XPERIENCE-0: SCALING LOCO-MANIPULATION WORLD MODELS ACROSS EMBODIMENTS WITH HUMAN EXPERIENCE

Anonymous authors
Paper under double-blind review

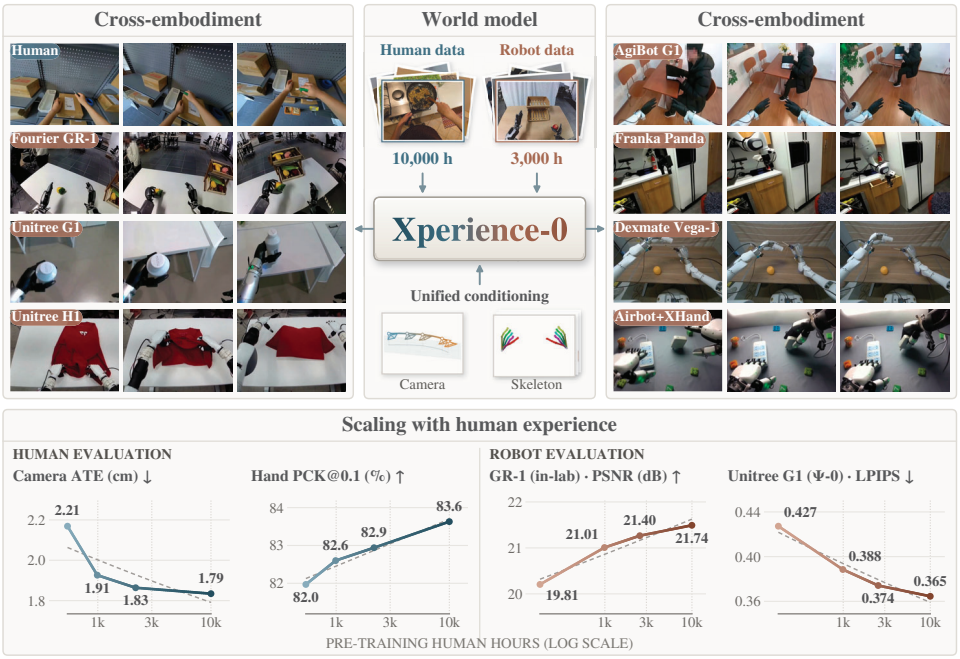


Figure 1: **Xperience-0: an egocentric world model for cross-embodiment loco-manipulation**, built on a unified kinematic representation and scaling with human experience.

ABSTRACT

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 **Xperience-0**, 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 derived from embodiment-specific forward kinematics. 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 on the DreamDojo, Ψ_0, ω-HOME, and Xp-Teleop benchmarks, 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.

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

1 INTRODUCTION

World models predict how environments evolve in response to actions (Ha & Schmidhuber, 2018; Bruce et al., 2024) and support decision-making by imagining future outcomes (Hafner et al., 2025), giving them considerable potential in embodied intelligence. However, most existing models (Wu & Gao, 2026; Xu et al., 2026; Alzayer et al., 2026) focus on fixed-base, gripper-based manipulation, leaving locomotion-induced viewpoint changes and dexterous hand–object interactions largely underexplored. Besides, generalizing across embodiments poses a further challenge. DreamDojo (Gao et al., 2026) replaces latent-action conditioning with robot-specific actions during post-training, and Cosmos 3 (Agarwal et al., 2026) 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 (Grauman et al., 2022; 2024; Hoque et al., 2026; Wang et al., 2023; Damen et al., 2022; AI, 2025; Liu et al., 2022) 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 **Xperience-0**, 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 and human review. 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 are derived from human pose annotations and robot forward kinematics, preserving embodiment-specific structure within a common image-aligned format. This conditioning interface remains consistent across human pretraining, human–robot midtraining, 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 Xperience-0 on the DreamDojo, Ψ_0, ω-HOME, AgiBot, and Xp-Teleop benchmarks, alongside held-out human videos. 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 video processed during pretraining 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 **Xperience-0**, 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 new human-in-the-loop pipeline that unifies collection, annotation, and quality checks across human and robot data, producing high-quality, full-modality annotations. 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 Xperience-0 achieves state-of-the-art cross-embodiment prediction, including on our long-horizon loco-manipulation **Xp-Teleop** benchmark. We further

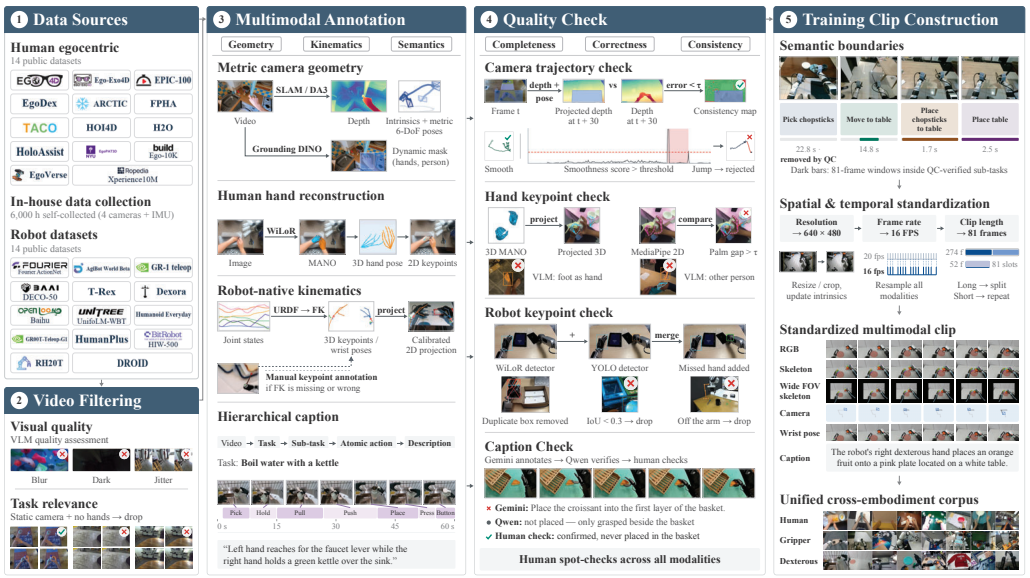


Figure 2: **Xp-Data curation pipeline.** Human and robot recordings undergo video filtering, multimodal annotation (camera trajectories, projected skeletons, hierarchical captions), and quality checks, yielding standardized training clips.

show that Xperience-0'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. Video world models support interactive environments (Parker-Holder et al., 2025) and autoregressive rollouts (Huang et al., 2025). In robotics, action conditioning ranges from embodiment-specific encodings (Wang et al., 2025a; Agarwal et al., 2026) and learned latent actions (Gao et al., 2026) to geometric renderings (Xu et al., 2026; Wu & Gao, 2026). AnyWorld (Chen et al., 2026) combines camera and skeleton controls for human-to-robot recomposition; Xperience-0 targets sustained loco-manipulation across humans, gripper robots, and dexterous humanoids through a shared interface for camera motion and embodiment-specific kinematics.

Human and robot datasets. Egocentric human datasets provide everyday activity and pose supervision (Grauman et al., 2022; Hoque et al., 2026), while robot demonstrations and teleoperation span gripper manipulation (O'Neill et al., 2024; Bu et al., 2025a) and humanoid loco-manipulation (Wei et al., 2026; Li et al., 2026b). Their heterogeneous motion, calibration, and language annotations motivate Xp-Data's unified training corpus, complemented by Xp-Teleop for evaluating long-horizon dexterous humanoid loco-manipulation.

Policy evaluation, planning, and learning. Predicted rollouts support policy evaluation (Li et al., 2025b; Guo et al., 2026) and model-based planning (Gao et al., 2026; Zhou et al., 2024). Prediction also contributes to policy learning: whereas vision-language-action models map observations and language to actions (Black et al., 2024), world action models couple action learning with future video or latent-state prediction (Ye et al., 2026; Yuan et al., 2026). We evaluate Xperience-0 for policy ranking, trajectory selection, and initialization of FastWAM for action learning (Sec. 5.3).

We summarize the closest connections above and provide an extended discussion in Appendix B.

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 (Grauman

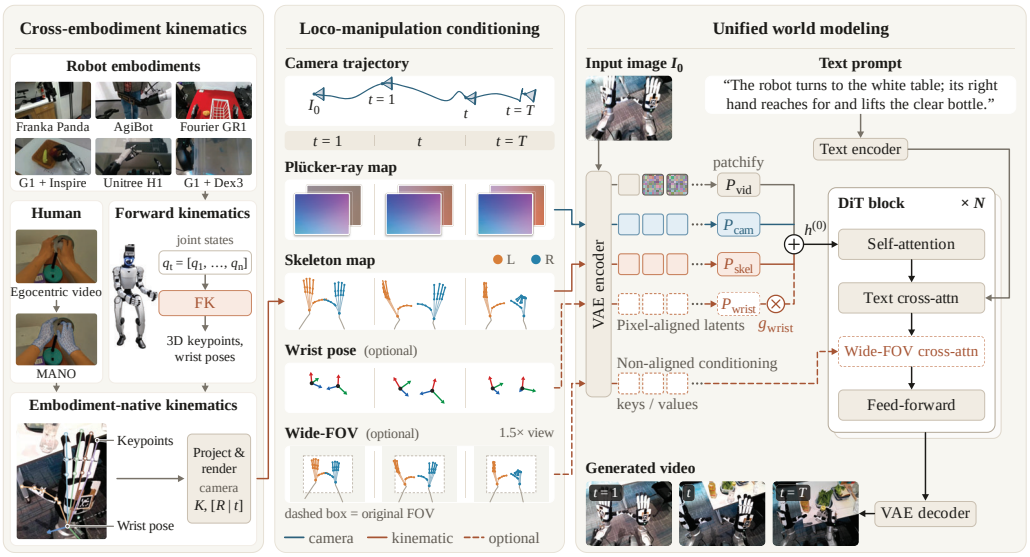


Figure 3: **Overview of Xperience-0. Left:** Human hand poses and robot forward kinematics provide embodiment-native keypoints and wrist poses for projection. **Middle:** Camera trajectories are encoded as Plucker-ray maps, while manipulation is represented by skeleton maps, optionally augmented with 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.

et al., 2022; 2024; Hoque et al., 2026; Wang et al., 2023; Damen et al., 2022; AI, 2025; Liu et al., 2022; Fang et al., 2024b; Fan et al., 2023; Kwon et al., 2021; Liu et al., 2024b; Garcia-Hernando et al., 2018; Punamiya et al., 2026; Ropedia, 2026; Bu et al., 2025a; Khazatsky et al., 2024; Fang et al., 2024a; Zhang et al., 2026b; Team & Mu, 2025; Bjorck et al., 2025; Fu et al., 2024; BitRobot et al., 2026; Li et al., 2026a; Niu et al., 2026; Hua et al., 2026; Unitree Robotics, 2026; Zhao et al., 2025) 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, we design a **human-in-the-loop data pipeline** that unifies data collection, multimodal annotation, and quality check (Fig. 2). The pipeline applies to both human and robot data and produces high-quality, full-modality annotations, including camera trajectories, hand skeletons, and captions. By integrating automated models with targeted human verification, it ensures annotation quality while substantially improving annotation efficiency. With this pipeline, we first clean and re-annotate 14 public egocentric human datasets and 13 public robot datasets spanning grippers and dexterous hands, converting heterogeneous raw recordings into training-ready data. To further push the boundary of model capabilities, we additionally collect over 6,000 hours of multi-view egocentric recordings captured in-house using a rig with four fisheye cameras with high-quality full-modality annotations. Together, these efforts yield Xp-Data, comprising over 10k hours of human data and 3k hours of robot data for model training, where every clip pairs egocentric video with temporally aligned camera trajectories, manipulator skeletons, and captions. In addition, Xp-Teleop provides about one held-out hour of long-horizon dexterous humanoid teleoperation with both locomotion and manipulation, in the same schema, recorded on an off-the-shelf rig (Unitree Robotics, 2024) (Sec. E; comparison in Table S2).

4 Model

Xperience-0 combines a unified kinematic control interface with a video diffusion model for cross-embodiment egocentric loco-manipulation prediction (Fig. 3). 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 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 $c_{1:T}$, Xperience-0 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

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

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, while robot keypoints are computed from joint states using embodiment-native URDF forward kinematics. For robot embodiment e, keypoint j is projected into the camera view as

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

where $\mathcal{F}_j^{(e)}$ maps joint states $\mathbf{q}_t$ to a keypoint position in the robot base frame, $(\mathbf{R}_t^{\text{c}\leftarrow\text{b}}, \mathbf{t}_t^{\text{c}\leftarrow\text{b}})$ transforms it into the current camera frame, and π denotes perspective division. We render the projected 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

Xperience-0 adopts a latent video diffusion transformer conditioned on the initial observation, text, and kinematic controls. 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 Xperience-0 in two stages: human pretraining 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 pretraining. 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 distill the 50-step teacher into a block-causal, 4-step student that streams video in real time while preserving the image, camera, skeleton, and text conditioning. The recipe has two stages: *(i)* block-causal autoregressive training with teacher forcing and error recycling, which feeds the model's own predicted drift back into the context so it learns to self-correct over long horizons; and *(ii)* asymmetric distribution-matching distillation (DMD) (Zhu et al., 2026b; Zhao et al., 2026) to a 4-step student. DMD contrasts the clean-latent predictions of a frozen real denoiser μ_ψ (the teacher, with classifier-free guidance) and an online fake denoiser μ_ϕ (a critic trained on the student's own 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 (Yin et al., 2024), and we apply several critic updates per generator update. Self-generated rollouts (Huang et al., 2025) together with a bounded KV cache (a local attention window plus an attention sink) make the per-block cost independent of horizon length, so the student streams fixed-size latent blocks in real time ($\approx$21 fps at 640×480) for minute-long interactive rollouts.

5 Experiments

5.1 Benchmark Results

Human egocentric video. We evaluate on 200 clips from EgoSuite (Lightwheel, 2026) and OpenAoE (Li et al., 2026c), two datasets unseen during training by both Xperience-0 and all compared methods. Xperience-0 achieves the lowest camera-trajectory errors, the highest hand-following accuracy, and the best visual quality (Table 1). 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. Xperience-0 outperforms

Table 1: **Evaluation on human video.** We evaluate camera following, hand following, and visual quality on 200 EgoSuite/OpenAoE clips. Both datasets are unseen during training by all models. Best results are in **bold**, and second best are underlined.

Model	Camera following			Hand following			Pixel quality		
	ATE↓	RPE↓	RRE↓	PCK@.1↑	PCKm@.1↑	Det↑	PSNR↑	SSIM↑	LPIPS↓
OSCAR (2B)	3.480	0.273	6.258	77.60	84.46	90.55	16.21	0.547	0.382
RynnWorld-Teleop (5B)	3.543	0.274	6.109	74.67	82.37	88.78	16.94	0.577	0.365
DreamDojo (14B)	2.652	0.227	3.730	60.25	72.61	80.00	17.76	0.601	0.344
Cosmos3-Nano (16B)	1.931	0.170	1.584	78.94	84.03	92.08	18.66	0.624	0.275
Cosmos3-Super (64B)	2.031	0.178	2.045	77.89	83.32	91.83	18.36	0.612	0.288
Xperience-0 (5B)	1.884	0.170	1.131	**83.85**	**88.56**	**93.78**	18.91	0.640	**0.245**
Xperience-0 (A14B)	**1.750**	**0.164**	**0.970**	81.38	86.67	92.62	**19.34**	**0.669**	0.246

Table 2: **Comparison on the DreamDojo benchmarks under different post-training settings.** Despite having only **5B** parameters, our model consistently outperforms comparable **14B** models 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
	Xperience-0 (5B)	20.854	0.764	0.213	20.569	0.778	0.208	19.045	0.760	0.217
	Xperience-0 (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
	Xperience-0 (5B)	21.021	0.786	0.206	20.776	0.801	0.199	19.290	0.787	0.207
	Xperience-0 (A14B)	**23.151**	**0.827**	**0.185**	**22.485**	**0.835**	**0.181**	**20.929**	**0.816**	**0.191**
100%	Cosmos-Predict2.5	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
	Xperience-0 (5B)	21.738	0.797	0.202	21.230	0.810	0.196	19.814	0.795	0.204
	Xperience-0 (A14B)	**23.217**	**0.828**	**0.186**	**22.731**	**0.845**	**0.177**	**21.086**	**0.826**	**0.187**

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 2).

