REAL WHERE IT MATTERS: SIM-AND-REAL CO-TRAINING FOR ROBOT WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Action-conditioned world models simulate future dynamics conditioned on robot actions, but their generalization is constrained by the limited coverage of real-world robot datasets. Existing datasets are dominated by successful human-teleoperated trajectories and shoulder-mounted exocentric viewpoints, leaving off-nominal interactions and broader camera viewpoints sparsely represented. Consequently, models trained on these data struggle to predict outcomes of non-optimal actions and generalize across camera viewpoints in the real world. In contrast, physics simulators can generate off-nominal interactions, such as retries and inter-object contacts, across diverse camera viewpoints at scale. In this paper, we investigate whether large-scale simulation data can complement finite real-world robot data for training action-conditioned world models for robotic manipulation. We jointly train a world model, dubbed **MixWorld**, on real ($\approx$ 95k episodes) and simulated ($\approx$ 550k episodes) trajectories, where real data ground predictions in the deployment domain while simulation provides substantially broader coverage of transition dynamics and camera viewpoints. Our results show that sim-and-real co-training substantially improves video prediction quality across computational metrics, model-based evaluations, and the comprehensive WorldArena benchmark. More importantly, these improvements translate into consistent gains in downstream applications, including policy evaluation and inference-time search. For policy evaluation, our world model improves Pearson correlation (r) by 109.3% and Spearman correlation (ρ) by 142.4%, while reducing Mean Maximum Rank Violation (MMRV) by 18.2% relative to the base world model. For *Best-of-N* inference-time search, our world model increases average success rates by 10.0%–33.3% over the base policy and 16.7%–30.0% over search with the base world model across the two tasks.

1 INTRODUCTION

Action-conditioned world models (Finn et al., 2016; Finn & Levine, 2017; Ha & Schmidhuber, 2018; Ebert et al., 2018) are learned simulators that predict future observations given robot action sequences. Modern high-fidelity world models offer a promising interface for scalable policy evaluation (Team et al., 2025; Tseng et al., 2026; Quevedo et al., 2026), inference-time search (Zhu et al., 2025; Assran et al., 2025; Qi et al., 2026), and synthetic data generation for policy learning (Yang et al., 2024b; Chandra et al., 2025; Jiang et al., 2026). Realizing these applications requires world models to faithfully predict how the physical world evolves under diverse robot actions, including off-nominal behaviors such as missed grasps, inter-object contacts, and retries. Achieving such accurate and controllable predictions requires diverse interaction training data that cover the range of state-action transitions encountered at deployment; otherwise, predictions may bias toward likely outcomes in the training distribution rather than faithfully reflecting the specified actions. This poses a **fundamental data challenge** since existing large-scale real-world robot datasets (Khazatsky et al., 2024; O'Neill et al., 2024) predominantly consist of successful task executions. Therefore, learning controllable world models demands broader state-action coverage of interaction dynamics than existing real-world robot datasets provide.

Action-conditioned world models (Finn et al., 2016; Finn & Levine, 2017; Ha & Schmidhuber, 2018; Ebert et al., 2018) are learned simulators that predict future observations given robot action sequences. Modern high-fidelity world models offer a promising interface for scalable policy evalu-

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.

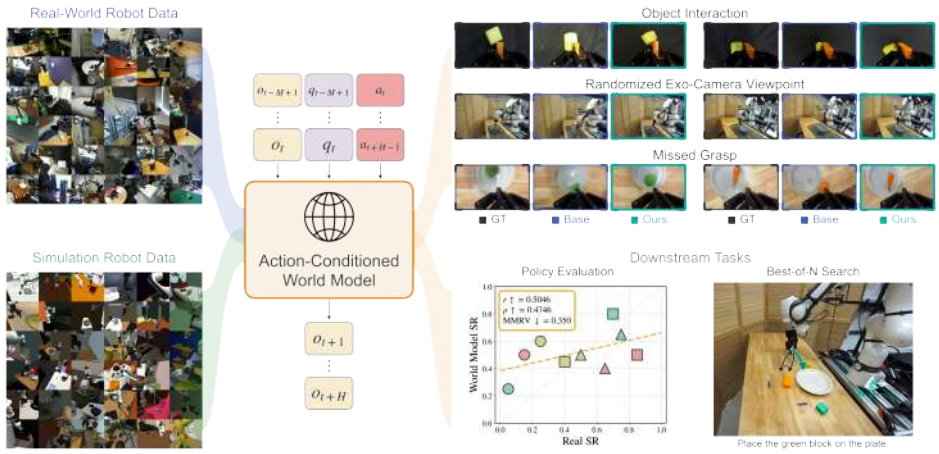


Figure 1: MixWorld is trained on a mixture of sim-and-real robot data and achieves improved performance on off-nominal transition dynamics prediction as demonstrated in the figure. MixWorld also outperforms the baseline Ctrl-World in downstream applications.

ation (Team et al., 2025; Tseng et al., 2026; Quevedo et al., 2026), inference-time search (Zhu et al., 2025; Assran et al., 2025; Qi et al., 2026), and synthetic data generation for policy learning (Yang et al., 2024b; Chandra et al., 2025; Jiang et al., 2026). These downstream tasks require world models to be controllable, i.e., the predictions must faithfully reflect the consequences of specified actions, including suboptimal, counterfactual, and failure-inducing behaviors. However, existing large-scale real-world robot data (Khazatsky et al., 2024; O'Neill et al., 2024) predominantly consist of successful task executions and therefore cover a narrow subset of state transitions. Models trained on such data receive limited supervision for off-nominal state transitions such as missed grasps, inter-object contacts, retries, and other outcomes induced by non-expert actions, and may bias predictions toward likely outcomes in the demonstration distribution. Therefore, learning controllable world models is faced with a fundamental data challenge and demands broader state-action coverage than real-world robot datasets provide (Bateman et al., 2026; Hansen & Wang, 2026).

