KINEWORLD: ACTION-INDUCED TRANSPORT FIELDS FOR EMBODIED WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Embodied world models predict the visual consequences of candidate actions before execution. However, existing action-conditioned world models often adopt uniformly weighted visual generation objectives that can be misaligned with embodied prediction needs. Even with explicit motion conditioning, these objectives can underemphasize spatially sparse changes that are critical to interaction. We propose **KineWorld**, a transport-aware world-modeling framework that extends robot kinematics from motion conditioning to the spatial allocation of generative supervision. **Kinematic Transport Lifting (KTL)** constructs renderer-derived, camera-aligned transport fields from commanded robot motion. **Transport-Aware World Diffusion (TAWD)** calibrates their motion support on the video-latent grid and reweights future-RGB flow matching through a normalized mixture of uniform and transport-focused distributions. We train KineWorld using ALOHA–AgileX bimanual manipulation data from **RoboTwin 2.0**. KineWorld achieves an **EWMScore-P of 68.95** in single-view evaluation and a **TWB-Score of 54.82** in multi-view evaluation. These results support a shift from appearance fitting toward action-consequence modeling for embodied decision-making.

1 INTRODUCTION

World Action Models (WAMs) enable robots to anticipate how the visual world will evolve under candidate actions before executing them in the physical environment. By generating action-conditioned future observations, these models can serve as predictive simulators for policy learning, action evaluation, planning, and synthetic experience generation. Recent advances in large-scale video generation further provide strong appearance and dynamics priors, making it possible to construct visually realistic robot rollouts from limited embodied data Zhu et al. (2024); Guo et al. (2025); Liao et al. (2025).

Existing world action models (WAMs) improve action responsiveness through numerical action tokens, learned latent codes, and spatially aligned conditions such as embodiment masks, action images, optical flow, and projected kinematic fields Zhu et al. (2024); Guo et al. (2025); Wang et al. (2026); Liao et al. (2025); Chen et al. (2026b). Yet motion conditioning does not itself alter the spatial weighting of the primary RGB-generation objective. When action-responsive regions occupy only a small part of the frame, uniform token averaging can bias optimization toward static appearance. Motion-sensitive auxiliary objectives and flow-branch reweighting address related concerns Zhu et al. (2024); Chen et al. (2026b). We further use robot kinematics to guide future-RGB supervision, aligning the model's learning priorities with action consequences. This raises a key question: **how can robot kinematics guide the learning of action consequences?**

Our key insight is that action conditioning and supervision allocation serve different purposes: the former specifies what motion to generate, whereas the latter determines which prediction errors receive greater weight during training. Robot kinematics provides a geometry-derived prior for this allocation by identifying where commanded robot motion is expected to appear in the camera view, without relying on generated future frames. This suggests using kinematic transport not only as a generation condition, but also as a spatial prior for the primary video-generation objective.

Building on this insight, we introduce **KineWorld**, a transport-aware training framework for action-conditioned world modeling. **Kinematic Transport Lifting (KTL)** constructs camera-aligned transport, while **Transport-Aware World Diffusion (TAWD)** calibrates its motion support on the

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.

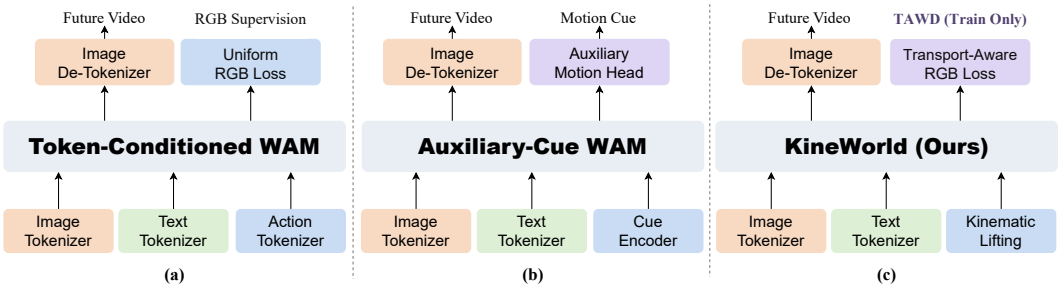

Figure 1: **(a)** Token-conditioned WAMs use numerical trajectories or learned action embeddings for video generation (Zhu et al., 2024; Guo et al., 2025; Liao et al., 2025), but typically apply *uniform RGB supervision*, potentially **underemphasizing interaction-critical local changes**. **(b)** Auxiliary-cue WAMs incorporate *motion-focused auxiliary cues*, such as masks (Chen et al., 2026c), action images (Zhen et al., 2026), optical flow (Chen et al., 2026b), or projected kinematic fields (Yang et al., 2026), to improve motion localization. These cues alone do not directly change the spatial weighting of the primary RGB objective, potentially **leaving motion localization and supervision priorities misaligned**. **(c)** KineWorld introduces *action-guided RGB supervision* by lifting camera-aligned kinematic transport onto the video latent grid and reweighting flow matching, enabling **action-consequence-oriented supervision** while preserving the total token weight.

video latent grid and converts the resulting support into a normalized mixture distribution for the RGB flow-matching objective. The resulting objective redistributes supervision while preserving the aggregate token weight, without adding inference parameters or sampling steps. This extends kinematic priors from generation conditions to learning guidance for interaction- and decision-oriented world prediction. Our main contributions are summarized as follows:

- **Action-Consequence-Oriented Learning.** We distinguish visual extent from interaction importance, motivating kinematics-guided learning of interaction-critical local changes.
- **KineWorld Framework.** We propose KineWorld, a transport-aware framework with KTL and TAWD. KTL constructs camera-aligned transport from the commanded rollout, while TAWD reweights RGB flow matching via a normalized mixture on the latent grid, preserving token weight and sampling.
- **Benchmark Evaluation.** Trained on RoboTwin 2.0 ALOHA–AgileX bimanual manipulation data (Chen et al., 2025), KineWorld achieves an EWMScore-P of 68.95 and a TWB-Score of 54.82 in single-view and multi-view evaluation, respectively. Our framework and evaluations provide the community with a reference for action-consequence-oriented world modeling.

