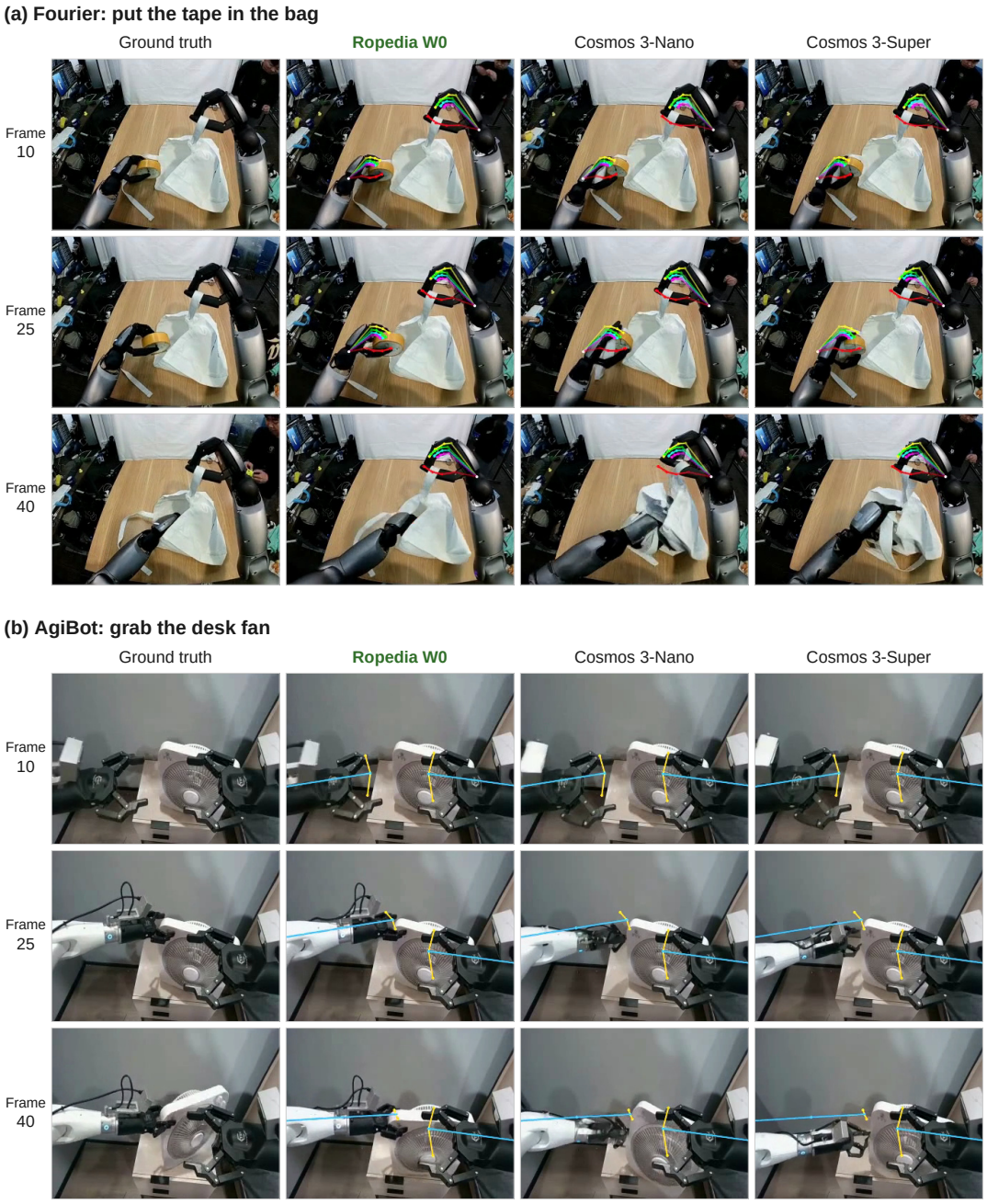


Figure S10 | **Robot-dataset comparisons (1/3).** (a) Fourier: putting tape in a bag; (b) AgiBot: grasping a desk fan. These are held-out clips: Ropedia W0 uses mid-training and Cosmos3 uses fine-tuning on the robot mixture. Colored skeletons are supplied input conditions overlaid on the generated frames.