Recent works address this problem by autonomously executing robot policies to collect interactions beyond successful human demonstrations (Yin et al., 2026; Guo et al., 2026b). While effective at diversifying experience, scaling data collection on physical robots remains intrinsically expensive: every trajectory consumes robot time and hardware resources, and broad exploration must contend with failure recovery. Simulation offers a complementary source of interaction experience that can be scaled without these constraints. Large-scale manipulation datasets can be efficiently generated in procedurally constructed environments (Raistrick et al., 2024; Kim et al., 2026), providing broad coverage of interaction dynamics across diverse actions, objects, and environments. Such simulation robot datasets have enabled robot manipulation policies that transfer zero-shot to the real world (Deng et al., 2025; Deshpande et al., 2026; Tian et al., 2026). Yet, whether interaction dynamics learned from large-scale simulation can transfer to improve real-world action-conditioned world models for robot manipulation remains largely unexplored.

In this work, we show that simulation can complement real-world robot data with off-nominal interaction dynamics that are sparsely represented in real-robot datasets, including missed grasps, retries, and inter-object contacts, while providing broader coverage of exocentric camera viewpoints. Our **key insight** is that simulation data provides scalable supervision for how diverse actions lead to diverse outcomes, while real-world data anchors the model to the visual appearance and real-world dynamics encountered at deployment. Together, the two complementary sources enable world models to learn from broader transition dynamics coverage while remaining grounded in the real world. Evidence for this principle has recently emerged in autonomous driving (Yang et al., 2025), where supplementing safe real-world vehicle trajectories with simulated non-expert and dangerous behaviors improves action controllability in real driving scenes. We investigate whether the same principle extends to world models for robot manipulation, a substantially more contact-rich setting.

Our contributions are summarized as below. We introduce MixWorld, an action-conditioned world model jointly trained on 95k real-world robot trajectories and 550k simulated robot trajecto-

ries that spans over 60k environments and 7k assets. To distill transition dynamics from simulation data while not deviating from the real-world visual appearance and dynamics, we propose two simple sim-and-real co-training schemes: naïve mixture and Low-Rank Adaptation (LoRA) finetuning. After evaluating on 3 datasets with PSNR, SSIM, LPIPS, FID, FVD, and the WorldArena benchmark (Shang et al., 2026), we find that MixWorld matches or outperforms baseline across all metrics. In downstream applications, MixWorld improves policy-evaluation correlations by 109.3% (r) and 142.4% (ρ), while reducing MMRV by 18.2% relative to Ctrl-World. For *Best-of-N* inference-time search, MixWorld improves task success by 10.0%–33.3% over the base policy and 16.7%–30.0% over search with Ctrl-World.

2 Related Works

See App. A.1 for a comprehensive related work section.

World Models for Robotics The development of generative modeling (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Rombach et al., 2022; Blattmann et al., 2023) has enabled the advance of high-fidelity action-conditioned world models (Guo et al., 2026c; Yang et al., 2024b; Zhu et al., 2025; NVIDIA et al., 2025; Hafner et al., 2025; Wang et al., 2026; Huang et al., 2026; Guo et al., 2026a; Ye et al., 2026b;a) that can be applied for diverse downstream applications, such as policy evaluation (Team et al., 2025; Zhu et al., 2025; Guo et al., 2026c; Tseng et al., 2026; Wang et al., 2026; Quevedo et al., 2026; Li et al., 2026; Yin et al., 2026; Jain et al., 2026; Lu et al., 2026), inference-time planning (Zhu et al., 2025; Assran et al., 2025; Qi et al., 2026; Jain et al., 2026), and synthetic data generation for policy improvement (Yang et al., 2024b; Chandra et al., 2025; Guo et al., 2026c; Yin et al., 2026; Jain et al., 2026; Guo et al., 2026b; Jiang et al., 2026; Zhu et al., 2026). However, world models (Guo et al., 2026c; Jain et al., 2026; Quevedo et al., 2026) that are trained on large-scale offline robot data (Khazatsky et al., 2024; Walke et al., 2023) show limited capabilities when predicting future frames of non-optimal or non-successful action trajectories, because the datasets bias to successful demonstrations by human teleoperators and rarely include counterfactual dynamics (Pitis et al., 2020). Although techniques such as diffusion forcing (Chen et al., 2024) and rotary position embedding (Su et al., 2024; Wei et al., 2025b) have been applied to improve the fidelity of world models (Jain et al., 2026), they do not address the fundamental problem of narrow data distribution. To broaden the data coverage of interaction dynamics, especially failure cases such as missed grasps and slips, recent works (Yin et al., 2026; Guo et al., 2026b) execute vision-language-action (VLA) policies (Black et al., 2025a) to iteratively collect online interactions between robots and the real world. Although they can obtain high-quality real-world counterfactual data to finetune the world model, collecting data on real robots is intrinsically expensive and time-consuming considering the robot time, hardware resources, and unpredictable failure modes that require human intervention. In contrast, we broaden transition dynamics by sourcing large-scale simulation data that contain counterfactual cases as well as a large variety of camera viewpoints surrounding the robot workspace.

Simulation for World Models Similar to the challenges faced by action-conditioned world models for robotic manipulation, world models for autonomous driving are constrained by the coverage of real-world driving data (Caesar et al., 2020; Sun et al., 2020). While large-scale real-world driving datasets have enabled increasingly capable generative world models (Hu et al., 2023; Wang et al., 2025; Russell et al., 2025; Gao et al., 2025), they inevitably underrepresent rare and safety-critical scenarios. ReSim (Yang et al., 2025) enriches real-world driving data (Dauner et al., 2024; Yang et al., 2024a) with non-expert data from driving simulator, i.e., CARLA (Dosovitskiy et al., 2017), and proposes multi-stage sim-and-real co-training to improve models' controllability. ReSim shares the most related philosophical idea to this paper, but we focus on co-training action-conditioned world models for robotic manipulation where there are richer and more complex contact dynamics and broader distribution of camera viewpoints. The most related work on training action-conditioned world model for robotic manipulation with simulation data is SimDist (Levy et al., 2026), which pre-trains world models on online interaction in simulation and perturbs expert action sequences to gain counterfactual experience. But their models are task-oriented and trained on setup-specific data. In contrast, our models are co-trained on large-scale real-world general robot data (Khazatsky et al., 2024) and simulation data with a large variety of objects and scenes (Deshpande et al., 2026).