2 Related Work

World Action Models. World Action Models (WAMs) increasingly leverage pretrained video generators to connect visual prediction with robot control (Guo et al., 2025; Liao et al., 2025). Existing approaches include action-conditioned video prediction (Zhu et al., 2024; Guo et al., 2025), joint video-action generation (Zhou et al., 2026), and action decoding from visual latent representations (Liao et al., 2025). Recent systems further broaden this paradigm. RealWonder connects physical actions to video generation through simulated RGB and flow (Liu et al., 2026), while CLAP integrates heterogeneous action representations for cross-embodiment learning (Liu & Shorinwa, 2026). However, advances in generative architectures and action interfaces do not inherently prioritize interaction-critical changes in the primary RGB objective. When these changes occupy few visual tokens, uniform weighting can underemphasize their contribution. KineWorld addresses this complementary problem through kinematics-guided allocation of future-RGB supervision.

Action Representations for WAMs. Existing WAMs encode robot actions as numerical tokens, learned latent codes, or spatially grounded visual cues (Zhu et al., 2024; Chen et al., 2026a; Zhou et al., 2026). Recent methods align actions with the visual domain through embodiment masks

(Chen et al., 2026c), multi-view action videos (Zhen et al., 2026), optical flow (Chen et al., 2026b), and camera-aligned kinematic fields (Yang et al., 2026). Beyond conditioning, FlowWAM reweights its flow-branch loss using reference-relative latent deviations (Chen et al., 2026b), while EA-WM uses frame-difference supervision to guide cross-stream fusion (Yang et al., 2026). In contrast, KineWorld uses kinematic transport to directly reweight future-RGB flow matching without introducing auxiliary objectives.

3 PRELIMINARIES

3.1 EMBODIED WORLD MODELING

An episode consists of RGB observations $\{I_t\}_{t=0}^{T}$, robot configurations $\{q_t\}_{t=0}^{T}$, a language instruction c, and an action trajectory $A = \{a_t\}_{t=0}^{T-1}$, where $I_t \in \mathbb{R}^{H\times W\times 3}$ and a_t controls the transition from t to $t+1$. In our bimanual setting, $a_t \in \mathbb{R}^{14}$ specifies an absolute 6-DoF end-effector pose target and one gripper command for each arm. The action is executed through an embodiment-specific controller that maps the commanded end-effector targets to robot-configuration transitions. An embodied world model learns $p_\theta(I_{1:T} \mid I_0, A, c)$, whose predictions should preserve both visual realism and action fidelity.

3.2 CONDITIONAL FLOW MATCHING

Given a K-frame clip $I_{0:K-1}$, a causal, first-frame-preserving video VAE produces $z^x = \mathcal{E}_{\mathrm{VAE}}(I_{0:K-1}) \in \mathbb{R}^{C_z\times Q\times H_z\times W_z}$. We partition it as $z^x = [z^{x,\mathrm{cond}}, z^{x,\mathrm{fut}}]$, where $z^{x,\mathrm{cond}} = z_0^x$ and $z^{x,\mathrm{fut}} = z_{1:Q-1}^x$.

Let $\tau \sim p(\tau)$, $\epsilon \sim \mathcal{N}(0,\mathbf{I})$, and $\sigma(\tau) \in [0,1]$ be a differentiable schedule with $\sigma(0)=0$ and $\sigma(1)=1$. Only future RGB latents are noised:

$$\begin{aligned} \tilde{z}_{\tau,\ell}^x &= \begin{cases} z_0^x, & \ell = 0, \\ (1-\sigma(\tau))z_\ell^x + \sigma(\tau)\epsilon_\ell, & 1 \le \ell \le Q-1, \end{cases} \\ v_{\tau,\ell}^\star &= \dot{\sigma}(\tau)(\epsilon_\ell - z_\ell^x), \qquad 1 \le \ell \le Q-1. \end{aligned} \tag{1}$$

For $\sigma(\tau)=\tau$, $v_{\tau,\ell}^\star = \epsilon_\ell - z_\ell^x$. Conditioned on language c and action representation ξ, the model predicts $\hat{v}_\theta = v_\theta(\tilde{z}_\tau^x, \tau \mid c, \xi)$. The future latent positions are

$$\mathcal{V} = \{(\ell,h,w) : 1 \le \ell \le Q-1,\ 1 \le h \le H_z,\ 1 \le w \le W_z\}, \qquad N = |\mathcal{V}|. \tag{2}$$

For $i = (\ell,h,w) \in \mathcal{V}$, the channel-averaged prediction error is

$$e_i(\theta) = \frac{1}{C_z}\sum_{\chi=1}^{C_z}\left(\hat{v}_{\theta,\chi i} - v_{\tau,\chi i}^\star\right)^2, \tag{3}$$

and standard conditional flow matching uniformly averages these errors:

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{I,c,\xi,\tau,\epsilon}\left[\frac{1}{N}\sum_{i\in\mathcal{V}} e_i(\theta)\right]. \tag{4}$$

The conditioning frame $\ell = 0$ remains clean and is excluded from the loss; generation integrates the learned vector field from $\tau = 1$ to $\tau = 0$.

