Learning Transferable Dynamics Priors from Action to World Modeling

Ze Huang[1*], Jiahui Zhang[1*], Hairuo Liu[2,3*], Chenxi Zhang[2], Ran Cheng[4], and Li Zhang[1,2†]

[1] School of Data Science, Fudan University, Shanghai, China
[2] Shanghai Innovation Institute, Shanghai, China
[3] Shanghai Jiao Tong University, Shanghai, China
[4] McGill University, Montreal, QC, Canada

https://github.com/LogosRoboticsGroup/A2World

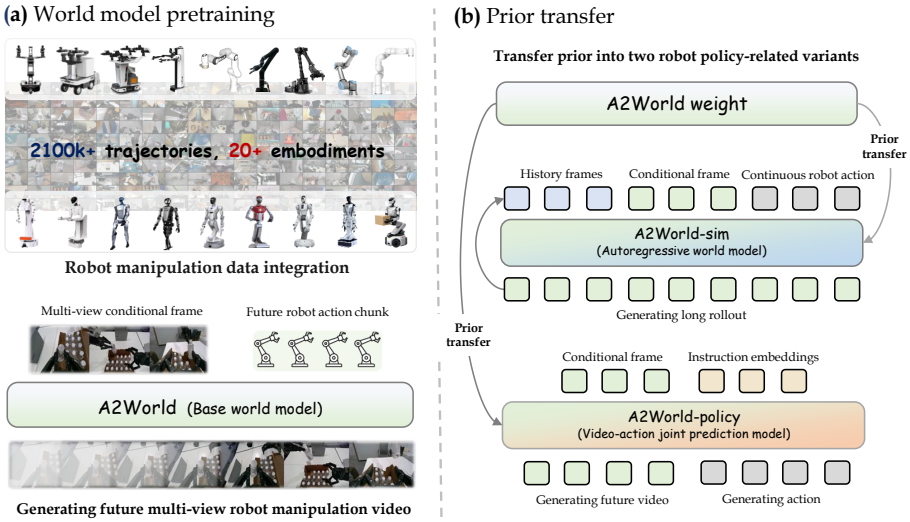

Fig. 1: We view action-conditioned world modeling as a transferable dynamics prior for robot learning. (a) We curate 2,100k+ robot manipulation trajectories spanning 20+ embodiments, and pretrain a DiT-based multi-view base world model (A2World) to predict future spatiotemporally consistent manipulation videos from an initial frame and a future action chunk. (b) The pretrained weights are then adapted as a reusable prior into two policy-related variants: A2World-sim, a long-horizon autoregressive simulator for policy evaluation, and A2World-policy, a video-action joint prediction model instantiated as an instruction-conditioned robot policy.

Abstract. We study action-conditioned world modeling as a scalable way to learn transferable dynamics priors for robot learning. By pretraining a model to predict how actions drive visual scene evolution, the resulting world model captures reusable interaction dynamics beyond appearance-level video generation. Concretely, we pretrain a multi-view interactive base diffusion world model, A2World, on large-scale

* Equal contribution. † Corresponding author: lizhangfd@fudan.edu.cn

robot manipulation data with real action annotations. We validate the learned dynamics priors from two complementary perspectives. First, we adapt A2World into a task- or scene-specialized real-world simulator, A2World-sim, whose long-horizon rollouts support simulator-based policy evaluation and scalable what-if analysis by replacing real-robot rollouts with world model rollouts. Second, starting from the same pretrained weights, we adapt A2World into a video-action joint prediction model, A2World-policy, that predicts actions under visual and instruction conditioning. Experiments across simulation benchmarks and real-robot settings demonstrate that action-conditioned world model pretraining yields transferable dynamics priors that benefit both simulator-centric and policy-centric robot learning.

Keywords: World model · Robot learning · Robot policy

1 Introduction

Recent robot learning methods increasingly build on video generative models [1, 6, 15, 47], and existing efforts have largely evolved along two directions. One line of work adapts these models into vision-language-action (VLA) policies [16, 18, 36, 38, 49, 50], with recent advances further moving toward joint video-action prediction [4, 29, 31, 32, 41, 55, 60]. The other line develops action-conditioned world models for data augmentation and policy evaluation [12, 13, 22, 24, 45, 61], and more recently combines them with reward models and reinforcement learning for policy post-training [25, 30, 58, 62].

Despite this progress, current approaches still underexploit robot-data pretraining as a source of transferable dynamics priors. Many works [13, 25, 29, 61] directly fine-tune from generic video-generation checkpoints, without large-scale pretraining on action-grounded robot data. Others [12, 36] pretrain on a single dataset, limiting the embodiments, camera setups, and motion patterns the model can absorb. Even large-scale efforts [4, 31, 58] are often optimized for a single downstream objective, rather than explicitly designed for transferable reuse across simulator-centric and policy-centric pipelines. As a result, the field has yet to establish a unified robot-data pretraining paradigm that learns reusable action-to-dynamics priors and reliably benefits both world model-based simulation and downstream policy learning.

In this paper, we study action-conditioned world modeling as a scalable route to learn transferable dynamics priors for robotics. The key intuition is that actions provide a natural causal supervision signal in manipulation: while objects, scenes, and viewpoints vary widely across datasets, the underlying interaction rules are governed by how actions induce state changes (e.g., contacts, grasps, pushes, and releases). Pretraining a model to predict visual scene evolution conditioned on actions therefore forces it to encode controllable, action-grounded dynamics beyond appearance-level video prediction, yielding reusable interaction priors. Leveraging the recent surge of openly available high-quality robot manipulation datasets [7, 21, 23, 46, 53], we pretrain a multi-view interactive diffusion world

model, A2World, directly on real robot action annotations. We do not rely on auxiliary latent-action models [4, 12, 55] to produce indirect pseudo labels, encouraging generalization across embodiments, camera setups, tasks, and motion patterns. We validate the value of this pretraining from two complementary perspectives that are central to robot learning. First, on the simulation side, we fine-tune A2World into a history-aware autoregressive world model, A2World-sim, and specialize it as a task- or scene-specific real-world simulator. A2World-sim supports long-horizon rollouts for simulator-based robot policy evaluation by replacing real-world rollouts with world model rollouts, reducing reliance on real-robot interaction during adaptation and assessment. Second, on the policy side, starting from the same pretrained weights, we transfer A2World into A2World-policy, a MoE-like video-action policy that shares attention across video and action tokens while keeping action-specific denoising branches, enabling action generation to reuse the pretrained visual dynamics prior while retaining dedicated modeling capacity.

The main contributions of this paper are as follows: **(i)** We introduce action-to-video world model pretraining as a robot-data-driven approach for learning transferable dynamics priors, and instantiate it with a multi-view interactive diffusion world model, A2World, pretrained on large-scale, high-quality manipulation data spanning diverse embodiments, tasks, environments, and object categories; **(ii)** We demonstrate that the learned dynamics prior can be reused for simulator-centric robot learning by adapting A2World into a history-aware autoregressive simulator, A2World-sim, enabling long-horizon action-conditioned rollouts for real-world simulation, simulator-based policy evaluation, and simulator-based post-training; **(iii)** We demonstrate that the same dynamics prior can also be reused for policy-centric robot learning by adapting A2World into a MoE-like video-action joint prediction model, A2World-policy, yielding a strong robot policy under visual and instruction conditioning; **(iv)** Across simulation benchmarks and our real-robot platform, we show that action-to-video pretraining produces markedly stronger transferable priors than text-conditioned video pretraining and other robot pretraining baselines, translating into consistent downstream gains and policy improvements across tasks.