3 PROBLEM FORMULATION

We study whether robot interactions from physics-based simulators can improve an action-conditioned world model's dynamics prediction in the real world. We consider two robot trajectory datasets $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{sim}}$, collected in real and simulated environments, respectively. At time t, the multi-view visual observation is $o_t = (I_t^{(1)}, \ldots, I_t^{(K)})$, where $I_t^{(k)}$ is the RGB image from camera k, and q_t denotes the robot's proprioceptive state. The temporally ordered observation–proprioception history is $\mathcal{H}_t = ((o_{\tau_1}, q_{\tau_1}), \ldots, (o_t, q_t))$, where $\tau_1 < \cdots < \tau_M = t$ are the selected frame indices. The future action sequence $a_{t+1:t+H}$ uses the same proprioceptive representation as q_t, with each a_{t+k} specifying the robot state used to condition prediction at frame $t+k$. Here, H is the prediction horizon of future frames. We learn the conditional distribution

$$p_{\boldsymbol{\theta}}(o_{t+1:t+H} \mid \mathcal{H}_t, a_{t:t+H-1}, \ell)\,, \tag{1}$$

where $o_{t+1:t+H} = (o_{t+1}, \ldots, o_{t+H})$ denotes the sequence of future multi-view observations, ℓ is an optional language instruction, and θ denotes the model parameters. The model jointly predicts future camera observations conditioned on the robot trajectory.

Each dataset yields training samples containing a context, a future action sequence, and the corresponding future visual observations. Let $\mathcal{L}(\theta; \mathcal{D})$ denote the video diffusion loss over training samples from dataset $\mathcal{D}$. We optimize

$$\mathcal{L}_{\text{total}}\left(\theta; \mathcal{D}_{\text{real}}, \mathcal{D}_{\text{sim}}\right) = \alpha\, \mathcal{L}\left(\theta; \mathcal{D}_{\text{sim}}\right) + (1-\alpha)\, \mathcal{L}\left(\theta; \mathcal{D}_{\text{real}}\right), \tag{2}$$

where $\alpha \in [0, 1]$ is the mixing coefficient that determines how training examples are allocated between the two data sources. Each training sample is drawn from $\mathcal{D}_{\text{sim}}$ with probability α or from $\mathcal{D}_{\text{real}}$ with probability $1-\alpha$ (Maddukuri et al., 2025). The average loss over these samples estimates the objective above.

Our evaluation target is the prediction quality of future frames in real environments, especially settings whose scene setup or interaction distributions differ from $\mathcal{D}_{\text{real}}$. Our *key assumption*, validated by prior work (Deshpande et al., 2026), is that simulation captures dynamics transferable to real-world interactions, allowing its broader coverage of state transitions to benefit real-world prediction despite the visual domain gap.

4 SIM-AND-REAL CO-TRAINING FOR ACTION-CONDITIONED MANIPULATION WORLD MODELS

We discuss the heterogeneous dataset composed of real-world robot data and simulation data, the action-conditioned world model architecture, and the sim-and-real co-training recipe.

4.1 HETEROGENEOUS DATASET

Real-World Data. We source our real-world robot data from the DROID dataset (Khazatsky et al., 2024), which contains human-teleoperated demonstrations collected using a standardized robot setup in everyday environments. It contains 95k trajectories including 76k successful episodes and 19k unsuccessful episodes, totaling 350 hours and spanning 564 scenes across 52 buildings. Each episode provides three synchronized camera views, including two external cameras and a wrist camera, alongside robot proprioception and language annotations. These varied collection settings capture diverse real-world appearances and physical interaction dynamics. Before sim-and-real co-training, we pretrain the action-conditioned world model on the DROID dataset.

Simulation Data. For large-scale diverse simulation data, we turn to MolmoBot-Data (Deshpande et al., 2026), specifically the Franka Pick-and-Place subset which targets the DROID platform. MolmoBot-Data is generated via large-scale procedural generation in MolmoSpaces (Kim et al., 2026). MolmoBot-Engine generates this data by sampling from a large set of multi-room environments, sampling manipulation tasks with those environments, and rolling out a scripted demonstrator while rejection filtering to ensure specific criteria are met. By implementing sophisticated action disturbances and aggressive randomization as well as recovery and retry behavior, MolmoBot-Engine

produces diverse interaction behavior ideal for large-scale training. In total, we use 2,638 hours of simulation data, spanning 550k episodes across over 60k environments and 7k assets.

We visualize the DROID and MolmoBot-Data datasets in Figure 4 in App. A.2 to qualitatively compare their exocentric camera poses distribution and demonstrate the counterfactual dynamics covered in simulation data but underrepresented in real-robot data.

4.2 Action-Conditioned World Model

World Model Architecture. MixWorld follows Ctrl-World (Guo et al., 2026c) and uses a Stable Video Diffusion (SVD) backbone (Blattmann et al., 2023) to jointly predict future multi-view observations conditioned on visual and proprioceptive history and robot actions, as shown in Fig. 2. A frozen variational autoencoder encodes each camera image into latents, which is concatenated across the three views along the height dimension. We denote the clean latents of the current anchor frame and future prediction frames by Z. During training, these latents are corrupted as $Z_\sigma = Z + \sigma E$, where $\sigma > 0$ is the noise level and E contains independent standard Gaussian noise. A spatiotemporal U-Net, denoted by F_θ, processes the noisy latents together with historical visual and proprioceptive features. Spatial attention captures dependencies across views, while temporal attention exchanges information across historical and prediction frames. The current observation also provides an image anchor, whose latent representation is repeated across the prediction window and concatenated with the noisy latents along the channel dimension. We instantiate proprioception using the end-effector Cartesian position, Euler-angle orientation, and gripper state. When enabled, a frozen CLIP text encoder (Radford et al., 2021) provides an instruction embedding that is expanded and added to each frame's conditioning embedding. We collectively denote these visual, proprioceptive, and language conditions by c. We train the model following an EDM-style (Karras et al., 2022) noise-weighted latent reconstruction objective,

$$\mathcal{L}_{\text{diff}}(\theta; \mathcal{D}_{\text{sim}}, \mathcal{D}_{\text{real}}) = \mathbb{E}_{\alpha,\sigma,E}\left[\frac{1+\sigma^2}{\sigma^2}\left\|\hat{Z}_\theta(Z_\sigma, \sigma, c) - Z\right\|_2^2\right], \tag{3}$$

Under the standard noise schedule, we sample $\log\sigma \sim \mathcal{N}(0.7, 1.6^2)$. This loss is applied to both real and simulated data through the co-training objective defined above.