4 KINEWORLD

4.1 OVERVIEW

KineWorld is a latent video world model that predicts future RGB observations from the current observation, a language instruction, and a candidate action sequence. A causal video VAE represents visual states in latent space, while a conditional flow model learns their action-conditioned dynamics. When projected robot motion occupies only a small part of the video, uniform token averaging

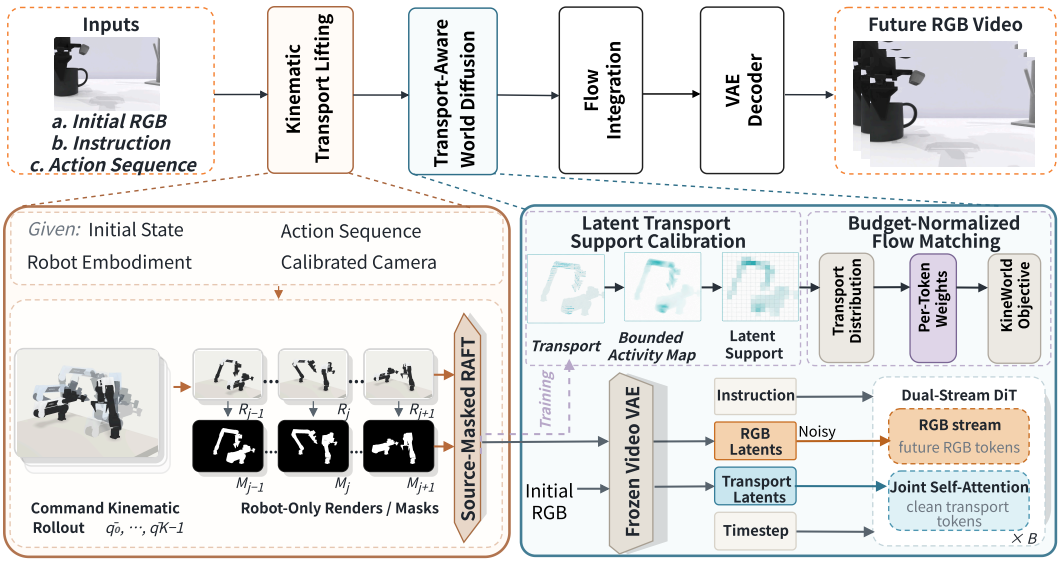

Figure 2: **Overview of KineWorld.** Kinematic Transport Lifting (KTL) converts the commanded robot rollout into camera-aligned transport and injects its latent representation as a clean structural condition. Transport-Aware World Diffusion (TAWD) calibrates the transport support and derives mean-one token weights for the flow-matching objective, emphasizing action-responsive regions under a fixed weight budget. The weighting branch is used only during training, while the inherited action expert remains unchanged.

assigns this region a correspondingly small share of the total token weight. KineWorld uses commanded kinematics to guide both generation and learning: camera-aligned transport conditions the predicted motion, while its spatial support directs a fixed share of future-RGB supervision toward regions of projected robot motion. KineWorld comprises two components. **Kinematic Transport Lifting (KTL)** constructs camera-aligned transport and injects its latent representation z^u into the video generator as a clean structural condition. **Transport-Aware World Diffusion (TAWD)** derives calibrated support d from the transport fields and constructs normalized token weights for the flow-matching objective. KTL supplies motion conditions during both training and inference, whereas TAWD modifies only training-time supervision.

For a training clip starting at dense state index s, we sample K visual keyframes with temporal stride δ and define $L = (K-1)\delta$:

$$\bar{A} = \{a_t\}_{t=s}^{s+L-1}, \qquad \bar{I}_j = I_{s+j\delta}, \qquad j = 0, \dots, K-1. \tag{5}$$

Starting from the initial robot configuration q_s, the commanded action sequence is executed through the embodiment-specific controller, $\hat{q}_0 = q_s, \hat{q}_{\tau+1} = F_{\mathcal{E}}(\hat{q}_\tau, a_{s+\tau}), \tau = 0, \dots, L-1$, and sampled at the same temporal stride as $\bar{q}_j = \hat{q}_{j\delta}, j = 0, \dots, K-1$. The resulting rollout depends only on the initial robot configuration, the commanded end-effector targets, and the embodiment-specific controller, without access to future RGB observations. Given $\bar{q}_0$, $\bar{A}$, the robot embodiment $\mathcal{E}$, and calibrated camera $\mathcal{C}$, Kinematic Transport Lifting constructs

$$U = \Gamma_{\mathcal{E},\mathcal{C}}(\bar{q}_0, \bar{A}) = \{u_j\}_{j=0}^{K-1}. \tag{6}$$

We use $K = 9$ and $\delta = 4$, corresponding to nine keyframes over 32 control transitions.

4.2 Kinematic Transport Lifting

Renderer-Derived Camera Transport. For every sampled robot configuration, a fixed renderer generates a robot-only RGB frame and its visible robot mask:

$$(R_j, \mathcal{M}_j) = \mathrm{Render}_{\mathcal{E},\mathcal{C}}(\bar{q}_j), j = 0, \dots, K-1, \tag{7}$$

where $R_j \in \mathbb{R}^{H\times W\times 3}$ and $\mathcal{M}_j \in \{0,1\}^{H\times W}$. We use a fixed background texture during robot-only rendering to stabilize correspondence estimation.

For each pair of consecutive rendered states, RAFT estimates forward displacement on the source image grid. The estimate is restricted to the visible source-robot support:

$$u_j = \begin{cases} 0, & j = 0, \\ \mathcal{M}_{j-1} \odot \mathrm{RAFT}(R_{j-1}, R_j), & 1 \le j \le K-1. \end{cases} \tag{8}$$

The binary mask is broadcast over the two displacement channels. Each $u_j : \Omega \to \mathbb{R}^2$ therefore defines a camera-aligned transport field over image domain Ω, where $u_j(p) = (u_{j,x}(p), u_{j,y}(p))$. RAFT (Teed & Deng, 2020) estimates image-plane correspondences between consecutive robot-only renders, and the source mask restricts the resulting field to visible robot pixels. We retain this render-and-estimate construction, keeping the transport-conditioning pipeline fixed when studying supervision allocation. The resulting u_j is therefore an image-based estimate of projected robot motion, rather than an analytic projection of surface trajectories.

