UMI-WM: INTERACTIVE WORLD MODELING FOR BIMANUAL MANIPULATION FROM WRIST CAMERAS

Anonymous authors
Paper under double-blind review

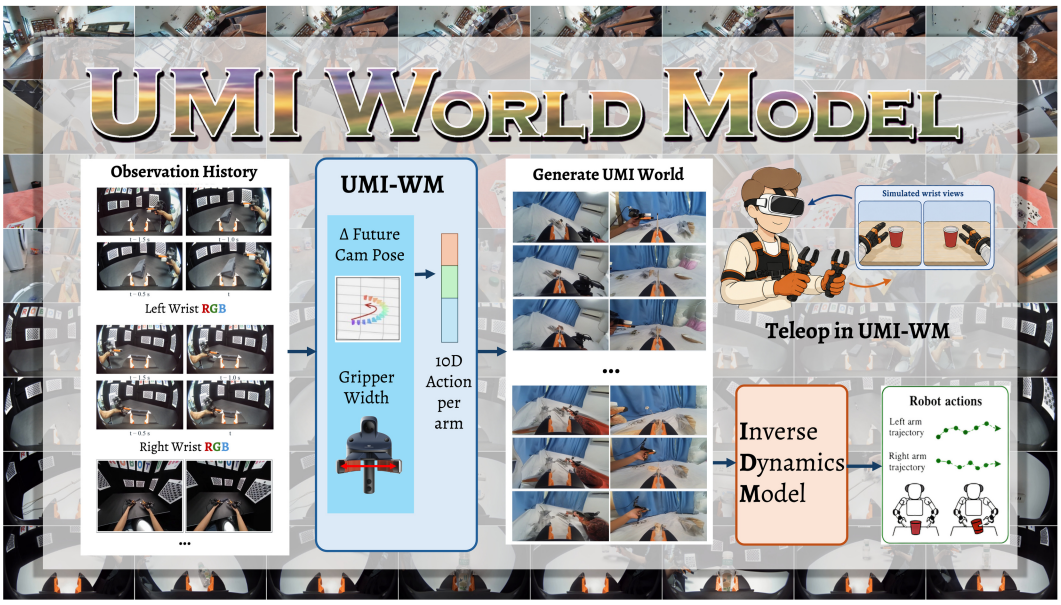

Figure 1: UMI-WM jointly generates both wrist-camera views conditioned on bimanual actions, and an inverse dynamics model (IDM) recovers actions from the generated videos for evaluation and robot execution. More results and demos are on our project page.

ABSTRACT

Universal Manipulation Interface (UMI) pairs wrist-camera videos with camera trajectories and gripper openings, providing scalable, robot-free data for learning how bimanual actions change a shared scene. Existing interactive world models, however, typically rely on text or single-camera control and struggle to follow each arm's actions while maintaining consistency between the two wrist views. We introduce UMI-WM, an action-conditioned bimanual world model that jointly generates both wrist views from observation history, per-arm relative camera poses, and gripper widths, with cross-view consistency, action following, and visual fidelity. UMI-WM injects each arm's actions into its own wrist stream, while sparse hub tokens exchange scene information between the two streams. Camera-aware geometry conditioning and camera-guided noise further support per-view motion control and cross-view geometric exchange. We adapt the model for block-causal generation and distill the 30-step Teacher into a Student with four denoising steps. We train UMI-WM on self-collected open-world and in-studio UMI recordings combined with public manipulation data. To evaluate action following, we use an inverse dynamics model (IDM) to recover relative camera poses and gripper widths from generated videos, and introduce Bimanual Action Fidelity (BAF) to measure how closely these recovered actions match the conditioning actions. UMI-WM achieves the best results on all five reference-video fidelity metrics and the best or competitive performance on VBench and WorldArena. The Teacher and Student outperform all compared baselines in action following, while the four-step Student generates paired wrist videos $6.9\times$ faster than the Teacher.

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

Recent interactive world models, including HY-WorldPlay (Sun et al., 2026), LingBot-World (Robbyant Team et al., 2026), Wonder (Xu et al., 2026), and Matrix-Game 3.5 (Qian et al., 2026), generate videos conditioned on camera poses, allowing users to control viewpoint motion through a scene. Extending this capability to bimanual manipulation requires modeling not only the motion of two independently controlled viewpoints but also how each end-effector's actions affect the shared environment. The Universal Manipulation Interface (UMI) (Chi et al., 2024) provides a natural interface for this extension. Its handheld grippers collect wrist-camera videos, relative end-effector trajectories, and gripper widths without requiring a robot. Because each wrist camera is rigidly mounted to its gripper, the relative camera poses constitute the end-effector pose component of the action, while a single additional dimension specifies the gripper opening. Together, these controls provide a compact bimanual action representation for jointly predicting both future wrist-camera views.

Extending interactive world models to bimanual manipulation raises three consistency requirements. First, cross-view consistency: the two independently controlled wrist views must show the same shared scene, so that changes caused by either hand appear in both. Second, spatial consistency within each view: the persistent gripper foreground must stay stable while the scene moves, whereas existing models let it drift. Third, action consistency: the generated viewpoint motion and gripper opening must follow the commanded per-arm camera pose and width, including small pose changes.

To address these challenges, we introduce **UMI-WM**, an interactive world model that jointly predicts both wrist videos conditioned on bimanual actions (Fig. 1). UMI-WM injects each arm's actions into its own wrist stream, exchanges shared-scene context between the two streams through hub tokens, and adds camera-aware geometry conditioning, so that each wrist's motion and the scene changes caused by either hand are modeled together; causal generation and distilling the 30-step Teacher into a four-step Student keep the rollout efficient. Training uses a unified data processing pipeline that integrates self-collected UMI demonstrations with public robot data through a shared camera-centric action representation. Finally, an inverse dynamics model (IDM) recovers relative camera poses and gripper widths from generated videos, and Bimanual Action Fidelity (BAF) measures how closely these recovered actions agree with the conditioning actions.

Our contributions are then summarized as follows:

- **UMI-WM.** We develop an interactive world model that jointly generates both wrist-camera views conditioned on per-arm camera poses and gripper widths, combining cross-view interaction and camera-aware conditioning with few-step causal generation.
- **Data pipeline and action-level evaluation.** We construct a three-tier data pyramid of public robot data, in-the-wild UMI recordings, and in-studio UMI recordings with a unified processing pipeline, and develop an IDM that recovers relative camera poses and gripper widths from video to compute Bimanual Action Fidelity (BAF).
- **Results.** On 320 held-out UMI episodes, UMI-WM leads all five reference-video fidelity metrics and is best or competitive on VBench and WorldArena. The Teacher and Student reach BAF scores of 76.36 and 71.55, respectively, outperforming all compared baselines, while the four-step Student achieves a $6.9\times$ speedup over the 30-step Teacher.

2 RELATED WORK

2.1 UMI GRIPPER AND UMI DATASETS

UMI (Chi et al., 2024) pairs wrist-camera videos, SLAM-reconstructed camera trajectories, and gripper widths for portable demonstration collection. Variants include FastUMI (Zhaxizhuoma et al., 2025), which simplifies hardware integration and pose tracking; DexUMI (Xu et al., 2025), which supports dexterous hands through a wearable exoskeleton; and MV-UMI (Rayyan et al., 2025), which adds third-person observations for broader scene context. Public datasets such as FastUMI-100K (Liu et al., 2025), RealOmni-Open (GenRobot, 2026), and Daimon-Infinity (Daimon Robotics, 2026) broaden the coverage of manipulation tasks and sensing modalities. We combine public manipulation data with self-collected UMI recordings to learn a bimanual world model conditioned on relative wrist-camera poses and gripper widths.

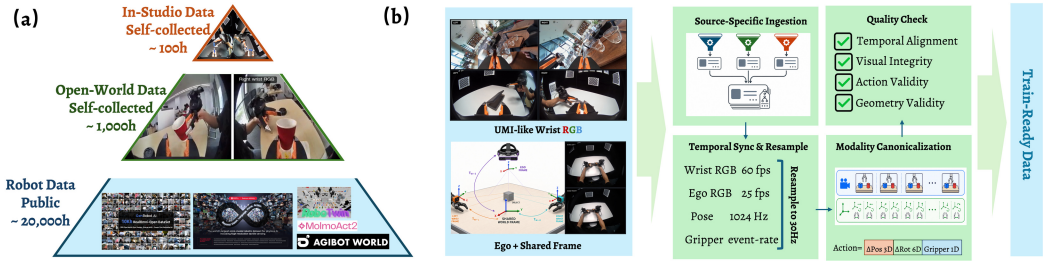

Figure 2: **UMI data and processing pipeline.** (a) A three-tier data pyramid combines public robot data with open-world and in-studio UMI recordings. (b) The pipeline performs temporal alignment, action canonicalization, and quality checks across data sources.

2.2 Camera-Controlled Interactive World Models

Camera-controlled world models generate observations along camera trajectories. CameraCtrl (He et al., 2025a), CamI2V (Zheng et al., 2024), and ReCamMaster (Bai et al., 2025a) introduce camera-pose conditioning into video generation, while HY-WorldPlay (Sun et al., 2026) and LingBot-World (Robbyant Team et al., 2026) support interactive exploration with persistent scene context. Extending this setting to multiple viewpoints requires modeling how independently controlled views evolve within a shared environment, as explored by SynCamMaster (Bai et al., 2025b), Gamma-World (Liu et al., 2026), and Khora (Zhao et al., 2026d). UMI-WM extends camera control to bimanual manipulation, where gripper actions change both wrist views and their shared scene.

2.3 World Models for Robotic Manipulation

World models as planners. World models generate visual plans and then translate them into robot actions. UniPi (Du et al., 2023), RoboDreamer (Zhou et al., 2024), and GE-Act 2.0 (AgiBot Research Team et al., 2026) couple visual planners with IDMs to recover actions. Large Video Planner (Chen et al., 2025a) similarly extracts actions from visual plans, while Goal-VLA (Chen et al., 2025b) derives target object poses from generated goal images. UMI-WM also uses an IDM to recover actions from video and defines BAF by comparing them with the conditioning actions.

World models as simulators. Genie Envisioner (Liao et al., 2026), GE-Sim 2.0 (Qiu et al., 2026), Interactive World Simulator (Wang et al., 2026b), and Ctrl-World (Guo et al., 2026) generate visual responses to robot actions for policy evaluation, while RynnWorld-Teleop (Zhao et al., 2026b) enables teleoperation within a generative world model. UMI-WM serves as an interactive simulator for bimanual manipulation, generating wrist-camera videos conditioned on per-arm relative camera poses and gripper widths. Appendix D reviews further related work.

3 UMI Data and Processing Pipeline

We organize public robot data and self-collected UMI recordings into a three-tier pyramid with a unified processing pipeline for world model training (Fig. 2).