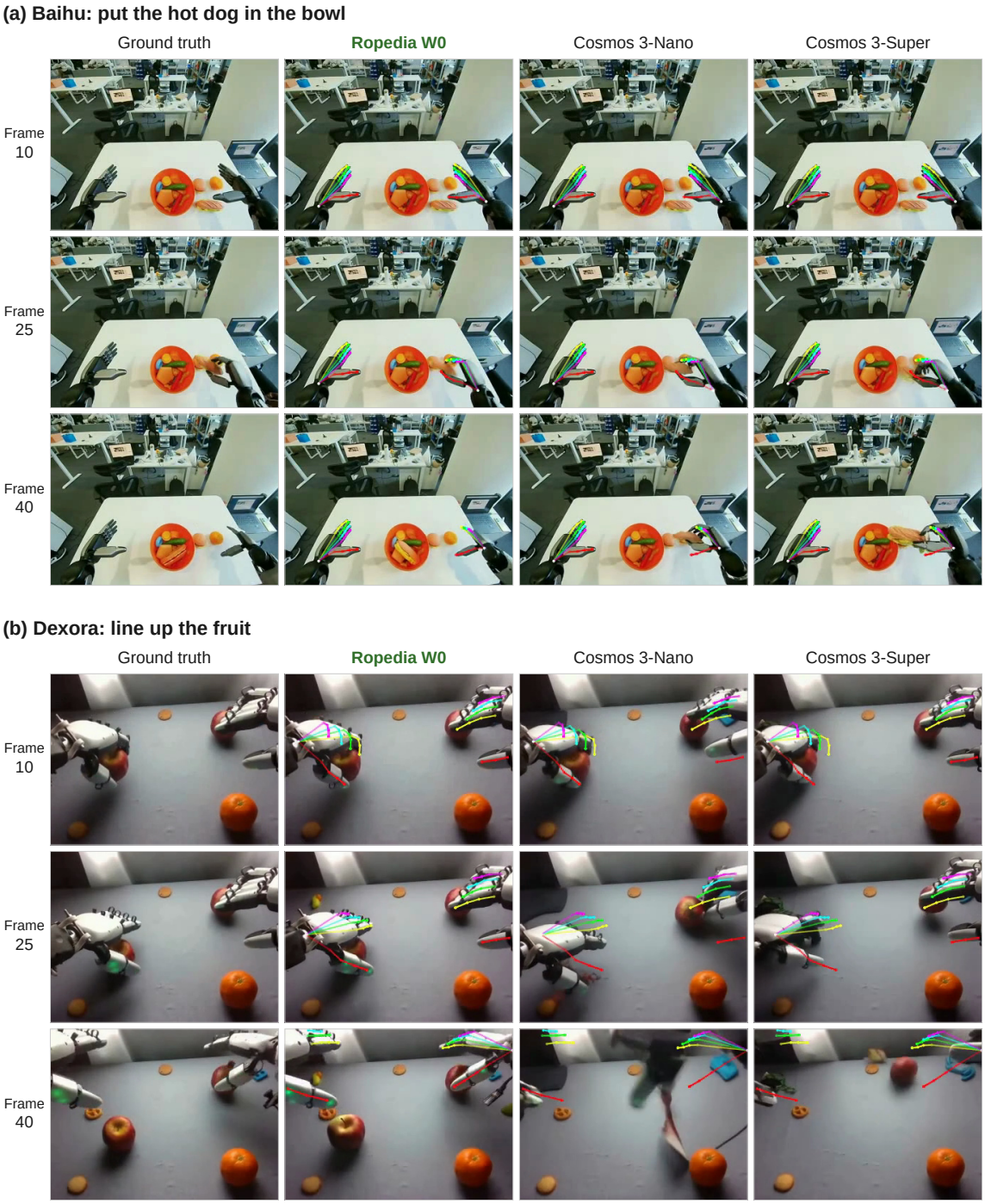


Figure S11 | **Robot-dataset comparisons (2/3).** (a) Baihu: placing a hot dog in a bowl; (b) Dexora: arranging fruit. Columns compare the ground truth, Ropedia W0, Cosmos3-Nano, and Cosmos3-Super at aligned frames. The held-out-clip setting and input skeleton overlays follow Fig. S10.