Ψ_0 and ω-HOME. On these dexterous-humanoid benchmarks (Table 3; breakdowns in Table S6), Xperience-0 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, A14B mid-training 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 S6). Qualitatively (Fig. S16), Xperience-0 follows the commanded hands and the changing viewpoint while the baselines drift or lose the arm.

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 3). Post-training lifts unseen scenes in step with seen ones (Table S6), indicating adaptation rather than memorization. Xperience-0 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 S7 and S8).

AgiBot. All compared checkpoints include AgiBot World data; on held-out gripper-manipulation windows, A14B midtrain leads all baselines and 5B is on par with the strongest (OSCAR), so the advantage is not limited to transfer settings (Table S10).

Single-checkpoint cross-embodiment prediction. A single mid-trained checkpoint predicts all eight embodiments on the cross-embodiment set with no post-training on these evaluation

Table 3: **Comparison on Ψ_0, ω-HOME, and Xp-Teleop under different post-training settings.** We evaluate no post-training, few-shot, and full-data adaptation. Best results are in **bold**, and second best are underlined.

Benchmark	Model	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
	Xperience-0 (5B)	15.22	0.594	0.385	15.59	0.603	0.369	16.41	0.628	0.325
	Xperience-0 (A14B)	**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
	Xperience-0 (5B)	17.82	0.714	0.277	18.30	0.721	0.259	19.25	0.742	0.228
	Xperience-0 (A14B)	**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
	Xperience-0 (5B)	14.98	0.489	0.320	15.52	0.509	0.293	15.89	0.517	0.279
	Xperience-0 (A14B)	**16.42**	**0.537**	**0.259**	**18.25**	**0.601**	**0.205**	**18.74**	**0.618**	**0.196**

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 S11). For a matched comparison, we fine-tune Cosmos 3 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 S5).

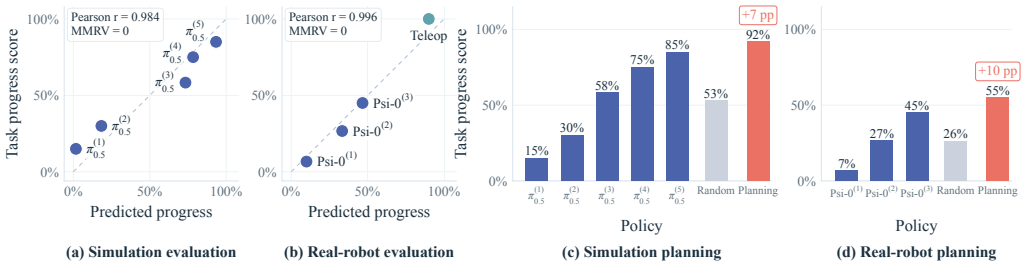
5.2 Ablation Studies and Analysis

Effectiveness of control conditions. We train separate variants without skeleton or camera conditioning and evaluate them on the cross-embodiment set (Table S11). 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 conditions, with camera conditioning particularly important for loco-manipulation.

Human pretraining and scaling. To isolate the effect of our pretraining, we post-train the same architecture initialized from either public video-generation weights ("official base") or our pretrained weights (Table S13). The gains are largest when data is scarce. With 1% of the post-training data, Xperience-0 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 advantage shrinks monotonically as target data grows—from 4.8–7.1 dB at 1% to ≈2.3 dB at 100% (Table S13)—so the human-video prior contributes most exactly when target data is scarce. Prediction quality also improves as we scale the amount of human-video pre-training data, showing a consistent, monotonic improvement across Ψ_0, Xp-Teleop, ω-HOME, held-out human video, and the DreamDojo GR-1 sets (Figs. S12–S14; Sec. G).

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 ($19.83 \rightarrow 21.98$, Table S11), $+2.3$ dB in-distribution on AgiBot ($18.60 \rightarrow 20.88$, Table S10), and, without any post-training, $+2.1/1.7/1.4$ dB on Ψ_0/ω-HOME/Xp-Teleop (Table 3). Model-size and data scaling are thus complementary, jointly driving the gains reported above.

Distillation. The distilled 4-step 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 6.3 s out to minute-long rollouts (Table S1; Sec. D).



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

Figure 4: **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.

5.3 Downstream Applications

We study four downstream uses of Xperience-0, spanning simulation and real robots: policy evaluation, model-based planning, world–action modeling, and live teleoperation.

Policy evaluation. Used as a proxy environment, Xperience-0 evaluates policy rollouts using a task-progress rubric following WorldEval (Li et al., 2025b) and SIMPLER (Li et al., 2024). Its scores closely match ground-truth evaluations, achieving Pearson correlations of $r = 0.984$ across five $\pi_{0.5}$ checkpoints in RoboTwin 2.0 (Chen et al., 2025) and $r = 0.996$ across three Ψ_0 checkpoints on a real Unitree G1 loco-manipulation task. Both evaluations yield MMRV $= 0$ and preserve the ground-truth checkpoint rankings (Fig. 4(a,b)), supporting the use of Xperience-0 for policy comparison and checkpoint selection.

Model-based planning. Following DreamDojo (Gao et al., 2026), Xperience-0 is used for closed-loop planning: it rolls out multiple candidate action sequences, an external value model scores the predicted rollouts, and the selected plan is executed in the environment. 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. 4(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 (Yuan et al., 2026) from our pretrained 5B world model and leaving its action head and training recipe unchanged. This improves action learning: trained on LIBERO (Liu et al., 2023) and evaluated directly on both LIBERO and the held-out LIBERO-Plus (Fei et al., 2025), it raises LIBERO-Plus success from 50.06% to 56.82% and improves every few-shot setting (Table S9), and on a real G1 loco-manipulation task it reaches 56.7% task progress vs. 33.3% for FastWAM, competitive with the VLA Ψ_0 (50.0%; Fig. S15).

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. S10; videos in the supplement).

6 Conclusion

We presented Xperience-0, 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 pretraining to post-training. Trained on Xp-Data with 10k hours of human video and 3k hours of robot data, Xperience-0 demonstrates strong zero-shot generalization, data-efficient adaptation, and competitive prediction across diverse benchmarks, including our new Xp-Teleop. Beyond video prediction, the model supports policy evaluation, model-based planning, and world–action modeling, while distillation enables real-time interactive simulation.

AI USE STATEMENT

Generative AI tools assisted with language polishing to improve the clarity and readability of the manuscript. The authors are responsible for the final content.

ETHICS STATEMENT

Xp-Teleop consists of dexterous-humanoid teleoperation recordings, in which a human operator controls the robot; the released data capture the robot embodiment. Our in-house egocentric collections record human participants, all of whom consented to data collection and to public release. The use and redistribution of third-party datasets remain subject to their original licenses and access conditions.

REPRODUCIBILITY STATEMENT

Dataset construction, annotation, and cross-embodiment retargeting are detailed in Sec. A. The model architecture, conditioning interface, and training configurations are specified in Sec. C, with the real-time distillation recipe in Sec. D and the teleoperation rig and demonstration collection in Sec. E. Evaluation benchmarks, splits, metrics, and baseline implementations are described in Sec. H.

REFERENCES

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

Build AI. Egocentric-10k, 2025. URL `https://huggingface.co/datasets/builddotai/Egocentric-10K`.

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

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

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

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

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801. IEEE, 2025.

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

BitRobot, Unitree, and Hugging Face. Hiw-500: Humanoids in-the-wild dataset for robot learning. `https://bitrobot-foundation.github.io/humanoids-in-the-wild-500-hours/`, 2026.

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

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

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

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

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

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025b.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. *Advances in Neural Information Processing Systems*, 38:93771–93784, 2026a.

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-anything: Simulation-ready physical 3d assets from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5839–5848, 2026b.

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-omni: Unified simulation-ready physical 3d generation for rigid, deformable, and articulated objects. *arXiv preprint arXiv:2605.21572*, 2026c.

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

Cheng Chen, Jerry Bai, Jiacheng Wei, Boyu Chen, Xiaoji Zheng, Fan Wu, Minghao Yang, Tianrun Chen, Ruibo Li, Xiaoyu Yue, et al. Anyworld: Factorized egocentric world models for cross-embodiment generalization. *arXiv preprint arXiv:2608.29242*, 2026.

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

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

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

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

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

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 653–660. IEEE, 2024a.

Irving Fang, Yuzhong Chen, Yifan Wang, Jianghan Zhang, Qiushi Zhang, Jiali Xu, Xibo He, Weibo Gao, Hao Su, Yiming Li, et al. Egopat3dv2: Predicting 3d action target from 2d egocentric vision for human-robot interaction. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3036–3043. IEEE, 2024b.

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

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

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

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

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

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 409–419. IEEE, 2018.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild. In *Forty-third International Conference on Machine Learning*, 2026.

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

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

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

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

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

D HO, J MONAS, JT REN, and C YU. 1x world model: evaluating bits, not atoms, 2025.

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

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

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

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

Qianxi Hua, Xinyue Li, Zheng Yan, Yang Li, Chi Zhang, Yongyao Li, and Yufei Liu. Vtouch++: A multimodal dataset with vision-based tactile enhancement for bimanual manipulation. *arXiv preprint arXiv:2604.20444*, 2026.

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

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

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

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Yuxin Jiang, Shengcong Chen, Siyuan Huang, Liliang Chen, Pengfei Zhou, Yue Liao, Xindong He, Chiming Liu, Hongsheng Li, Maoqing Yao, et al. Enerverse-ac: Envisioning embodied environments with action condition. *arXiv preprint arXiv:2505.09723*, 2025.

Zhennan Jiang, Kai Liu, Yuxing Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Zhengtao Zhang, Chao Yu, Haoran Li, and Dongbin Zhao. World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. *IEEE Robotics and Automation Letters*, 2026.

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

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

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

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10118–10128. IEEE, 2021.

Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, et al. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. *arXiv preprint arXiv:2510.00406*, 2025a.

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

Xukun Li, Yu Sun, Lei Zhang, Bosheng Huang, Yibo Peng, Yuan Meng, Haojun Jiang, Shaoxuan Xie, Guocai Yao, Alois Knoll, et al. Deco: Decoupled multimodal diffusion transformer for bimanual dexterous manipulation with a plugin tactile adapter. *arXiv preprint arXiv:2602.05513*, 2026a.

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

Zhe Li, Zhenzhe Zhang, Yangyang Wei, Wenjie Zhang, Xichen Yuan, Peiyuan Zhi, Gen Li, Xinying Guo, Fengjie Gao, Jianfei Yang, et al. ω-0: A latent predictive world action model for concurrent humanoid loco-manipulation. *arXiv preprint arXiv:2608.06375*, 2026b.

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

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

Lightwheel. Egosuite-open100k. `https://huggingface.co/collections/LightwheelAI/egosuite-open100k`, 2026.

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

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

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

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024a.

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

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751. IEEE, 2024b.

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

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025.

Jiangran Lyu, Kai Liu, Xuheng Zhang, Haoran Liao, Yusen Feng, Wenxuan Zhu, Tingrui Shen, Jiayi Chen, Jiazhao Zhang, Yifei Dong, et al. Lda-1b: Scaling latent dynamics action model via universal embodied data ingestion. *arXiv preprint arXiv:2602.12215*, 2026.

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

MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958. IEEE, 2026.

Dantong Niu, Zhuoyang Liu, Zekai Wang, Boning Shao, Zhao-Heng Yin, Anirudh Pai, Yuvan Sharma, Stefano Saravalle, Ruijie Zheng, Jing Wang, et al. T-rex: Tactile-reactive dexterous manipulation. *arXiv preprint arXiv:2606.17055*, 2026.

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

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

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

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 4, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254. IEEE, 2025.

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

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

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

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

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

Ropedia. Xperience-10m: A large-scale egocentric multimodal dataset with structured 3d/4d annotations, 2026. Dataset.

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

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

Fourier ActionNet Team and Yao Mu. Actionnet: A dataset for dexterous bimanual manipulation. *Fourier ActionNet dataset release*, 2025.

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

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

Unitree Robotics. XR-Teleoperate: An open-source teleoperation framework and data collection toolkit for embodied intelligence. `https://github.com/unitreerobotics/xr_teleoperate`, 2024. Accessed: 2026-02.

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

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, et al. Bridgedata v2: A dataset for robot learning at scale. In *Conference on robot learning*, pp. 1723–1736. PMLR, 2023.

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

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

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

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

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

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20213–20224. IEEE, 2023.

Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12713–12724. IEEE, 2025b.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, et al. Ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.

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

Zhuoyuan Wu and Jun Gao. Oscar: Omni-embodiment action-conditioned world model for robotics. *arXiv preprint arXiv:2606.04463*, 2026.

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

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

Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. Kinema4d: Kinematic 4d world modeling for spatiotemporal embodied simulation. *arXiv preprint arXiv:2603.16669*, 2026.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025a.

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

Xiuyu Yang, Bohan Li, Shaocong Xu, Nan Wang, Chongjie Ye, Zhaoxi Chen, Minghan Qin, Yikang Ding, Zheng Zhu, Xin Jin, et al. Orv: 4d occupancy-centric robot video generation. *arXiv preprint arXiv:2506.03079*, 2025b.

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

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

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

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

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

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

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

Zongzheng Zhang, Jingrui Pang, Zhuo Yang, Kun Li, Minwen Liao, Saining Zhang, Guoxuan Chi, Jinbang Guo, Huan-ang Gao, Modi Shi, et al. Dexora: Open-source vla for high-dof bimanual dexterity. *arXiv preprint arXiv:2605.18722*, 2026b.

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

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

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

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

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

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844. IEEE, 2025b.

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

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

A DATA

This appendix provides additional details omitted from the main Data section. Throughout, an *episode* (robot) or *clip* (human) is a contiguous source recording, a *window* is an 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 (Grauman et al., 2022; 2024; Hoque et al., 2026; Wang et al., 2023; Damen et al., 2022; AI, 2025; Liu et al., 2022; Fang et al., 2024b; Fan et al., 2023; Kwon et al., 2021; Liu et al., 2024b; Garcia-Hernando et al., 2018; Punamiya et al., 2026; Ropedia, 2026) 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 (Bu et al., 2025a; Khazatsky et al., 2024; Fang et al., 2024a; Zhang et al., 2026b; Team & Mu, 2025; Bjorck et al., 2025; Fu et al., 2024; BitRobot et al., 2026; Li et al., 2026a; Niu et al., 2026; Hua et al., 2026; Unitree Robotics, 2026; Zhao et al., 2025) 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 (Lin et al., 2025). We also use Grounding DINO (Liu et al., 2024a) to identify dynamic regions, such as hands and people, for exclusion from subsequent camera-trajectory consistency checks. For kinematics, we reconstruct human hands with WiLoR (Potamias et al., 2025) and project the resulting 3D MANO (Romero et al., 2017) keypoints into the image. Robot keypoints and wrist poses are derived from joint states through URDF-based forward kinematics and projected where camera calibration is available. Since many public robot datasets lack reliable camera calibration, we manually annotate or correct image-space keypoints when accurate projection is unavailable. For semantics, we generate captions at the task, subtask, and atomic-action levels, describing the acting hands, manipulated objects, and interactions.

Quality checks. We verify annotation completeness, correctness, and cross-modal consistency. Camera trajectories are checked for depth-reprojection consistency and temporal smoothness, following (Wang et al., 2026a). Human hand annotations are checked both geometrically and semantically: projected 3D keypoints are compared against independent 2D detections from MediaPipe (Lugaresi et al., 2019) and a fine-tuned MMPose (MMPose Contributors, 2020) 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 annotated robot-hand data, which recover missed hands and remove duplicate or misaligned detections. Captions generated by Gemini (Comanici et al., 2025) are verified by Qwen (Bai et al., 2023), and discrepancies are resolved by human review. Finally, human spot checks across all modalities catch 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 (Comanici et al., 2025) merges the keyframe captions within a window into one window-level caption. Each clip provides synchronized RGB, projected skeletons, camera parameters, and a caption (available for all but a small fraction of robot windows), plus wide-FOV skeletons and wrist poses where available, in a format shared across human and robot data.