3.1 UMI Data

Public manipulation data. Public UMI datasets, including RealOmni-Open and Daimon-Infinity, provide UMI recordings aligned to a common action representation. MolmoAct2 (Fang et al., 2026a) and AgiBot World 2026 (AgiBot World Team, 2026) contribute robot wrist-camera videos with actions converted to the same representation. RoboTwin (Chen et al., 2026b) and RoboLab (Yang et al., 2026) provide synthetic wrist-camera videos with relative camera poses and gripper widths.

Open-world UMI recordings. We collect open-world bimanual demonstrations with paired DM-DataClaw devices (Appendix A.1, Fig. 7). DM-DataClaw is a tweezer-style UMI gripper; each device records wrist-camera video, tracks camera pose through SLAM, and measures gripper width. The two camera trajectories remain in separate local frames.

In-studio UMI recordings. Our studio aligns both wrist trajectories in a shared world frame. AprilGrid geometry (Furgale et al., 2013) and calibrated intrinsics yield synchronized inter-wrist poses for cross-view geometric supervision (Appendix A.1, Fig. 8). The public RealOmni-Open and

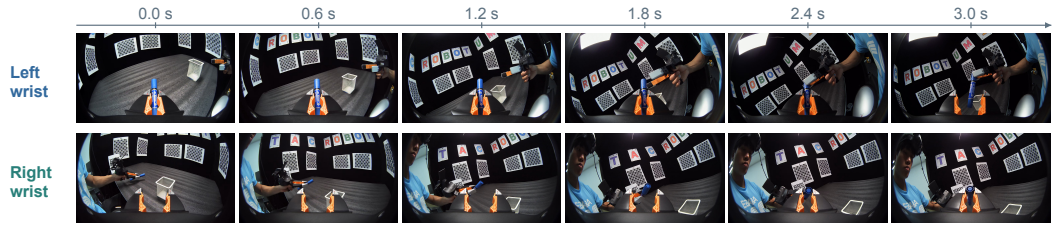

Figure 3: **In-studio bimanual recording.** Synchronized left- and right-wrist views of a marker handover, sampled every 0.6 s along a shared timeline. AprilGrid targets support camera registration.

Daimon-Infinity recordings do not provide this shared-frame registration. We record synchronized wrist and stereo egocentric RGB videos; UMI-WM uses only the two wrist views with camera poses and gripper widths as conditioning actions (Fig. 3).

3.2 Processing Pipeline and Action Representation

We convert heterogeneous data sources into a common wrist-camera observation–action schema by synchronizing video, camera poses, and gripper states at 30 Hz. We retain available calibration for image rectification and geometry-aware modeling, reject invalid samples, and divide trajectories into aligned training windows. Appendix A.2 details ingestion, resampling, rectification, and validation.

To remove dependence on global coordinates and support open-world recordings whose wrists are tracked in separate local frames, we express each trajectory relative to the first frame t_0 of its window. For arm $i \in \{L, R\}$ with camera pose $(\mathbf{R}_t^i, \mathbf{p}_t^i)$ in its available reference frame, we jointly define the relative pose and 10D action as

$$\begin{aligned} \Delta\mathbf{p}_t^i &= (\mathbf{R}_{t_0}^i)^\top(\mathbf{p}_t^i - \mathbf{p}_{t_0}^i), \qquad \Delta\mathbf{R}_t^i = (\mathbf{R}_{t_0}^i)^\top\mathbf{R}_t^i, \\ \mathbf{a}_t^i &= \left[(\Delta\mathbf{p}_t^i)^\top, \operatorname{rot6d}(\Delta\mathbf{R}_t^i)^\top, g_t^i\right]^\top \in \mathbb{R}^{10}, \qquad \operatorname{rot6d}(\mathbf{R}) = \left[\mathbf{R}_{:,1}^\top, \mathbf{R}_{:,2}^\top\right]^\top. \end{aligned} \tag{1}$$

We use the continuous 6D rotation representation of Zhou et al. (2019). For tweezer-style devices, g_t^i is the recorded opening angle, termed gripper width and used without temporal differencing.

4 UMI-WM: Interactive Multi-View World Modeling

4.1 Action Conditioning

Let $\mathcal{O}_{1:H}$ be the paired-wrist observation history, $\mathcal{A} = \{\mathbf{a}_{1:T}^L, \mathbf{a}_{1:T}^R\}$ the per-arm actions (Eq. 1), and $\mathcal{G}$ the available camera intrinsics and extrinsics. We initialize a video DiT from Matrix-Game 3.5 (Qian et al., 2026). A frozen VAE encodes each view; a shared temporal transformer encodes the aligned per-arm actions. Conditional prediction and view-specific action injection are

$$p_\theta\left(\mathcal{O}_{H+1:T}^{L,R} \mid \mathcal{O}_{1:H}, \mathcal{A}, \mathcal{G}\right), \qquad \widetilde{\mathbf{h}}_{v,n,s}^{(\ell)} = \mathbf{h}_{v,n,s}^{(\ell)} + \gamma_\ell \mathbf{u}_n^{\pi(v)}, \tag{2}$$

where $\mathcal{O}_{H+1:T}^{L,R}$ denotes the paired future wrist videos; $\mathbf{h}_{v,n,s}^{(\ell)}$ is the visual token at view v, latent time index n, spatial position s, and layer ℓ; $\pi(v)$ maps the view to its associated arm; $\mathbf{u}_n^{\pi(v)}$ is the encoded action state; and γ_ℓ is a learned scalar gate. The action state is broadcast over that view's spatial tokens before each DiT block, while the hub exchanges information across the two arms.

4.2 Multi-View Modeling and Causal Generation

SRAE and Sparse Hub Attention. Adapting Simplex Rotary Agent Encoding (SRAE) and Sparse Hub Attention from Gamma-World (Liu et al., 2026), we encode view, role, and actor identities separately and exchange information between wrist streams through learned hub tokens. Actor identity routes the appropriate per-arm action state in Eq. 2, while the hub receives a validity-masked summary of both arms, preserving view-specific control and shared-scene context. Appendix B.1 details tokenization, rotary identity construction, attention connectivity, and action routing.

Camera-aware geometry conditioning. We combine CameraNoise (Zhao et al., 2026a) with ray-based and relative-pose conditioning inspired by PRoPE (Li et al., 2025b) and Matrix-Game

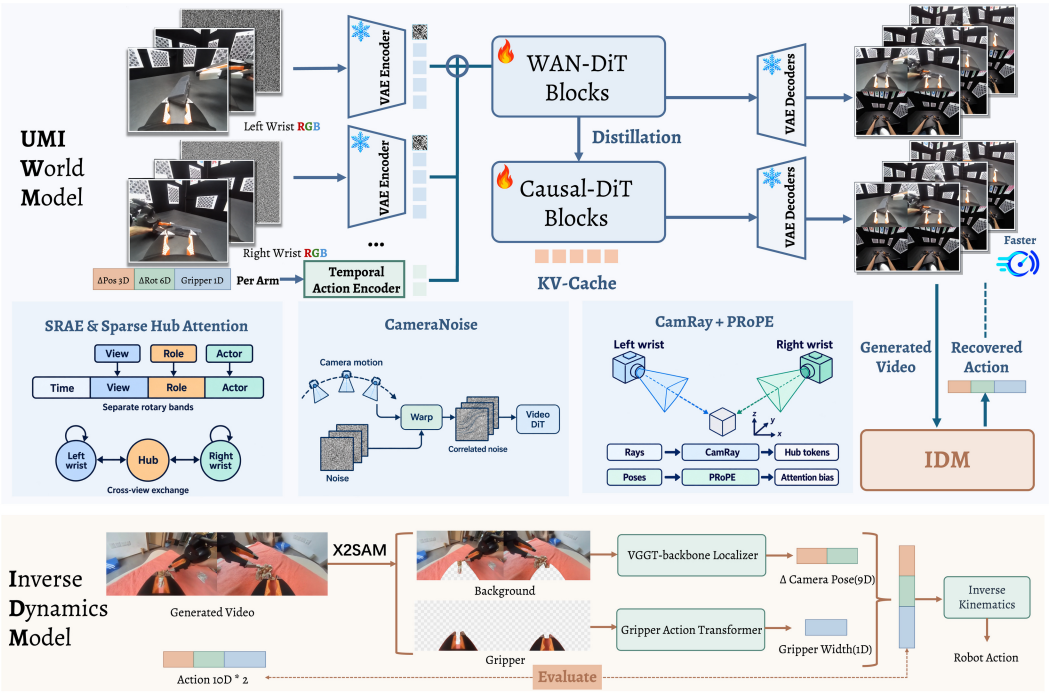

Figure 4: **UMI-WM overview.** SRAE and Sparse Hub Attention coordinate wrist streams; CameraNoise supplies motion-aware noise; CamRay and PRoPE provide geometric conditioning. Distillation enables four-step causal generation. The IDM recovers actions from generated videos for BAF evaluation and robot execution via inverse kinematics.

3.5 (Qian et al., 2026), coupling per-view motion control with cross-view geometric exchange. This interface supports both local-frame and calibrated shared-frame trajectories; Appendix B.2 details the noise construction, geometric features, and attention biases.

Progressive training and causal generation. Training follows the data pyramid from broad single-wrist coverage to joint dual-wrist modeling and calibrated geometry refinement. Following CausVid (Yin et al., 2025) and Self Forcing (Huang et al., 2025c), we then convert the model to block-causal generation and distill it to four denoising steps for interactive rollout. Appendix B.3 details the training stages, objectives, causal masks, distillation, and inference.

4.3 Inverse Dynamics Model and Bimanual Action Fidelity

Inverse dynamics model. Similar to VPT (Baker et al., 2022), our IDM recovers per-arm relative camera poses and gripper widths from generated wrist videos (Fig. 4). X2SAM (Wang et al., 2026a) masks separate the gripper foreground from the scene background. For pose recovery, a trainable temporal head processes frozen VGGT (Wang et al., 2025) background features. For gripper estimation, a temporal transformer processes fixed RGB crops that preserve local appearance, using the foreground masks as auxiliary inputs. We train both branches on real UMI videos from the training split only, with supervision from recorded translations, rotations, gripper openings, and temporal changes in opening; the IDM training set shares no episodes with the 320 held-out test episodes used in Section 5. At each conditioning timestamp, the predicted pose relative to the initial frame and the gripper width form the 10D action in Eq. 1.

Bimanual action fidelity. For arm $i \in \{L, R\}$ and component $k \in \{\text{pos}, \text{rot}, \text{grip}\}$, let E_k^i be the mean error between conditioning and video-recovered actions. We define

$$\text{BAF} = 100 \exp\left[-\frac{\lambda}{6} \sum_{i \in \{L,R\}} \sum_{k} \left[\frac{E_k^i}{\tau_k} - 1\right]_+\right], \tag{3}$$

where $[x]_+ = \max(x, 0)$. Higher scores indicate closer action agreement, reaching 100 when all six mean errors lie within their tolerance bands. We use $\tau = (5\,\mathrm{cm}, 5^\circ, 0.1\,\mathrm{rad})$ and $\lambda = 0.25$ for all models; Appendix B.4 details component errors and score aggregation over valid predictions.