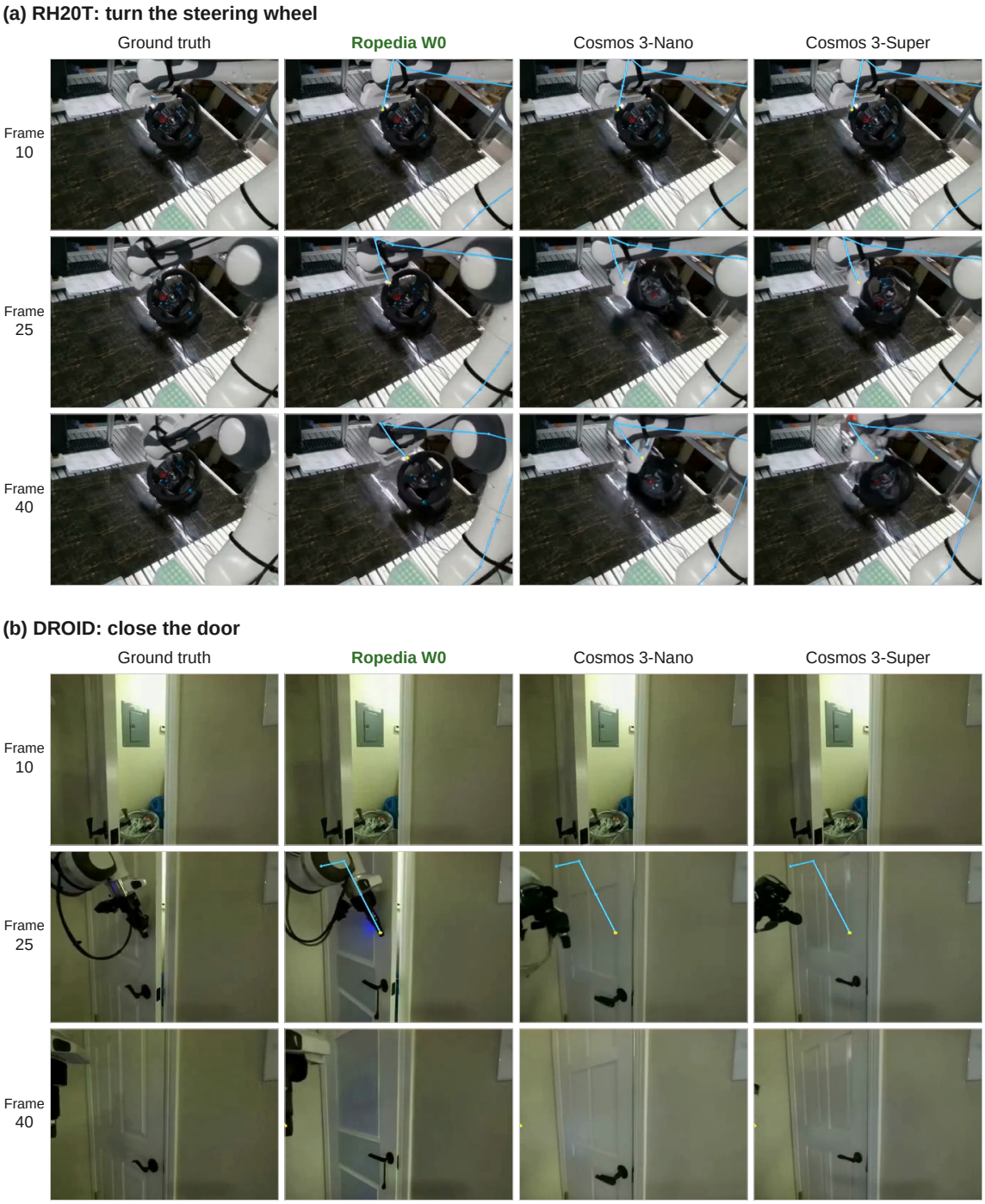


Figure S12 | **Robot-dataset comparisons (3/3).** (a) RH20T: turning a steering wheel; (b) DROID: closing a door. These examples extend the comparison to gripper motion and interactions with articulated objects. The held-out-clip setting and input skeleton overlays follow Fig. S10.

E.2. Zero-Shot Loco-Manipulation

Figures S13 and S14 compare zero-shot predictions on household tasks from ω-HOME [65], a held-out loco-manipulation dataset not used in training by the compared models. The examples show generated hand motion, object interactions, and scene appearance against time-aligned ground-truth frames.