Latent Dynamics Consistency Loss. The diffusion objective supervises reconstruction of individual video latents but does not explicitly constrain how the predicted latent state evolves over time. Inspired by latent overshooting that regularizes latent dynamics by encouraging consistency across multiple prediction horizons (Hafner et al., 2019), and temporal consistency objectives for driving world models (Yang et al., 2025), we additionally supervise latent changes over multiple temporal intervals. Specifically, we match the temporal differences between predicted and ground-truth latents:

$$\mathcal{L}_{\text{dyn}} = \mathbb{E}_\alpha\left[\frac{1}{J}\sum_{j=1}^{J}\frac{1}{H+1-j}\sum_{i=0}^{H-j}\left\|(\hat{Z}_{i+j} - \hat{Z}_i) - (Z_{i+j} - Z_i)\right\|_2^2\right], \tag{4}$$

where Z_i and $\hat{Z}_i$ denote the ground-truth and reconstructed multi-view latents at frame i, respectively. The index $i = 0$ corresponds to the current anchor frame, and $i = 1, \ldots, H$ to future frames. The maximum interval $J \leq H$ controls the temporal range of the loss. We omit the denoiser's parameters and conditioning arguments for readability. The expectation follows the same sim–real sampling and noise process as the denoising objective. The resulting training objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}}, \tag{5}$$

where $\mathcal{L}_{\text{diff}}$ denotes the mixed-domain noise-weighted denoising loss and λ_{dyn} controls the strength of temporal supervision.

4.3 Sim-And-Real Co-Training

Our primary goal is to determine whether co-training on simulation and real-world robot data improves the prediction quality of action-conditioned world models. A key design choice remaining as a question is how to incorporate simulation data into training. Our sim-and-real co-train starts from a checkpoint pretrained on DROID dataset only, and we consider two strategies, described below.

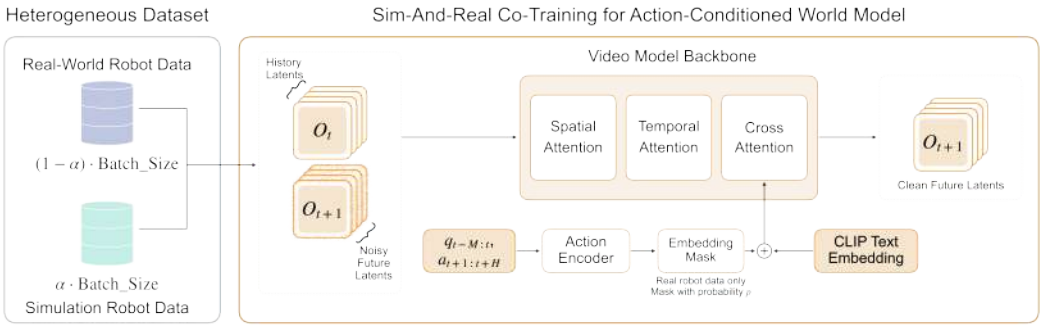


Figure 2: Architecture of action-conditioned world model.

Naïve Mixture. Prior work in policy learning has demonstrated that co-training robot policies on simulation and real-world data using simple mixture can improve policy performance without architecture adaptation for multiple domains (Maddukuri et al., 2025; Wei et al., 2025a). We adopt this approach but for the training of world models. Specifically, the mixing coefficient $\alpha \in [0, 1]$ controls the expected proportion of simulation samples in each minibatch as formulated in Sec. 3, while the realized proportion varies with sampling. Both domains use the same noise-level sampling distribution and noise-weighted denoising objective.

Low-Rank Adaptation of Denoiser. Since MixWorld is first pre-trained on DROID dataset, which is the target domain for deployment, we consider Low-Rank Adaptation (LoRA) (Hu et al., 2022) for sim-and-real co-training to prevent the model capacity from deviating wildly towards the simulation domain. Let D_{θ_0} denote the pretrained denoiser, the spatiotemporal UNet whose transformer layers each contain spatial and temporal self-attention together with spatial and temporal cross-attention over the action-conditioning tokens. We write $\mathcal{W}_{\text{attn}} \subset \theta_0$ for the set of query, key, value, and output projections of all attention operators. We freeze θ_0 and adapt each projection $W \in \mathcal{W}_{\text{attn}}$ by a rank-r residual,

$$\tilde{W} = W + \frac{\alpha}{r} BA, \qquad A \in \mathbb{R}^{r \times d_{\text{in}}},\ B \in \mathbb{R}^{d_{\text{out}} \times r}, \tag{6}$$

with B initialized to zero so that the adapted model coincides with D_{θ_0} at the start of training. The trainable parameters are the low-rank factors together with the action encoder, $\phi = \{(A_W, B_W)\}_{W \in \mathcal{W}_{\text{attn}}} \cup \psi$. The objective is unchanged from full fine-tuning and the overall loss is minimized over ϕ alone.

Our hypothesis is that the information transferable from simulation to the real domain factorizes along this partition: scene dynamics—how spatial positions within a frame, frames across time, and per-frame action tokens are mixed—reside in the attention projections and are shared across domains, whereas appearance resides predominantly in the convolutional layers, is domain-specific, and is already well specified for the real domain by θ_0. Therefore, the constrained update allows the model to learn the action-conditioned dynamics present in simulation while structurally preventing simulation appearance from contaminating real-domain generation.