5 EXPERIMENTS

We compare UMI-WM with baselines (Table 1), probe its components (Table 2), and evaluate Bimanual Action Fidelity (BAF) on GT and generated wrist videos (Table 3).

5.1 EXPERIMENTAL SETUP

Test data. We evaluate UMI-WM on 320 UMI test episodes comprising 1,008 paired-wrist windows, each containing 189 frames with per-arm controls at 30 Hz. Task descriptions are taken from existing annotations or generated from wrist-camera videos before evaluation. Scores are averaged over two generation seeds with equal episode weights.

Evaluation metrics. We evaluate visual fidelity using PSNR, SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), FID (Heusel et al., 2017), and FVD (Unterthiner et al., 2018), video quality using VBench (Huang et al., 2024) and WorldArena (Shang et al., 2026) (Appendix C.1), and action following using BAF (Appendix C.2). We compare generated frames with the corresponding ground-truth frames, excluding the initial conditioning frame; FID and FVD use one selected window per episode. We report paired-wrist FPS for generation speed.

Table 1: **Comparison with baseline world models.** Bold/underline: best/second-best. GT is unranked; – means not reported. Gamma-World is fine-tuned on our UMI training split. VBench: SC/BC subject/background consistency, MS motion smoothness, IQ imaging quality. WorldArena: JS JEPA similarity, PS photometric smoothness, PV perspectivity. FPS: paired frames/s.

	Reference and distribution fidelity					**VBench**				**WorldArena**			**Eff.**
Model	PSNR↑	SSIM↑	LPIPS↓	FID↓	FVD↓	SC↑	BC↑	MS↑	IQ↑	JS↑	PS↑	PV↑	FPS↑
GT video	–	–	–	–	–	0.860	0.926	0.984	0.716	1.000	3.377	3.750	–
Text-conditioned world models													
Wan2.2 TI2V-5B	8.345	0.312	0.715	99.76	3477.06	0.824	0.903	**0.988**	0.604	0.059	1.827	**3.875**	0.18
Cosmos Predict2.5	7.759	0.308	0.685	98.62	967.40	0.688	0.850	0.974	0.631	0.284	1.747	3.500	0.10
MiniMax H3	9.477	0.371	0.577	58.91	687.94	0.768	0.874	0.982	0.609	0.080	1.772	3.375	0.70
Action-conditioned world model													
Gamma-World	12.264	0.548	0.397	40.30	331.25	0.874	0.933	0.984	0.655	0.382	4.236	3.375	0.55
Camera-controlled world models													
HY-WorldPlay 5B	8.599	0.338	0.606	68.15	638.08	0.825	0.899	0.968	0.705	0.135	2.296	3.250	0.81
LingBot-World-Fast	7.660	0.270	0.752	128.16	2190.36	0.627	0.855	0.958	0.464	0.179	1.669	3.250	2.32
UMI-WM (ours)													
UMI-WM Teacher	**12.513**	0.576	**0.332**	**33.52**	**232.53**	**0.885**	**0.938**	0.984	**0.707**	0.386	**4.339**	3.500	0.26
UMI-WM Student	12.334	**0.580**	0.336	37.89	284.68	0.875	0.933	0.984	0.655	**0.408**	4.300	3.125	1.81

5.2 BASELINE COMPARISONS

Table 1 compares UMI-WM with text-, action-, and camera-conditioned baselines. Text-conditioned models such as MiniMax-H3 (MiniMax, 2026) receive task descriptions and initial wrist images; camera-controlled models additionally receive mapped wrist trajectories. Both groups generate wrists separately. Gamma-World is fine-tuned on our UMI training split and uses paired 20D actions encoding each arm's relative camera pose and gripper width (Eq. 1).

Figure 5 compares UMI-WM with Cosmos Predict2.5, MiniMax-H3, and HY-WorldPlay 5B at 1, 3, and 5 s. Teacher and Student maintain the backpack appearance and paired-wrist layout more consistently across the sequence. Further qualitative comparisons appear in Appendix C.1.

Generation quality. The Teacher achieves the best PSNR, LPIPS, FID, and FVD, and the Student the best SSIM. Gamma-World, the strongest baseline, reaches an FVD of 331.25 and an LPIPS of 0.397. Across the reported VBench and WorldArena metrics, at least one UMI-WM variant ranks



Figure 5: **Paired-wrist qualitative comparison.** Rows show GT and models; columns show three timestamps, each pairing left and right wrist views. Red boxes mark baseline failures.

first or ties for second. The gains in frame reconstruction and distributional fidelity demonstrate UMI-WM's advantage in wrist-view prediction over the compared baselines.

Excluded benchmark metrics. We also evaluated WorldModelBench (Li et al., 2025a), iWorld-Bench (Fang et al., 2026b), and EWMBench (Hu et al., 2025), but omit them from the main comparison because their tested metrics were unreliable on our wrist videos: static or noisy controls could receive favorable scores, pose estimates were unstable, and the released gripper tracker often failed. Appendix C.1 summarizes these checks and the resulting metric selection.

Generation efficiency. Timings use one A100-80GB GPU and include conditioning, denoising, and decoding for both wrists. The four-step Student reaches 1.81 paired frames/s versus 0.26 for the 30-step Teacher, a 6.9× speedup, for joint generation of 189-frame wrist videos. LingBot-World-Fast reaches 2.32 paired frames/s by generating each wrist separately. The Student combines strong visual prediction quality with substantially faster paired-wrist generation than the Teacher.

5.3 Geometry Ablations and Sampling Sensitivity

Geometry probes. Removing CamRay with geometric hub routing or PRoPE lowers BAF from 76.36 to 46.79 or 47.26, respectively, while also worsening LPIPS and imaging quality. Both variants start from the Stage 2 checkpoint at step 15,000 and train the retained adapters, with the backbone and action conditioner frozen. Within these probes, the full geometry configuration achieves the best visual and action fidelity, supporting our design for geometry conditioning.

Student sampling. With the checkpoint, observations, actions, seeds, and horizons fixed, four steps achieve the best LPIPS of 0.336 with BAF 71.55 at 1.81 paired frames/s. Two steps increase FPS to 3.31 but reduce BAF to 53.82; eight steps raise BAF to 76.12 and improve imaging quality but halve FPS to 0.94. This supports four steps as our default balance between action fidelity and speed.

Table 2: **Geometry probes and sampling sensitivity.** Bold/underline: best/second-best values within the Student sampling block for each metric.

Variant	LPIPS ↓	Imaging quality ↑	BAF ↑	Paired-wrist FPS ↑
UMI-WM Teacher	0.332	0.707	76.36	0.26
Geometry probes				
w/o CamRay + geometric Hub	0.456	0.654	46.79	0.65
w/o PRoPE	0.434	0.655	47.26	0.81
Student sampling				
Student, 2 steps	0.416	0.638	53.82	**3.31**
Student, 4 steps (default)	**0.336**	<u>0.655</u>	<u>71.55</u>	<u>1.81</u>
Student, 8 steps	<u>0.385</u>	**0.673**	**76.12**	0.94

5.4 Bimanual Action Fidelity Analysis

Table 3 reports per-wrist motion errors (Appendix B.4) and BAF (Eq. 3). All models share the same frozen IDM and query timestamps. We compute BAF per window and IDM seed, then average seeds within windows and windows within episodes, weighting episodes equally. The Teacher and Student reach BAF 76.36 and 71.55, above adapted Gamma-World at 52.93, the best camera-controlled baseline at 41.26, and the static reference at 18.78. UMI-WM leads most error columns; Matrix-Game 3.0 has lower right-wrist position error (15.59 cm versus the Teacher's 15.63 cm). Text-conditioned scores remain near the static reference. These results support UMI-WM's advantage in generating wrist-view motion that agrees with the reference actions for both arms.

BAF adds a monotonic action-fidelity scale to the visual and distributional metrics above, anchored by GT at 97.72 and static at 18.78: higher scores closer to GT indicate stronger agreement with conditioning actions, whereas lower scores closer to static indicate weaker controllable motion. Scores below static may indicate a collapsed video with little effective information; for camera- and text-conditioned baselines, BAF measures agreement with recorded actions (Appendix C.2).

Table 3: **Bimanual Action Fidelity comparison.** Bold/underline: best/second-best among generated models; GT and static controls are references. Gamma-World is fine-tuned on our UMI training split; all entries use the shared 320-episode test set and the same frozen IDM.

	Left wrist (L)			Right wrist (R)			
Video source / model	$E^L_{\text{pos}}\downarrow$ (cm)	$E^L_{\text{rot}}\downarrow$ (°)	$E^L_{\text{grip}}\downarrow$ (rad)	$E^R_{\text{pos}}\downarrow$ (cm)	$E^R_{\text{rot}}\downarrow$ (°)	$E^R_{\text{grip}}\downarrow$ (rad)	BAF ↑
GT video (shared test)	3.89	3.45	0.0275	4.52	4.10	0.0231	97.72
Static conditioning frame (control)	23.01	77.21	0.0762	34.43	85.07	0.0915	18.78
Text-Conditioned World Models							
Wan2.2 TI2V-5B	22.69	77.28	0.1011	34.46	85.02	0.1168	18.64
Cosmos Predict2.5-2B	32.75	85.62	0.1571	35.45	88.04	0.1625	17.54
MiniMax-H3	21.19	83.12	0.1059	39.02	101.95	0.1085	15.20
Cosmos3-Nano-Policy-DROID (I2V)	28.46	106.11	0.1250	37.84	75.43	0.1568	14.78
RynnWorld-Teleop (text+image SFT)	22.93	77.81	0.1037	33.59	84.26	0.1273	18.94
Action-Conditioned World Models							
Genie-Envisioner (GE-Sim)	21.63	72.97	0.0855	35.84	79.42	0.0621	20.57
RynnWorld-Teleop (control band)	22.41	77.50	0.1171	33.88	85.14	0.1224	18.72
Gamma-World	19.42	25.27	0.0808	24.39	28.70	0.0771	52.93
Camera-Controlled World Models							
HY-WorldPlay 5B	16.66	42.11	0.1264	25.33	59.51	0.1726	36.53
LingBot-World-Fast	38.71	56.68	0.1411	30.74	65.96	0.1444	24.19
DreamX-World-5B-Cam	49.53	102.66	0.1249	28.56	70.63	0.1208	14.89
Matrix-Game 2.0	22.72	51.33	0.1349	23.78	57.52	0.1070	32.36
Matrix-Game 3.0	20.33	37.53	0.1424	**15.59**	55.24	0.1712	41.26
YUME 1.5	23.18	67.70	0.1048	31.83	74.95	0.1511	22.51
Infinite-World	19.88	63.08	0.1345	31.62	66.66	0.1685	25.23
UMI-WM Teacher	**12.49**	<u>6.66</u>	<u>0.0285</u>	<u>15.63</u>	**18.88**	<u>0.0188</u>	**76.36**
UMI-WM Student	<u>14.44</u>	**6.16**	**0.0270**	17.91	<u>24.78</u>	**0.0165**	<u>71.55</u>

We map each baseline's available controls to the common per-arm action space before applying the same frozen IDM and BAF computation. In Table 3, "I2V" means image-to-video input, "text+image SFT" is RynnWorld-Teleop's supervised text-and-image route, and "GE-Sim" is Genie-Envisioner's rendered-control interface. RynnWorld-Teleop's "control band" renders mapped controls as a dual-hand skeleton video (Appendix C.2).