Fig. S1 summarizes the curation funnel described below.

A.1 ADDITIONAL CORPUS ANALYSIS

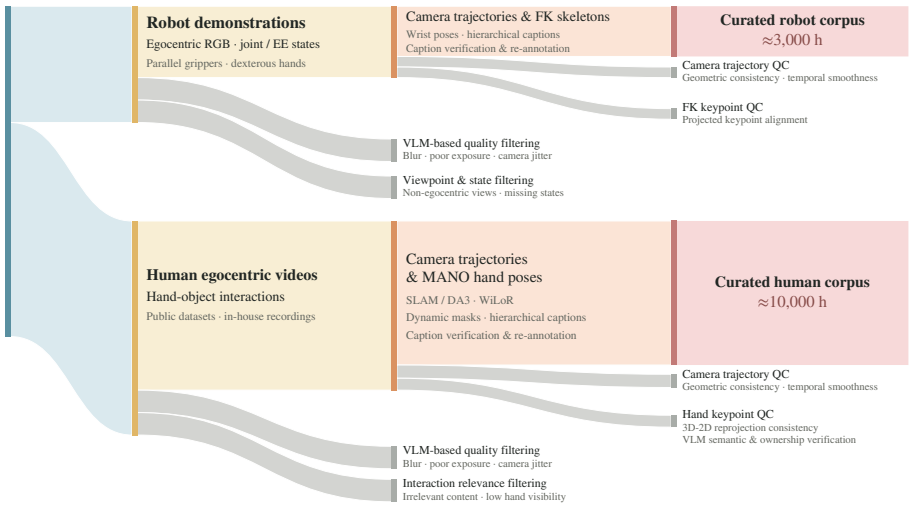


Figure S1: **Data curation funnel for Xp-Data.** Robot and human egocentric recordings pass through source screening, motion annotation, and quality control; gray bands mark data removed at each step, and the surviving streams form the unified corpus of ∼3,000 robot hours and ∼10,000 human hours.

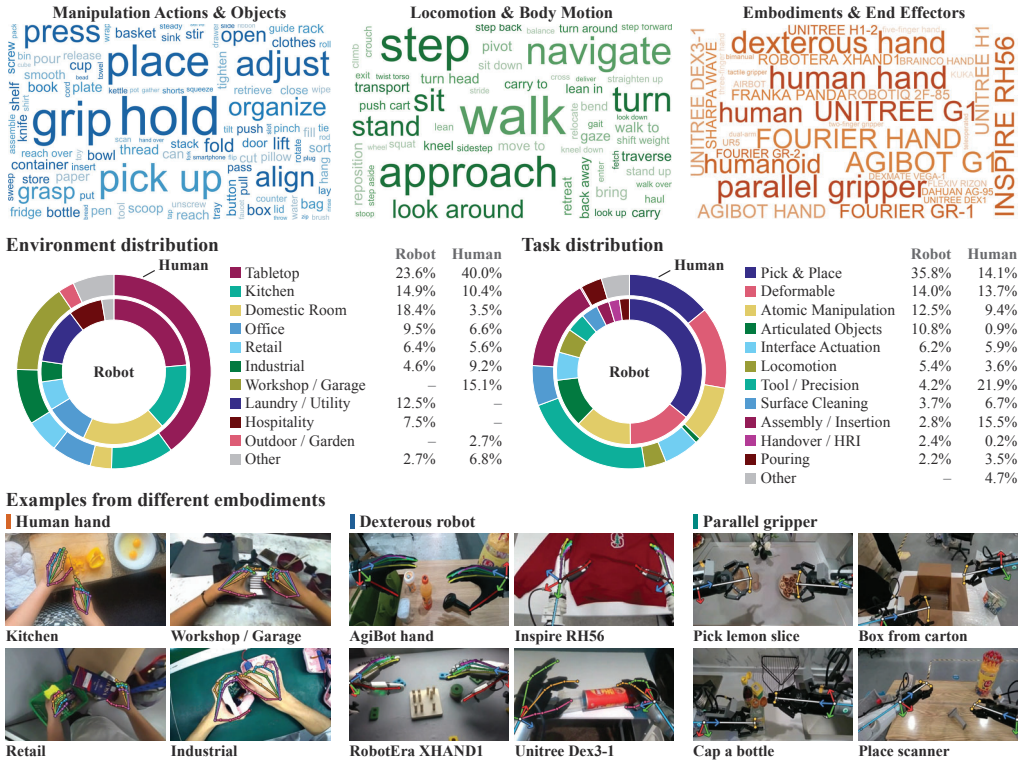


Figure S2: **Overview of Xp-Data.** Top: caption terms. Middle: environment and task distributions for robot (inner ring) and human (outer ring) data. Bottom: unified skeleton annotations across human hands, dexterous robot hands, and parallel grippers.

A.2 Qualitative Examples

Figures S3–S6 visualize the corpus. Figures S3 and S5 show example frames from the human and robot portions, illustrating the coverage of environments, tasks, objects, and embodiments. Figures S4 and S6 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.

B Extended Related Works

Robotic video world models. Early world models learned compact latent dynamics for control (Ha & Schmidhuber, 2018; Hafner et al., 2025); large video generators yield interactive worlds (Bruce et al., 2024; Parker-Holder et al., 2024; 2025) and ego-motion-controlled driving simulators (Hu et al., 2023; Russell et al., 2025; Gao et al., 2024). Complementary approaches generate simulation-ready, physically grounded 3D assets and scenes (Cao et al., 2026a;b;c). Distillation and self-rollout training convert video generators into causal few-step models (Yin et al., 2025; Huang et al., 2025; Cui et al., 2026; Liu et al., 2025a), augmented with long-horizon memory (Hong et al., 2025; Sun et al., 2025). Robotic world models learn video dynamics from robot trajectories (Yang et al., 2023; Zhu et al., 2025b; Jiang et al., 2025; Liao et al., 2026; Guo et al., 2026; Zhen et al., 2025; Xu et al., 2026; HO et al., 2025), often adapting a video foundation model (Ali et al., 2025; Wan et al., 2025). Numerical action tokens require embodiment-specific encoding (Wang et al., 2025a; Guo et al., 2026); latent actions inferred from video (Bruce et al., 2024; Ye et al., 2025; Gao et al., 2025; Garrido et al., 2026; Gao et al., 2026) make unlabeled footage usable, but may entangle actions with other visual changes (Zhang et al., 2026a); spatially aligned renderings provide explicit geometric conditioning (Jiang et al., 2025; Xu et al., 2026; Yang et al., 2025b; Wang et al., 2025b; Wu & Gao, 2026). Cross-embodiment models use embodiment-specific action heads, projections, or adaptation (Wang et al., 2025a; Gao et al., 2026; Agarwal et al., 2026), a single rendered interface (Wang et al., 2025b; Wu & Gao, 2026), or factorized human-to-robot recomposition (Chen et al., 2026). The last, AnyWorld (Chen et al., 2026), 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 (Bar et al., 2025; Hu et al., 2023) and in egocentric predictors conditioned on or jointly modeling full-body pose (Bai et al., 2026; Pallotta et al., 2026; Xiu et al., 2026). Xperience-0 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 (Grauman et al., 2022; Damen et al., 2022; Grauman et al., 2024); pose-annotated collections add body or hand pose (Ma et al., 2024; Liu et al., 2022; Hoque et al., 2026; Punamiya et al., 2026), and portable rigs bring hand pose into everyday scenes (Wang et al., 2024; Liu et al., 2025b). Aggregated and in-the-wild demonstrations (O'Neill et al., 2024; Khazatsky et al., 2024; Walke et al., 2023; Fang et al., 2024a; Wu et al., 2024) and large teleoperation sets (Bu et al., 2025a) 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 (Zhao et al., 2025; Team & Mu, 2025; Qiu et al., 2025; Wei et al., 2026; Li et al., 2026b). Evaluation data include held-out robot video clips (Zhu et al., 2025b; Gao et al., 2026; Wu & Gao, 2026), multimodal world-model benchmarks (Agarwal et al., 2026), and benchmarks measuring visual quality, physical consistency or embodied utility (Yue et al., 2025; Motamed et al., 2026). 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 (Li et al., 2024) to video world models through recorded-action rollouts (Zhu et al., 2025b; Li et al., 2025b; Wu & Gao, 2026) or closed-loop policy interaction (Quevedo et al., 2026; Guo et al., 2026; Tseng et al., 2025; Team et al., 2025). For planning, world models score or refine action proposals using predicted outcomes (Qi et al., 2026; Du & Song, 2026; Jain et al., 2026; Gao et al., 2026), in pixel or learned latent spaces (Zhou et al., 2024; Assran et al., 2025), and support reinforcement fine-tuning and synthetic data generation (Zhu et al., 2026a; Jiang et al., 2026; Li et al., 2025a; Xiao et al., 2025; Jang et al., 2025). Vision-language-action models map observations and language to actions using pretrained vision-language models (Brohan et al., 2023; Kim et al., 2024; Black et al., 2024;

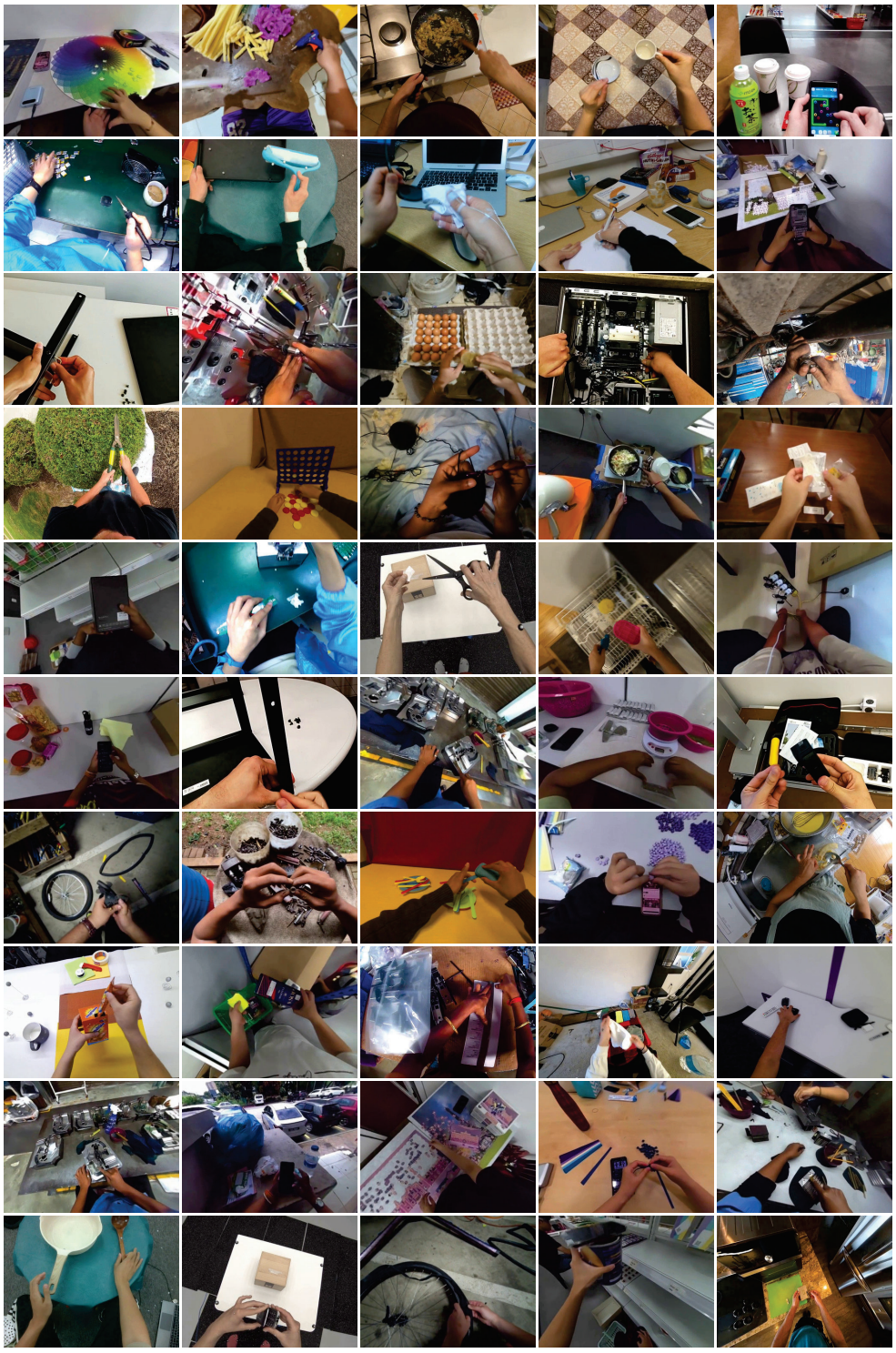


Figure S3: **Example frames from the egocentric human data in Xp-Data.** The grid illustrates the diversity of environments (kitchens, workshops, industrial floors, offices, and domestic rooms), tasks (cooking, repair, assembly, crafting, and games), and viewpoints covered by the human portion of the corpus.

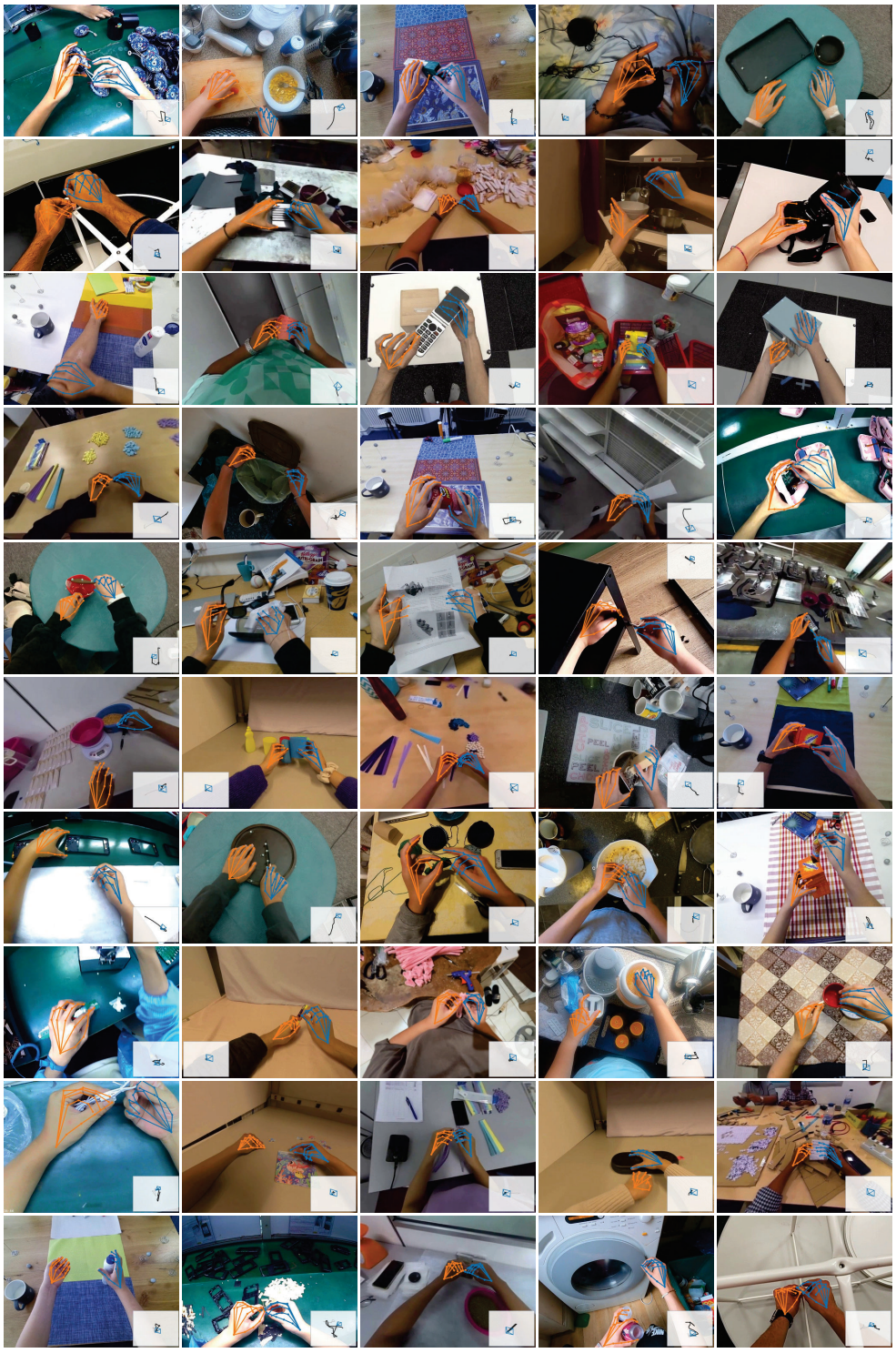


Figure S4: **Motion annotations on the egocentric human data.** Projected 21-point hand skeletons are overlaid on each frame, with the two hands drawn in different colors; the inset shows the camera trajectory of the surrounding window.