2 Related works

Action-conditioned robot world models Action-conditioned robot world models [12, 13, 22, 24, 25, 30, 36, 43, 45, 51, 54, 58, 61, 62] condition on low-level robot signals, such as end-effector pose changes or joint-angle deltas, to generate future manipulation videos. Early applications of such models focused on data augmentation for improving VLA training [13], or on serving as safe and scalable validators for VLA policies without requiring real-robot execution [36, 45]. More recent work has also started to improve their generalization ability across unseen tasks and environments [12]. A subset of world model-related works further treats the learned world model as a simulator, and combines it with reward models and reinforcement learning algorithms for VLA policy post-training [25, 30, 54, 58, 62].

Recent efforts have moved beyond validation-only settings (e.g., simply replacing the simulator [37,39] in standard benchmarks with a learned world model), and toward more practical pipelines that couple real-world world model simulation with policy improvement [58,62].

World-action models Instead of predicting control commands directly from a single observation, world-action models [18,29,31,32,36,41,55,56] predict how the scene will evolve and then convert those predicted visual changes into executable actions. Cosmos Policy [29] illustrates how video models can be adapted for control by fine-tuning them to produce action-relevant predictions. DreamZero [55] further reveals the potential of using video models as base models for zero-shot transfer to action generation. LingBot-VA [31] interleaves video and action tokens in an autoregressive diffusion policy and introduces efficient mechanisms for closed-loop control. In contrast, we start from an action-conditioned video world model and show that its pretrained dynamics prior transfers to a strong policy with only minimal modifications. This transfer avoids a separate action-only pretraining stage and requires only lightweight policy-specific adaptations beyond joint video–action modeling.

3 Methodology

3.1 A2World: the base robot world model

Our base world model (A2World, Fig. 2.(a)) is designed as an action-conditioned generative world model to learn transferable dynamics priors. Given a conditioning frame o_t and a chunk of future actions $a_{t+1:t+k}$, it forecasts future observations $o_{t+1:t+k}$:

$$\text{A2World} : p\left(o_{t+1:t+k} \mid o_t, a_{t+1:t+k}\right). \tag{1}$$

Action conditioning Concretely, the action chunk a is encoded by an MLP, $e = \text{MLP}(a)$, and the resulting action embedding is added to the diffusion timestep embedding used by every DiT block: $\tilde{\tau}(\sigma) = \tau(\sigma) + e$. $\tau(\sigma)$ denotes the diffusion timestep embedding followed by an MLP projection, and $\tilde{\tau}(\sigma)$ is then consumed by adaptive layer normalization to produce dynamic modulation parameters (scale, shift, and gate).

In this base model, we disable other conditioning inputs by setting the conditional feature to zero, i.e., $\mathbf{c} = \mathbf{0}$, so the model relies solely on action-conditioned temporal modulation.

Specifically, A2World follows a standard latent video diffusion pipeline and operates in the continuous latent space produced by the WAN2.1 tokenizer [47]. It injects timestep conditions into each DiT block [42] via adaptive layer normalization [2]. We train A2World with the EDM denoising score matching objective [26] to recover the clean latent sequence from its corrupted version:

$$\mathcal{L}_{\text{A2World}}(\sigma) = \mathbb{E}_{\mathbf{z},\mathbf{n}}\left[\left\|\text{A2World}\left(\mathbf{z}+\mathbf{n}; \sigma, \mathbf{c}=\mathbf{0}, a\right) - \mathbf{z}\right\|_2^2\right], \tag{2}$$

where $\mathbf{z}$ is a clean VAE-encoded latent video, $\mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ is i.i.d. Gaussian noise, and σ is the noise level.

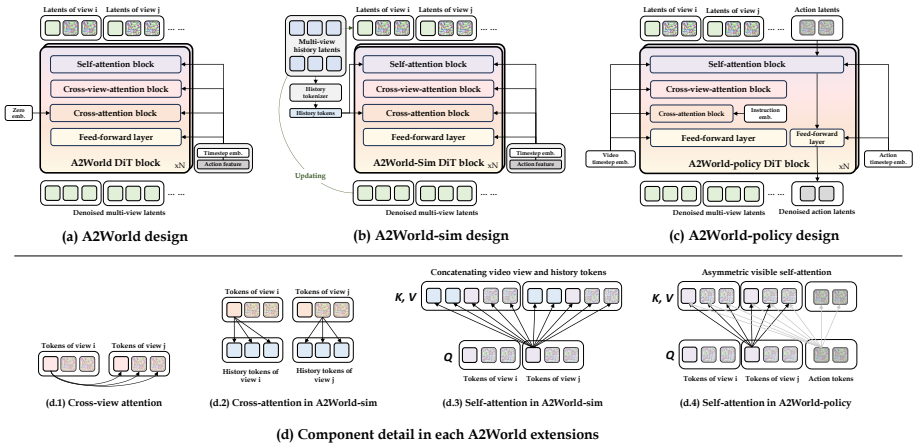

Fig. 2: Detailed module design of our A2World series.

Multi-view generation Most robot manipulation setups naturally provide multi-view observations, typically including multiple first- or third-view cameras. Thus, our A2World is implemented to generate multi-view videos jointly. We denote the multi-view observation as $o_t = \left\{ o_t^{(v)} \right\}_{v=1}^{V}$, where v indexes the camera view. We pack the V views into the temporal dimension and run a single latent video diffusion forward pass: let $\mathbf{z}_{\text{mv}}$ denote the clean multi-view latent sequence, which we reshape as $\mathbf{z}_{\text{mv}} \in \mathbb{R}^{\mathbf{B}\times\mathbf{C}\times(\mathbf{V}\cdot\mathbf{T})\times\mathbf{H}\times\mathbf{W}}$, so that multi-view generation is processed as temporally concatenated video generation. To identify different cameras, we introduce learnable view embeddings $\epsilon_{\text{view}}(v) \in \mathbb{R}^{\mathbf{d_e}}$ for different views (e.g., first-view vs. third-view), broadcast them over the corresponding spatial-temporal latent grids, and concatenate them to the latent input tokens before patch embedding along the channel dimension:

$$\tilde{\mathbf{z}}_{\text{mv}}^{(v)} = \text{concat}\left(\mathbf{z}_{\text{mv}}^{(v)}, \epsilon_{\text{view}}(v)\right), \quad \tilde{\mathbf{z}}_{\text{mv}} \in \mathbb{R}^{\mathbf{B}\times(\mathbf{C}+\mathbf{d_e})\times(\mathbf{V}\cdot\mathbf{T})\times\mathbf{H}\times\mathbf{W}}. \tag{3}$$

This provides explicit camera identity to the DiT. In addition, to enforce cross-view consistency, we insert cross-view attention modules (Fig. 2.(d.1)) into each DiT block, where tokens from one view attend to tokens from the other views:

$$\tilde{\mathbf{z}}_{\text{mv}}^{(w)} \leftarrow \tilde{\mathbf{z}}_{\text{mv}}^{(w)} + \text{CrossViewAttn}\left(\tilde{\mathbf{z}}_{\text{mv}}^{(w)}, \tilde{\mathbf{z}}_{\text{mv}}^{(u)}, u \neq w\right). \tag{4}$$

This module encourages spatiotemporally consistent rollouts across views, while preserving view-specific details.

For notational convenience, we use $\mathbf{z}$ to denote the multi-view latent tokens, and write $\mathbf{z}^v$ when distinguishing them from action latents in the remainder.

A2World pretraining We pretrain A2World on curated high-quality robot manipulation datasets to learn transferable dynamics priors across diverse tasks and environments. To enable action-conditioned pretraining across embodiments, we

Table 1: Curated and preprocessed datasets used for pretraining in A2World and fine-tuning in A2World-sim and A2World-policy.