Video-Native Transport Condition. For each pixel p, we define the transport magnitude and direction as $m_j(p) = \|u_j(p)\|_2$ and $\phi_j(p) = \mathrm{atan2}(u_{j,y}(p), u_{j,x}(p))$, respectively. We encode direction as hue and clipped magnitude as saturation:

$$h_j(p) = \frac{\phi_j(p) + \pi}{2\pi}, \eta_j(p) = \mathrm{clip}\left(\frac{m_j(p)}{M_{\mathrm{px}}}, 0, 1\right), G_j(p) = \mathrm{HSV2RGB}\big(h_j(p), \eta_j(p), 1\big). \tag{9}$$

Here, $M_{\mathrm{px}} > 0$ is a fixed displacement-clipping threshold. Zero displacement has zero saturation and is therefore mapped to white.

The RGB and transport clips are encoded separately:

$$z^x = \mathcal{E}_{\mathrm{VAE}}(\bar{I}_{0:K-1}), \quad z^u = \mathcal{E}_{\mathrm{VAE}}(G_{0:K-1}), \quad z^x, z^u \in \mathbb{R}^{C_z \times Q \times H_z \times W_z}. \tag{10}$$

The encoded transport z^u serves as a structural condition for RGB generation. TAWD computes its supervision weights directly from the transport fields u_j, rather than from their VAE latents. This separates supervision allocation from possible compression artifacts in the transport representation, although such artifacts may still affect the conditioning stream.

Clean Transport Conditioning. KTL supplies the transport-conditioning stream of the Wan2.2-based generator. The RGB latent follows the flow-matching path defined in Equation 1, while the transport latent remains clean: $\tilde{z}^u_{\tau,\ell} = z^u_\ell$ for $\ell = 0, \ldots, Q-1$. The model therefore receives a clean first RGB latent frame, noised future RGB latents, and a clean transport sequence:

$$\hat{v}_\theta = v_\theta\big(\tilde{z}^x_\tau, \tilde{z}^u_\tau, \tau \mid c\big). \tag{11}$$

Here, *clean transport* means that diffusion noise is not added to the transport input. Its hidden representations can nevertheless interact with the RGB stream throughout denoising.

RGB and transport latents are patchified separately, augmented with stream identifiers, and assigned aligned spatiotemporal positions. Let X^x_b and X^u_b denote the RGB and transport token sequences entering Transformer block b, respectively. Their joint self-attention is written as

$$H_b = [X^x_b; X^u_b], \Delta H_b = \mathrm{MSA}_b(H_b), [\Delta X^x_b; \Delta X^u_b] = \mathrm{Split}(\Delta H_b), b = 1, \ldots, B. \tag{12}$$

The two streams interact through joint self-attention throughout RGB denoising. TAWD leaves this conditioning architecture unchanged and modifies only the training objective. The computational cost of transport conditioning, including joint attention, is accounted for separately in Appendix A.6.

4.3 Transport-Aware World Diffusion

Latent Transport Support Calibration. Raw transport magnitude is not directly comparable across trajectories because it varies with camera scale, action amplitude, visible robot area, and temporal stride. We therefore construct a bounded activity map directly from the source-masked transport:

$$r_j(p) = \mathrm{clip}\left(\frac{\|u_j(p)\|_2}{M_{\mathrm{px}}}, 0, 1\right), \qquad j = 0, \ldots, K-1. \tag{13}$$

Because u_j has already been restricted by $\mathcal{M}_{j-1}$ in Equation 8, the visibility mask is not applied again. In particular, $r_0 = 0$ follows from $u_0 = 0$.

A fixed, non-learned operator $\mathcal{P}_{\text{lat}}$ performs spatiotemporal average pooling according to the video VAE compression ratios and aligns the activity sequence with the RGB latent grid:

$$d_{\ell hw} = \text{sg}\left[\mathcal{P}_{\text{lat}}(r_{0:K-1})_{\ell hw}\right], \qquad d_{0hw} = 0. \tag{14}$$

Here, $d \in [0,1]^{Q \times H_z \times W_z}$ and $\text{sg}(\cdot)$ denotes stop-gradient. The index ℓ denotes latent time, whereas j denotes an input keyframe. Computing d from raw transport rather than from the learned transport latent z^u provides an explicit zero-motion reference and prevents the transport encoder from adapting the supervision weights.

Budget-Normalized Flow Matching. We reuse the future-token set $\mathcal{V}$ from Equation 2. For a flattened future latent position $i = (\ell, h, w) \in \mathcal{V}$, let d_i denote the corresponding calibrated transport support and define $S_d = \sum_{i \in \mathcal{V}} d_i$. The calibrated support is converted into a sample-wise transport distribution with an explicit zero-motion fallback:

$$\rho_i^{\text{tr}} = \begin{cases} \dfrac{d_i}{S_d}, & S_d > \varepsilon_{\text{sup}}, \\ \dfrac{1}{N}, & S_d \le \varepsilon_{\text{sup}}, \end{cases} \qquad \sum_{i \in \mathcal{V}} \rho_i^{\text{tr}} = 1, \tag{15}$$

where $\varepsilon_{\text{sup}} > 0$ is a small numerical threshold.

For a non-negative transport-strength parameter β, we define $\alpha_\beta = \beta/(1+\beta) \in [0,1)$ and construct the per-token weights:

$$w_i = N\left[(1-\alpha_\beta)\frac{1}{N} + \alpha_\beta \rho_i^{\text{tr}}\right], \qquad \frac{1}{N}\sum_{i \in \mathcal{V}} w_i = 1. \tag{16}$$

The uniform component preserves scene-wide supervision, while the transport component reallocates an α_β fraction of the aggregate token weight according to projected robot-motion support. When $\beta = 0$, $w_i = 1$, exactly recovering uniform flow matching. When no valid transport is present, the fallback in Equation 15 also produces uniform weights.