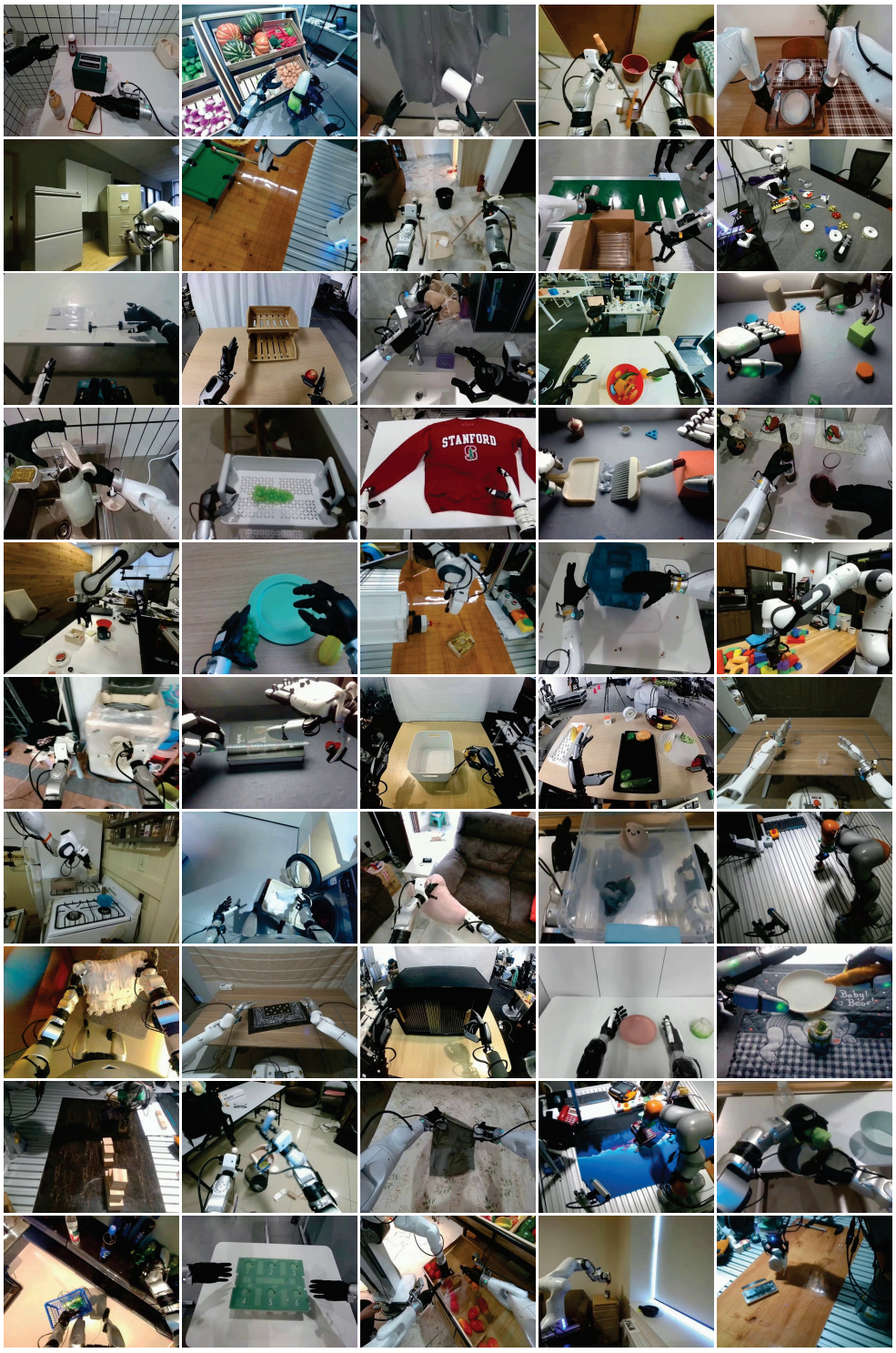


Figure S5: **Example frames from the robot data in Xp-Data.** The grid spans dexterous five-finger hands, three-finger hands, and parallel grippers on single-arm, dual-arm, and humanoid platforms, across household, retail, industrial, and tabletop scenes.



Figure S6: **Motion annotations on the robot data.** Skeletons derived from embodiment-native forward kinematics are projected onto each frame, covering dexterous hands and grippers; the inset shows the camera trajectory of the surrounding window.

Intelligence et al., 2025; Bjorck et al., 2025), increasingly trained with human video (Ye et al., 2025; Bu et al., 2025b;a; Yang et al., 2025a; Luo et al., 2025; Kareer et al., 2025; Wei et al., 2026), whereas world action models couple future video or latent-state prediction with action learning (Cheang et al., 2024; Zhu et al., 2025a; Hu et al., 2024; Wang et al., 2025a; Liao et al., 2026; Ye et al., 2026; Agarwal et al., 2026; Bi et al., 2026; Lyu et al., 2026; Li et al., 2026b; Wang et al., 2026b; Yuan et al., 2026). Xperience-0 supports these applications in Sec. 5; as a world action model, it emits the shared kinematic representation using shared weights across embodiments.

C PRELIMINARIES

Action-conditioned dynamics modeling. A world model predicts how an environment evolves under an agent's actions. Let $s_t \in \mathcal{S}$ denote the environment state and $a_t \in \mathcal{A}$ the action at step t. Under a Markov formulation, the dynamics are described by a stochastic transition kernel:

$$s_{t+1} \sim p\left(\cdot \mid s_t, a_t\right), \qquad p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S}), \tag{6}$$

where $\Delta(\mathcal{S})$ denotes the space of probability distributions over $\mathcal{S}$. Repeated application of this transition defines a distribution over future state trajectories. Xperience-0 models these dynamics in egocentric observation space: given an initial visual observation and a sequence of factorized loco-manipulation controls (Sec. 4.1), it predicts the corresponding future video. The controls specify camera and manipulator motion, allowing the model to capture their combined effects on future observations.

Latent video diffusion backbones. We instantiate Xperience-0 at two scales using the Wan2.2 family of pretrained video diffusion models (Wan et al., 2025). Wan2.2-TI2V-5B uses a dense diffusion transformer with approximately 5B parameters. Wan2.2-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 (Peebles & Xie, 2023). 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 (Lipman et al., 2022): 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.

D DISTILLATION AND INTERACTIVE SIMULATION DETAILS

This appendix illustrates the distillation procedure and the interactive rollout mechanism summarized in Sec. 4.4.

Distillation recipe. Distillation proceeds in two stages that both retain the camera–skeleton conditioning of the base model. *(i) Block-causal autoregressive training* uses fixed-size latent blocks with within-block bidirectional and cross-block causal attention; each block is predicted from lightly noised ground-truth context (teacher forcing), and an error-recycling buffer feeds the model's own time-bucketed prediction errors back into the context, so that training exposes the drift encountered at inference. The resulting 50-step model serves as the teacher. *(ii) DMD* distills the teacher into a 4-step student with three networks initialized from it—a trainable generator, a frozen real denoiser (the teacher, with classifier-free guidance), and an online fake denoiser (a critic). Training and inference share a bounded KV cache (a local window plus an attention sink), so memory and per-block latency do not grow with rollout length.

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 ($\approx$21 fps), compared with about 41 s for the 50-step teacher—a $7\times$ end-to-end speed-up—and streams its first block within roughly 1.4 s.

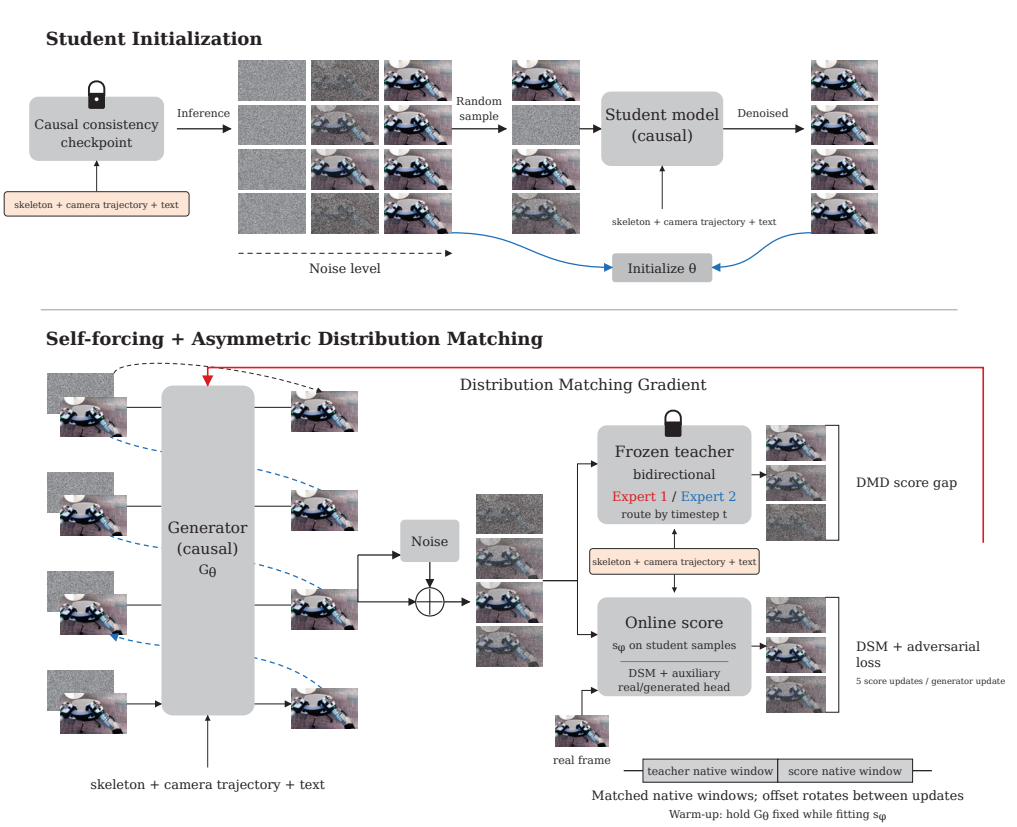


Figure S7: **Distillation of Xperience-0 into a few-step causal student.** Top: the block-causal student is trained with teacher forcing and error recycling, then distilled to 4 steps. Bottom: asymmetric distribution matching—the causal generator G_θ produces samples from its own rollouts, while a frozen real denoiser (the teacher, with classifier-free guidance) and an online fake denoiser (a critic) provide the distribution-matching gradient that updates the generator.

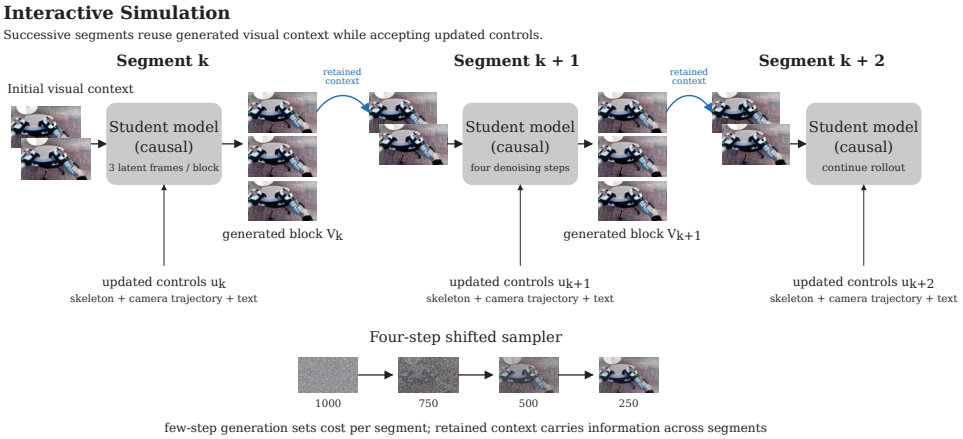


Figure S8: **Interactive simulation with the distilled student.** Successive segments reuse the previously generated visual context while accepting updated controls (skeleton, camera trajectory, and text). Each block is produced by the few-step causal student, so the per-segment cost is bounded while 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**

Table S2: **Comparison of humanoid teleoperation datasets** across capabilities, task horizon, and available modalities.

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	✓	✓	✓	✓	✓	✓

Qualitative comparison. Fig. S9 shows ground truth, teacher, and student frames sampled across a GR1 rollout. The 4-step student preserves the arm and hand configuration, object layout, and scene appearance of the 50-step teacher.

E Additional Details for Xp-Teleop

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



Figure S9: **Qualitative teacher vs. student rollout** on a GR1 test clip. Columns: ground truth, 50-step teacher, 4-step student; rows are chronological keyframes.

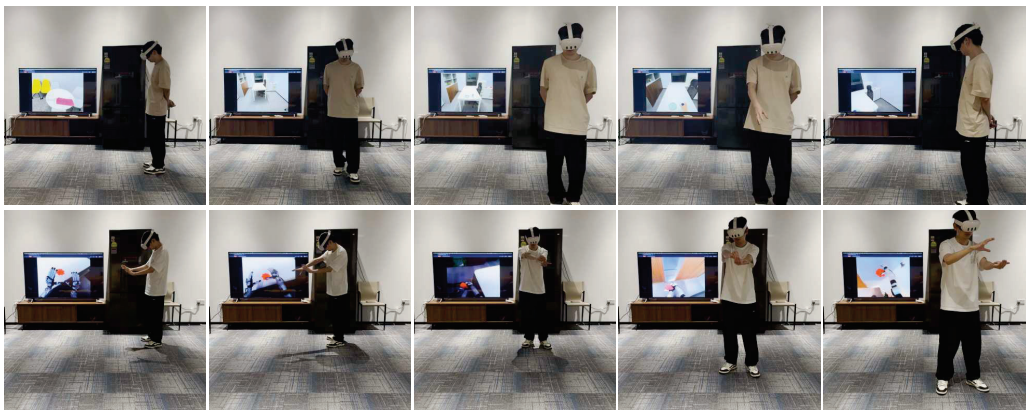


Figure S10: **Live teleoperation with Xperience-0.** On-site recordings show the headset user and the monitor displaying model-generated feedback. Five chronological keyframes show interactions with human (top) and robot (bottom) embodiments.