Dataset	Trajectories	Embodiment	Data usage	Used in
AgiBot [7]	1003k	Genie-1	Pretraining	A2World
DROID [27]	76k	Franka	Pretraining	A2World
OPEN-X [40]	19k	UR5, WidowX, X-ARM, Jaco2, Franka	Pretraining	A2World
InternData-A1 [46]	630k	Genie-1, ARX LIFT-2, Franka, Agilex Split Aloha	Pretraining	A2World
InternData-M1-Agilex [21]	105k	Dual-arm Franka	Pretraining	A2World
RoboCoin [53]	246k	15 types including AirBot MMK2, Unitree G1edu-u3, etc.	Pretraining	A2World
Galaxea [23]	77k	Galaxea R1 Lite	Pretraining	A2World
LIBERO [37]	6k	Franka	Fine-tuning	A2World-sim, A2World-policy
LIBERO-Plus-Spatial [11]	3.7k	Franka	Evaluation	A2World-sim, A2World-policy
RoboNet [10]	100k	Sawyer, Franka, WidowX, KUKA, Baxter, Fetch	Fine-tuning	A2World-sim
Custom data	2k	Dual-arm Flexiv Rizon 4S with Robotiq-2F-85 gripper	Fine-tuning	A2World-sim, A2World-policy

unify all actions into a shared dual-arm format, where each arm is represented by 7 dimensions (end-effector pose and gripper state). For single-arm robots, the missing arm is zero-padded. Tab. 1 summarizes the datasets used. Since different datasets often have different camera setups and view conventions, directly mixing them can introduce conflicts for multi-view generation. To mitigate this, we adopt a dataset-consistent batching strategy, i.e., each mini-batch is sampled from a single dataset only.

3.2 A2World-sim: adapting A2World into a long-horizon simulator

A2World is pretrained to learn transferable action-to-dynamics priors from large-scale robot data. To reuse this prior for long-horizon simulation, we transfer A2World into a history-aware autoregressive world model, A2World-sim (Fig. 2.(b)), which conditions on past observations and sequentially rolls out future observations under actions:

$$\text{A2World-sim} : p\left(o_{t+1:t+k} \mid o_t,\ a_{t+1:t+k},\ \mathcal{H}_{t-1}\right), \quad \mathcal{H}_{t-1} = o_{1:t-1}. \tag{5}$$

Pose-guided history sampling Given a history sequence and robot actions, we sample a compact set of frames that covers the executed motion trajectory. Specifically, we compute a weighted arc-length from relative actions and sample frames uniformly along it (Alg. 1). This preserves key states, e.g., turning points, under a fixed history budget. For clarity, we illustrate the single-arm variant here, and the dual-arm case is analogous and provided in the Appendix.

History injection After selecting a compact history subset, we inject history into the DiT in two complementary ways. First, the sampled history latent sequence is tokenized and mapped into history tokens:

$$\mathbf{H} = \text{Tok}_{\text{hist}}(\mathcal{H}[\mathcal{S}]) \in \mathbb{R}^{B \times N_h \times D}. \tag{6}$$

These tokens replace the generic cross-attention condition (set to zero in pretraining) and are attended by target video latent tokens (Fig. 2.(d.2)):

$$\mathbf{z} \leftarrow \mathbf{z} + \text{CrossAttn}\left(\mathbf{z}, \mathbf{H}\right). \tag{7}$$

Second, we additionally inject the same history tokens into the self-attention memory path by projecting them to key and value memories and concatenating

Algorithm 1 Arc-uniform history sampling

Require: History length T_h; relative actions $\{\Delta x_r\}_{r=1}^{T_h-1}$ with $\Delta x_r = [\Delta p_r, \Delta\theta_r]$; budget m; weights $w_t, w_r > 0$
Ensure: Sampled history indices $\mathcal{S}$
Choose anchor indices r_s (earliest valid frame) and $r_e = T_h$ (latest frame)
Compute weighted step lengths: $d_r \leftarrow \sqrt{w_t\|\Delta p_r\|_2^2 + w_r\|\Delta\theta_r\|_2^2}$ for $r = 1, \ldots, T_h - 1$
Compute cumulative arc-length: $A_{r_s} \leftarrow 0$; $A_r \leftarrow \sum_{q=r_s}^{r-1} d_q$ for $r = r_s + 1, \ldots, r_e$
Initialize $\mathcal{S} \leftarrow \{r_s, r_e\}$ and set $n_{\text{mid}} \leftarrow m - 2$
for $s = 1$ to n_{mid} **do**
    $\bar{A}_s \leftarrow A_{r_s} + \frac{s}{m-1}(A_{r_e} - A_{r_s})$
    $\hat{r} \leftarrow \arg\min_{r \in (r_s, r_e)} |A_r - \bar{A}_s|$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{\hat{r}\}$
end for
return chronological indices in $\mathcal{S}$

them with the current latent self-attention (Fig. 2.(d.3)). This provides a global history memory injection, allowing target latent tokens to directly interact with compact historical states inside self-attention.

Autoregressive generation Given an initial observation o_t and future actions, we roll out A2World-sim in a chunk-wise autoregressive manner. At each step, the model predicts a future chunk $o_{t+1:t+k}$ conditioned on the current frame, history memory, and action chunk, and the generated frames are then appended to the history buffer and reused as the condition for the next rollout step. This converts the short-horizon predictor into a long-horizon action-conditioned simulator.

To improve long-horizon stability, we adopt a Self-forcing-style [19] training strategy: during training, the model is periodically conditioned on its own generated frames, rather than always using ground-truth frames. Unlike teacher-forcing distillation, we do not require a separate teacher model, since under action conditioning and a given initial frame, the future trajectory is largely determined by the underlying dynamics. Self-forcing therefore directly exposes the model to its own rollout errors and trains it to recover from them.

3.3 A2World-policy: adapting A2World into a robot policy

To transfer the pretrained dynamics prior to the policy setting, we adapt A2World into a MoE-like vision-action joint prediction model, A2World-policy (Fig. 2.(c)), which jointly models future observations and actions in a single generative process. This adaptation instantiates the world model as an instruction-conditioned robot policy that takes a language instruction l and an initial frame o_t as input:

$$\text{A2World-policy} : p\left(o_{t+1:t+k}, a_{t+1:t+k} \mid o_t, l\right). \tag{8}$$

We encode the instruction l with a pretrained T5 text encoder [44] and use the resulting token embeddings $\mathbf{h}_l$ as the cross-attention context in each DiT block. We model future observation latents and future action tokens jointly under the same conditional context given the initial frame and instruction, so that the pretrained world model can be transferred to both simulator-oriented and policy-oriented downstream tasks with one initialization.

Joint diffusion formulation Let $\mathbf{z}^v$ and $\mathbf{z}^a$ denote clean video and action latents, respectively. We add Gaussian perturbations independently:

$$\mathbf{z}^v_{\sigma_v} = \mathbf{z}^v + \mathbf{n}^v, \quad \mathbf{n}^v \sim \mathcal{N}(0, \sigma_v^2 \mathbf{I}), \qquad \mathbf{z}^a_{\sigma_a} = \mathbf{z}^a + \mathbf{n}^a, \quad \mathbf{n}^a \sim \mathcal{N}(0, \sigma_a^2 \mathbf{I}), \tag{9}$$

where σ_v and σ_a are the modality-specific noise levels. In our shared-timestep training variant, a single base noise level σ_{base} is sampled and then scaled per modality $\sigma_v = \alpha_v \sigma_{\text{base}}$, $\sigma_a = \alpha_a \sigma_{\text{base}}$, which improves video-action alignment while preserving modality-specific noise scales.