Using the token error $e_i(\theta)$ defined in Equation 3, the KineWorld objective is

$$\mathcal{L}_{\text{Kine}} = \mathbb{E}_{\bar{I}_{0:K-1}, \bar{A}, c, \tau, \epsilon}\left[\frac{1}{N}\sum_{i \in \mathcal{V}} \text{sg}(w_i) e_i(\theta)\right]. \tag{17}$$

Because token weights are normalized to unit mean, transport magnitude and support only redistribute supervision. The calibrated map highlights projected robot-motion support rather than object motion or all downstream physical effects, while the uniform component preserves supervision over other scene regions.

Auxiliary Action Objective. We retain the inherited action head for checkpoint-matched comparisons. Let θ_{wm} and θ_{act} denote the disjoint parameter sets of the world generator and action head, respectively. The action head operates on stopped world-model features:

$$\hat{A} = g_{\theta_{\text{act}}}\left(\text{sg}\big(H_{\theta_{\text{wm}}}\big)\right). \tag{18}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{Kine}}(\theta_{\text{wm}}) + \lambda_{\text{act}} \mathcal{L}_{\text{act}}\left(\theta_{\text{act}}; \text{sg}\big(H_{\theta_{\text{wm}}}\big)\right). \tag{19}$$

Because the action head receives stopped features and uses a disjoint parameter set, $\nabla_{\theta_{\text{wm}}} \mathcal{L}_{\text{act}} = 0$. The auxiliary branch is unchanged across all checkpoint-matched comparisons and is not part of the KineWorld-specific optimization.

Training and Inference. During training, TAWD derives the calibrated support d from the renderer-derived transport and converts it into ρ^{tr} and w for reweighting the flow-matching objective. At inference time, the support calibration and loss-weighting branch is not required. The clean transport latent z^u remains as a structural condition, while the learned vector field is integrated from $\tau = 1$ to $\tau = 0$ to obtain future RGB latents, which are decoded into future video frames.

Table 1: **Single-View Reference Profiles.** EWM denotes EWMScore-P (Shang et al., 2026); metric definitions and aggregation appear in Appendix A.3.1. These aggregates do not constitute a shared RoboTwin 2.0 evaluation. Source status is documented in Appendix A.3.

Method	Overall	Generation Quality			Embodied Capability		
	EWM ↑	Vis. ↑	Mot. ↑	Cont. ↑	Phys. ↑	3D ↑	Ctrl. ↑
JF_World (DreamX Team et al., 2026)	64.85	66.52	30.66	56.91	69.27	97.88	88.06
BWM-Super (BWM Team, 2026)	64.30	67.20	30.16	58.36	64.15	97.35	87.19
BWM-Turbo (BWM Team, 2026)	63.96	66.99	30.10	57.22	64.46	97.74	86.05
FlowWAM-FiveAges (Chen et al., 2026b)	63.87	66.60	29.92	54.47	65.97	98.48	88.09
WoVR_Plus (Jiang et al., 2026)	57.89	58.37	30.52	57.91	47.95	85.31	80.75
WorldScape-0.2 (Fang et al., 2026)	57.55	58.00	30.84	55.41	47.61	88.49	79.15
PF-Cosmos (Zhang et al., 2026)	51.46	48.23	29.57	50.88	33.76	88.18	70.99
Wan2.2 (Wan Team, 2025)	49.21	49.55	26.89	51.51	29.24	81.31	66.57
GigaWorld-0 (GigaWorld Team et al., 2025)	48.06	46.09	30.37	46.49	31.35	76.10	68.57
Vidar (Feng et al., 2025)	47.13	47.61	29.13	44.76	30.02	74.63	66.60
IRASim (Zhu et al., 2024)	44.97	46.25	27.86	41.39	30.28	70.62	63.11
KineWorld (Ours)	68.95	67.09	33.36	69.01	76.49	98.53	87.94

Table 2: **Single-View Local Diagnostics.** Reported video-quality measurements; metric scales and evaluation scope are detailed in Appendix A.3.

Method	AES ↑	IMG ↑	BG ↑	SC ↑	PC ↑	INT ↑	PER ↑	IF ↑
Wan2.2-TI2V-5B (Wan Team, 2025)	42.10	44.42	85.17	**80.90**	0.96	68.68	88.52	76.00
KineWorld	**45.26**	**61.21**	**87.19**	75.53	**16.43**	**79.20**	**97.30**	**85.84**

Table 3: **Multi-View Reference Profiles.** TWB-Score is the TriWorldBench overall score (TriWorldBench Team, 2026); definitions appear in Appendix A.3.2. Bold marks column maxima. These profiles are not a shared RoboTwin 2.0 evaluation. Components appear in Table A.11.

Method	Overall	Embodied Fidelity			Generation Quality		
	TWB-Score ↑	Tri. ↑	Task ↑	Phys. ↑	Mot. ↑	Temp. ↑	Vis. ↑
BWM (BWM Team, 2026)	**65.54**	81.87	**86.05**	60.40	41.29	62.81	**31.42**
WoVR_Plus (Jiang et al., 2026)	65.39	**82.37**	82.80	57.12	**43.68**	64.10	31.14
DreamDojo (Gao et al., 2026)	51.72	69.63	56.24	43.84	27.96	60.84	21.02
Motus (Bi et al., 2025)	42.35	66.70	49.69	34.60	24.63	26.56	16.26
Genie Envisioner (Liao et al., 2025)	40.73	62.39	33.17	54.00	20.17	32.18	17.46
Ctrl-World (Guo et al., 2025)	42.35	57.42	43.72	34.67	29.28	46.17	16.64
KineWorld (Ours)	54.82	66.36	50.72	**63.24**	32.28	**69.90**	29.15

Table 4: **Core Module Ablation.** EWM denotes EWMScore-P. NFE counts denoiser evaluations.