5.5 Applications

VR teleoperation in the world model. Figure 6 demonstrates UMI-WM as a visual environment for bimanual teleoperation. The interface maps tracked hand motion and finger separation to per-arm camera-pose and gripper controls, then jointly generates both wrist views as visual feedback for subsequent commands. Hand tracking and a UMI gripper overlay visualize the input mapping, while the IDM recovers generated actions for inverse-kinematics mapping to robot configurations.

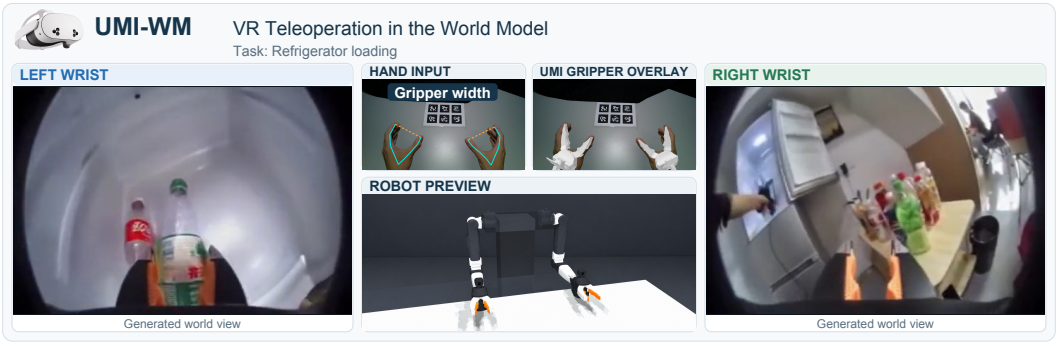

Figure 6: **VR teleoperation in the world model.** Interface for a refrigerator-loading rollout with hand tracking, a UMI gripper overlay, generated wrist views, and a robot preview.

Robot action planning. To validate the interface, we execute a UMI-WM rollout on the physical robot. After interactive approval, the IDM recovers relative wrist-camera poses and gripper widths. A calibrated mapping converts the poses into end-effector targets for inverse kinematics, while the recovered widths drive the grippers. The matched physical and wrist-view sequences from initial approach to completed transport are shown in Appendix C.3 (Fig. 11).

Visual domain randomization. Similar to ReShoot (Kim et al., 2026), UMI-WM regenerates paired wrist videos after consistent appearance edits to both initial frames, keeping per-arm trajectories and gripper controls fixed. The edits expand the visual coverage of the demonstrations, and BAF decides which edited rollouts join policy training alongside the originals. Fig. 12 in Appendix C.3 shows hue, object-color, and background edits to both initial frames, the rollout regenerated from each, and its BAF relative to the unedited rollout.

Data quality screening. UMI-WM also audits recorded data: the frozen IDM recovers actions from raw wrist-camera frames, and segments where they persistently depart from the recorded SLAM trajectories and gripper controls are flagged for manual review, trajectory re-estimation, or temporal realignment. Fig. 13 in Appendix C.3 shows per-timestep discrepancy traces for a clean recording and two corrupted controls, screening scores over the test windows, and a flagged real recording whose recorded and recovered wrist translations and 3D paths diverge.

6 Conclusion

We introduced UMI-WM, an interactive world model that jointly predicts wrist-camera videos from per-arm camera motion and gripper controls. A unified UMI data pipeline, cross-view information exchange, and camera-aware conditioning support few-step causal generation. On the shared 320-episode evaluation, the Teacher leads PSNR, LPIPS, FID, and FVD, while the Student leads SSIM and reaches 1.81 paired-wrist FPS, a 6.9$\times$ speedup. Their BAF scores of 76.36 and 71.55 lie between the GT reference at 97.72 and the static reference at 18.78, and exceed all compared baselines on the higher-is-better action-fidelity scale. Removing camera-aware conditioning in the geometry probes lowers both imaging quality and BAF, supporting its role in action following. Together with the physical execution demonstration, these results connect wrist-view prediction, action recovery, and robot planning through a shared control interface.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, support literature retrieval and discovery, and assist research ideation and execution, including experiment planning, code development and debugging, data-processing utilities, and figure preparation. The authors reviewed suggestions against cited sources, implementation code, experimental records, and visual results and take responsibility for the final manuscript and its claims.

ETHICS STATEMENT

Our UMI data were collected internally for research on visual world modeling. Wrist-camera and egocentric recordings underwent privacy screening before use, and the paper reports aggregate evaluations together with selected task-scene visualizations. Public datasets and third-party assets are used only under their respective licenses and permissions.

REPRODUCIBILITY STATEMENT

Sections 3–5 describe the data sources and processing, coordinate and action conventions, model components, staged training, IDM and BAF, and evaluation protocols. The appendices expand the model objectives, geometry-probe and Gamma-World training settings, benchmark configurations, metric aggregation, validity handling, and evaluator limitations.

REFERENCES

AgiBot Research Team, Renhang Liu, Wenzhi Zhao, Zhuo Yang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Guanghui Ren, Youlun Peng, Rongjun Jin, Nan Wang, Sukai Wang, Xindong He, Jinyuan Feng, Ziyu Xiong, Linqing Zhong, Yifei Wei, Feng Han, Long Zhang, Da Huang, Nanshu Zhao, Chenghao Yin, Mo Wu, Zhaodong Yan, Kongtao Hu, Yuxiang Yan, Aogelijiang Niyazi, Yu Fang, Jia Zeng, Lizhu Meng, Daizhen Lv, Haoyu Cao, Zhiwen Hou, Lianjin Ye, Yuehan Niu, Zhikai Cai, Xuan Hu, Hui Min, Xiongfeng Cai, Yue Liao, Jing Wu, Soujanya Poria, Ye Li, Sanping Zhou, and Maoqing Yao. GE-Act 2.0: Pretraining and scaling a world-action model for robotic manipulation, 2026. URL https://arxiv.org/abs/2609.05588.

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

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J. Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in Atari. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6bdde0373d53d4a501249547084bed43-Abstract-Conference.html.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B. Lindell, and Sergey Tulyakov. AC3D: Analyzing and improving 3D camera control in video diffusion transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pp. 22875–22889. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02130.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative

rendering from a single video. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025*, pp. 14834–14844. IEEE, 2025a. doi: 10.1109/ICCV51701.2025.01376.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. SynCamMaster: Synchronizing multi-camera video generation from diverse viewpoints. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=m8Rk3HLGFx.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video PreTraining (VPT): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9c7008aff45b5d8f0973b23e1a22ada0-Abstract-Conference.html.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott E. Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning, ICML 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Ryan D. Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael S. Ryoo, Paul E. Debevec, and Ning Yu. Go-with-the-Flow: Motion-controllable video diffusion models using real-time warped noise. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pp. 13–23. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00011.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 4724–4733. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.502.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion Forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2aee1c4159e48407d68fe16ae8e6e49e-Abstract-Conference.html.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T. Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, Vincent Sitzmann, and Yilun Du. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025a.

Haonan Chen, Bangjun Wang, Jingxiang Guo, Tianrui Zhang, Yiwen Hou, Xuchuan Huang, Chenrui Tie, and Lin Shao. World4Omni: A zero-shot framework from image generation world model to robotic manipulation. *arXiv preprint arXiv:2506.23919*, 2025b.

Jason Chen, I-Chun Arthur Liu, Gaurav S. Sukhatme, and Daniel Seita. CRAFT: Video diffusion for bimanual robot data generation. *arXiv preprint arXiv:2604.03552*, 2026a.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026b.

Yushuo Chen, Xiaoyu Shi, Xiaoshi Wu, Xintao Wang, Pengfei Wan, and Yebin Liu. HandsOnWorld: Unconstrained egocentric video generation with camera-disentangled hand control. *arXiv preprint arXiv:2607.02075*, 2026c.

Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-TeleVision: Teleoperation with immersive active visual feedback. In *Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2729–2749. PMLR, 2024. URL https://proceedings.mlr.press/v270/cheng25b.html.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal Manipulation Interface: In-the-wild robot teaching without in-the-wild robots. In *Robotics: Science and Systems XX*, 2024. doi: 10.15607/RSS.2024.XX.045.

Daimon Robotics. Daimon-Infinity, 2026. URL https://modelscope.cn/datasets/daimonrobotics/Daimon-Infinity. Dataset resource; accessed on September 10, 2026.

Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer, 2024. URL https://oasis-model.github.io/.

Runyu Ding, Yuzhe Qin, Jiyue Zhu, Chengzhe Jia, Shiqi Yang, Ruihan Yang, Xiaojuan Qi, and Xiaolong Wang. Bunny-VisionPro: Real-time bimanual dexterous teleoperation for imitation learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2025*, pp. 12248–12255. IEEE, 2025. doi: 10.1109/IROS60139.2025.11247017.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, Geng Li, Jifan Li, Ruimin Lin, Qingfeng Shi, Bingze Song, Lei Sun, Jing Tang, Ruitian Tian, Jun Wang, Jiahong Wu, Pengfei Zhang, Shen Zhang, and Jiashu Zhu. DreamX-World 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/1d5b9233ad716a43be5c0d3023cb82d0-Abstract-Conference.html.

Haoquan Fang, Jiafei Duan, Donovan Clay, Sam Wang, Shuo Liu, Weikai Huang, Xiang Fan, Wei-Chuan Tsai, Shirui Chen, Yi Ru Wang, Shanli Xing, Jaemin Cho, Jae Sung Park, Ainaz Eftekhar, Peter Sushko, Karen Farley, Angad Wadhwa, Cole Harrison, Winson Han, Ying-Chun Lee, Eli VanderBilt, Rose Hendrix, Suveen Ellawela, Lucas Ngoo, Joyce Chai, Zhongzheng Ren, Ali Farhadi, Dieter Fox, and Ranjay Krishna. MolmoAct2: Action reasoning models for real-world deployment. *arXiv preprint arXiv:2605.02881*, 2026a.

Jianjie Fang, Yingshan Lei, Qin Wan, Ziyou Wang, Yuchao Huang, Yongyan Xu, Baining Zhao, Weichen Zhang, Chen Gao, Xinlei Chen, and Yong Li. iWorld-Bench: A benchmark for interactive world models with a unified action generation framework. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026b.

Paul Timothy Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2013*, pp. 1280–1286. IEEE, 2013. doi: 10.1109/IROS.2013.6696514.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi "Jim" Fan. DreamDojo: A generalist robot world model from large-scale human videos. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026.

GenRobot. 10Kh RealOmni-Open Dataset, 2026. URL https://www.genrobot.ai/data/open-dataset. Dataset resource; accessed on September 10, 2026.