MoE-like video-action blocks Let $\mathbf{z}^\ell_v$ and $\mathbf{z}^\ell_a$ denote video and action tokens at block ℓ. An A2World-policy block shares one self-attention module across modalities (Fig. 2.(d.4)), with modality-specific AdaLN and MLP branches. We refer to this as MoE-like because video and action share the attention expert for interaction, while each modality keeps its own lightweight denoising branch. Video tokens are updated by each attention layer. Action tokens attend to both video and action tokens via the shared self-attention. Both streams then pass through their AdaLN and MLP branches, and the final action sequence is predicted by a linear head on the last-layer action tokens.

Training objective and guided inference The model predicts clean video and action latents jointly:

$$(\hat{\mathbf{z}}^v, \hat{\mathbf{z}}^a) = \text{A2World-policy}(\mathbf{z}^v + \mathbf{n}^v, \mathbf{z}^a + \mathbf{n}^a; \sigma_v, \sigma_a, \mathbf{c} = \mathbf{h}_l)\,, \tag{10}$$

with a weighted joint denoising objective:

$$\mathcal{L}_{\text{A2World-policy}}(\sigma_v, \sigma_a) = \mathbb{E}_{\mathbf{z}^v, \mathbf{z}^a, \mathbf{n}^v, \mathbf{n}^a}\left[\mathbf{w}(\sigma_v)\|\hat{\mathbf{z}}^v - \mathbf{z}^v\|_2^2 + \lambda_a\, \mathbf{w}(\sigma_a)\|\hat{\mathbf{z}}^a - \mathbf{z}^a\|_2^2\right]. \tag{11}$$

At inference, we use modality-wise classifier-free guidance:

$$\hat{\mathbf{z}}^m_{\text{cfg}} = \hat{\mathbf{z}}^m_{\text{u}} + s_m(\hat{\mathbf{z}}^m_{\text{c}} - \hat{\mathbf{z}}^m_{\text{u}})\,, \quad m \in \{v, a\}, \tag{12}$$

where s_v and s_a can be set separately for video and action, enabling controllable trade-offs between visual fidelity and action accuracy. When a single guidance scalar γ is used, we set $s_v = s_a = 1 + \gamma$.

4 Experiments

4.1 Experimental setups

World model training setups Our A2World initializes its base DiT backbone from the Cosmos-Predict2-2B-Video2World checkpoint [1], while the additional conditioning and multi-view modules are properly initialized, resulting in a total model size of 2.5B parameters. A2World conditions on the initial frame and, given a 20-step action chunk, generates the next 20 frames. We pretrain the A2World on a mixture of processed robot manipulation datasets [7,21,23,27,40,46,53], with a total of 2156k trajectories (Tab. 1). For pretraining, we train all model parameters with 64 H200 GPUs, using a batch size of 12 per GPU and

gradient accumulation of 4 for 2 epochs. For fine-tuning A2World-sim, we use 8 H200 GPUs with a batch size of 24 and a task-dependent number of steps. Both stages use the fused Adam optimizer with a learning rate of 1e-4 and weight decay of 0.1. For history-aware mechanism, we set history length $T_h = 20$, $w_r = 0.3$, $w_t = 1.0$. Other parameter settings are detailed in the Appendix.

A2World-policy training setups We initialize the A2World-policy from the pretrained A2World, and initialize the cross-attention from the Cosmos-Predict2-2B checkpoint since A2World pretraining sets the cross-attention conditioning to zero. We initialize action-specific modules by copying the corresponding video-branch parameters from A2World for stable joint fine-tuning. The resulting A2World-policy has 3.0B parameters. We train with 32 H200 GPUs, global batch size 256, and learning rate 1e-4, using 20k steps for LIBERO and real-robot fine-tuning. For OOD policy evaluation, we fine-tune A2World-policy variants on LIBERO for 24k steps and evaluate on LIBERO-Plus Spatial.

Custom real-robot dataset collection We constructed a dual-arm manipulation platform based on Flexiv robots, following the setup of Toyota Research Institute (TRI) [3], and conducted data collection through VR teleoperation. The dataset comprises 5 tasks, i.e., *insert RAM module*, *flip small box*, *toggle power switch*, *lift box high*, *put chain in the box*. These tasks involve challenging scenarios such as rich-contact manipulation, articulated objects, and deformable objects. To ensure data quality, we performed careful curation by selecting episodes with minimal idle frames and high task completion rates. Details are shown in Fig. 3.

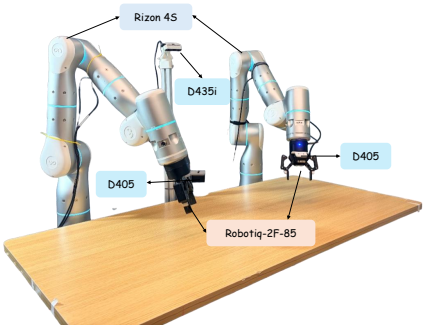

Fig. 3: Our real-robot platform. Two Flexiv arms with Robotiq-2F-85 grippers are mounted symmetrically at 45° facing the tabletop. The vision system uses one front-view Intel RealSense D435i and two wrist-mounted D405 cameras (480×640, 30fps).

4.2 Base world model capability demonstration

For our pretrained A2World, we qualitatively demonstrate strong action-conditioned multi-view rollouts. Fig. 4 shows DROID [27] rollouts. From the same initial observation, A2World can be steered to grasp different objects and can also simulate failures under appropriate actions, suggesting it models action-to-visual dynamics beyond success-only generation. Fig. 5 further shows precise full-DoF control on RoboCoin [53] with consistent predictions across heterogeneous views. Finally, Fig. 6 presents out-of-distribution (OOD) rollouts on RoboMind [52] and VIOLA [63], where scenes and camera setups are unseen during pretraining. A2World remains coherent, indicating robust generalization and transferable dynamics priors that we exploit in downstream experiments.

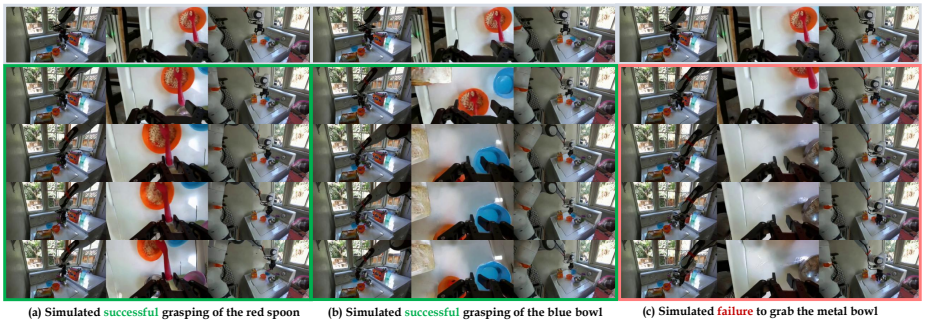

Fig. 4: Rollout presentation on DROID [27] using A2World. Given an initial frame, A2World can freely simulate grasping different objects in the scene, and can also reproduce a failed grasp attempt under the corresponding action. This suggests that A2World learns to respond to action inputs rather than merely generating 'successful' outcomes, which is crucial for downstream counterfactual evaluation.

Fig. 5: Full-DoF control of robotic arm on RoboCoin [53] using A2World. Such scripted control never appears in pretraining data, yet A2World can faithfully simulate the resulting scene navigation, indicating strong counterfactual controllability.