Core Module		Retained Video Dimensions							Action Response (%)		Efficiency		
KTL	TAWD	EWM ↑	Visual ↑	Motion ↑	Content ↑	Physics ↑	3D ↑	Control ↑	INT ↑	TA ↑	Params (B) ↓	NFE ↓	FPS ↑
–	–	49.21	49.55	26.89	51.51	29.24	81.31	66.57	44.32	14.15	5.000	50	2.80
✓	–	65.99	64.46	32.39	66.39	69.40	95.95	84.73	76.69	62.11	5.001	25	3.01
✓	✓	68.95	67.09	33.36	69.01	76.49	98.53	87.94	82.40	70.57	5.001	25	3.02

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Dataset and Task Definition. We consider action-conditioned video prediction using RoboTwin 2.0 Aloha–AgileX Clean-50 demonstrations (Chen et al., 2025). The collection contains 1,800 development-training, 200 validation, and 500 held-out episodes, the combined training set contains the first two partitions. Single-view comparisons, multi-view comparisons, and local generation diagnostics are reported separately.

Implementation Details. We initialize the video generator from Wan2.2-TI2V-5B (Wan Team, 2025). KTL estimates camera-aligned transport with torchvision RAFT (Teed & Deng, 2020) and encodes it through the video VAE, while TAWD reweights only the future-RGB objective without adding sampling steps. Appendix A.5 describes the training procedure and controlled-comparison settings; Appendix A.6 provides parameter and runtime accounting.

Table 5: **KTL Ablation.** Three-view validation diagnostic on 100 episodes (300 videos per variant). Dataset scope and metric definitions appear in Appendix A.8.1.

Method	Overall Fidelity				View PSNR (dB)			Boundary PSNR (dB)	
	PSNR ↑	SSIM ↑	N-MAE ↓	Mot. Corr. ↑	Head ↑	Left ↑	Right ↑	First ↑	Last ↑
KineWorld (white wrist flow)	35.35	0.9622	16.532	0.9132	36.10	36.13	33.82	42.01	32.32
KineWorld (all-view RAFT)	**42.76**	**0.9860**	**5.713**	**0.9953**	**41.88**	**43.28**	**43.13**	**42.65**	**42.76**

Table 6: **TAWD Ablation.** EWM denotes EWMScore-P. Both variants start from the same step-315 checkpoint and are trained for 50 updates with the same data order, seeds, KTL condition, and inference and evaluation settings; only β changes. Intervals use 20,000 episode-paired bootstrap resamples over $M = 1000$ complete episodes (Appendix A.7).

Matched Continuation	β	Video Metrics				
		EWM ↑	Motion ↑	TA ↑	INT ↑	Visual ↑
Uniform weighting	0	66.10	32.50	62.80	77.00	64.80
TAWD	1	**67.85**	**33.35**	**68.40**	**81.20**	**65.40**
Δ (TAWD − Uniform)	–	+1.75	+0.85	+5.60	+4.20	+0.60
95% paired CI	–	[+1.10, +2.37]	[+0.25, +1.39]	[+3.20, +7.90]	[+2.10, +6.00]	[−0.10, +1.20]

Table 7: **Model Efficiency and Inference Speed.** Comparison of model size and inference efficiency across representative world models. KineWorld uses precomputed kinematic transport for the reported inference speed.

Method	Type	Efficiency		Output	Hardware
		Params (B) ↓	FPS ↑		
IRASim-XL (Zhu et al., 2024)	Robot WAM	0.679	0.53	288 × 512	1×A100
Wan2.2-TI2V-5B (Wan Team, 2025; Zhao et al., 2026)	General I2V	5.000	2.80	480 × 832	1×H100
RynnWorld-Teleop-Causal (Zhao et al., 2026)	Robot WAM	5.000	40.00	480 × 832	1×H100
GE-Sim 2.0 (Qiu et al., 2026)	Robot WAM	2.000	10.87	–	1×H100
KineWorld (Ours)	Robot WAM	5.001	3.02	640 × 480	1 HCU

5.2 Reference Profiles and Local Diagnostics

Single-View Result Profiles. Table 1 summarizes reported single-view system profiles; Table A.2 provides all 15 component scores. The KineWorld candidate profile has an EWMScore-P of 68.95 and Trajectory Accuracy of 70.57. These profiles provide system-level context rather than a common-training comparison on the Clean-50 held-out set.

Reported Local Diagnostics. Table 2 reports the local diagnostic profiles for Wan2.2-TI2V-5B and KineWorld. KineWorld has higher reported Image Quality (61.21 vs. 44.42), Interaction Quality (79.20 vs. 68.68), and Instruction Following (85.84 vs. 76.00), but lower Subject Consistency (75.53 vs. 80.90). These measurements are analyzed separately from the aggregate profile in Table 1, preserving their original evaluation scope and score scales.

Multi-View Result Profiles. Table 3 reports a TWB-Score of 54.82 for KineWorld, placing it third among the seven listed profiles. KineWorld has the highest Physical and 3D Coherence (63.24) and Temporal Consistency (69.90), exceeding the next-best reported values by 2.84 and 5.80 points, respectively. Tri-View Consistency (66.36) and Task Alignment (50.72) remain below the leading entries, leaving room to improve cross-view and task fidelity.

5.3 Ablation Study

Core Module Ablation. Table 4 shows that KTL alone greatly improves all metrics over the baseline. EWM rises from 49.21 to 65.99, Physics from 29.24 to 69.40, and TA from 14.15 to 62.11, while NFE drops from 50 to 25. Adding TAWD further boosts EWM to 68.95, Physics to 76.49, INT to 82.40, and TA to 70.57, with almost unchanged parameters and slightly higher FPS. These results demonstrate that KTL and TAWD are complementary core modules that deliver substantial gains in action response, physical plausibility, and generation efficiency at negligible overhead.