Jingxiang Guo, Jiayu Luo, Zhenyu Wei, Yiwen Hou, Zhixuan Xu, Xiaoyi Lin, Chongkai Gao, and Lin Shao. TelePhantom: A user-friendly teleoperation system with virtual assistance for enhanced effectiveness. *arXiv preprint arXiv:2412.13548*, 2024.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. MineWorld: a real-time and open-source interactive world model on Minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Oxh5CstDJU.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025a. URL https://openreview.net/forum?id=Z4evOUYrk7.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. CameraCtrl II: Dynamic scene exploration via camera-controlled video diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025*, pp. 13416–13426. IEEE, 2025b. doi: 10.1109/ICCV51701.2025.01246.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025c.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Yue Hu, 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.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixé, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026.

Siyuan Huang, Liliang Chen, Pengfei Zhou, Shengcong Chen, Yue Liao, Zhengkai Jiang, Yue Hu, Peng Gao, Hongsheng Li, Maoqing Yao, and Guanghui Ren. EnerVerse: Envisioning embodied future space for robotics manipulation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/360052c2c6d0c8ec24c476d43236ab25-Abstract-Conference.html.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025c. URL http://papers.nips.cc/paper_files/paper/2025/hash/f4823f831af67a3ef15e41a85434422a-Abstract-Conference.html.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pp. 21807–21818. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02060.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, Loic Magne, Ajay Mandlekar, Avnish Narayan, You Liang Tan, Guanzhi Wang, Jing Wang, Qi Wang, Yinzhen Xu, Xiaohui Zeng, Kaiyuan Zheng, Ruijie Zheng, Ming-Yu Liu, Luke Zettlemoyer, Dieter Fox, Jan Kautz, Scott Reed, Yuke Zhu, and Linxi Fan. DreamGen: Unlocking generalization in robot learning through neural trajectories. In *Conference on Robot Learning, CoRL 2025*, 2025.

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

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

Chiyoung Kim, Min Sung Choi, Jinho Ju, Chanhoe Gu, Donghwan Hwang, Wonseok Choi, Woongsun Jeon, and Minhyeok Lee. ReShoot: Generative visual domain randomization of recorded robot demonstrations for visuomotor policy learning, 2026. URL https://arxiv.org/abs/2609.19661.

Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas J. Guibas, and Gordon Wetzstein. Collaborative Video Diffusion: Consistent multi-video generation with camera control. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/1d49235669869ab737c1da9d64b7c769-Abstract-Conference.html.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World action models beyond pixels with in-the-wild egocentric human data. *arXiv preprint arXiv:2607.08436*, 2026.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025a. URL http://papers.nips.cc/paper_files/paper/2025/hash/4ec03ed08a3fcb59e1c815b5598beff1-Abstract-Datasets_and_Benchmarks_Track.html.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025b. URL http://papers.nips.cc/paper_files/paper/2025/hash/17a7075094632c88cccdd86270ad715b-Abstract-Conference.html.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. In *Robotics: Science and Systems XXI*, 2025c.

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*, 2025d.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Hu Yue, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie Envisioner: A unified world foundation platform for robotic manipulation. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Fangfu Liu, Kai He, Tianchang Shen, Tianshi Cao, Sanja Fidler, Yueqi Duan, Jun Gao, Igor Gilitschenski, Zian Wang, and Xuanchi Ren. Gamma-World: Generative multi-agent world modeling beyond two players. *arXiv preprint arXiv:2605.28816*, 2026.

Kehui Liu, Zhongjie Jia, Yang Li, Zhaxizhuoma, Pengan Chen, Song Liu, Xin Liu, Pingrui Zhang, Haoming Song, Xinyi Ye, Nieqing Cao, Zhigang Wang, Jia Zeng, Dong Wang, Yan Ding, Bin Zhao, and Xuelong Li. FastUMI-100K: Advancing data-driven robotic manipulation with a large-scale UMI-style dataset. *arXiv preprint arXiv:2510.08022*, 2025.

Yawen Luo, Xiaoyu Shi, Jianhong Bai, Menghan Xia, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. CamCloneMaster: Enabling reference-based camera control for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers, SA Conference Papers 2025*, pp. 20:1–20:10. ACM, 2025. doi: 10.1145/3757377.3763841.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026*, 2026.

MiniMax. MiniMax H3: An open model breaking the boundaries between tasks and modalities, 2026. URL https://www.minimax.io/blog/minimax-h3.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Kl
'ar, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025a.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, Prithvijit Chattopadhyay, Mike Chen, Yongxin Chen, Yu Chen, Shuai Cheng, Yin Cui, Jenna Diamond, Yifan Ding, Jiaojiao Fan, Linxi Fan, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Ruiyuan Gao, Yunhao Ge, Jinwei Gu, Aryaman Gupta, Siddharth Gururani, Imad El Hanafi, Ali Hassani, Zekun Hao, Jacob Huffman, Joel Jang, Pooya Jannaty, Jan Kautz, Grace Lam, Xuan Li, Zhaoshuo Li, Maosheng Liao, Chen-Hsuan Lin, Tsung-Yi Lin, Yen-Chen Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Yifan Lu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Seungjun Nah, Yashraj Narang, Abhijeet Panaskar, Lindsey Pavao, Trung Pham, Morteza Ramezanali, Fitsum Reda, Scott Reed, Xuanchi Ren, Haonan Shao, Yue Shen, Stella Shi, Shuran Song, Bartosz Stefaniak, Shangkun Sun, Shitao Tang, Sameena Tasmeen, Lyne Tchapmi, Wei-Cheng Tseng, Jibin Varghese, Andrew Z. Wang, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Jiashu Xu, Dinghao Yang, Xiaodong Yang, Haotian Ye, Seonghyeon Ye, Xiaohui Zeng, Jing Zhang, Qinsheng Zhang, Kaiwen Zheng, Andrew Zhu, and Yuke Zhu. World simulation with video foundation models for physical AI. *arXiv preprint arXiv:2511.00062*, 2025b.

Edwin Olson. AprilTag: A robust and flexible visual fiducial system. In *IEEE International Conference on Robotics and Automation, ICRA 2011*, pp. 3400–3407. IEEE, 2011. doi: 10.1109/ICRA.2011.5979561.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation, ICRA 2024*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pp. 4172–4182. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00387.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-Game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

Zezhong Qian, Xiaowei Chi, Yuming Li, Shizun Wang, Zhiyuan Qin, Xiaozhu Ju, Sirui Han, and Shanghang Zhang. WristWorld: Generating wrist-views via 4D world models for robotic manipulation. *arXiv preprint arXiv:2510.07313*, 2025.

Boxiang Qiu, Liliang Chen, Yue Liao, Nan Wang, Lintao Wang, Jiayi Luo, Wenzhi Zhao, Shengcong Chen, Di Chen, Ye Li, Chen Gao, Shuicheng Yan, Si Liu, Maoqing Yao, and Guanghui Ren. GE-Sim 2.0: A roadmap towards comprehensive closed-loop video world simulators for robotic manipulation. *arXiv preprint arXiv:2605.27491*, 2026.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026.

Omar Rayyan, John Abanes, Mahmoud Hafez, Anthony Tzes, and Fares J. Abu-Dakka. MV-UMI: A scalable multi-view interface for cross-embodiment learning. *arXiv preprint arXiv:2509.18757*, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. GEN3C: 3D-Informed world-consistent video generation with precise camera control. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pp. 6121–6132. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00574.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026.

Shilin Shan, Chuhao Zhou, Ruize Wang, Xinyan Chen, Xiangyu Chen, Xinyu Zhou, Boyu Ma, Iris Yuxuan Hu, Jingliang Li, Celeste Yuxuan Hu, Geng Li, Guohao Chen, Tianrui Zhu, Zhe Li, Yanjie Ze, Haoran Geng, Zhiyang Dou, Jianxin Bi, Yuejiang Liu, Jianshu Zhou, Jiachen Li, Paul Liang, Tatsuya Harada, Robert Katzschmann, Harold Soh, Na Li, Edward Johns, Danica Kragic, Jan Peters, Wojciech Matusik, Masayoshi Tomizuka, Jitendra Malik, and Jianfei Yang. Learning Physical Interaction: A survey of tactile- and force-aware robot learning. *arXiv preprint arXiv:2608.07558*, 2026.

Yu Shang, Xin Zhang, Yinzhou Tang, Lei Jin, Chen Gao, Wei Wu, and Yong Li. RoboScape: Physics-informed embodied world model. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/5c1ff00b27ba052039bb41531236baac-Abstract-Conference.html.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. WorldArena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

Yichao Shen, Fangyun Wei, Zhiying Du, Yaobo Liang, Yan Lu, Jiaolong Yang, Nanning Zheng, and Baining Guo. VideoVLA: Video generators can be generalizable robot manipulators. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/89a3b655a8b68ae1c76b768152c9c19d-Abstract-Conference.html.

Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1985*, pp. 245–254. ACM, 1985. doi: 10.1145/325334.325242.

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. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*, pp. 2818–2826. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.308.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Hao Wang, Limeng Qiao, Chi Zhang, Lin Ma, Guanglu Wan, Xiangyuan Lan, and Xiaodan Liang. X2SAM: Any segmentation in images and videos. *arXiv preprint arXiv:2605.00891*, 2026a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pp. 5294–5306. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00499.

Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. In *Robotics: Science and Systems XXII*, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers, SIGGRAPH 2024*. ACM, 2024. doi: 10.1145/3641519.3657518.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, Yidan Xietian, Jiangbo Pei, Liang Hu, Boyi Jiang, Hua Xue, Zidong Wang, Haofeng Sun, Wei Li, Wanli Ouyang, Xianglong He, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026c.

Isaac Ronald Ward, Michelle Ho, Houjun Liu, Aaron O. Feldman, Joseph A. Vincent, Liam Kruse, Sean Cheong, Duncan Eddy, Mykel J. Kochenderfer, and Mac Schwager. Foundational world models accurately detect bimanual manipulator failures. In *IEEE International Conference on Robotics and Automation, ICRA 2026*, 2026.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. DayDreamer: World models for physical robot learning. In *Conference on Robot Learning, CoRL 2022*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2226–2240. PMLR, 2022. URL `https://proceedings.mlr.press/v205/wu23c.html`.

Ruiqi Wu, Xuanhua He, Cheng Meng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M. Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026.

Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. In *Conference on Robot Learning, CoRL 2025*, 2025.

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

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=sFyTZEqmUY`.

Xuning Yang, Rishit Dagli, Alex Zook, Hugo Hadfield, Ankit Goyal, Stan Birchfield, Fabio Ramos, and Jonathan Tremblay. RoboLab: A high-fidelity simulation benchmark for analysis of task generalist policies. In *Robotics: Science and Systems XXII*, 2026.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/54dcf25318f9de5a7a01f0a4125c541e-Abstract-Conference.html.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pp. 6613–6623. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.00632.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pp. 22963–22974. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02138.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. TrajectoryCrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025*, pp. 100–111. IEEE, 2025. doi: 10.1109/ICCV51701.2025.00017.