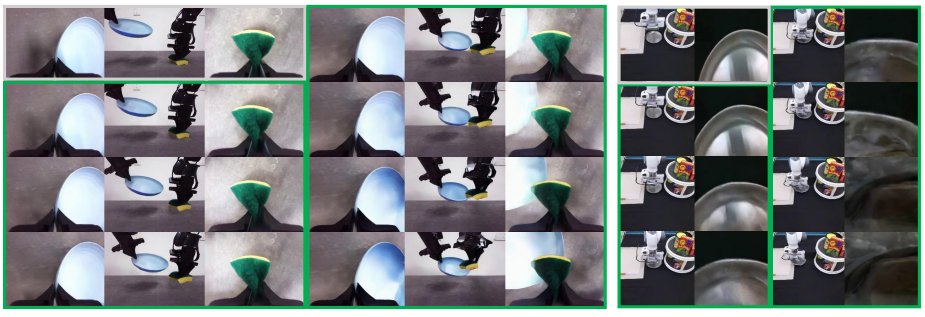

Fig. 6: Out-of-distribution rollouts on RoboMind [52] (left) and VIOLA [63] (right) using A2World. For both RoboMind and VIOLA, the scenes and camera setups are unseen during pretraining.

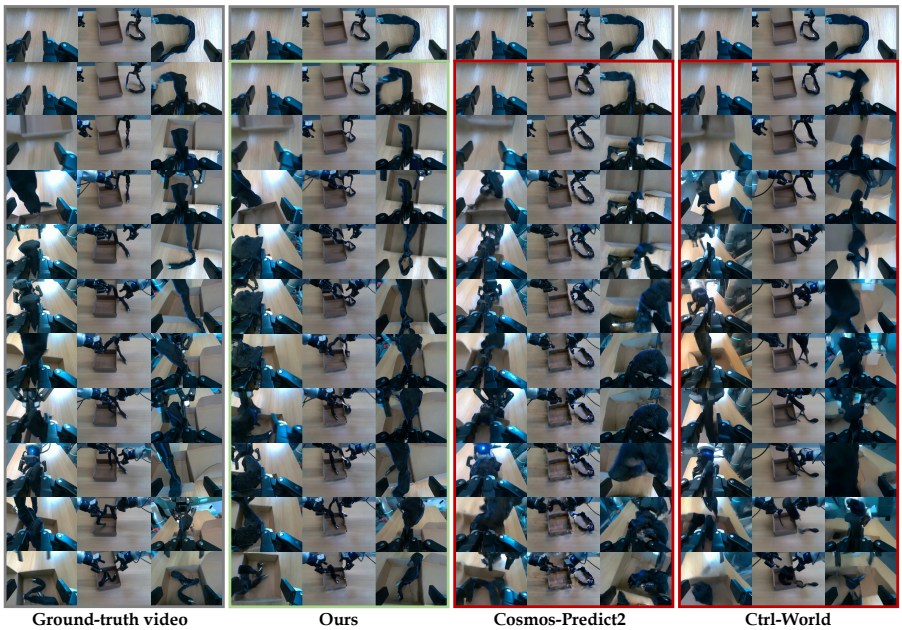

Fig. 7: Qualitative long-rollout generation on *Put chain in the box*. We autoregressively generate long-horizon videos from same initial frame and real robot actions (frames every 2s). Baselines drift after ∼6s and quickly collapse, while our model follows the actions more faithfully, completes the task, and maintains high visual quality.

Table 2: Rollout quality comparison on simulation and real-robot datasets.

Methods	LIBERO [37]					Real Robot (Flexiv Rizon 4S)				
	PSNR ↑	SSIM ↑	tSSIM ↑	$\overline{\text{EPE}}$ ↓	$\overline{\cos}$ ↑	PSNR ↑	SSIM ↑	tSSIM ↑	$\overline{\text{EPE}}$ ↓	$\overline{\cos}$ ↑
Cosmos-Predict2 [1]	25.36	.8792	.7631	.4009	.5755	24.99	.8355	.7009	.8010	.2289
Ctrl-World [13]	23.60	.8632	.7445	.6827	.3730	21.91	.8422	**.7158**	.9689	.1195
Prophet [58]	26.12	.8887	.7789	.3667	.5932	<u>25.55</u>	<u>.8454</u>	.7102	<u>.7439</u>	<u>.2650</u>
A2World-sim-T-pre	<u>26.18</u>	<u>.8892</u>	<u>.7794</u>	<u>.3533</u>	<u>.5973</u>	24.64	.8198	.6411	.9328	.2293
A2World-sim	**26.64**	**.8957**	**.7862**	**.3498**	**.6045**	**25.95**	**.8511**	<u>.7139</u>	**.7218**	**.2784**

4.3 Fine-tuned world model evaluation

Rollout quality evaluation We fine-tune A2World into A2World-sim on the LIBERO [37] and our custom data for 30k steps, and evaluate generation quality. We compare against strong pretrained baselines: **(i)** Cosmos-Predict2 [1], using the Cosmos-Predict2-2B-Video2World-480p-10fps model, which is pretrained on broad web-scale data with text conditioning; **(ii)** Ctrl-World [13], pretrained on DROID with multi-view modeling; **(iii)** Prophet [58], which is pretrained with action conditioning but operates in a single-view setting; **(iv)** a text-conditioned A2World variant pretrained on the same robot data, which replaces action conditioning with T5-based text-conditioning during pretraining, denoted as **T-pre**. Beyond standard video-generation metrics, we additionally evaluate action-conditioning fidelity using the optical flow based metric ($\overline{\text{EPE}}$, $\overline{\cos}$) [58], which measures whether the induced motion in generated videos matches the input

Table 3: Video prediction evaluation on RoboNet [10].

Methods	FVD↓	PSNR↑	SSIM↑	LPIPS↓
MaskViT [14]	211.7	20.4	67.1	17.0
iVideoGPT [51]	197.9	23.8	80.8	14.7
SAMPO [48]	175.3	**25.3**	**84.7**	12.3
A2World-sim	**146.1**	24.1	81.6	**8.9**

Table 4: OOD simulator evaluation on LIBERO-Plus Spatial.

Methods	PSNR↑	SSIM↑	tSSIM↑	EPE↓	cos↑
DreamDojo [12]	23.79	.8571	**.7778**	.2738	.2107
A2World-sim	**25.91**	**.8719**	.7401	**.1301**	**.2761**

actions. Tab. 2 reports the results. Our A2World-sim consistently achieves the best rollout quality across both LIBERO and real-robot data, improving not only appearance metrics but also action-faithfulness. Together with the qualitative comparison (Fig. 7), these results indicate more accurate dynamics modeling under action-conditioned rollouts. We also fine-tune A2World-sim for 60k steps and evaluate on RoboNet [10], achieving strong video prediction performance against autoregressive baselines (Tab. 3).

OOD simulator evaluation To further test simulator transfer under distribution shift, we fine-tune A2World-sim on LIBERO and evaluate rollouts on LIBERO-Plus Spatial. Tab. 4 shows that action-conditioned A2World pretraining improves action-faithfulness metrics over DreamDojo [12]. The supplement further provides a qualitative OOD comparison, where DreamDojo drifts toward the training-domain appearance while A2World-sim better preserves the novel scene and interaction dynamics.

World model as real-world simulator We evaluate whether A2World-sim can serve as a real-world simulator for policy evaluation on our robot setup. Following prior protocols [12, 13, 34, 35], we compare real-world policy success rates with those from closed-loop rollouts in A2World-sim (Fig. 8). We run each policy in closed loop inside A2World-sim by autoregressively rolling out observations conditioned on the policy's action chunks. Actions are generated at 30 fps and downsampled to 10 fps to match A2World-sim. Success rates are estimated from ∼25 real rollouts and 64 simulator rollouts per policy, with outcomes verified manually. A2World-sim shows strong agreement with the real world (Spearman $\rho = 0.916$, Pearson $r = 0.965$, $R^2 = 0.930$, $N = 8$), indicating it is a faithful simulator for policy evaluation.