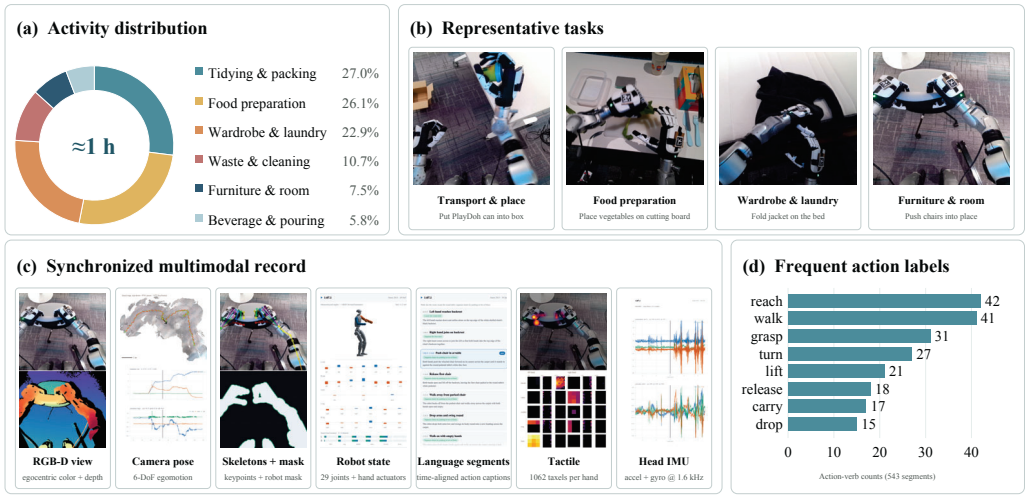


Figure S11: **Overview of Xp-Teleop.** (a) Activity distribution. (b) Representative long-horizon episodes. (c) Synchronized multimodal records (RGB-D, camera poses, skeletons/masks, robot states, language, tactile, head IMU). (d) Common action-verb frequencies.

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 S2 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×).

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. Fixed-base datasets (Unitree G1 Dex1/Dex3, PH2D, RoboMIND Tien Kung, GR00T-Teleop-G1) 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 sub-task 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.

Modality criteria. The modality columns describe information available in the analyzed dataset release, not signals that could potentially be reconstructed from the robot hardware. *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. A release qualifies when it provides per-frame joint information (*WB joints*) together with sufficient camera specifications (*Cal.*) Combined with the publicly available URDF of the robot, one can project the relevant finger-to-arm kinematic structure into the egocentric image.

RGB@fps reports the released image resolution and frame rate. Dpt. denotes a released depth stream. IMU is marked available only when explicit inertial-sensor signals are provided; estimated root or torso poses alone do not qualify. WB joints requires recorded whole-body joint states covering both upper and lower limbs, including the legs. Arm- or hand-only states, estimated root or torso poses, and base commands alone do not satisfy this criterion. The entry describes recording coverage and does not require the recorded leg joints to move substantially. Tac. denotes tactile sensing. Cal. requires both camera intrinsics and camera extrinsics; camera poses without intrinsics do not qualify. Kpts. denotes temporally aligned keypoints.

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.

Dataset-comparison protocol. Table S2 compares dexterous humanoid teleoperation datasets using episode-level task-horizon statistics and modalities available in the analyzed releases. The tabletop group contains recordings with the robot's legs locked, whereas the loco-manipulation group contains recordings in which locomotion is commanded.

Task-horizon metrics. All three horizon entries are means over the analyzed episodes. Duration is the episode length in seconds. Loco. range is the diagonal length, in meters, of the axis-aligned bounding box enclosing the available camera or robot trajectory for an episode. Subtasks is the number of instruction-level task stages in an episode. Because subtask boundaries follow the wording and granularity of each release's instructions, this metric should be interpreted as an annotation-level measure of task horizon rather than a platform-independent action count.

Dataset-specific qualifications. Unitree Dex1 satisfies WB joints only in v3; v2 contains arm and gripper states only. Although v3 supports base commands, observed leg-joint variation is below 0.03 rad, so its recorded locomotion range is treated as zero. Ψ_0 supports locomotion but does not release leg-joint states in the analyzed data. RoboMIND entries refer to its Xsens dexterous-hand subset, which provides depth but no hand-joint states. Humanoid Everyday horizon values refer to its loco-manipulation subset; tactile data are available only for the G1/Dex3 embodiment. The analyzed ω-HOME files contain no depth stream. RGB and keypoint entries use the audited values reported in Table S2.

F Supplementary Evaluation on Human Egocentric Videos

Table S3 reports pooled results on 200 clips (100 per dataset) from EgoSuite (Lightwheel, 2026) and OpenAoE (Li et al., 2026c).

Table S3: **Results on a 200-clip EgoSuite/OpenAoE evaluation set** (100 clips each). Best bold, second-best underlined; protocol in Sec. F.

(a) Camera and hand action following

Method	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
Xperience-0 (5B), pretrain	1.793	0.165	1.078	**32.7**	83.63	**88.66**	93.15
Xperience-0 (5B), midtrain	1.884	0.170	1.131	33.2	**83.85**	88.56	**93.78**
Xperience-0 (A14B), pretrain	**1.542**	**0.152**	0.980	36.0	81.27	86.85	92.20
Xperience-0 (A14B), midtrain	1.750	0.164	**0.970**	36.6	81.38	86.67	92.62

(b) Video prediction and generation quality

Method	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
Xperience-0 (5B), pretrain	18.90	0.6409	0.2444	93.64	75.00	40.73	73.0	189.9
Xperience-0 (5B), midtrain	18.91	0.6402	0.2453	93.71	72.00	41.15	70.1	190.1
Xperience-0 (A14B), pretrain	**19.77**	**0.6749**	**0.2381**	93.55	76.50	40.20	**64.0**	**155.9**
Xperience-0 (A14B), midtrain	19.34	0.6691	0.2462	93.45	**79.00**	40.23	72.4	168.0

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 (Lin et al., 2025). Hand metrics use MediaPipe (Lugaresi et al., 2019) 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.

G Scaling Analysis

Figures S12–S14 report how Xperience-0's prediction quality improves as we scale the amount of human-video pre-training data, on held-out human video (Fig. S12), the DreamDojo cross-evaluation sets (Fig. S13), and the Ψ_0/Xp-Teleop/ω-HOME transfer benchmarks (Fig. S14). 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; the largest point slightly exceeds the 10k-hour corpus because pre-training continues a little into a second epoch. For held-out human video the checkpoint

Table S4: **Per-dataset results on EgoSuite and OpenAoE.** Each set contains 100 clips from Table S3. 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
Xperience-0 (5B), pretrain	1.012	20.45	109.8	2.574	17.35	140.0
Xperience-0 (5B), midtrain	1.038	20.38	<u>103.0</u>	2.729	17.43	139.4
Xperience-0 (A14B), pretrain	**0.872**	**21.48**	**94.8**	**2.212**	<u>18.07</u>	**125.8**
Xperience-0 (A14B), midtrain	<u>0.988</u>	<u>20.51</u>	114.4	<u>2.511</u>	**18.17**	135.8

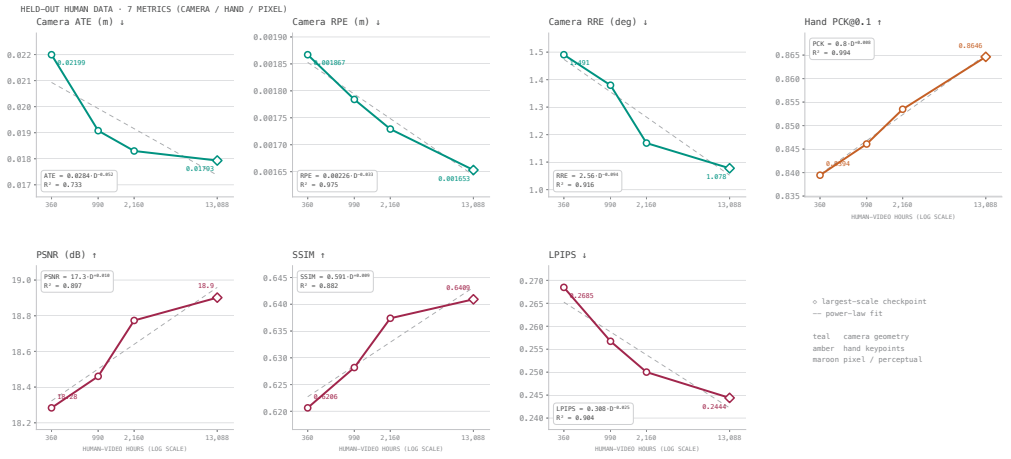


Figure S12: **Scaling on held-out human video.** Zero-shot prediction on 200 held-out egocentric clips across seven metrics—camera geometry (ATE, RPE, RRE via DA3-Streaming pose recovery), hand keypoints (PCK@0.1 at hand-detection confidence 0.5), and pixel quality (PSNR, SSIM, LPIPS)—versus the amount of human-video pre-training data. Dashed lines indicate the fitted trend.

is evaluated zero-shot, whereas for the DreamDojo and transfer benchmarks it is post-trained on the target with the same recipe before evaluation.

H Experimental Setup Details

Benchmarks. We evaluate action-conditioned video prediction on five benchmarks spanning distinct embodiments and data sources: (i) the DreamDojo benchmarks (Gao et al., 2026), three Fourier GR-1 robot evaluation sets from DreamDojo—In-lab (100 clips), EgoDex (65 clips), and DreamDojo-HV (68 clips); (ii) Ψ_0 (Wei et al., 2026), 100 evaluation windows of dexterous humanoid manipulation (50 seen-task, 50 unseen-task); (iii) ω-HOME (Li et al., 2026b), 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. E), 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 (Bu et al., 2025a), 50 held-out windows of gripper-based manipulation, where all compared models include AgiBot data in training.

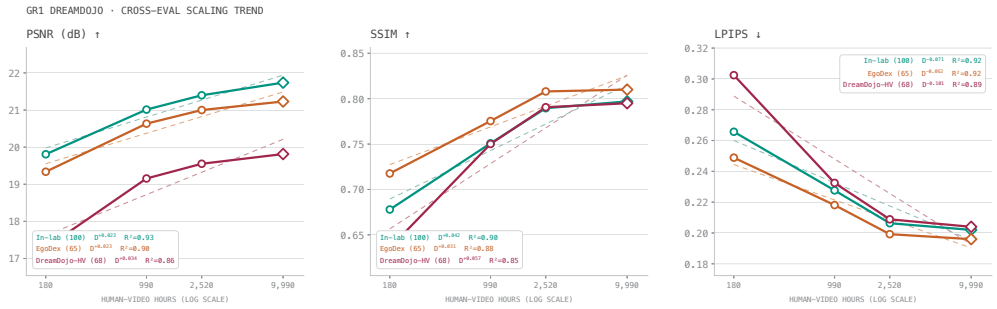


Figure S13: **Cross-evaluation scaling on the DreamDojo benchmark.** PSNR, SSIM, and LPIPS on the In-lab, EgoDex, and DreamDojo-HV evaluation sets versus the amount of human-video pre-training data, under an identical GR-1 post-training budget; the largest-scale point uses Xperience-0 (5B).

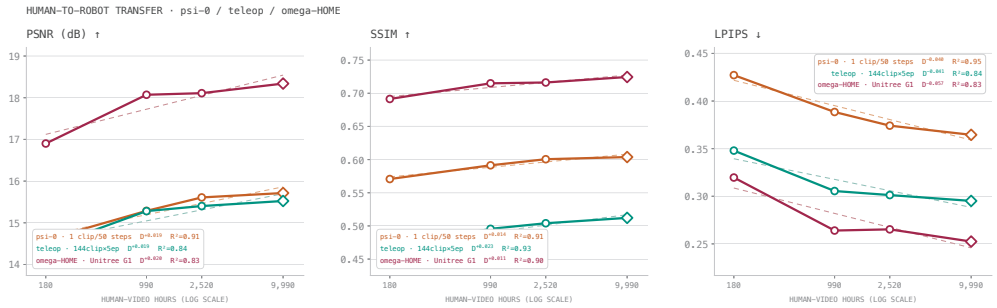


Figure S14: **Human-to-robot transfer scaling.** PSNR, SSIM, and LPIPS on Ψ_0, Xp-Teleop, and ω-HOME versus the amount of human-video pre-training data (log scale); each point is a pre-training checkpoint post-trained on the target benchmark before evaluation. Dashed lines indicate the fitted trend.

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 Xperience-0: *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.

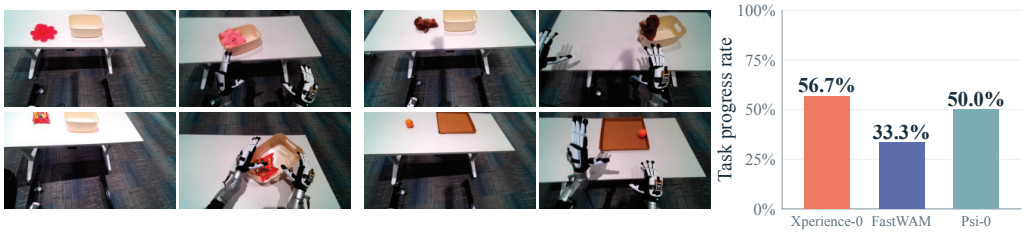
Baselines. We compare against generalist robot world models with different action interfaces: DreamDojo (2B and 14B), which conditions on continuous latent actions pretrained from large-scale human video, and Cosmos3 (Nano, 16B; Super, 64B), which treats action as a first-class modality with per-domain action encoders. 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 pretraining. All baselines are post-trained with their official recipes.

Table S5: **World-model evaluation on Ψ_0, ω-HOME, and Xp-Teleop** under different post-training settings. Best bold, second-best underlined.

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
Xperience-0 (5B), pretrain	15.17	0.587	0.394	15.59	0.606	0.364	16.26	0.625	0.330
Xperience-0 (5B), midtrain	15.22	0.594	0.385	15.59	0.603	0.369	16.41	0.628	0.325
Xperience-0 (A14B), pretrain	15.87	0.614	0.368	17.83	0.664	0.272	18.94	0.692	0.241
Xperience-0 (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
Xperience-0 (5B), pretrain	17.35	0.698	0.295	18.35	0.728	0.250	19.02	0.738	0.233
Xperience-0 (5B), midtrain	17.82	0.714	0.277	18.30	0.721	0.259	19.25	0.742	0.228
Xperience-0 (A14B), pretrain	18.40	0.723	0.265	20.43	0.768	0.197	21.49	0.785	0.175
Xperience-0 (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
Xperience-0 (5B), pretrain	14.52	0.465	0.336	15.50	0.518	0.296	15.90	0.520	0.277
Xperience-0 (5B), midtrain	14.98	0.489	0.320	15.52	0.509	0.293	15.89	0.517	0.279
Xperience-0 (A14B), pretrain	14.95	0.496	0.343	17.98	0.600	0.215	18.26	0.601	0.209
Xperience-0 (A14B), midtrain	**16.42**	**0.537**	**0.259**	**18.25**	**0.601**	**0.205**	**18.74**	**0.618**	**0.196**

I Full Benchmark Numbers

Table S5 lists the complete per-metric numbers behind Table 3, including the pretrain/midtrain contrast for both model scales. Table S6 gives the seen/unseen breakdown for all three benchmarks. Table S9 gives the full per-perturbation world–action results on LIBERO-Plus, and Table S12 the full AgiBot comparison including VBench video-quality metrics. Tables S7 and S8 report long-horizon rollout comparisons (≈32 s zero-shot on ω-HOME and ≈30 s on Xp-Teleop) against the baselines.



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

Figure S15: **World–action modeling.** (a) Real-robot executions of the policy initialized from Xperience-0, shown as four chronological pairs of recorded keyframes with different objects, containers, and placements. (b) Aggregate real-robot task progress for Xperience-0 initialization, FastWAM, and Ψ_0.