Qihang Zhang, Lin Li, Luyao Zhang, Shuai Yang, Yiming Luo, Shuaiting Li, Ruilin Wang, Junke Wang, Jiahao Shao, Gangwei Xu, Jiaming Zhou, Yishu Shen, Yudong Jin, Fangyi Xu, Shuailei Ma, Jiaqi Liao, Guanxing Lu, Zifan Shi, Yongkun Wen, Yujie Zhao, Weixuan Tang, Xinyang Wang, Chaojian Li, Jiapeng Zhu, Ka Leong Cheng, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Native video-action pretraining for generalizable robot control. *arXiv preprint arXiv:2607.08639*, 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068.

Haoyu Zhao, Jiaxi Gu, Haoran Chen, Qingping Zheng, Yejin Choi, Hongyi Yang, Junqi Cheng, Yuang Zhang, Zenghui Lu, Huan Yu, Jie Jiang, Peng Shu, Zuxuan Wu, and Yu-Gang Jiang. CameraNoise: Enabling faithful camera control in video diffusion through geometry-flow-guided noise warping. In *Forty-third International Conference on Machine Learning, ICML 2026*, 2026a.

Haoyu Zhao, Xingyue Zhao, Hangyu Li, Biao Gong, Kehan Li, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. RynnWorld-Teleop: An action-conditioned world model for digital teleoperation. *arXiv preprint arXiv:2607.06558*, 2026b.

Haoyu Zhao, Zihao Zhao, Tianyu Deng, Ziqin Xu, Zihao Zhang, Xudong Wang, Jinxiang Guo, Chen Gao, Ziyi Ye, Yeying Jin, Jiaxi Gu, Zuxuan Wu, and Shuicheng Yan. Can MiniMax-H3 reason about the physical world? an evaluation of omni-modal generative model, 2026c. URL https://arxiv.org/abs/2609.18323.

Renjie Zhao, Yuxiang Wu, Mingyu Zhang, Jiaxin Li, Sisi Li, He Li, Yimin Sheng, Tianxi Tan, Zhenkai Zhang, Jiao Liang, Jianyi Zhu, and Yong-Lu Li. Population-scalable multi-agent world modeling. *arXiv preprint arXiv:2608.08600*, 2026d.

Zhaxizhuoma, Kehui Liu, Chuyue Guan, Zhongjie Jia, Ziniu Wu, Xin Liu, Tianyu Wang, Shuai Liang, Pengan Chen, Pingrui Zhang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. FastUMI: A scalable and hardware-independent universal manipulation interface with dataset. In *Conference on Robot Learning, CoRL 2025*, 2025.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Zirui Zheng, Jiaqian Yu, Xiongfeng Peng, Jun Shi, Mingyi Li, Chao Zhang, Weiming Li, Dong Wang, Huchuan Lu, and Xu Jia. Mem-World: Memory-augmented action-conditioned world models for persistent robot manipulation. In *Conference on Robot Learning, CoRL 2026*, 2026.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *Forty-first International Conference on Machine Learning, ICML 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61885–61896. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/zhou24f.html.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pp. 5745–5753. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00589.

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. In *Robotics: Science and Systems XXI*, 2025a.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025*, pp. 9834–9844. IEEE, 2025b. doi: 10.1109/ICCV51701.2025.00917.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*, 2026.

A Data Collection and Processing

A.1 UMI Data Collection

DM-DataClaw is the tweezer-style UMI gripper used in our collection: its measured jaw opening supplies the gripper-width component of Eq. 1. The camera and jaws are rigidly coupled, so wrist-view motion and jaw state share a device timeline. Figure 7 shows the camera and tactile hardware together with a representative handheld session. Tactile and force sensing provide complementary information for contact-rich manipulation (Shan et al., 2026).

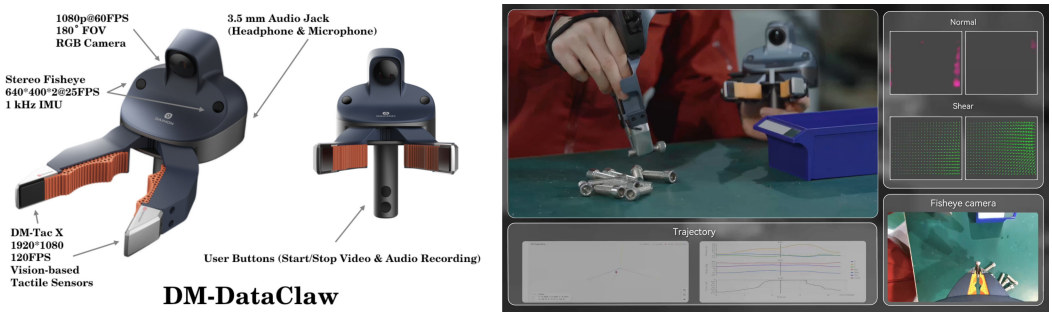

Figure 7: **DM-DataClaw hardware and data collection.** Device specifications and a handheld collection example, adapted from Daimon-Infinity (Daimon Robotics, 2026).

Collection environment. Figure 8 shows the studio enclosure. AprilGrid targets (Furgale et al., 2013), built from AprilTag markers (Olson, 2011), span multiple viewing directions to preserve geometric constraints as the operator moves within the booth.



Figure 8: **UMI data collection booth.** Studio enclosure and calibration targets.

Stereo egocentric views. The synchronized eye-camera pair provides wider and more persistent AprilGrid visibility for multi-frame registration of all camera trajectories in a shared world frame (Fig. 9); fixed extrinsics then recover both eye-camera poses.

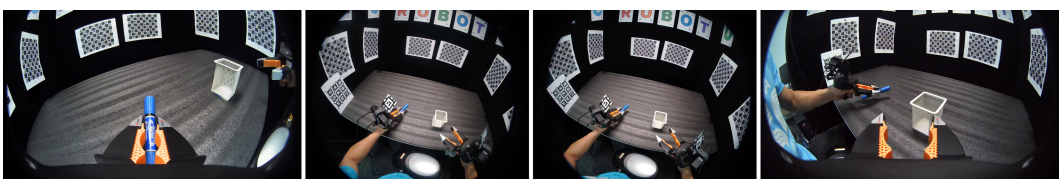

Figure 9: **Four-camera studio recording.** Left wrist, left eye, right eye, and right wrist.

A.2 Processing Pipeline Details

Source-specific ingestion. We convert heterogeneous recordings into a common episode format containing wrist-camera videos, pose trajectories, and gripper states, together with available egocentric observations and calibration metadata. Source-specific adapters standardize field conventions while preserving source identifiers, episode indices, and video-segment mappings, allowing each processed sample to be traced back to its original recording.

Temporal alignment. We align video and state streams using their timestamps and resample them onto a common 30 Hz timeline. Translations are linearly interpolated, rotations use spherical linear interpolation (SLERP; Shoemake, 1985), and gripper readings are aligned through zero-order hold. Video streams are resampled by dropping or repeating frames to match the aligned state sequences.

Action serialization. We apply Eq. 1 after temporal resampling, assigning each retained video frame one pose–gripper tuple. Source and segment mappings remain attached to the converted sequence so it can be checked against the original streams before training windows are emitted.

Image rectification and calibration. We rectify fisheye wrist images to pinhole views at training resolution and retain their intrinsics, hand–eye extrinsics, and available shared-frame registration.

Quality checks. We validate temporal coverage and frame counts, video decodability, pose and action validity, and camera calibration. Checks reject missing required values, non-finite states, invalid rotations, and inconsistent episode indexing. Geometry validation additionally checks camera intrinsics, rigid transforms, and black borders introduced by rectification.

B Model and Training Details

B.1 Action, Identity, and Hub Encoding

Video and action tokenization. A frozen video VAE encodes each view independently. A shared patch embedding maps the latents to tokens while retaining their view and temporal indices. We sample each arm's action sequence at uniformly spaced temporal indices to match the number of video latent frames, then encode the sampled actions with a shared temporal transformer.

Factorized identity encoding. We factorize token identity into view, role, and actor components, encoded in separate rotary bands alongside the temporal coordinates. We adapt Simplex Rotary Agent Encoding (SRAE) from Gamma-World (Liu et al., 2026) to encode view identity, while role and actor encodings specify the stream category and its associated actor.

For a pool of $N > 1$ view identities, each active view receives a distinct simplex vertex in $d_{\mathrm{id}} \geq N - 1$ dimensions. The vertices and rotary phase offsets are

$$\mathbf{s}_v = \sqrt{\frac{N}{N-1}}\,\mathbf{Q}\left(\mathbf{e}_v - \frac{1}{N}\mathbf{1}\right), \qquad \boldsymbol{\phi}_v = \alpha \mathbf{s}_v, \quad v = 1, \ldots, N, \tag{4}$$

where $\mathbf{e}_v \in \mathbb{R}^N$ is a one-hot vector and $\mathbf{Q}$ is a fixed isometric embedding of the zero-sum subspace into $\mathbb{R}^{d_{\mathrm{id}}}$. The scalar $\alpha > 0$ scales the phase offsets. The unit vertices satisfy $\mathbf{s}_v^\top \mathbf{s}_{v'} = -1/(N-1)$ for $v \neq v'$, giving equal pairwise distances. For the two wrist views used here, $N = 2$ gives opposite simplex vertices, with one phase vector assigned to each wrist-camera stream.

Each coordinate of $\boldsymbol{\phi}_v$ controls a rotary pair in the view-identity band. Role and actor phases occupy their respective bands, with actor identities determining the action routing in Eq. 2. Spatial and temporal indices are retained alongside the factorized identities throughout the DiT.

Sparse Hub Attention. Following Gamma-World (Liu et al., 2026), we append a small set of learned hub tokens at every latent time. Visual tokens attend only within their own stream and to the hub, whereas hub tokens attend to all wrist streams and other hub tokens. Successive layers therefore exchange cross-view information through a view–hub–view pathway without requiring direct all-to-all visual attention. We additionally condition the hub on the validity-masked mean of the encoded per-arm action states, so the shared representation receives both arms' controls while actor identity preserves view-specific routing. This hub pathway also supports open-world samples whose wrist trajectories remain in separate local coordinate frames.

B.2 CAMERA-AWARE MODELING

Camera-guided noise warping. We adopt CameraNoise (Zhao et al., 2026a) to encode relative camera poses into temporally correlated noise. For each view, calibrated intrinsics and consecutive relative camera transforms define geometry-guided reprojection flow on the latent grid using a fixed pseudo-depth. We apply the official noise warping implementation and mix the resulting noise with independent Gaussian noise using variance normalization. Each view uses its own local pose chain, supporting both open-world and in-studio recordings.