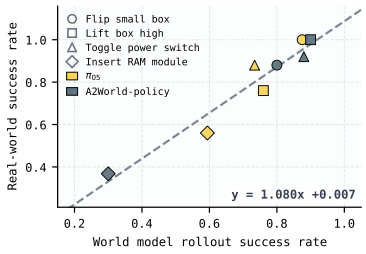

Fig. 8: Simulator consistency on real-robot tasks. Real-world success rates correlate strongly with A2World-sim rollout success rates across policies and tasks.

4.4 Vision-action joint prediction evaluation

In this section, we evaluate A2World-policy obtained by directly transferring A2World with only minimal policy-specific changes, without any separate action-only pretraining. We initialize the policy largely from the pretrained video world

Table 5: Success rate evaluation results on LIBERO [37].

Methods	Spatial	Object	Goal	Long	Average
Diffusion Policy [9]	78.3	92.5	68.3	50.5	72.4
4D-VLA [57]	88.9	95.2	90.9	79.1	88.6
Dita [17]	97.4	94.8	93.2	83.6	92.3
π_0 [5]	96.8	98.8	95.8	85.2	94.2
UniVLA [8]	96.5	96.8	95.6	92.0	95.2
$\pi_{0.5}$ [20]	**98.8**	98.2	98.0	92.4	96.9
OpenVLA-OFT [28]	97.6	98.4	97.9	94.5	97.1
CogVLA [33]	98.6	98.8	96.6	95.4	97.4
Cosmos Policy [29]	98.1	**100.0**	98.2	97.6	98.5
A2World-policy	98.2	99.2	**98.6**	**98.2**	**98.6**

Table 6: OOD policy evaluation on LIBERO-Plus Spatial.

Methods	Bg.	Cam.	Lang.	Light	Layout	Init	Noise	Avg.
Cosmos Policy [29]	89.1	75.5	94.4	98.6	94.3	53.1	96.0	85.6
FastWAM [56]	65.9	7.4	72.1	90.8	59.5	42.3	32.8	51.5
π_0 [5]	95.0	70.7	67.9	92.8	94.0	49.1	87.7	78.6
X-VLA [59]	**100.0**	24.5	82.1	**100.0**	93.8	91.4	63.5	77.7
GE-Act [36]	89.1	72.3	90.8	**100.0**	89.6	75.7	**97.7**	87.5
A2World-policy-C-init	83.3	53.2	90.5	96.6	89.9	77.1	74.1	80.2
A2World-policy-T-pre	89.9	73.1	93.3	99.3	94.0	92.9	60.7	85.8
A2World-policy-A-pre	88.0	74.2	**100.0**	99.7	**94.5**	**93.1**	80.9	88.5
A2World-policy-P-pre	89.1	**75.8**	**100.0**	**100.0**	92.2	91.4	72.9	**88.6**

model weights, fine-tune it only on downstream robot data, and observe strong performance in both simulation benchmarks and real-robot evaluations.

Evaluation on LIBERO We evaluate A2World-policy as a direct policy on LIBERO [37] under the standard 4-suite protocol. As shown in Tab. 5, A2World-policy achieves an overall success rate of 98.6%, outperforming strong baselines with the best average performance.

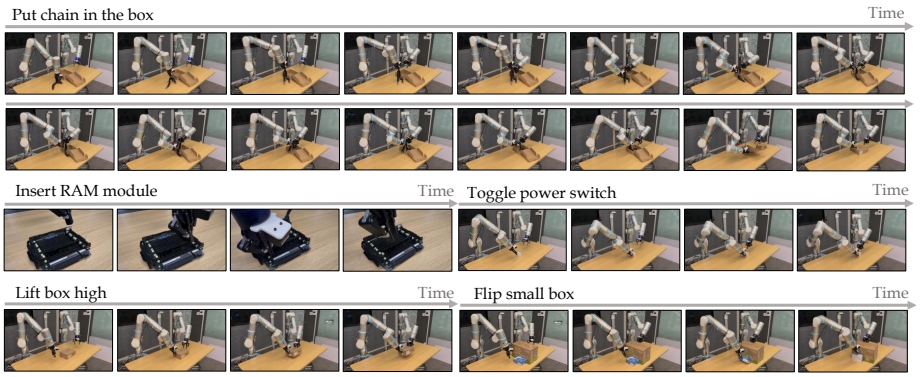

Fig. 9: Real-robot execution results of A2World-policy. We visualize executions on five real-robot tasks. A2World-policy achieves high success rates across five tasks. In contrast, the baselines [20, 31] often struggle on longer, harder tasks with complex objects (e.g., the chain task), leading to incomplete or unstable executions.

OOD policy evaluation We evaluate OOD policy transfer by fine-tuning on LIBERO and testing on LIBERO-Plus Spatial [11], which introduces diverse visual, linguistic, and dynamics shifts. We compare **C-init** (directly initializing from Cosmos-Predict2), **T-pre** (text-conditioned world-model pretraining on the same robot data), **A-pre** (our action-to-video A2World pretraining), and **P-pre** (policy-targeted text-video-action pretraining following Eq. 8). As shown in Tab. 6, A-pre reaches 88.5% average success, clearly outperforming C-init (80.2%) and T-pre (85.8%). It is also essentially on par with P-pre (88.6%), although P-pre uses a downstream-matched pretraining target. This indicates that action-to-video pretraining already captures the dynamics prior needed for policy transfer. More importantly, unlike policy-targeted pretraining, the same

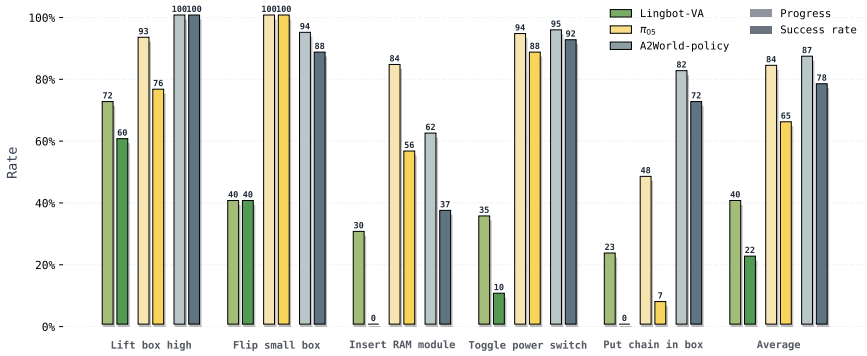

Fig. 10: Detailed real-robot evaluation results. Our A2World-policy exceeds all baselines in terms of overall task success rate.

Table 7: Results of A2World-policy variants on LIBERO.

Settings	Overall succ.
Text-conditioned Cosmos init (C-init)	97.0
Text-conditioned A2World pretrain (T-pre)	97.4
Action-video A2World pretrain (A-pre)	98.6
Policy-targeted pretrain (P-pre)	**98.8**
Video frozen; only shared self-attn trainable	86.2

Table 8: Ablation on history sampling on LIBERO.

Methods	PSNR↑	SSIM↑	tSSIM↑	EPE↓	cos↑
No history	25.41	.8806	.7663	.3969	.5778
Sliding window	25.63	.8840	.7699	.3900	.5853
Pose-guided (ours)	26.64	.8957	.7862	.3498	.6045

A-pre checkpoint is naturally reusable as a long-horizon simulator (Sec. 4.3), providing a dual-use prior for both simulator and policy adaptation.

Evaluation on real-robot We finally evaluate A2World-policy on a real-robot suite covering diverse contact-rich manipulation, including lifting or reorientation, precision insertion, switch or hinge interactions, and deformable-object handling. All scores are obtained via third-party evaluation by data-collection operators under a standardized protocol. Fig. 10 reports task progress and final success, and Fig. 9 visualizes executions (progress definition in the Appendix). A2World-policy consistently outperforms strong recent baselines, including $\pi_{0.5}$ [20] and LingBot-VA [31], with the largest gains on the most challenging long-horizon, contact-rich tasks where baselines often stall early.