We additionally apply domain-asymmetric action dropout. The action embedding is masked with probability 0.5 for real trajectories but never for simulation data, so that the unconditional video prior is estimated from real data alone while simulation contributes only to the action-conditioned term, restricting its influence to dynamics rather than appearance.

5 Experiments

We evaluate MixWorld's and Ctrl-World's prediction quality as well as efficacy on downstream tasks. We design our experiments to answer the following research questions: **Q1:** Can sim-and-real co-training improve world models' video prediction quality on different metrics? **Q2:** How does sim-and-real co-training affect models' ability for policy evaluation? **Q3:** Does inference-time search with our world model improve a base policy?

Table 1: World model prediction quality comparison across different out-of-distribution (OOD) datasets. Results are evaluated on exocentric and wrist views. Best scores are shown in **bold** and second-best results are underlined. WEAVER shows wrist view only because it predicts an exocentric view and a wrist camera view.

View	Model	Eval Set 1 (Guo et al., 2026c)					Eval Set 2 (Jain et al., 2026)					Our Eval Set				
		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	FVD ↓
Exo	Ctrl-World	23.75	0.845	0.081	22.94	203.6	21.37	0.828	0.097	26.77	234.6	19.77	0.734	0.124	38.49	600.9
	Naïve Mix (**Ours**)	24.71	0.859	0.073	21.92	195.2	**22.45**	**0.840**	**0.085**	27.18	234.4	**21.08**	**0.760**	**0.100**	**33.67**	**338.2**
	LoRA w/ DCL (**Ours**)	**24.84**	**0.861**	**0.071**	**21.88**	**190.6**	22.37	0.839	0.087	**26.03**	**210.1**	20.55	0.751	0.107	34.65	391.5
Wrist	Ctrl-World	18.90	0.696	0.277	30.48	324.4	17.93	0.714	0.264	19.37	162.4	19.03	0.642	0.309	22.09	220.8
	WEAVER	17.50	0.640	0.332	41.88	561.9	17.23	0.682	0.296	25.20	222.9	18.42	0.583	0.361	50.12	478.5
	Naïve Mix (**Ours**)	19.56	0.713	0.252	29.47	324.1	17.93	0.717	0.256	19.67	157.6	**20.30**	**0.674**	**0.273**	20.19	223.9
	LoRA w/ DCL (**Ours**)	**19.60**	**0.713**	**0.247**	**29.24**	**310.3**	**18.15**	**0.722**	**0.253**	**18.71**	**150.5**	19.80	0.662	0.278	**20.08**	**218.7**

Table 2: Average scores across 3 evaluation datasets in WorldArena benchmark.

Model	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Background Consistency	Photometric Consistency
Ctrl-World	47.0	36.8	83.4	36.3	33.4	69.0	68.4	70.7	39.2
Naïve Mix (**Ours**)	**48.3**	36.9	86.5	**38.8**	33.8	68.9	**75.8**	**78.2**	**43.7**
LoRA w/ DCL (**Ours**)	48.1	**37.0**	**89.0**	38.5	**34.0**	**69.7**	74.3	76.8	41.6

Model	Physics Adherence		3D Accuracy		Controllability		EWM Score
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	
Ctrl-World	60.1	29.7	95.1	81.9	54.2	87.7	59.5
Naïve Mix (**Ours**)	**62.0**	**39.3**	**95.3**	**83.3**	57.0	**87.8**	**62.4**
LoRA w/ DCL (**Ours**)	61.8	38.1	95.1	83.0	**57.2**	87.7	62.1

5.1 Experimental Setup

Hardware Setup. Our evaluation follows the DROID robot platform (Khazatsky et al., 2024), which includes a Franka FR3 arm installed with 2F-85 Robotiq Gripper, 2 × ZED 2i exocentric cameras and a ZED Mini wrist camera. The overall setup is visualized in Fig. 5 in App. A.3.

Implementation Details. The backbone is a 1.3B-parameter SVD 3D UNet plus a 2.1M action encoder. VAE and CLIP encoder are frozen. Naïve mixture trains the entire model, and LoRA finetuning freezes this UNet and trains rank-64 adapters on the attention maps, keeping the action encoder trainable, for 28.7M updated parameters in total ($\approx 1.9\%$ of the UNet). We use image resolution 320×192. We initialize from the official Ctrl-World checkpoint pretrained on DROID and co-train on DROID and MolmoBot for 100k steps with a batch size of 128 on 8 × H200 GPUs for 2–3 days. We use a learning rate of 1×10^{-5} for full-model co-training. For LoRA fine-tuning, we use a learning rate of 1×10^{-4} for the LoRA attention parameters and 1×10^{-5} for the rest trainable parameters. LoRA fine-tuning additionally uses the dynamics consistency loss with $\lambda_{\text{dyn}} = 0.1$ and $J = 4$, together with 50% action dropout on real samples.

For evaluation, we adopt the official checkpoints from baselines Ctrl-World (Guo et al., 2026c) and WEAVER (Jain et al., 2026). For our models, we take checkpoint at 65k training steps from naïve mixture co-training with α being 0.9 and checkpoint at 70k training steps from LoRA finetuning with α being 0.5.

5.2 Can Sim-And-Real Co-Training Improve World Model's Prediction Quality

Evaluation Data. To mitigate potential bias associated with any single evaluation dataset, we evaluate the prediction quality of world models on trajectories drawn from 3 independently sourced evaluation sets: Ctrl-World (Guo et al., 2026c), WEAVER (Jain et al., 2026), and an additional set of trajectories collected by ourselves. Ctrl-World contributes 30 trajectories spanning 5 tasks: *Drawer*, *PnP*, *Wipe Table*, *Laptop*, and *Towel Fold*. Each task contains 5 trajectories, except for a task with 10 trajectories. WEAVER contributes 100 trajectories across 5 tasks, with 20 trajectories per task: *Stack Bowls*, *PnP Bag*, *PnP Marker*, *PnP Towel*, and *Pour Beans*. In addition, we collect 60 PnP trajectories via teleoperation in the DROID setup, with 20 trajectories for each of 3 conditions that are underrepresented in the other two datasets, including *success under randomized exocentric*

camera viewpoints, *missed grasps*, and *inter-object interactions*. We generate a total of 1,920 clips from the 3 datasets for trajectory replay prediction and evaluation.