Geometry-aware conditioning. Inspired by camera conditioning in PRoPE (Li et al., 2025b) and Matrix-Game 3.5 (Qian et al., 2026), we combine ray-based features with a learned relative projective attention bias. For calibrated views in a shared frame, each ray comprises its camera origin and unit viewing direction. Given intrinsics $\mathbf{K}_v$, camera-to-world pose $(\mathbf{R}_t^v, \mathbf{p}_t^v)$, and homogeneous pixel coordinate $\bar{\mathbf{x}}$, the six-dimensional ray feature is

$$\mathbf{r}_t^v(\bar{\mathbf{x}}) = \left[\mathbf{p}_t^v;\ \frac{\mathbf{R}_t^v \mathbf{K}_v^{-1}\bar{\mathbf{x}}}{\|\mathbf{R}_t^v \mathbf{K}_v^{-1}\bar{\mathbf{x}}\|_2}\right] \in \mathbb{R}^6. \tag{5}$$

We encode the ray features and aggregate them into hub tokens using geometric affinities. An MLP maps relative camera transforms and intrinsics to per-head attention biases for within-view temporal interactions, while the hub exchanges information across views. These geometry branches operate on samples with valid calibration and shared-frame registration.

B.3 TRAINING AND DISTILLATION

Progressive teacher training. We train the bidirectional teacher in three stages using the data pyramid in Section 3. Stage 1 uses all three tiers for single-wrist generation with per-arm action conditioning. Stage 2 trains joint dual-wrist generation on open-world and in-studio recordings, introducing CameraNoise and Sparse Hub Attention. Stage 3 uses only calibrated in-studio dual-wrist recordings, adding ray-based conditioning and relative projective attention biases.

Training objective. For clean latents $\mathbf{z}_0$ and noise $\boldsymbol{\epsilon}$, we form $\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_0 + \sigma\boldsymbol{\epsilon}$ for $\sigma \in [0,1]$ and minimize the flow-matching objective (Lipman et al., 2023) over future tokens,

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\|v_\theta(\mathbf{z}_\sigma, \sigma; \mathcal{C}) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\|_{\mathcal{M}}^2\Big], \tag{6}$$

where $\mathcal{C} = (\mathcal{O}_{1:H}, \mathcal{A}, \mathcal{G})$ collects the conditions in Eq. 2, and $\|\cdot\|_{\mathcal{M}}^2$ averages over the valid future tokens indexed by $\mathcal{M}$. Stage 1 uses independent Gaussian noise; Stages 2 and 3 use CameraNoise for samples with valid camera geometry. We keep observation latents clean and optimize the DiT and active conditioning modules with the VAE frozen.

Causal adaptation. We adapt the teacher with block-causal attention, restricting visual and hub tokens to the current and preceding blocks while preserving view–hub connectivity. The action encoder uses causal masking. We train with Eq. 6 on a sampled active block, using clean preceding latents as context, to obtain a causal teacher for distillation.

Few-step distillation. We distill the causal teacher into a few-step student by matching denoised endpoints. Both models receive the same initial noise, preceding latents, and action horizon. Let $F_\theta^{(K)}$ denote a K-step Euler rollout for the active block. We match their denoised endpoints with

$$\mathcal{L}_{\mathrm{distill}} = \mathbb{E}\left[\left\| F_{\theta_{\mathrm{S}}}^{(K_{\mathrm{S}})}(\boldsymbol{\epsilon}; \mathcal{C}) - \mathrm{sg}\Big[F_{\theta_{\mathrm{T}}}^{(K_{\mathrm{T}})}(\boldsymbol{\epsilon}; \mathcal{C})\Big]\right\|_{\mathcal{M}}^2\right], \qquad K_{\mathrm{S}} < K_{\mathrm{T}}, \tag{7}$$

where θ_{S} and θ_{T} denote the student and frozen teacher parameters, respectively, and $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator. We use $K_{\mathrm{T}} = 30$ and $K_{\mathrm{S}} = 4$, optimizing only the student DiT through its final denoising step. Student-generated preceding blocks provide rollout context during training.

At inference, the Student accepts per-arm actions for the next block and jointly generates the wrist views in four denoising steps. The completed block is appended to a bounded latent history containing both wrist streams, which provides visual context for subsequent predictions. New actions are supplied block by block and routed to their respective wrist streams, while hub tokens maintain cross-view exchange throughout the rollout.

B.4 BAF Component Errors and Aggregation

For each arm i, the mean component errors over valid evaluation frames are

$$E^i_{\text{pos}} = \frac{1}{|\mathcal{V}^i_{\text{pos}}|} \sum_{t \in \mathcal{V}^i_{\text{pos}}} \left\| \widehat{\Delta \mathbf{p}}^i_t - \Delta \mathbf{p}^i_t \right\|_2,$$

$$E^i_{\text{rot}} = \frac{1}{|\mathcal{V}^i_{\text{rot}}|} \sum_{t \in \mathcal{V}^i_{\text{rot}}} \arccos\left(\frac{\text{tr}((\widehat{\Delta \mathbf{R}}^i_t)^\top \Delta \mathbf{R}^i_t) - 1}{2} \right), \tag{8}$$

$$E^i_{\text{grip}} = \frac{1}{|\mathcal{V}^i_{\text{grip}}|} \sum_{t \in \mathcal{V}^i_{\text{grip}}} \left| \hat{g}^i_t - g^i_t \right|,$$

where $(\Delta \mathbf{p}^i_t, \Delta \mathbf{R}^i_t, g^i_t)$ is the conditioning action and $(\widehat{\Delta \mathbf{p}}^i_t, \widehat{\Delta \mathbf{R}}^i_t, \hat{g}^i_t)$ is its IDM estimate, with poses relative to the initial frame. Equation 3 aggregates the six component errors. The parameters in Section 4.3 define the tolerance bands and decay rate. We average scores over videos with all six errors available. Recovered and conditioning poses use the same coordinate convention and scale; GT videos and baselines lacking full UMI controls use recorded actions as the reference.

C Experimental Details

C.1 Baseline Comparisons

Baseline sources. Text-conditioned baselines are Wan2.2 TI2V-5B (Team Wan et al., 2025), Cosmos Predict2.5-2B (NVIDIA et al., 2025b), MiniMax-H3 (MiniMax, 2026; Zhao et al., 2026c), Cosmos3-Nano-Policy-DROID (NVIDIA, 2026), and RynnWorld-Teleop (Zhao et al., 2026b) (text+image SFT). Action-conditioned baselines are Genie-Envisioner (Liao et al., 2026), RynnWorld-Teleop (control band), and Gamma-World (Liu et al., 2026). Camera-controlled baselines are HY-WorldPlay 5B (Sun et al., 2026), LingBot-World-Fast (Robbyant Team et al., 2026), DreamX-World-5B-Cam (DreamX Team et al., 2026), Matrix-Game 2.0 (He et al., 2025c) and 3.0 (Wang et al., 2026c), YUME 1.5 (Mao et al., 2026), and Infinite-World (Wu et al., 2026).

Evaluation protocol. Each valid interval starts at a stride of 189 frames, including the final valid start. Task descriptions are fixed before generation, with missing descriptions obtained from wrist videos using Volcano Engine's LAS service. Frame metrics use the 188 predicted frames at 320×384 after the conditioning frame, aligned to reference timestamps. We average over frames, wrists, and seeds within each window, then over windows with equal episode weights.

Reference and distribution metrics. FID uses Inception-v3 (Szegedy et al., 2016) frame features and FVD uses I3D (Carreira & Zisserman, 2017) video features separately for each wrist and seed. Both distribution metrics use one fixed-duration, fixed-rate, fixed-resolution window per episode. For FVD, pixels are normalized once and resized to 224×224 without extractor rescaling.

Excluded benchmark metrics. We also evaluated WorldModelBench (Li et al., 2025a), iWorld-Bench (Fang et al., 2026b), and EWMBench (Hu et al., 2025), but omit them from the main comparison. WorldModelBench was insensitive to swapped task descriptions on one of two paired windows and returned the same visual-quality answer for original and corrupted videos. iWorld-Bench trajectory scores changed across repeated ViPE (Huang et al., 2025a) pose estimates and accepted a frozen-frame control, while EWMBench's released gripper tracker produced no valid tracks at its default thresholds and retained large localization errors after lowering them. We therefore retain only the validated VBench and WorldArena dimensions.

C.2 Bimanual Action Fidelity Analysis

Evaluation details. Generated videos use seeds 17 and 29 for IDM inference, with relative-pose supervision and local geometric refinement; the gripper estimator is fixed. Component errors use meters and radians before scoring (Appendix B.4). Geometry probes and Student sampling use separate evaluation runs from the main comparison.

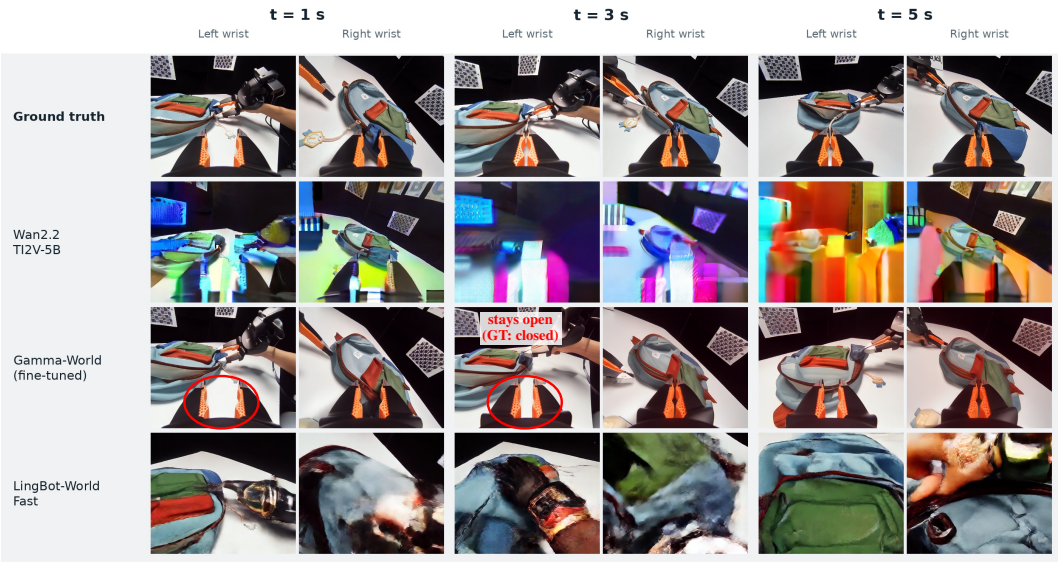

Figure 10: **Additional temporal comparison.** Additional baselines on the paired-wrist sequence in Figure 5. Red circles mark Gamma-World's gripper staying open as GT closes.

Rendered-control adapters. Genie-Envisioner uses the Cosmos2-based GE-Sim v0.1 world model. GE-Sim receives projected end-effector markers colored by gripper opening and six-channel camera raymaps, then jointly generates 189-frame wrist videos at 30 Hz.

RynnWorld-Teleop. For the control entry, we render the mapped UMI pose-and-gripper sequence frame by frame as a synchronized 21-point dual-hand skeleton video, pass it through the Wan2.2 VAE, and add the resulting control features to patch embeddings at scale 0.109; the initial wrist image supplies the visual condition for each view. This route has no numerical action interface.

Gamma-World fine-tuning. We initialize from the existing student checkpoint at iteration 6,400 and reinitialize the action encoders. Each arm supplies window-relative translation, 6D rotation, and gripper opening (10D; Eq. 1), giving 20D total. We train on paired rectified 189-frame videos at 30 Hz and 320×384 resolution for 3,000 iterations.