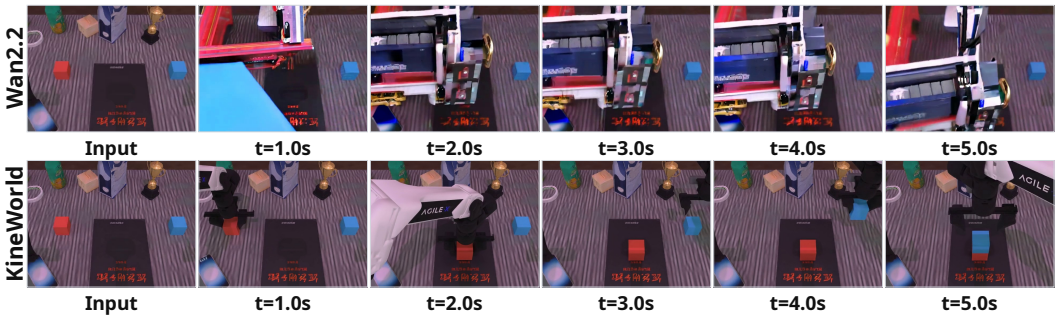

Figure 3: **Retained qualitative comparison between Wan2.2 and KineWorld.** Decoded episode-550 frames at $t = 1.0$, 2.0, 3.0, 4.0, and 5.0 s from 24-fps videos. This is a descriptive rollout comparison, not an action- or checkpoint-matched ablation.

KTL Design. Table 5 reports a fixed-checkpoint diagnostic on 100 three-view episodes (300 videos per variant). All-view RAFT guidance improves all metrics over white wrist-flow placeholders, raising PSNR from 35.35 to 42.76 dB, SSIM from 0.9622 to 0.9860, and motion correlation from 0.9132 to 0.9953, while reducing N-MAE from 16.532 to 5.713. Gains are consistent across views and boundary frames. These results validate KTL as effective camera-aligned transport for motion conditioning.

TAWD Design. Table 6 compares uniform CFM ($\beta = 0$) with TAWD ($\beta = 1$) under matched 50-update continuations from the same step-315 checkpoint on 1,000 paired episodes, sharing all settings except the spatial allocation of a fixed token weight. TAWD improves EWMScore-P by 1.75, Trajectory Accuracy by 5.60, Interaction Quality by 4.20, and Motion Quality by 0.85, while Visual Quality remains inconclusive. These results validate TAWD as an effective mechanism for reallocating future-RGB supervision toward action-responsive regions.

Model Size and Inference Speed. Table 7 reports model size and speed. KineWorld uses 5.001B parameters (KTL adds 1.19M, TAWD none) and reaches 3.02 FPS per HCU worker with precomputed transport. Baselines are literature-reported under method-specific settings, not local speed tests, and are not directly comparable across hardware and inference settings.

5.4 Visualization

Figure 3 qualitatively compares Wan2.2 and KineWorld rollouts on episode 550, showing decoded frames at $t = 1.0 - 5.0$ s from 24-fps videos. KineWorld preserves morphology and layout and yields action-responsive, temporally consistent rollouts, advancing action-consequence modeling.

6 Conclusion

We presented KineWorld, a kinematics-guided framework for action-conditioned video prediction. KTL constructs camera-aligned transport, and TAWD calibrates its support and uses it to redistribute future-RGB supervision between action-responsive regions and the full scene. The objective preserves the total token weight and recovers uniform supervision under zero motion, without adding inference parameters or denoising steps. Trained on ALOHA–AgileX bimanual manipulation data from **RoboTwin 2.0**, KineWorld achieves an **EWMScore-P of 68.95** in single-view evaluation and a **TWB-Score of 54.82** in multi-view evaluation. These results highlight the potential of kinematics-guided supervision to advance world models from appearance fitting toward action-consequence modeling for embodied decision-making.

Limitation and Future Work. KineWorld relies on calibrated robot and camera models, and its transport support does not explicitly capture object motion or contact dynamics. Future work will investigate calibration-robust transport estimation, object-aware supervision, and contact-aware world modeling.

AI USE DISCLOSURE

Generative AI tools assisted with research framing, manuscript editing, code inspection, experimental-protocol design, literature discovery, and reference and figure formatting. AI-assisted editing did not generate experimental measurements. The authors are responsible for the scientific claims, numerical results, citations, and final manuscript.

REPRODUCIBILITY STATEMENT

We provide a complete specification of the data splits, training and inference procedures, evaluation protocols, and statistical analysis to support reproduction of KineWorld. The matched comparison settings and objective definitions are described in the main text. We commit to publicly releasing the code, checkpoints, and data manifests upon acceptance.

REFERENCES

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model, 2025. URL https://arxiv.org/abs/2512.13030.

BWM Team. BWM: A low-cost high-fidelity world simulator for robot learning, 2026. URL https://arxiv.org/abs/2607.29302.

Jialei Chen, Kai Wang, Kang Chen, Shuaihang Chen, Feng Gao, Wenhao Tang, Zhiyuan Li, Weilin Liu, Zhuyu Yao, Boxun Li, Yuanbo Xu, and Chao Yu. LaWAM: Latent world action models for efficient dynamics-aware robot policies, 2026a. URL https://arxiv.org/abs/2606.15768.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, 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, 2025. URL https://arxiv.org/abs/2506.18088.

Yixiang Chen, Peiyan Li, Yuan Xu, Qisen Ma, Jiabing Yang, Kai Wang, Jianhua Yang, Dong An, He Guan, Gaoteng Liu, Jianlou Si, Jun Huang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. FlowWAM: Optical flow as a unified action representation for world action models, 2026b. URL https://arxiv.org/abs/2607.13017.