Table S6: **Seen/unseen breakdown on Ψ_0, ω-HOME, 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
Xperience-0 (5B), pretrain	15.34	15.01	0.33	15.78	15.39	0.39	16.67	15.85	0.82
Xperience-0 (5B), midtrain	15.45	14.99	0.46	15.75	15.44	0.31	16.68	16.13	0.55
Xperience-0 (A14B), pretrain	15.96	15.78	0.18	17.90	17.77	0.13	19.24	18.64	0.60
Xperience-0 (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
Xperience-0 (5B), pretrain	18.18	16.52	1.66	18.87	17.83	1.04	19.79	18.26	1.53
Xperience-0 (5B), midtrain	18.52	17.13	1.39	18.89	17.71	1.18	20.00	18.50	1.50
Xperience-0 (A14B), pretrain	18.87	17.93	0.94	21.06	19.81	1.25	22.97	20.01	2.96
Xperience-0 (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
Xperience-0 (5B), pretrain	15.40	13.64	1.76	16.37	14.63	1.74	16.76	15.05	1.71
Xperience-0 (5B), midtrain	15.84	14.11	1.73	16.37	14.68	1.69	16.75	15.03	1.72
Xperience-0 (A14B), pretrain	16.09	13.80	2.29	18.85	17.11	1.74	19.18	17.34	1.84
Xperience-0 (A14B), midtrain	**17.25**	**15.59**	1.66	**19.15**	**17.35**	1.80	**19.61**	**17.87**	1.74

Table S7: **Long-horizon rollout on ω-HOME** ($\approx$32 s, zero-shot; 12 chained windows). Overall metrics are frame-averaged; per-segment PSNR at $\approx$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
Xperience-0 (5B), pretrain	13.64	0.651	0.516	23.92	12.57	10.35
Xperience-0 (5B), midtrain	13.93	0.664	0.506	26.33	12.74	10.43
Xperience-0 (A14B), pretrain	15.20	0.696	0.478	25.03	14.03	12.40
Xperience-0 (A14B), midtrain	**16.48**	**0.738**	**0.376**	29.33	15.20	12.35

J Additional Qualitative Comparisons

We provide qualitative comparisons against the baselines across embodiments. In every figure all methods receive the same initial frame and conditioning (converted to each method's native interface); rows are the ground truth and each compared method, and columns are chronological time points.

J.1 Dexterous Humanoid (ω-HOME)

The headline loco-manipulation comparison is Fig. S16: a $\approx$32 s zero-shot rollout in which the robot grabs a beverage, walks to the fridge, and stows it. The additional $\approx$32 s walking rollouts (Figs. S17–S21) are likewise zero-shot chains (no method post-trained on ω-HOME), each com-

Table S8: **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
Xperience-0 (5B), midtrain	19.8	45.3%	65.2	10.83	0.336	0.573	0.892	0.469	0.70
Xperience-0 (A14B), midtrain	**12.0**	**6.9%**	**19.5**	**13.47**	**0.433**	**0.444**	**0.906**	**0.510**	0.70

Table S9: **World–action modeling on LIBERO and LIBERO-Plus.** Success rates (%) with FastWAM or Xperience-0 init (LIBERO-Plus held out); best bold. Improvement is the absolute gain, relative gain over FastWAM in gray.

Initialization	LIBERO	LIBERO-Plus							
		Camera	Robot	Language	Lighting	Background	Noise	Layout	**Overall**
≈43 demonstrations per task (full data)									
FastWAM	96.95	17.32	44.32	67.47	79.96	52.04	38.66	60.92	50.06
Xperience-0	**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 demonstrations per task									
FastWAM	87.30	**18.32**	23.10	43.59	55.27	29.46	**28.36**	42.36	33.58
Xperience-0	**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 demonstrations per task									
FastWAM	60.00	7.57	12.58	25.31	**51.32**	**19.05**	15.12	27.15	21.44
Xperience-0	**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%)**

bining locomotion (the robot walks to the workspace) with manipulation; Xperience-0 follows both throughout, whereas the baselines drift or collapse over the rollout. Fig. S22 additionally contrasts zero-shot and post-trained predictions on a manipulation task.

J.2 Cross-Embodiment Robots

J.3 Human Egocentric Video

J.4 Few-Shot Data Efficiency on GR-1

K WorldArena Comparison against BLM

We additionally compare Xperience-0 (A14B) against the Boundless World Model (BLM) on WorldArena Track 1 (RoboTwin 2.0 (Chen et al., 2025) `fixed_scene_task` test set, 1000 episodes). Both models use their official code and public checkpoints, run in the same environment and scored by the identical official WorldArena pipeline over the same 1000 episodes at 640×480; both receive the same Real-ESRGAN super-resolved first frame ("BLM + SR"), and we report our original-length run.

Xperience-0 leads on 12 of the 15 metrics and by $+5.9$ EWMScore over BLM + SR (Table S14). 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. S29 shows this qualitatively.

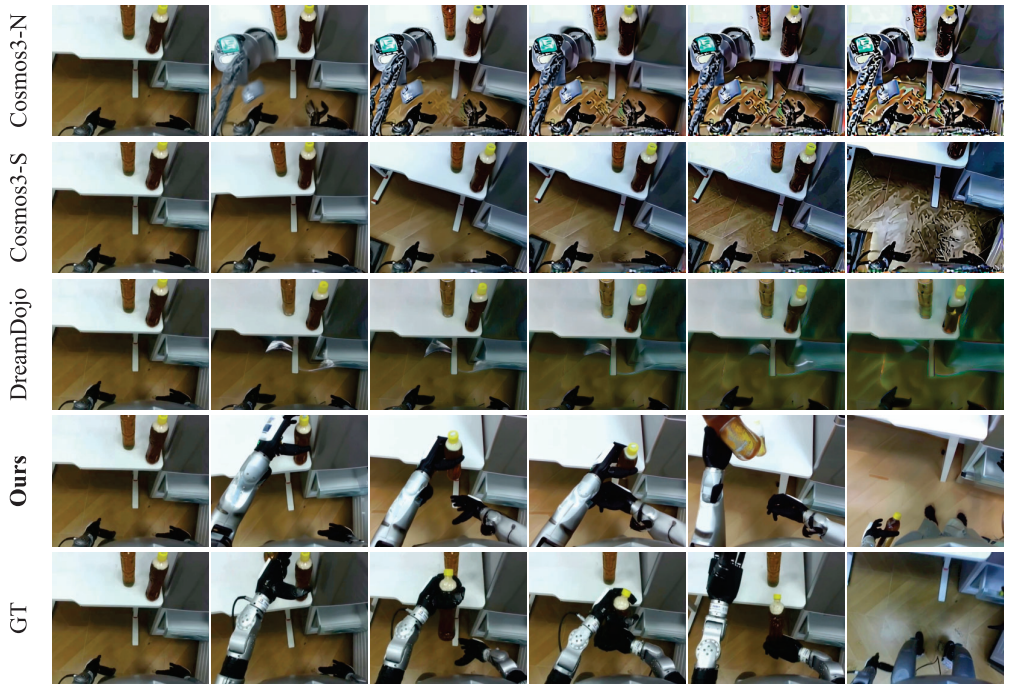


Figure S16: **Qualitative loco-manipulation comparison** (ω-HOME, a held-out Unitree G1 loco-manipulation benchmark not used in training): a $\approx$32 s zero-shot rollout in which the robot grabs a beverage, walks to the fridge, and stows it (no method post-trained on ω-HOME). Rows are the ground truth and each method (Xperience-0 in bold); columns are time. Xperience-0 tracks both the locomotion (the changing viewpoint) and the manipulation and stays close to the ground truth, whereas the baselines drift or collapse.



Figure S17: **Loco-manipulation rollout: clean the bed.** The robot walks to the bed and cleans it.



Figure S18: **Loco-manipulation rollout: clean the bed** (second episode).

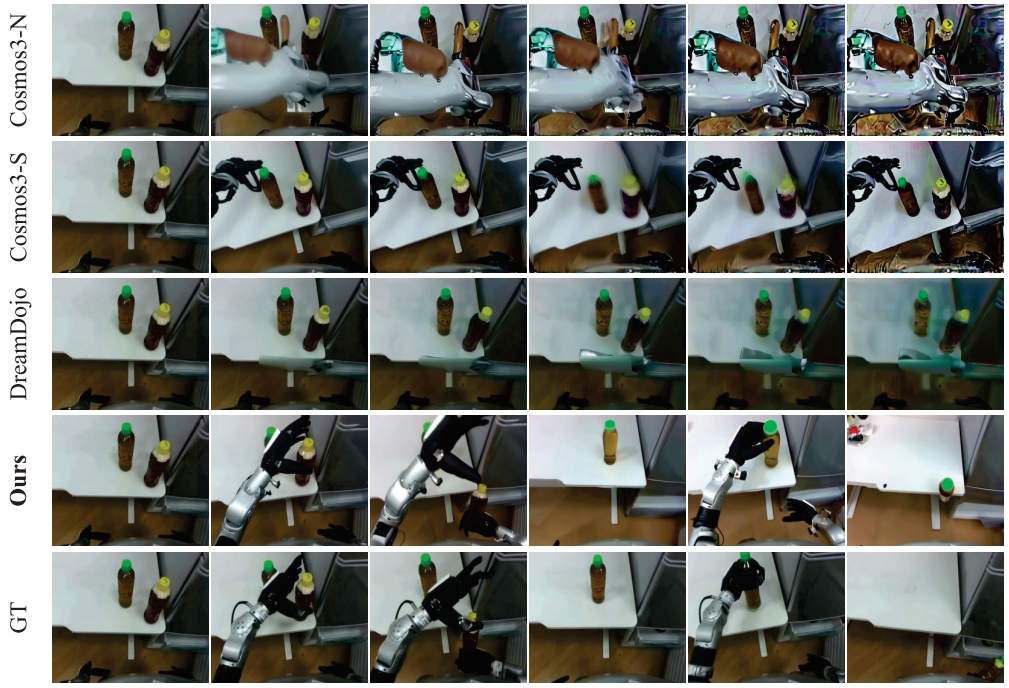


Figure S19: **Loco-manipulation rollout: stow a beverage.** The robot grabs a beverage, walks to the fridge, and places it on the lower shelf (second episode).

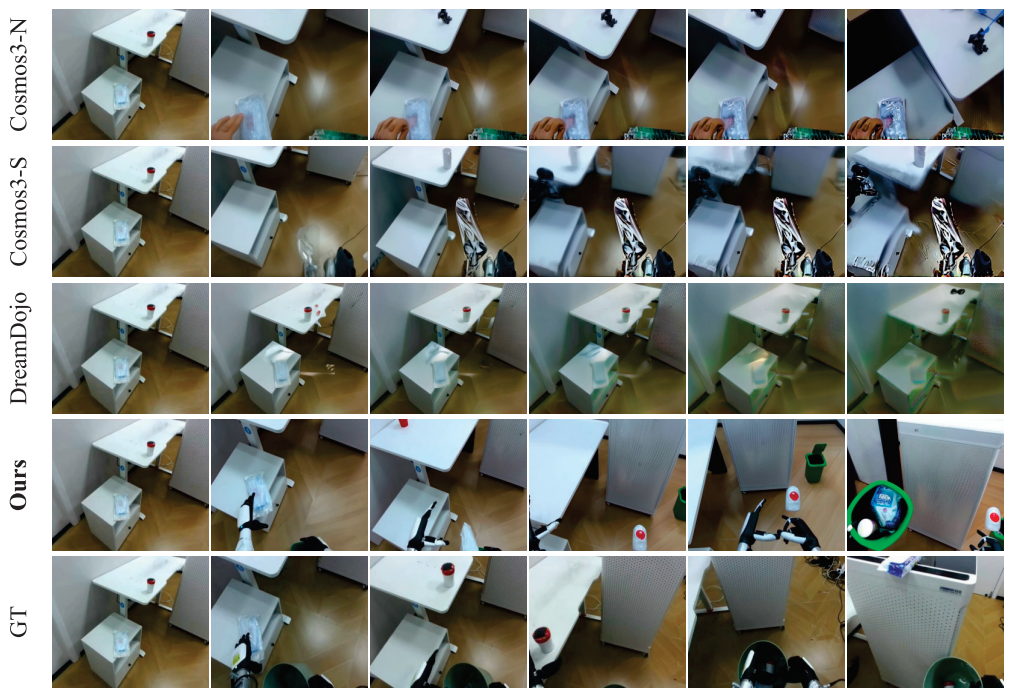


Figure S20: **Loco-manipulation rollout: pick up garbage.** The robot walks around and picks up garbage.

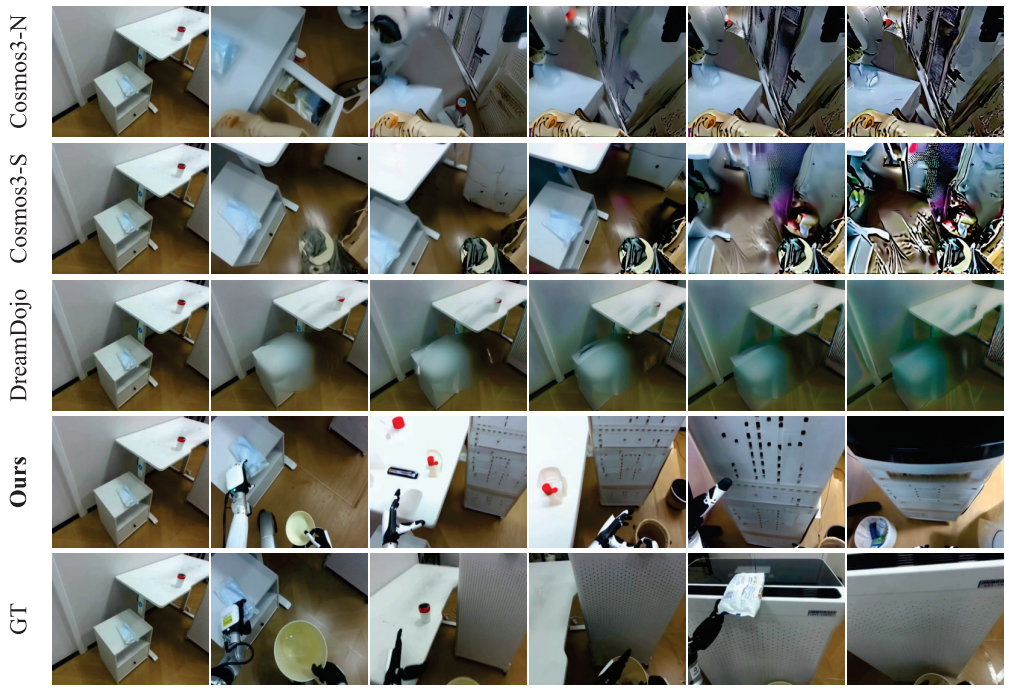


Figure S21: **Loco-manipulation rollout: pick up garbage** (second episode).

Table S10: **Comparison on AgiBot.** VBench metrics in Table S12.

Model	PSNR↑	SSIM↑	LPIPS↓
DreamDojo (2B)	18.26	0.750	0.218
DreamDojo (14B)	18.30	<u>0.755</u>	<u>0.210</u>
Cosmos3-Nano (16B)	17.70	0.742	0.230
Cosmos3-Super (64B)	17.28	0.732	0.241
OSCAR (2B)	18.56	0.746	0.217
Xperience-0 (5B)	<u>18.60</u>	0.750	0.214
Xperience-0 (A14B)	**20.88**	**0.790**	**0.154**

Table S11: **Cross-embodiment set** (400 zero-shot windows over eight embodiments). Cosmos3 is fine-tuned on the same robot mixture that Xperience-0's mid-training already covers, and neither is post-trained on these windows (Xperience-0 is thus *zero-shot*); each ablation variant is trained without one conditioning channel. Best in bold.

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

L Wide-FOV Skeleton Conditioning

The head camera on many humanoid and tabletop rigs crops the manipulator: on the WorldArena (RoboTwin 2.0) test set the in-frame (narrow) skeleton touches the image border in 99.4% of frames, and a $1.5\times$ field-of-view render raises in-frame manipulator coverage from 0.37 to 0.56. A narrow skeleton alone is therefore ambiguous about the arm configuration whenever the arm enters, leaves, or works at the frame edge. The optional wide-FOV branch introduced in Sec. 4.2 feeds a $1.5\times$ render (with the narrow field of view marked) through per-block cross-attention, so the model can see the off-frame portion of the arm.

Figs. S30 and S31 illustrate the effect when the arm works at or beyond the frame edge: the narrow-only model reconstructs the off-frame arm less accurately, while the wide branch recovers it to match the ground truth. Each column is labeled on the figure (prediction *without* the wide-FOV branch, *with* it, and *ground truth*); the wide skeleton is drawn on the $1.5\times$ canvas and the yellow box marks the narrow head-camera field of view. We also report a per-episode border-band keypoint error.