C.3 Applications

Physical robot execution. Figure 11 pairs the external-camera view of the physical robot with the corresponding wrist-view preview for the rollout in Section 5.5, at five timestamps sampled from initial approach through completed object transport.

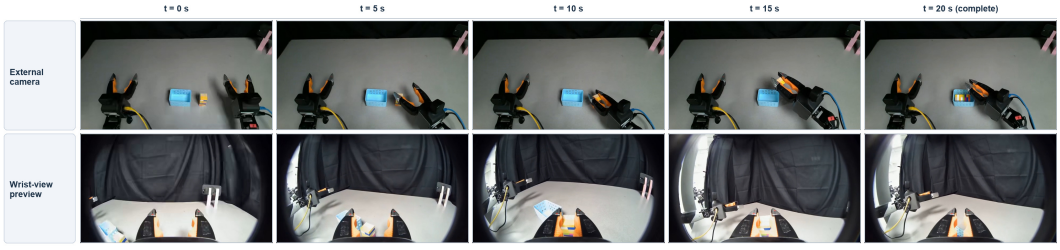

Figure 11: **Physical robot execution and wrist-view preview.** The top row shows external-camera images; the bottom row shows the corresponding wrist-view preview.

Visual domain randomization. Figure 12 applies three edits to both initial frames (hue and saturation shift, object color inversion, background replacement) and regenerates the 189-frame rollout under unchanged controls. BAF (Section 4.3) screens them: hue and object-color edits stay

within one point of the original (87.0 and 88.4 versus 88.1), whereas background replacement lowers BAF by 3.6–6.8 points across three windows and is flagged before policy training.

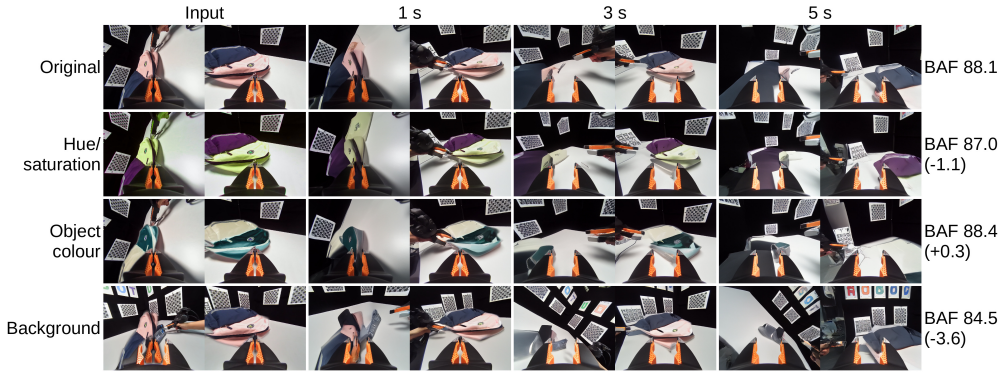

Figure 12: **Visual domain randomization.** Rows: original and three initial-frame edits with their BAF; columns: input and generated frames at 1, 3, and 5 s, each pairing left and right wrists.

Data quality screening. Figure 13 applies the frozen IDM to raw recordings and compares the recovered actions with the recorded SLAM trajectories and gripper controls, normalized by the BAF tolerances of Section 4.3; a segment is flagged when the discrepancy stays above three tolerances for at least 0.8 s. Synthetic time-shift, trajectory-scale, and frozen-rotation corruptions are separated from clean recordings with AUROC 0.98, 0.98, and 0.90, with no false alarms on clean recordings.

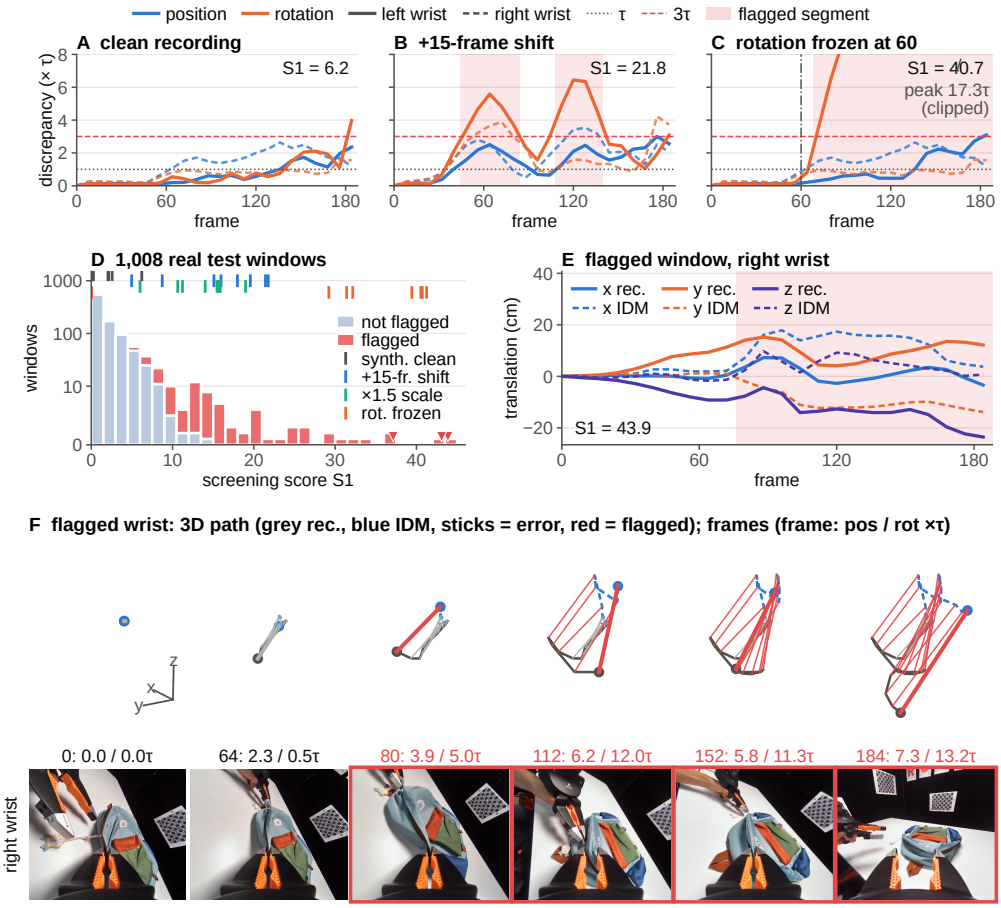

Figure 13: **Data quality screening.** (A–C) Per-timestep IDM-versus-recorded discrepancy for a clean recording and two corrupted controls with flagged segments shaded. (D) Screening scores over 1,008 test windows. (E) A flagged real window: recorded versus recovered right-wrist translation. (F) 3D path of the flagged wrist, recorded versus recovered with error sticks in red, and its frames.

D Additional Related Work

Camera-controlled and multi-view video generation. Camera control in video diffusion has been extended to object and camera motion (Wang et al., 2024), dynamic scenes (He et al., 2025b), analysis of camera conditioning in video transformers (Bahmani et al., 2025), and 3D-informed generation (Ren et al., 2025). ReCamMaster (Bai et al., 2025a) and TrajectoryCrafter (Yu et al., 2025) re-render a given video along new camera trajectories, and CamCloneMaster (Luo et al., 2025) transfers camera motion from a reference video. SynCamMaster (Bai et al., 2025b) and Collaborative Video Diffusion (Kuang et al., 2024) jointly generate several camera views of a shared scene with cross-view consistency. In UMI-WM, each view instead follows its own gripper-mounted trajectory and gripper width, so the two views move independently while observing a shared scene. Warped-noise motion control (Burgert et al., 2025) underlies the camera-guided noise warping adopted in Appendix B.2. Interactive world models for games (Alonso et al., 2024; Decart et al., 2024; Guo et al., 2025; He et al., 2025c) and general interactive world models such as Astra (Zhu et al., 2026) similarly generate observations from user controls.

Autoregressive and few-step video generation. UMI-WM builds on the diffusion transformer (DiT) architecture (Peebles & Xie, 2023). Diffusion Forcing (Chen et al., 2024) enables causal rollout with per-frame noise levels. CausVid (Yin et al., 2025) converts bidirectional video diffusion models into fast autoregressive generators with distribution matching distillation (Yin et al., 2024b;a), and Self Forcing (Huang et al., 2025c) conditions training on the model's own rollouts. The block-causal adaptation and few-step student in Appendix B.3 follow this line, matching denoised endpoints (Eq. 7) with student-generated preceding blocks as context.

Action-conditioned robot world models. UniSim (Yang et al., 2024) learns an interactive real-world simulator from diverse data, IRASim (Zhu et al., 2025b) generates robot videos from frame-level trajectories, Cosmos (NVIDIA et al., 2025a) provides world foundation models for physical AI, and Vid2World (Huang et al., 2026) converts pretrained video diffusion models into interactive world models. Closer to our setting, Ctrl-World (Guo et al., 2026) predicts multi-view observations, including wrist views, for policy evaluation and improvement; Mem-World (Zheng et al., 2026) adds memory for persistent manipulation, and WristWorld (Qian et al., 2025) generates wrist views from anchor views. These models typically train on large robot datasets such as Open X-Embodiment (Open X-Embodiment Collaboration, 2024) and DROID (Khazatsky et al., 2024), whereas UMI-WM conditions each wrist view on its own relative camera pose and gripper width.

Uses of robot world models. World models support policy evaluation (Quevedo et al., 2026; Li et al., 2025d), bimanual failure detection (Ward et al., 2026), embodied future prediction (Huang et al., 2025b; Jiang et al., 2025; Shang et al., 2025), bimanual data generation (Chen et al., 2026a), and learning from human videos (Gao et al., 2026; Chen et al., 2026c).

Joint video–action and latent world models. Joint video–action models (Li et al., 2025c; Zhu et al., 2025a; Shen et al., 2025; Ye et al., 2026; Zhang et al., 2026; Yan et al., 2026; Li et al., 2026) predict future observations or world representations together with actions for policy learning. Latent world models instead learn dynamics in latent space for control and planning, as in DayDreamer (Wu et al., 2022), TD-MPC2 (Hansen et al., 2024), and V-JEPA 2 (Assran et al., 2025).

Actions from video. Inverse dynamics models label unlabeled video with actions, as in VPT (Baker et al., 2022), while Genie (Bruce et al., 2024) learns latent actions without labels. DreamGen (Jang et al., 2025) applies an IDM to world-model rollouts to obtain pseudo-actions for policy training. BAF uses the IDM to compare recovered actions with the conditioning actions.

Teleoperation interfaces. Immersive VR teleoperation systems such as Open-TeleVision (Cheng et al., 2024) and Bunny-VisionPro (Ding et al., 2025) give the operator real-time feedback from the robot, and TelePreview (Guo et al., 2024) previews a virtual arm before commands are executed, whereas the application in Section 5 renders both wrist views from the world model.