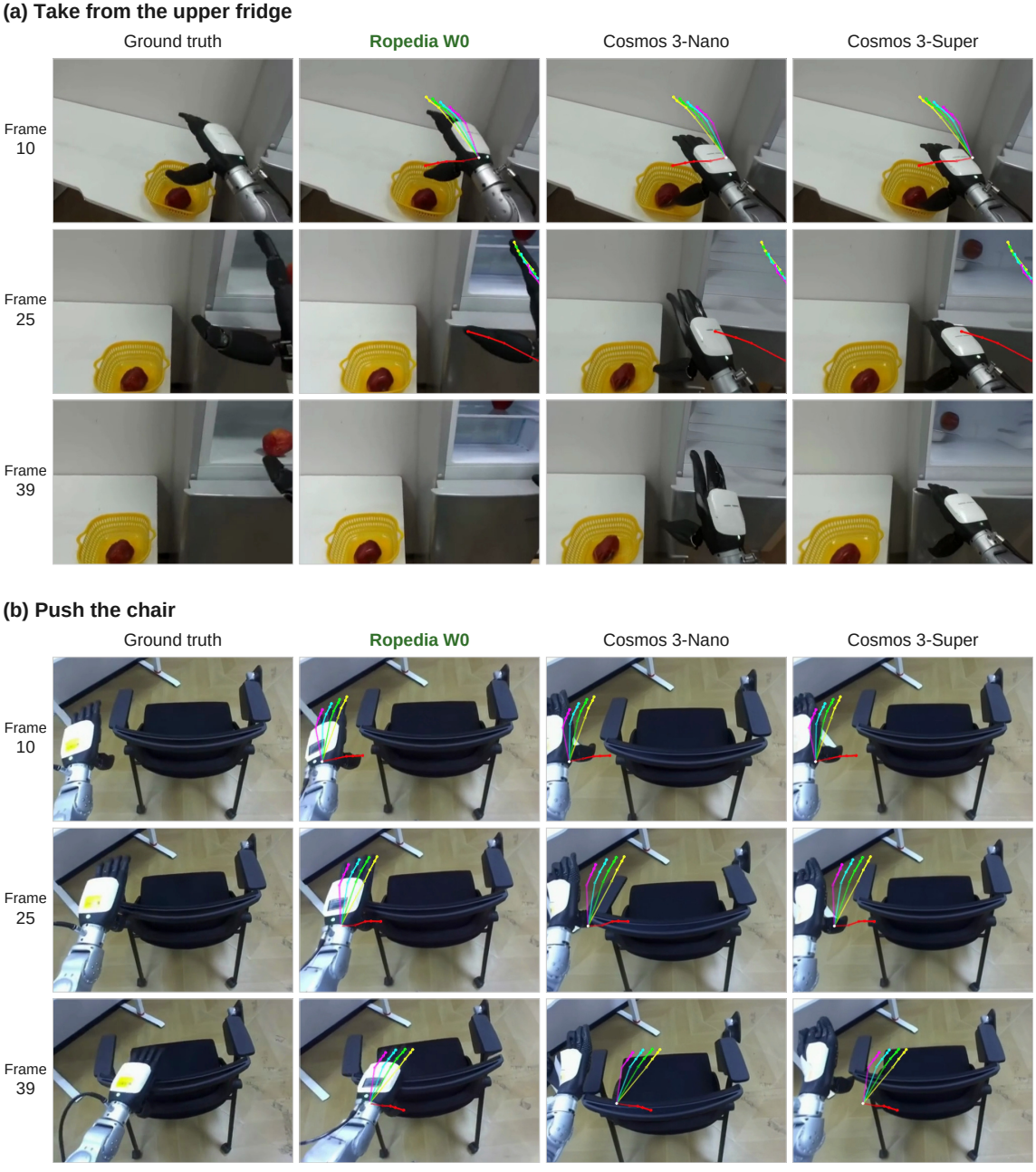


Figure S13 | **Zero-shot comparisons on ω-HOME (1/2).** (a) Reaching into the upper fridge; (b) pushing a chair. The compared models have not been trained on ω-HOME. Rows show aligned frames from the same clip; colored skeletons are input conditions overlaid on each generated video.