M Limitations and Future Work

Very-long-horizon rollouts. The distilled student sustains real-time rates and stays stable well past a minute (Sec. D). Stronger memory mechanisms or periodic re-grounding are promising directions for extending stable interactive simulation to even longer horizons.

Scaling. 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 pre-training data (Sec. G), suggesting that scaling data further and model size, and broadening embodiment and task coverage, are promising directions toward more capable and reliable embodied world models.

Table S12: **Comparison on AgiBot** (50 held-out gripper windows). VBench consistency, aesthetics, and dynamic degree (Dyn. best near GT). 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
Xperience-0 (5B), midtrain	<u>18.60</u>	0.750	0.214	**0.9437**	**0.9573**	0.4378	0.60
Xperience-0 (A14B), midtrain	**20.88**	**0.790**	**0.154**	0.9368	0.9525	<u>0.4407</u>	0.64

Table S13: **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
	Xperience-0	**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
	Xperience-0	**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
	Xperience-0	**21.74**	**0.797**	**0.202**	**21.23**	**0.810**	**0.196**	**19.81**	**0.795**	**0.204**

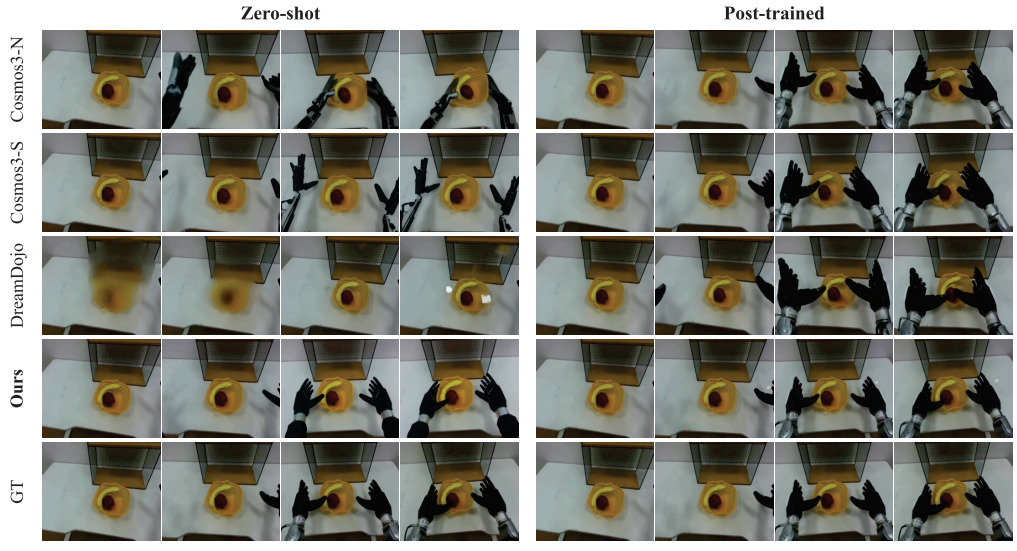


Figure S22: **ω-HOME manipulation task** (fruit-bucket arrangement). Left: zero-shot; right: after post-training. Xperience-0 follows the commanded hands and matches the ground truth in both settings, whereas the baselines lose the arm or drift.

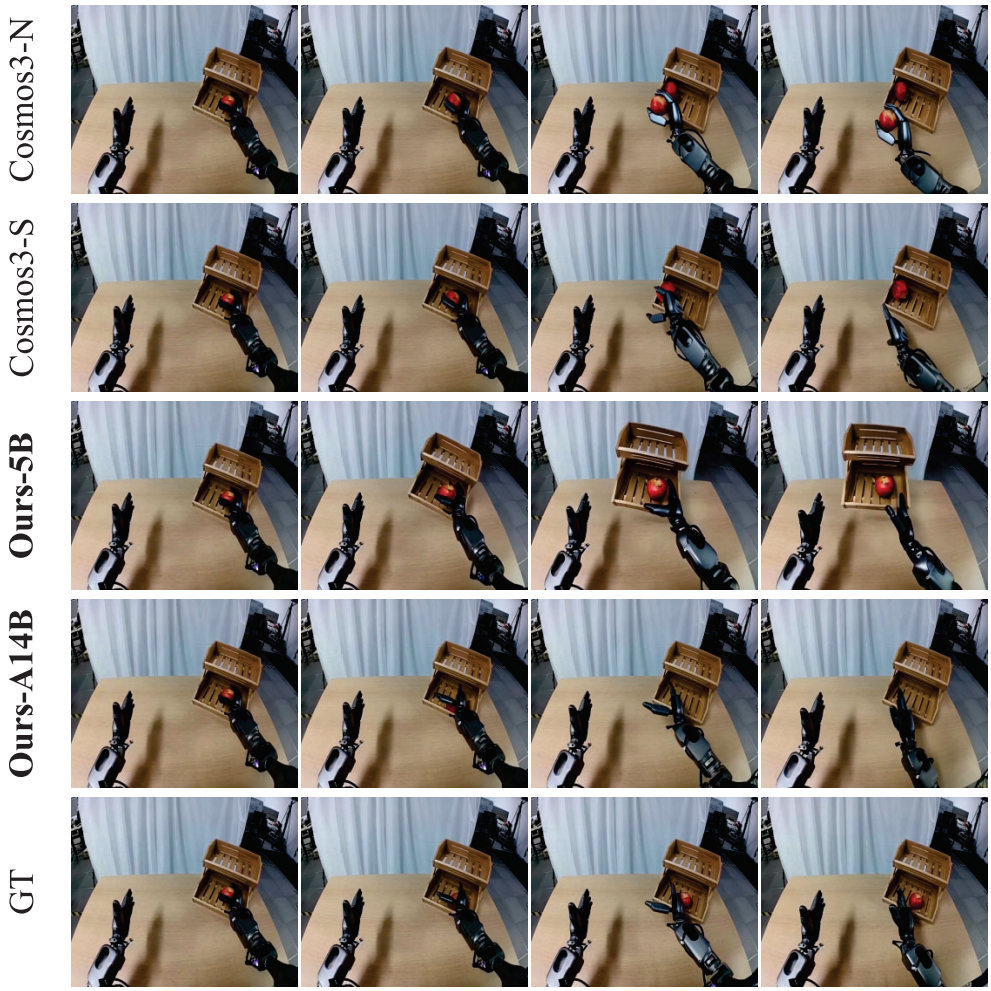


Figure S23: **Cross-embodiment prediction (GR-1).** Xperience-0 (A14B and 5B) is evaluated *zero-shot* (no post-training on this GR-1 set), whereas Cosmos3 is *fine-tuned* on the same robot mixture. Xperience-0's zero-shot predictions still track the ground truth more closely than the fine-tuned baselines.

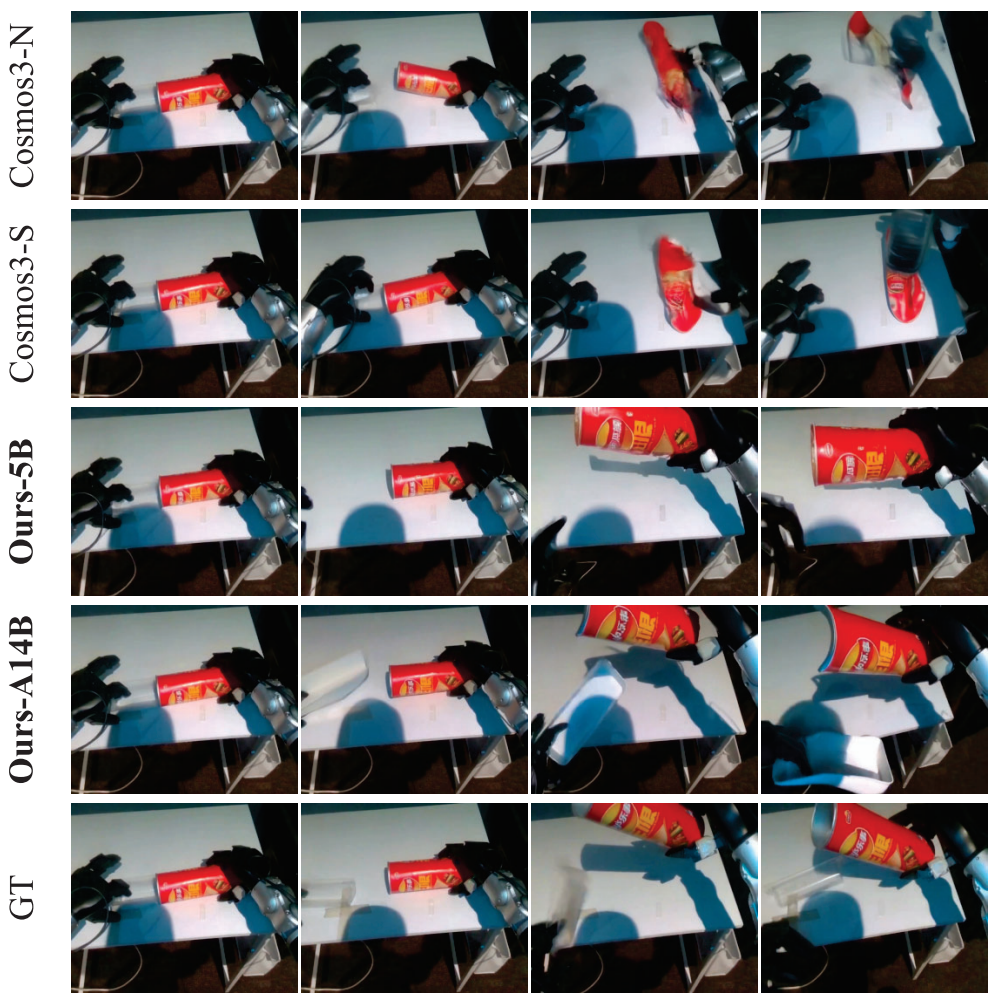


Figure S24: **Cross-embodiment prediction (Ψ_0).** Both Xperience-0 (A14B and 5B) and Cosmos3 are trained on the same robot mixture (Xperience-0 via mid-training, Cosmos3 via fine-tuning); neither is post-trained on Ψ_0. Xperience-0 reproduces the manipulation more faithfully without ever training on the Ψ_0 dataset.

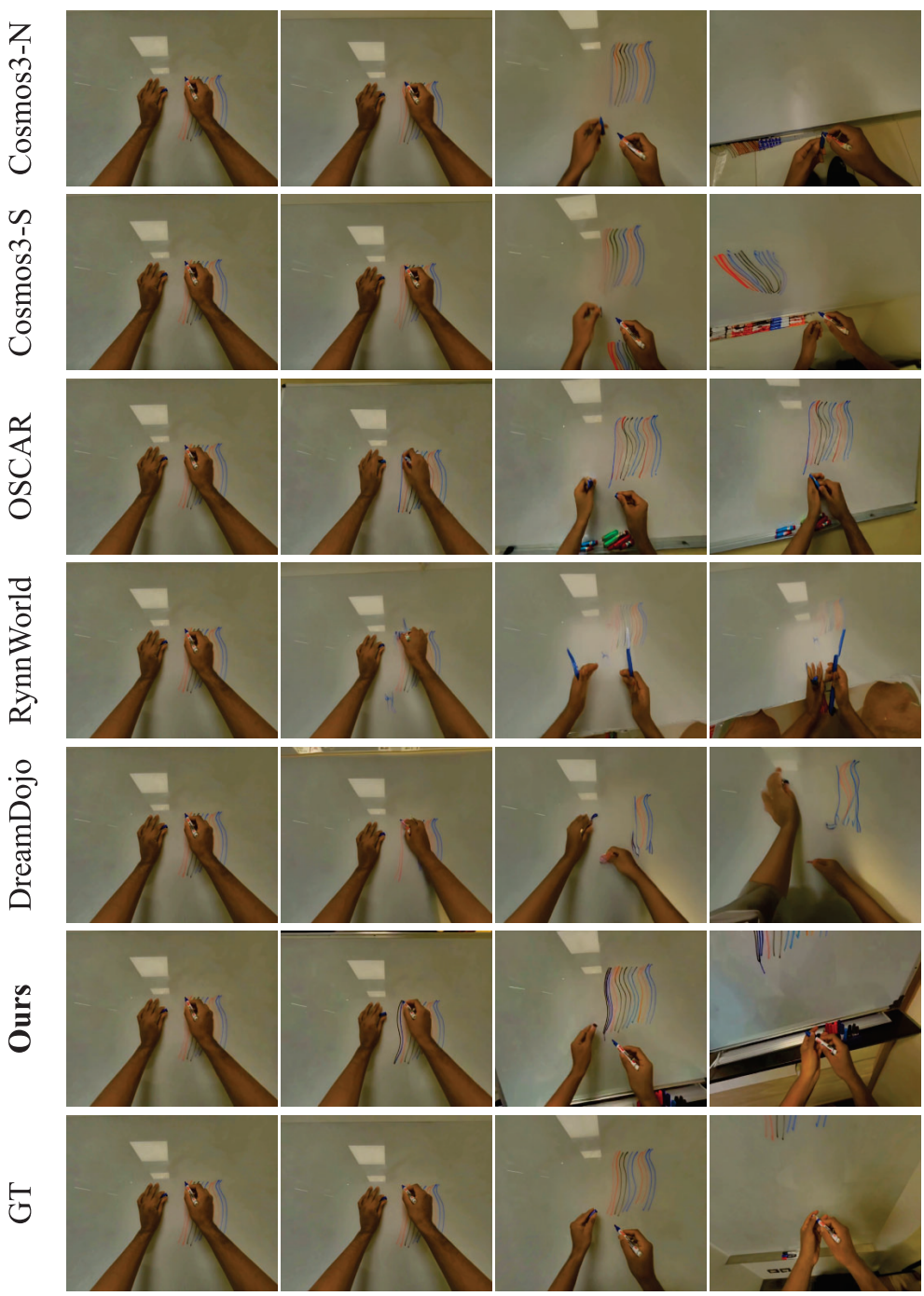


Figure S25: **Human egocentric video (overall fidelity), zero-shot.** Xperience-0 stays closest to the ground truth across the rollout, while the baselines drift or degrade.