Yixiang Chen, Peiyan Li, Jiabing Yang, Keji He, Xiangnan Wu, Yuan Xu, Kai Wang, Jing Liu, Nianfeng Liu, Yan Huang, and Liang Wang. BridgeV2W: Bridging video generation models to embodied world models via embodiment masks, 2026c. URL https://arxiv.org/abs/2602.03793.

DreamX Team, Rui Chen, Xiangxiang Chu, Geng Li, Jifan Li, Qingfeng Shi, Datao Tang, Jing Tang, Jun Wang, and Pengfei Zhang. DreamX-Phi 1.0: Action-conditioned video world model for robotic manipulation, 2026. URL https://arxiv.org/abs/2608.13489.

Jianjie Fang, Yongyan Xu, Ziyou Wang, Chen Gao, Yuchao Huang, Zhaolu Wang, Rongze Tang, Mingyuan Jia, Baining Zhao, Weichen Zhang, Xin Zhang, Haisheng Su, Yu Shang, Wei Wu, Xinlei Chen, and Yong Li. Worldscape-MoE: A unified mixture-of-experts world model for scalable heterogeneous action control, 2026. URL https://arxiv.org/abs/2607.03964.

Yao Feng, Hengkai Tan, Xinyi Mao, Guodong Liu, Shuhe Huang, Chendong Xiang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist bimanual manipulation, 2025. URL https://arxiv.org/abs/2507.12898.

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 Fan. DreamDojo: A generalist robot world model from large-scale human videos, 2026. URL https://arxiv.org/abs/2602.06949.

GigaWorld Team, Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Haoyun Li, Jiagang Zhu, Kerui Li, Mengyuan Xu, Qiuping Deng, Siting Wang, Wenkang Qin, Xinze Chen, Xiaofeng Wang, Yankai Wang, Yu Cao, Yifan Chang, Yuan Xu, Yun Ye, Yang Wang, Yukun Zhou, Zhengyuan Zhang, Zhehao Dong, and Zheng Zhu. GigaWorld-0: World models as data engine to empower embodied ai, 2025. URL https://arxiv.org/abs/2511.19861.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation, 2025. URL https://arxiv.org/abs/2510.10125.

Zhennan Jiang, Shangqing Zhou, Yutong Jiang, Zefang Huang, Mingjie Wei, Yuhui Chen, Tianxing Zhou, Zhen Guo, Hao Lin, Quanlu Zhang, Yu Wang, Haoran Li, Chao Yu, and Dongbin Zhao. WoVR: World models as reliable simulators for post-training vla policies with rl, 2026. URL https://arxiv.org/abs/2602.13977.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, 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, 2025. URL https://arxiv.org/abs/2508.05635.

Kechen Liu and Ola Shorinwa. CLAP: Cross-embodiment video world models are zero-shot physical simulators, 2026. URL https://arxiv.org/abs/2608.27406.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. RealWonder: Real-time physical action-conditioned video generation, 2026. URL https://arxiv.org/abs/2603.05449.

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, 2026. URL https://arxiv.org/abs/2605.27491.

Yu Shang, Yinzhou Tang, Yiding Ma, Zhuohang Li, Lei Jin, Weikang Su, Xin Jin, Zhaolu Wang, Ziyou Wang, Xin Zhang, Haisheng Su, Weizhen He, Wei Wu, Haoyi Duan, Gordon Wetzstein, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, Chen Gao, and Yong Li. WorldArena 2.0: Extending embodied world model benchmarking on modality, functionality and platform, 2026. URL https://arxiv.org/abs/2605.17912.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419, 2020.

TriWorldBench Team. TriWorldBench: A benchmark evaluating triple-view embodied world models. Official benchmark and public leaderboard, 2026. URL https://www.triworldbench.com/. Accessed 2026-09-03.

Wan Team. Wan: Open and advanced large-scale video generative models, 2025. URL https://arxiv.org/abs/2503.20314.

Linbo Wang, Yupeng Zheng, Qiang Chen, Shiwei Li, Yichen Zhang, Zebin Xing, Qichao Zhang, Xiang Li, Deheng Qian, Pengxuan Yang, Yihang Dong, Ce Hao, Xiaoqing Ye, Junyu Han, Yifeng Pan, and Dongbin Zhao. Latent-WAM: Latent world action modeling for end-to-end autonomous driving, 2026. URL https://arxiv.org/abs/2603.24581.

Zhaoyang Yang, Yurun Jin, Lizhe Qi, Cong Huang, and Kai Chen. EA-WM: Event-aware generative world model with structured kinematic-to-visual action fields, 2026. URL https://arxiv.org/abs/2605.06192.

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, Ruihua Zhang, Enze Xie, Ming-Yu Liu, and Daquan Zhou. PhysisForcing: Physics reinforced world simulator for robotic manipulation, 2026. URL https://arxiv.org/abs/2606.28128.

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, 2026. URL https://arxiv.org/abs/2607.06558.

Haoyu Zhen, Zixian Gao, Qiao Sun, Yilin Zhao, Yuncong Yang, Yilun Du, Tsun-Hsuan Wang, Yi-Ling Qiao, and Chuang Gan. Action images: End-to-end policy learning via multiview video generation, 2026. URL https://arxiv.org/abs/2604.06168.

Pengfei Zhou, Shengcong Chen, Di Chen, Jiaxu Wang, Rongjun Jin, Bingwen Zhu, Yike Pan, Songen Gu, Kuanning Wang, Shufeng Nan, Xingyu Qiu, Chenhao Qiu, Pu Yang, Yunuo Cai, Jianxiong Gao, Yifan Li, Yanwei Fu, Xiangyu Yue, Zhi Chen, and Jianlan Luo. τ_0-WM: A unified video-action world model for robotic manipulation, 2026. URL https://arxiv.org/abs/2606.01027.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: Learning interactive real-robot action simulators, 2024. URL https://arxiv.org/abs/2406.14540.