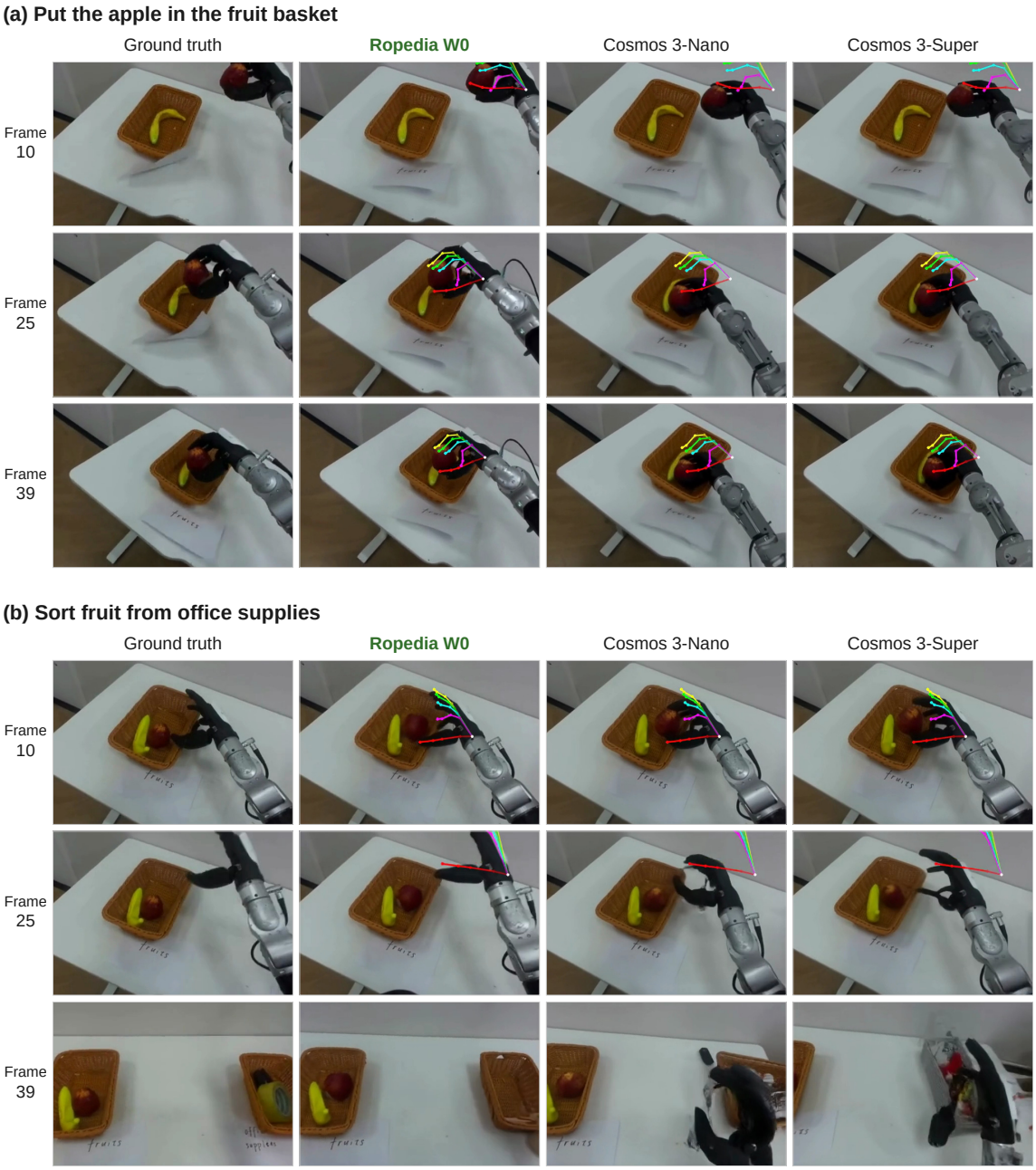


Figure S14 | **Zero-shot comparisons on ω-HOME (2/2).** (a) Putting an apple in a fruit basket; (b) sorting fruit from office supplies. Ground-truth frames provide a reference for the generated hand and object configurations. The setting and input skeleton overlays follow Fig. S13.

E.3. Post-Training and Data Efficiency

Adaptation on ω-HOME. Figure S15 compares zero-shot and post-trained predictions against the baselines on a manipulation task.