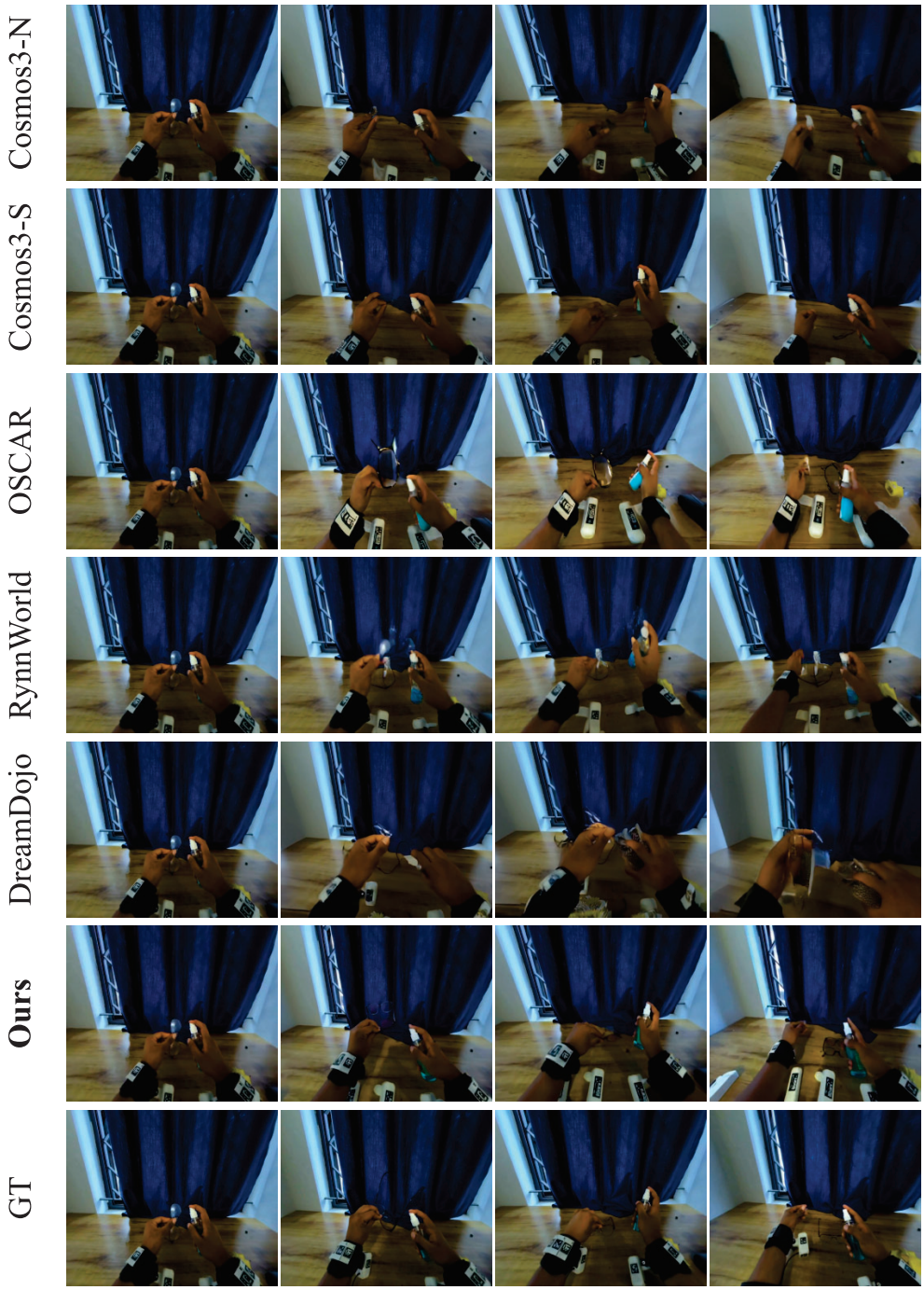


Figure S26: **Human egocentric video (large egomotion), zero-shot.** Under substantial head motion, Xperience-0 follows the commanded camera trajectory and hands, whereas the baselines drift or freeze.



Figure S27: **Few-shot data efficiency on GR-1 (In-lab).** Both models are post-trained on 1% of the data; Xperience-0 initializes from our pretrained base and Wan2.2-A14B from public video-generation weights. Our base yields markedly sharper, more faithful predictions.

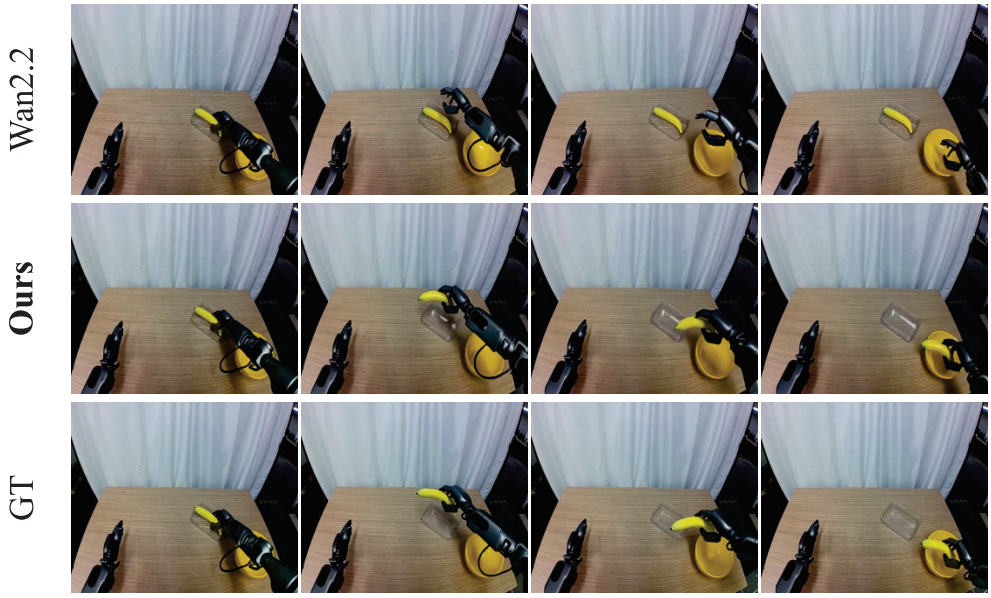


Figure S28: **Few-shot data efficiency on GR-1 (EgoDex).** Same 1% budget; initializing from our pretrained base preserves object and hand detail that the public base loses.

Table S14: **WorldArena Track 1 comparison against BLM** (RoboTwin 2.0, 1000 episodes; identical official pipeline and inputs for both). EWMScore is the mean of the 15 metrics; best in bold.

Metric	BLM (default)	BLM + SR	Xperience-0 (A14B)
EWMScore	59.20	61.01	**66.92**
Image Quality	41.45	51.54	**53.57**
Aesthetic Quality	42.51	**43.09**	41.04
JEPA Similarity	91.41	90.93	**96.00**
Dynamic Degree	16.63	17.49	**25.48**
Flow Score	7.87	8.49	**15.75**
Motion Smoothness	66.45	66.13	**67.55**
Subject Consistency	64.69	68.08	**79.51**
Background Consistency	69.28	71.78	**84.17**
Photometric Consistency	**39.68**	35.95	27.98
Interaction Quality	67.16	70.22	**77.42**
Trajectory Accuracy	33.62	34.64	**63.78**
Depth Accuracy	92.95	94.71	**98.56**
Perspectivity	91.96	96.04	**97.72**
Instruction Following	72.46	76.22	**85.50**
Semantic Alignment	**89.88**	89.79	89.73

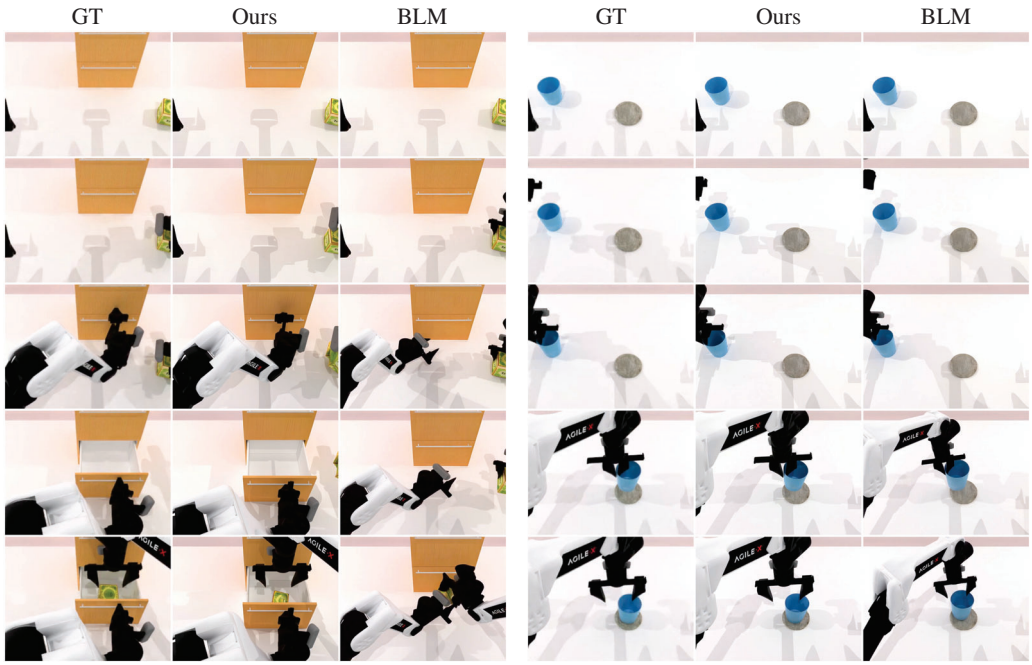


Figure S29: **Qualitative comparison on WorldArena** (left: drawer + box; right: cup onto coaster). Each panel: five chronological frames (rows); columns are *GT* | *Xperience-0 (A14B)* | *BLM* (Table S14). Xperience-0 follows the commanded arm trajectory and completes the task like the ground truth (opening the drawer and placing the box, precisely seating the cup), whereas BLM's arm drifts and under-moves (it never opens the drawer).

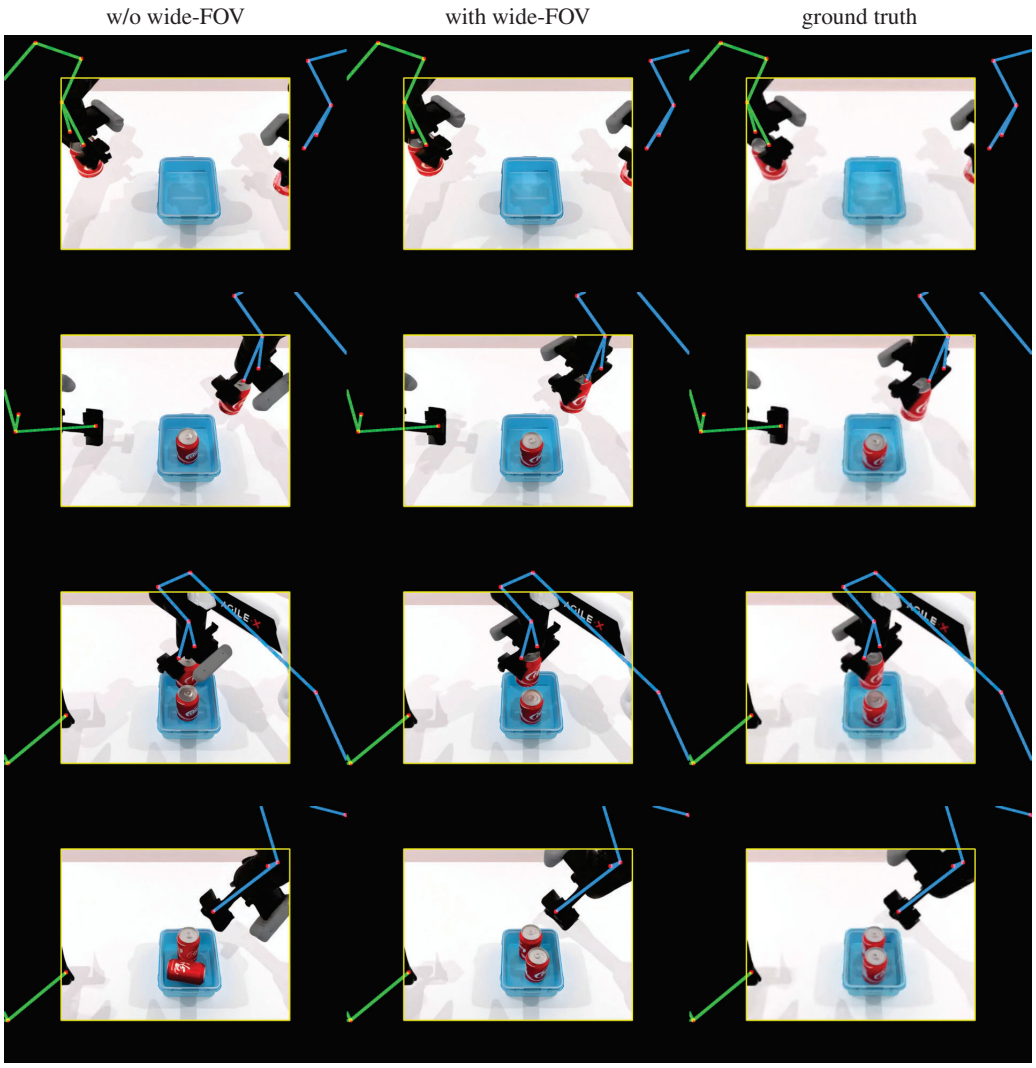


Figure S30: **Wide-FOV skeleton branch, example 1** (four moments in time order; columns labeled above). The narrow-only model places the off-frame arm less accurately, whereas the wide branch recovers it to match the ground truth. Here the outer border-band keypoint error drops from 8.9 to 6.9 px, and the wide branch is closer to the ground truth on 91% of frames.

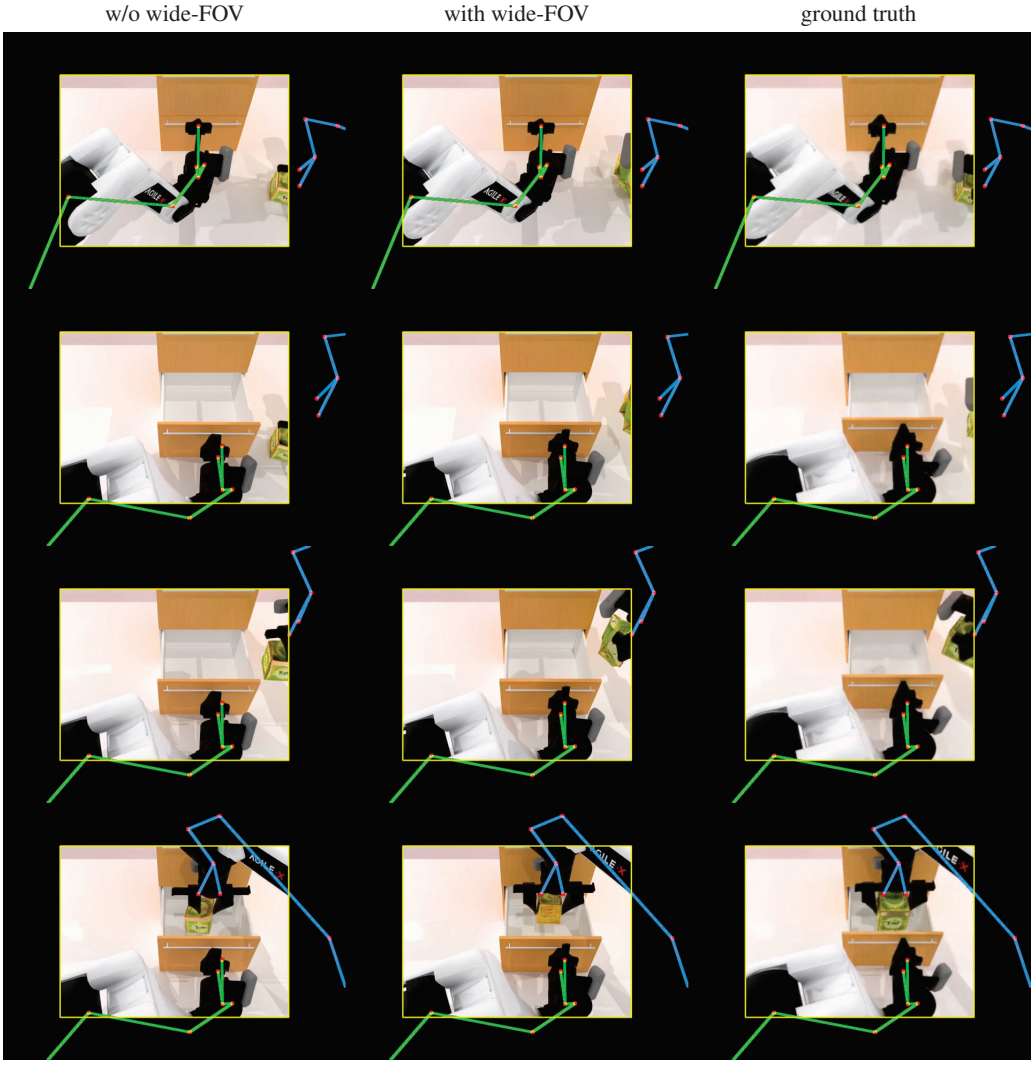


Figure S31: **Wide-FOV skeleton branch, example 2** (four moments in time order; columns labeled above). As the arm reaches beyond the frame, the wide branch again recovers its off-frame pose. Here the border-band keypoint error drops from 14.0 to 11.7 px, with the wide branch closer to the ground truth on 93% of frames.