4.5 Ablations and discussions

Ablation on history sampling We ablate the effect of history sampling on video generation quality by comparing our pose-guided history sampling against: **(i)** a standard sliding-window memory strategy that keeps the most recent frames; **(ii)** no history injection. As reported in Tab. 8, pose-guided sampling consistently yields better rollout quality, indicating that selecting motion-informative histories is important for stable and history-faithful generation.

Pretraining variants Tab. 7 compares policy variants after LIBERO fine-tuning. Policy-targeted pretraining gives the best in-domain average (98.8%), while our action-video A2World pretraining is very close (98.6%). Compared with C-init and T-pre, A-pre benefits from a less ambiguous action-conditioned objective: given the current observation and future actions, the future visual

transition is largely determined, whereas text instructions can correspond to many valid action sequences. This stronger action-to-dynamics prior explains why A-pre transfers better to policy learning than generic video initialization or text-conditioned pretraining. Together with the OOD results in Tab. 6, this suggests that a policy-specific pretraining objective mainly provides a small endpoint-specific gain, whereas action-to-video pretraining offers nearly the same policy transfer while also serving as the simulator initialization.

Coupling between video modeling and action learning We observe a consistent positive coupling between video prediction and action generation during training. In Fig. 11, each point is a validation checkpoint, where video consistency on future prediction and a normalized action quality score are plotted on the x-axis and y-axis, respectively (definitions in the Appendix). Better video consistency co-occurs with better action quality, and full joint training reaches a stronger upper-right frontier than freezing the video branch. Here, the video-frozen variant (86.2% in Tab. 7) freezes video-specific prediction modules, while keeping the action branch and shared transformer layers trainable. Overall, the shared representations learned for forecasting and control reinforce each other, improving both video and action.

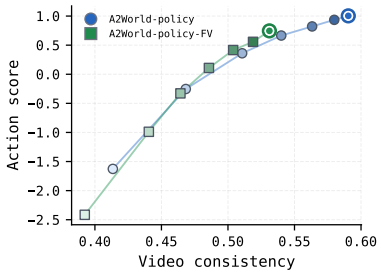

Fig. 11: Video–action coupling during A2World-policy training. Improving video prediction consistently correlates with better action generation.

5 Conclusions

We argue that action-conditioned world modeling is a scalable approach for learning transferable dynamics priors. By leveraging actions as causal supervision, large-scale action-to-video pretraining distills reusable interaction knowledge that generalizes across tasks and environments. We instantiate this idea with A2World, a multi-view diffusion world model pretrained on diverse real-robot data, and show that its learned prior can be adapted to both A2World-sim for long-horizon rollout and policy evaluation, and A2World-policy for instruction-conditioned control. Across simulation benchmarks and real-robot experiments, action-to-video pretraining consistently yields a stronger transferable prior than text-conditioned or task-specific robot pretraining.

Acknowledgements

This work was supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123004), Ningbo grant (2025Z038), and National Natural Science Foundation of China (Grant No. 62376060).

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint (2025) 2, 8, 11
2. Ali, A., Bai, J., Bala, M., Balaji, Y., Blakeman, A., Cai, T., Cao, J., Cao, T., Cha, E., Chao, Y.W., et al.: World simulation with video foundation models for physical ai. arXiv preprint (2025) 4
3. Barreiros, J., Beaulieu, A., Bhat, A., Cory, R., Cousineau, E., Dai, H., Fang, C.H., Hashimoto, K., Irshad, M.Z., Itkina, M., et al.: A careful examination of large behavior models for multitask dexterous manipulation. arXiv preprint (2025) 9
4. Bi, H., Tan, H., Xie, S., Wang, Z., Huang, S., Liu, H., Zhao, R., Feng, Y., Xiang, C., Rong, Y., et al.: Motus: A unified latent action world model. arXiv preprint (2025) 2, 3
5. Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: pi0: A vision-language-action flow model for general robot control. arXiv preprint (2024) 13
6. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint (2023) 2
7. Bu, Q., Cai, J., Chen, L., Cui, X., Ding, Y., Feng, S., Gao, S., He, X., Hu, X., Huang, X., et al.: Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. arXiv preprint (2025) 2, 6, 8
8. Bu, Q., Yang, Y., Cai, J., Gao, S., Ren, G., Yao, M., Luo, P., Li, H.: Univla: Learning to act anywhere with task-centric latent actions. arXiv preprint (2025) 13
9. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. IJRR (2025) 13
10. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C.: Robonet: Large-scale multi-robot learning. arXiv preprint (2019) 6, 12
11. Fei, S., Wang, S., Shi, J., Dai, Z., Cai, J., Qian, P., Ji, L., He, X., Zhang, S., Fei, Z., et al.: Libero-plus: In-depth robustness analysis of vision-language-action models. arXiv preprint (2025) 6, 13
12. Gao, S., Liang, W., Zheng, K., Malik, A., Ye, S., Yu, S., Tseng, W.C., Dong, Y., Mo, K., Lin, C.H., et al.: Dreamdojo: A generalist robot world model from large-scale human videos. arXiv preprint (2026) 2, 3, 12
13. Guo, Y., Shi, L.X., Chen, J., Finn, C.: Ctrl-world: A controllable generative world model for robot manipulation. ICLR (2026) 2, 3, 11, 12
14. Gupta, A., Tian, S., Zhang, Y., Wu, J., Martín-Martín, R., Fei-Fei, L.: Maskvit: Masked visual pre-training for video prediction. arXiv preprint (2022) 12
15. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint (2024) 2
16. He, H., Bai, C., Pan, L., Zhang, W., Zhao, B., Li, X.: Learning an actionable discrete diffusion policy via large-scale actionless video pre-training. NeurIPS (2024) 2
17. Hou, Z., Zhang, T., Xiong, Y., Duan, H., Pu, H., Tong, R., Zhao, C., Zhu, X., Qiao, Y., Dai, J., et al.: Dita: Scaling diffusion transformer for generalist vision-language-action policy. In: ICCV (2025) 13