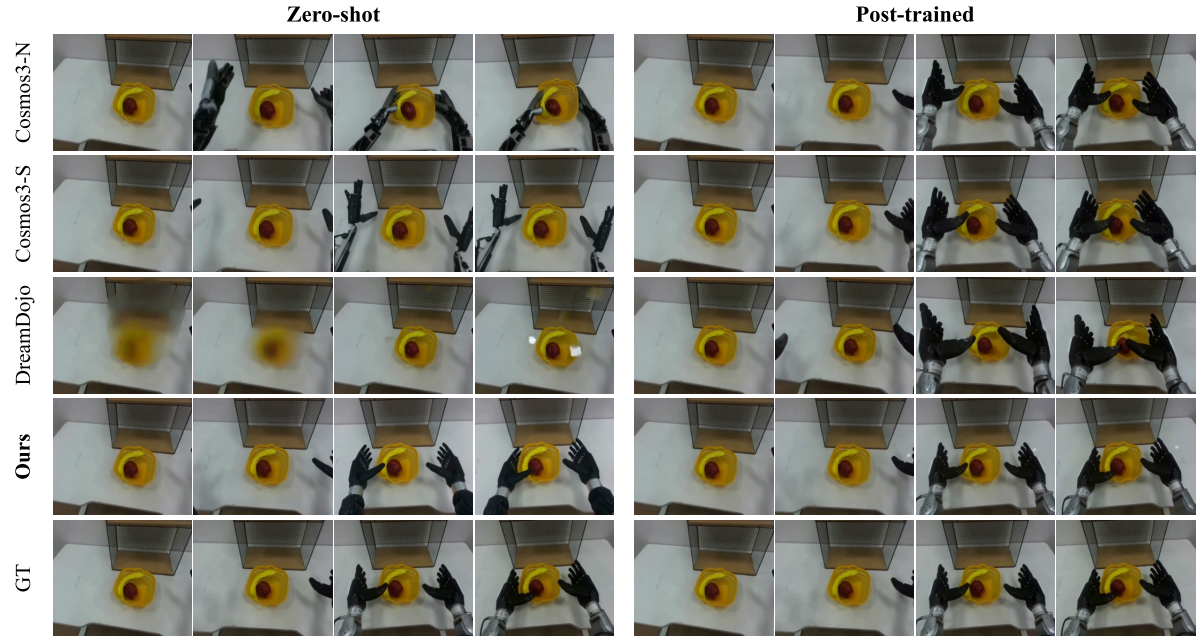


Figure S15 | **ω-HOME [65] manipulation task** (fruit-bucket arrangement). Left: zero-shot; right: after post-training. Ropedia W0 follows the commanded hands and matches the ground truth in both settings, whereas the baselines lose the arm or drift.

Pretrained initialization under limited data. We probe data efficiency by post-training on only 1% of the GR-1 data from two initializations: our pretrained base versus public video-generation weights (Wan2.2-I2V-A14B). Across both the In-lab (Fig. S16) and EgoDex (Fig. S17) settings, initializing from our pretrained base yields markedly sharper predictions and preserves object and hand detail that the public base loses.



Figure S16 | **Few-shot data efficiency on GR-1 (In-lab).** Both models are post-trained on 1% of the data; Ropedia W0 initializes from our pretrained base and Wan2.2-I2V-A14B [108] from public video-generation weights. Our base yields markedly sharper, more faithful predictions.

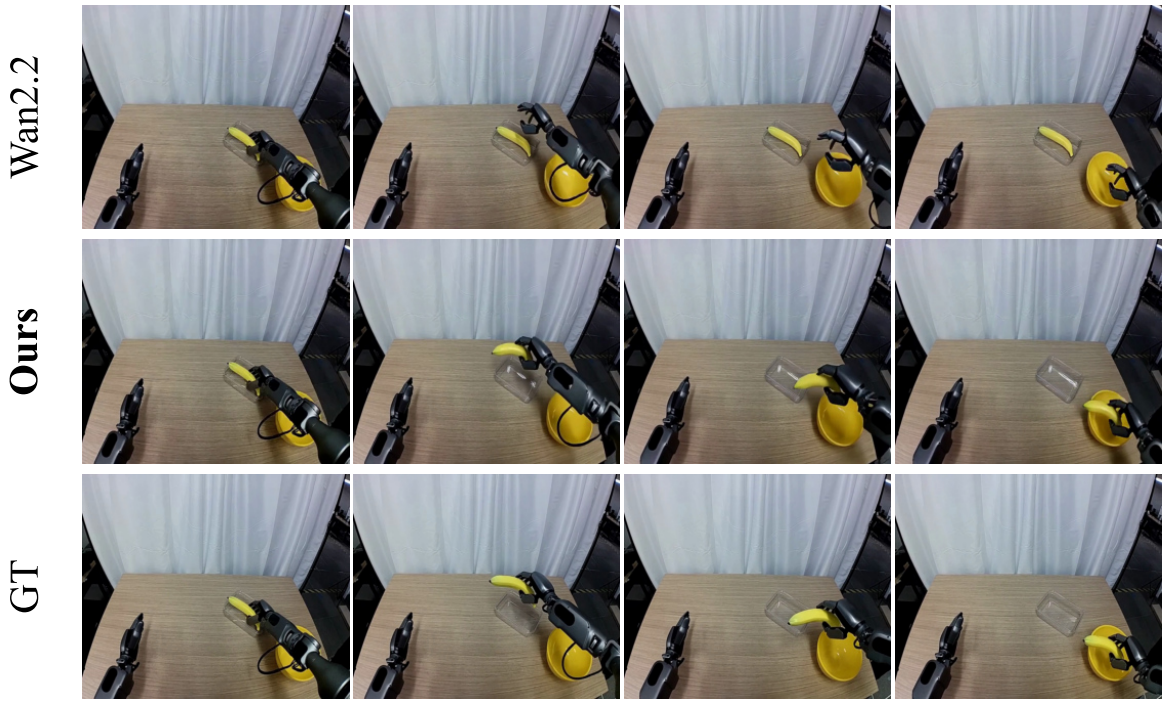


Figure S17 | **Few-shot data efficiency on GR-1 (EgoDex [46]).** Same 1% budget; initializing from our pretrained base preserves object and hand detail that the public base loses.

Post-training data budgets. Figure S18 shows the same model under zero-shot, few-shot, and one-epoch target-data post-training on two further ω-HOME tasks.