Evaluation Metrics. We evaluate video prediction quality using frame-level metrics (PSNR (Hore & Ziou, 2010), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018)), image-distribution-level metrics (FID (Heusel et al., 2017)), and video-distribution metrics (FVD (Unterthiner et al., 2018)). We additionally evaluate with WorldArena (Shang et al., 2026), a unified benchmark for embodied world models. Since quantitative metrics may not align effectively with human preference, we additionally perform Arena-style A/B human preference comparison evaluation (Chiang et al., 2024) with 5 unique annotators and 510 comparisons, reporting an Arena style human preference ELO score using Bradley Terry preference model (Bradley & Terry, 1952). See App. A.5.

Quantitative Metrics Results. Table 1 shows that sim-and-real co-training consistently improves prediction quality across evaluation sources and camera views. LoRA with DCL outperforms baseline on all five metrics in every setting and is strongest overall on the Ctrl-World and WEAVER datasets. When evaluating on our trajectories, naïve mix achieves the best PSNR, SSIM, and LPIPS in both views. Video-level gains (FVD) are more pronounced in exocentric views, including a 43.7% FVD reduction with naïve mix on our trajectories. Table 2 shows that both co-training strategies also improve holistic world-model quality, raising the overall EWM score from 59.5 to 62.4 (naïve mix) and 62.1 (LoRA with DCL). The largest gain is in trajectory accuracy, where co-training improves the SAM3-tracked end-effector score by 32% relative to the baseline (29.7 $\rightarrow$ 39.3), together with substantial improvements in subject consistency (+7.4), background consistency (+7.5), and JEPA similarity (up to 89.0 with LoRA with DCL). See per-dataset WorldArena scores in App. A.4

5.3 World Model for Policy Evaluation

To test MixWorld's utility for policy evaluation, we compare it to real-world policy evaluations and report Pearson's r, rank-order correlation (Spearman's ρ), and Mean Maximum Rank Violation (MMRV) between real and synthetic rollouts for both MixWorld and Ctrl-World.

Policies. We evaluate 3 different generalist VLA policies: π_0-DROID (Black et al., 2025b), π_0-FAST-DROID (Pertsch et al., 2025), and $\pi_{0.5}$-DROID (Black et al., 2025a) from the π series of models, trained on large-scale real data and post-trained on DROID (Khazatsky et al., 2024).

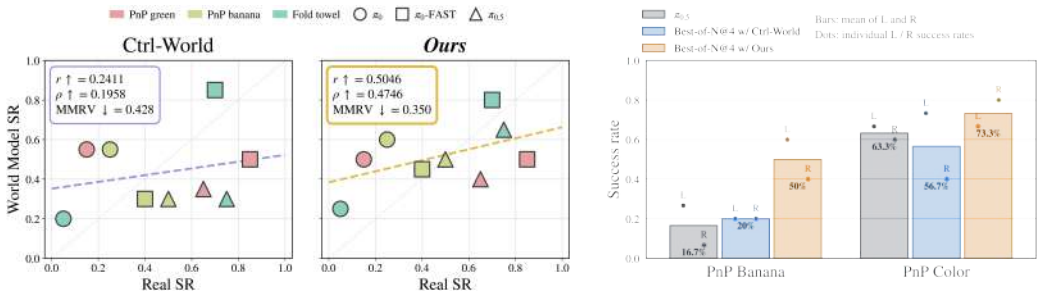
Tasks. We evaluate each of the 3 policies on 3 tasks: *PnP Banana*, which requires a robot to pick up a banana and place it on a plate while ignoring a similarly-colored distractor object, *PnP Block*, which requires a robot to pick up the green block while ignoring other colors and place it on the plate, requiring language following capabilities, and *Fold Towel*, which requires a robot to fold a towel in half. Full methodology details are described in App. A.6.

World Model Rollouts. Given the videos from all 3 cameras (left, right, wrist) from the real-world rollouts, we take the first frame from each camera along with the corresponding proprioceptive observation and use them to initialize matched policy rollouts in both MixWorld and Ctrl-World.

Results. As shown in Fig. 3, policy evaluations in MixWorld have stronger ranker-order correlation and consistency with real-world evaluations, compared to Ctrl-World. MixWorld and real-world evaluations are correlated with Pearson's $r = 0.50$, Spearman's $\rho = 0.47$, and MMRV $= 0.35$. This is much better representative of real-world evaluations than Ctrl-World, which is correlated with the real world at Pearson's $r = 0.24$, Spearman's $\rho = 0.20$, and MMRV $= 0.43$. Evidently, mixing in synthetic data provides myriad counterfactuals and diverse actions, allowing for a greater faithfulness in policy outcomes to the real world. In particular, we note that despite mixing in synthetic data that focuses on rigidbody interaction, the *Fold Towel* task is highly correlated between MixWorld and the real world, indicating that synthetic data is not deleterious to a wider variety of tasks than simply pick and place.

5.4 World Model for Inference-Time Search

Inference-Time Search. We combine a stochastic policy, an action-conditioned world model, and a reward model for inference-time search (Zhang & Shah, 2026) via *Best-of-N* selection (Lightman et al., 2024). Specifically, at each decision step, we sample $N = 4$ candidate action chunks using



(a) **Policy evaluation.** World-model versus real-robot success rates across three policies and three tasks. Insets report Pearson's r (↑), Spearman's ρ (↑), and mean maximum rank violation (MMRV ↓) (Li et al., 2024).

(b) **Inference-time search.** Real-robot success rates of $\pi_{0.5}$ and Best-of-N@4 using Ctrl-World or our world model. Each method is evaluated on 15 episodes per camera view per task.

Figure 3: Comparison of base world model and our world model's performance in downstream tasks: world model as policy evaluator (3a) and inference-time search via *Best-of-N* (3b).