18. Hu, Y., Guo, Y., Wang, P., Chen, X., Wang, Y.J., Zhang, J., Sreenath, K., Lu, C., Chen, J.: Video prediction policy: A generalist robot policy with predictive visual representations. arXiv preprint arXiv:2412.14803 (2024) 2, 4
19. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint (2025) 7
20. Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al.: pi0.5: a vision-language-action model with open-world generalization. arXiv preprint (2025) 13, 14
21. InternData-M1 Contributors: Interndata-m1 (2025), https://github.com/InternRobotics/InternManip, accessed 27 Jun 2026 2, 6, 8
22. Jang, J., Ye, S., Lin, Z., Xiang, J., Bjorck, J., Fang, Y., Hu, F., Huang, S., Kundalia, K., Lin, Y.C., et al.: Dreamgen: Unlocking generalization in robot learning through video world models. arXiv preprint (2025) 2, 3
23. Jiang, T., Yuan, T., Liu, Y., Lu, C., Cui, J., Liu, X., Cheng, S., Gao, J., Xu, H., Zhao, H.: Galaxea open-world dataset and g0 dual-system vla model. arXiv preprint (2025) 2, 6, 8
24. Jiang, Y., Chen, S., Huang, S., Chen, L., Zhou, P., Liao, Y., He, X., Liu, C., Li, H., Yao, M., et al.: Enerverse-ac: Envisioning embodied environments with action condition. arXiv preprint arXiv:2505.09723 (2025) 2, 3
25. Jiang, Z., Liu, K., Qin, Y., Tian, S., Zheng, Y., Zhou, M., Yu, C., Li, H., Zhao, D.: World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. arXiv preprint (2025) 2, 3
26. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. NeurIPS (2022) 4
27. Khazatsky, A., Pertsch, K., Nair, S., Balakrishna, A., Dasari, S., Karamcheti, S., Nasiriany, S., Srirama, M.K., Chen, L.Y., Ellis, K., et al.: Droid: A large-scale in-the-wild robot manipulation dataset. arXiv preprint (2024) 6, 8, 9, 10
28. Kim, M.J., Finn, C., Liang, P.: Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint (2025) 13
29. Kim, M.J., Gao, Y., Lin, T.Y., Lin, Y.C., Ge, Y., Lam, G., Liang, P., Song, S., Liu, M.Y., Finn, C., et al.: Cosmos policy: Fine-tuning video models for visuomotor control and planning. arXiv preprint (2026) 2, 4, 13
30. Li, H., Ding, P., Suo, R., Wang, Y., Ge, Z., Zang, D., Yu, K., Sun, M., Zhang, H., Wang, D., et al.: Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators. arXiv preprint (2025) 2, 3
31. Li, L., Zhang, Q., Luo, Y., Yang, S., Wang, R., Han, F., Yu, M., Gao, Z., Xue, N., Zhu, X., Shen, Y., Xu, Y.: Causal world modeling for robot control. arXiv preprint (2026) 2, 4, 13, 14
32. Li, S., Gao, Y., Sadigh, D., Song, S.: Unified video action model. arXiv preprint (2025) 2, 4
33. Li, W., Zhang, R., Shao, R., He, J., Nie, L.: Cogvla: Cognition-aligned vision-language-action model via instruction-driven routing & sparsification. arXiv preprint (2025) 13
34. Li, X., Hsu, K., Gu, J., Pertsch, K., Mees, O., Walke, H.R., Fu, C., Lunawat, I., Sieh, I., Kirmani, S., et al.: Evaluating real-world robot manipulation policies in simulation. arXiv preprint (2024) 12
35. Li, Y., Zhu, Y., Wen, J., Shen, C., Xu, Y.: Worldeval: World model as real-world robot policies evaluator. arXiv preprint (2025) 12
36. Liao, Y., Zhou, P., Huang, S., Yang, D., Chen, S., Jiang, Y., Hu, Y., Cai, J., Liu, S., Luo, J., et al.: Genie envisioner: A unified world foundation platform for robotic manipulation. arXiv preprint (2025) 2, 3, 4, 13

37. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. NeurIPS (2023) 4, 6, 11, 13
38. Luo, Y., Du, Y.: Grounding video models to actions through goal conditioned exploration. arXiv preprint (2024) 2
39. Mees, O., Hermann, L., Rosete-Beas, E., Burgard, W.: Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. RA-L (2022) 4
40. O'Neill, A., Rehman, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., et al.: Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In: ICRA (2024) 6, 8
41. Pai, J., Achenbach, L., Montesinos, V., Forrai, B., Mees, O., Nava, E.: mimic-video: Video-action models for generalizable robot control beyond vlas. arXiv preprint (2025) 2, 4
42. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023) 4
43. Quevedo, J., Sharma, A.K., Sun, Y., Suryavanshi, V., Liang, P., Yang, S.: Worldgym: World model as an environment for policy evaluation. arXiv preprint (2025) 3
44. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research (2020) 7
45. Team, G.R., Choromanski, K., Devin, C., Du, Y., Dwibedi, D., Gao, R., Jindal, A., Kipf, T., Kirmani, S., Leal, I., et al.: Evaluating gemini robotics policies in a veo world simulator. arXiv preprint arXiv:2512.10675 (2025) 2, 3
46. Tian, Y., Yang, Y., Xie, Y., Cai, Z., Shi, X., Gao, N., Liu, H., Jiang, X., Qiu, Z., Yuan, F., et al.: Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. arXiv preprint (2025) 2, 6, 8
47. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint (2025) 2, 4
48. Wang, S., Tian, J., Wang, L., Liao, Z., Li, J., Dong, H., Xia, K., Zhou, S., Tang, W., Gang, H.: Sampo: Scale-wise autoregression with motion prompt for generative world models. arXiv preprint (2025) 12
49. Wen, Y., Lin, J., Zhu, Y., Han, J., Xu, H., Zhao, S., Liang, X.: Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. NeurIPS (2024) 2
50. Wu, H., Jing, Y., Cheang, C., Chen, G., Xu, J., Li, X., Liu, M., Li, H., Kong, T.: Unleashing large-scale video generative pre-training for visual robot manipulation. arXiv preprint (2023) 2
51. Wu, J., Yin, S., Feng, N., He, X., Li, D., Hao, J., Long, M.: ivideogpt: Interactive videogpts are scalable world models. In: NeurIPS (2024) 3, 12
52. Wu, K., Hou, C., Liu, J., Che, Z., Ju, X., Yang, Z., Li, M., Zhao, Y., Xu, Z., Yang, G., et al.: Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. arXiv preprint (2024) 9, 10
53. Wu, S., Liu, X., Xie, S., Wang, P., Li, X., Yang, B., Li, Z., Zhu, K., Wu, H., Liu, Y., et al.: Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. arXiv preprint (2025) 2, 6, 8, 9, 10

54. Xiao, J., Yang, Y., Chang, X., Chen, R., Xiong, F., Xu, M., Zheng, W.S., Zhang, Q.: World-env: Leveraging world model as a virtual environment for vla post-training. arXiv preprint (2025) 3
55. Ye, S., Ge, Y., Zheng, K., Gao, S., Yu, S., Kurian, G., Indupuru, S., Tan, Y.L., Zhu, C., Xiang, J., Malik, A., Lee, K., Liang, W., Ranawaka, N., Gu, J., Xu, Y., Wang, G., Hu, F., Narayan, A., Bjorck, J., Wang, J., Kim, G., Niu, D., Zheng, R., Xie, Y., Wu, J., Wang, Q., Julian, R., Xu, D., Du, Y., Chebotar, Y., Reed, S., Kautz, J., Zhu, Y., Fan, L.J., Jang, J.: World action models are zero-shot policies. arXiv preprint (2026) 2, 3, 4
56. Yuan, T., Dong, Z., Liu, Y., Zhao, H.: Fast-wam: Do world action models need test-time future imagination? arXiv preprint (2026) 4, 13
57. Zhang, J., Chen, Y., Xu, Y., Huang, Z., Zhou, Y., Yuan, Y.J., Cai, X., Huang, G., Quan, X., Xu, H., et al.: 4d-vla: Spatiotemporal vision-language-action pretraining with cross-scene calibration. NeurIPS (2025) 13
58. Zhang, J., Huang, Z., Gu, C., Ma, Z., Zhang, L.: Reinforcing action policies by prophesying. arXiv preprint (2025) 2, 3, 4, 11
59. Zheng, J., Li, J., Wang, Z., Liu, D., Kang, X., Feng, Y., Zheng, Y., Zou, J., Chen, Y., Zeng, J., et al.: X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. arXiv preprint (2025) 13
60. Zhu, C., Yu, R., Feng, S., Burchfiel, B., Shah, P., Gupta, A.: Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. arXiv preprint (2025) 2
61. Zhu, F., Wu, H., Guo, S., Liu, Y., Cheang, C., Kong, T.: Irasim: A fine-grained world model for robot manipulation. In: ICCV (2025) 2, 3
62. Zhu, F., Yan, Z., Hong, Z., Shou, Q., Ma, X., Guo, S.: Wmpo: World model-based policy optimization for vision-language-action models. arXiv preprint (2025) 2, 3, 4
63. Zhu, Y., Joshi, A., Stone, P., Zhu, Y.: Viola: Imitation learning for vision-based manipulation with object proposal priors. In: CoRL (2023) 9, 10