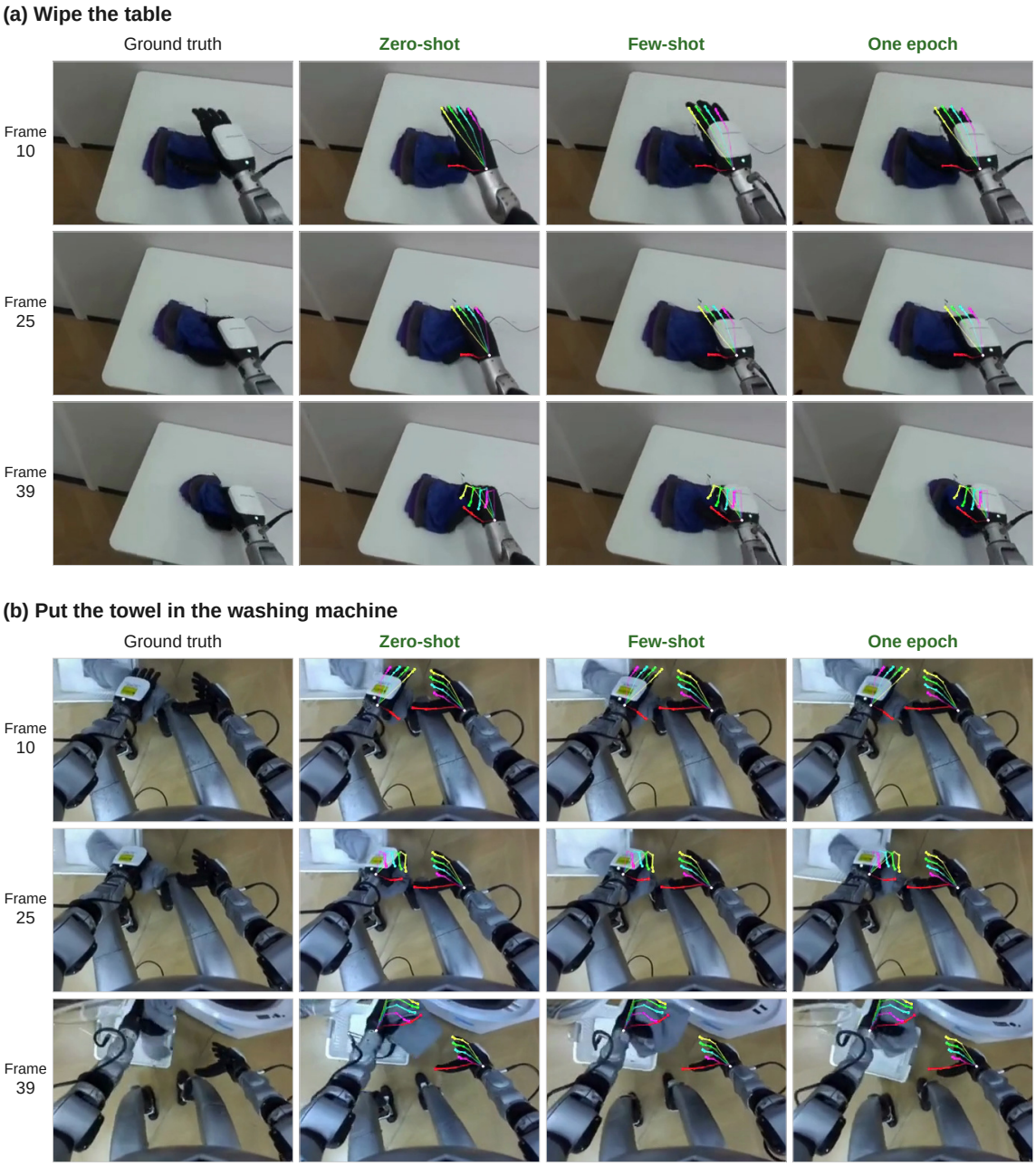


Figure S18 | **Post-training examples on ω-HOME.** (a) Wiping a table; (b) putting a towel in a washing machine. Columns show ground truth and Ropedia W0 under the three settings labeled in the project-page demonstrations: zero-shot, few-shot, and one epoch of target-data post-training. Colored skeletons are overlaid input conditions. Each row uses the same frame index across settings.

E.4. Additional Long-Horizon Rollouts

Figure S19 shows two additional 32-second generated sequences with their supplied camera and hand conditions. These examples visualize rollout continuity; the quantitative long-horizon comparisons are reported in Sec. D.3.

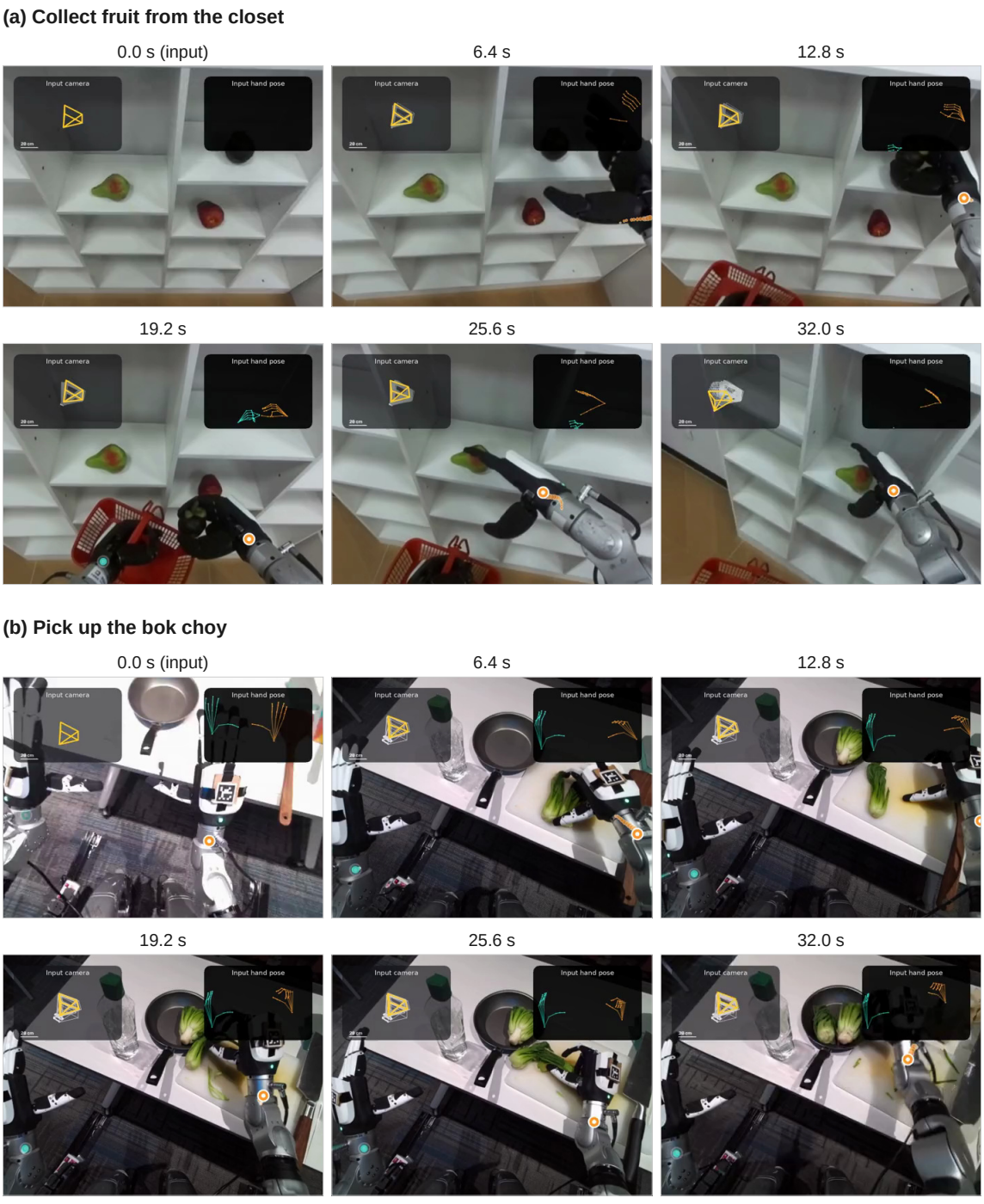


Figure S19 | **Additional 32-second generated rollouts.** (a) Collecting fruit from a closet; (b) picking up bok choy. Each sequence starts with the input image at 0 s, followed by generated frames sampled every 6.4 s, read left to right and then top to bottom. The camera and hand-pose insets display supplied conditions.

F. Limitations and Future Work

1. The distilled student supports real-time, minute-long rollouts (Sec. B.2); stronger memory mechanisms or periodic re-grounding are promising directions for extending stable interactive simulation to longer horizons.

2. We study up to a 14B-active model and scale the amount of human-video pre-training data while holding the data mixture fixed; prediction quality improves monotonically with the amount of data (Sec. C.2), suggesting that scaling data and model size further, and broadening embodiment and task coverage, are promising directions toward more capable and reliable embodied world models.

3. Our kinematic interface captures camera and manipulator motion but not physical contact or object dynamics; coupling the world model with simulation-ready, physically grounded 3D assets and scenes [18–20] is a promising direction toward modeling the underlying dynamics rather than only their visual consequences.