$\pi_{0.5}$-DROID (Black et al., 2025a), each consisting of 15 predicted action steps, and roll out each candidate with the world model to predict the corresponding future frames. We then use Robometer (Liang et al., 2026) to rank the predicted trajectories and select the highest-scoring candidate. The robot executes the first 12 action steps of the selected candidate and repeat the selection procedure afterwards. See Fig. 6 for detailed explanation of one iteration of *Best-of-N* selection in App. A.7.

$\pi_{0.5}$-DROID takes 1 exocentric view and 1 wrist-camera view as input, while the world model conditions on and predicts 2 exocentric views (Exo-Left and Exo-Right) and 1 wrist-camera view. We evaluate the policy separately using Exo-Left and Exo-Right as its exocentric input. For each setting, Robometer scores the world model prediction corresponding to the same exocentric view used by the policy. We conduct 15 episodes per exocentric view, for a total of 30 episodes per task.

Tasks. We evaluate the base policy and inference-time search with both MixWorld and Ctrl-World on two tasks selected from Sec. 5.3: *PnP Banana* and *PnP Block*. Although these tasks are shared with the policy evaluation, we re-evaluate the base policy alongside both search methods under the same experimental conditions to ensure a controlled comparison. Note that the task objects' initial configurations and exocentric camera poses are different from those used in Sec. 5.3, although the overall task stays the same.

Results. As shown in Fig. 3b, inference-time search with MixWorld improves task-level success rates by 10.0%–33.3% over the base policy and by 16.7%–30.0% over search with Ctrl-World. Across exocentric viewpoints, our method matches or exceeds the base policy for both Exo-Left and Exo-Right on both tasks. In contrast, search with Ctrl-World does not consistently improve the base policy, but rather slightly reduces average success on *PnP Block* from 63.3% to 56.7%, whereas our method increases it to 73.3%.

6 Discussion

We propose MixWorld, an action-conditioned world model that is jointly trained on real-world robot data and large-scale simulation robot data, along with two simple sim-and-real co-training schemes for world models. The resulting improvements on world models extend beyond video prediction quantitative metrics, human preference evaluations, to downstream applications of policy evaluation and inference-time search.

Our study has several limitations. The simulation data considered here focus primarily on rigid-object pick-and-place interactions. Extending the tasks and object materials in simulators further benefits world models' capability. Moreover, unlike in sim-and-real co-training in policy learning, we have not experimented across different ratios of simulation and real robot data.

AI USE STATEMENT

In this work, we used generative AI tools for writing/debugging code and polishing writing. This includes tools and scripts to process and convert datasets, which were reviewed and vetted by authors, and the outputted datasets were also qualitatively reviewed via visualizations and sanity checks. No other required disclosure use-cases were accomplished via generative AI tools, including experimental design or analysis, citations, etc.

REFERENCES

Lars Ankile, Anthony Simeonov, Idan Shenfeld, Marcel Torne, and Pulkit Agrawal. From imitation to refinement - residual rl for precise assembly. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 01–08, 2025. doi: 10.1109/ICRA55743.2025.11127442.

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

Samuel M. Bateman, Tenny Yin, Chongyi Zheng, Lei Huang, Brian Wang, , Jaime Eysenbach, Benjamin Fernández Fisac, and Dhruv Shah. Training controllable world models via action influence maximization. In *Proceedings of The 10th Conference on Robot Learning (CoRL)*, 2026.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 17–40, 27–30 Sep 2025a.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Lucy Xiaoyang Shi, Laura Smith, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A Vision-Language-Action Flow Model for General Robot Control. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025b. doi: 10.15607/RSS.2025.XXI.010.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023.

Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, Sergey Levine, and Vincent Vanhoucke. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4243–4250, 2018. doi: 10.1109/ICRA.2018.8460875.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2020. doi: 10.1109/CVPR42600.2020.01164.

Akshay L Chandra, Iman Nematollahi, Chenguang Huang, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. Diwa: Diffusion policy adaptation with world models. *Conference on Robot Learning (CoRL)*, 2025.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024. doi: 10.52202/079017-0759.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 21–27 Jul 2024.

Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. Navsim: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 28706–28719, 2024. doi: 10.52202/079017-0902.

Matt Deitke, Winson Han, Alvaro Herrasti, Aniruddha Kembhavi, Eric Kolve, Roozbeh Mottaghi, Jordi Salvador, Dustin Schwenk, Eli VanderBilt, Matthew Wallingford, Luca Weihs, Mark Yatskar, and Ali Farhadi. Robothor: An open simulation-to-real embodied ai platform. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3161–3171, 2020. doi: 10.1109/CVPR42600.2020.00323.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. Procthor: Large-scale embodied ai using procedural generation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5982–5994, 2022. doi: 10.52202/068431-0433.

Shengliang Deng, Mi Yan, Songlin Wei, Haixin Ma, Yuxin Yang, Jiayi Chen, Zhiqi Zhang, Taoyu Yang, Xuheng Zhang, Heming Cui, Zhizheng Zhang, and He Wang. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data. In *9th Annual Conference on Robot Learning*, 2025.

Abhay Deshpande, Maya Guru, Rose Hendrix, Snehal Jauhri, Ainaz Eftekhar, Rohun Tripathi, Max Argus, Jordi Salvador, Haoquan Fang, Matthew Wallingford, Wilbert Pumacay, Yejin Kim, Quinn Pfeifer, Ying-Chun Lee, Piper Wolters, Omar Rayyan, Mingtong Zhang, Jiafei Duan, Karen Farley, Winson Han, Eli Vanderbilt, Dieter Fox, Ali Farhadi, Georgia Chalvatzaki, Dhruv Shah, and Ranjay Krishna. Molmob0t: Large-scale simulation enables zero-shot manipulation. *In 9th Annual Conference on Robot Learning (CoRL)*, 2026.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78 of *Proceedings of Machine Learning Research*, pp. 1–16, 13–15 Nov 2017.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2786–2793, 2017. doi: 10.1109/ICRA.2017.7989324.

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29, 2016.

Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28135–28144, October 2025.

Jun Guo, Xiaojian Ma, Yikai Wang, Min Yang, Huaping Liu, and Qing Li. Flowdreamer: A rgb-d world model with flow-based motion representations for robot manipulation. *IEEE Robotics and Automation Letters*, 11(3):2466–2473, 2026a. doi: 10.1109/LRA.2026.3653273.

Yanjiang Guo, Tony Lee, Lucy Xiaoyang Shi, Jianyu Chen, Percy Liang, and Chelsea Finn. VLAW: Iterative co-improvement of vision-language-action policy and world model. In *Forty-third International Conference on Machine Learning*, 2026b.

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

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565, 2019.

Danijar Hafner, Wilson Yan, and Timothy P. Lillicrap. Training agents inside of scalable world models. *CoRR*, abs/2509.24527, 2025. doi: 10.48550/ARXIV.2509.24527.

Nicklas Hansen and Xiaolong Wang. Hallucination in world models is predictable and preventable, 2026.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *Proceedings of the 2010 20th International Conference on Pattern Recognition*, ICPR '10, pp. 2366–2369, USA, 2010. ISBN 9780769541099. doi: 10.1109/ICPR.2010.579.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20765–20779, June 2026.

Arnav Kumar Jain, Yilin Wu, Jesse Farebrother, Gokul Swamy, and Andrea Bajcsy. Weaver: Efficient world models for robot video prediction. *CoRR*, abs/2606.13672, 2026.

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

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022.

Alexander Khazatsky, Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, et al. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.120.

Yejin Kim, Wilbert Pumacay, Omar Rayyan, Max Argus, Winson Han, Eli Vanderbilt, Jordi Salvador, Abhay Deshpande, Rose Hendrix, Snehal Jauhri, Shuo Liu, Nur Muhammad Mahi Shafiullah, Maya Guru, Ainaz Eftekhar, Karen Farley, Donovan Clay, Jiafei Duan, Arjun Guru, Piper Wolters, Alvaro Herrasti, Ying-Chun Lee, Georgia Chalvatzaki, Yuchen Cui, Ali Farhadi, Dieter Fox, and Ranjay Krishna. MolmoSpaces: Large-Scale Open Ecosystem for Robot Manipulation and Navigation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.091.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. AI2-THOR: An Interactive 3D Environment for Visual AI. *ArXiv*, abs/1712.05474, 2017.

Yu Lei, Minghuan Liu, Abhiram Maddukuri, Zhenyu Jiang, and Yuke Zhu. A mechanistic analysis of sim-and-real co-training in generative robot policies. In *International Conference on Machine Learning (ICML)*, 2026.

Jacob Levy, Tyler Westenbroek, Kevin Huang, Fernando Palafox, Patrick Yin, Shayegan Omidshafiei, Dong-Ki Kim, Abhishek Gupta, and David Fridovich-Keil. Simulation Distillation: Pretraining World Models in Simulation for Rapid Real-World Adaptation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.017.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Yaxuan Li, Junjie Wen, Zhongyi Zhou, Yefei Chen, Chaomin Shen, Yaxin Peng, and Yichen Zhu. Scaling real-world robot policy evaluation via discrete diffusion world model. In *Forty-third International Conference on Machine Learning*, 2026.

Anthony Liang, Yigit Korkmaz, Jiahui Zhang, Minyoung Hwang, Abrar Anwar, Sidhant Kaushik, Aditya Shah, Alex S. Huang, Luke Zettlemoyer, Dieter Fox, Yu Xiang, Anqi Li, Andreea Bobu, Abhishek Gupta, Stephen Tu, Erdem Biyik, and Jesse Zhang. Robometer: Scaling General-Purpose Robotic Reward Models via Trajectory Comparisons. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.140.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Susie Lu, Haonan Chen, Weirui Ye, and Yilun Du. Driftworld: Fast world modeling through drifting. *arXiv preprint arXiv:2607.15065*, 2026.

Abhiram Maddukuri, Zhenyu Jiang, Lawrence Yunliang Chen, Soroush Nasiriany, Yuqi Xie, Yu Fang, Wenqi Huang, Zu Wang, Zhenjia Xu, Nikita Chernyadev, Scott Reed, Ken Goldberg, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Sim-and-Real Co-Training: A Simple Recipe for Vision-Based Robotic Manipulation. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.109.

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *7th Annual Conference on Robot Learning*, 2023.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024.

NVIDIA, Arslan Ali, Junjie Bai, Maciej Bala, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, et al. Open x-embodiment: Robotic learning datasets and rt-x models : Open x-embodiment collaboration0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient Action Tokenization for Vision-Language-Action Models. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.012.

Silviu Pitis, Elliot Creager, and Animesh Garg. Counterfactual data augmentation using locally factored dynamics. *Advances in Neural Information Processing Systems*, 33:3976–3990, 2020.

Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Inference-time enhancement of generative robot policies via predictive world modeling. *IEEE Robotics and Automation Letters*, 11 (5):5534–5541, 2026. doi: 10.1109/LRA.2026.3673995.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 18–24 Jul 2021.

Alexander Raistrick, Lingjie Mei, Karhan Kayan, David Yan, Yiming Zuo, Beining Han, Hongyu Wen, Meenal Parakh, Stamatis Alexandropoulos, Lahav Lipson, Zeyu Ma, and Jia Deng. Infinigen indoors: Photorealistic indoor scenes using procedural generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21783–21794, June 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, Los Alamitos, CA, USA, June 2022.

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

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Xin Wang, Wenwu Zhu, and Yong Li. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026. ISSN 2666-6510. doi: https://doi.org/10.1016/j.aiopen.2026.07.002.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.127063.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurélien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2443–2451, 2020. doi: 10.1109/CVPR42600.2020.00252.

GR Team, K Choromanski, C Devin, Y Du, D Dwibedi, R Gao, A Jindal, T Kipf, S Kirmani, I Leal, et al. Evaluating gemini robotics policies in a veo world simulator. *arXiv preprint arXiv:2512.10675*, 2025.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, Yaping Li, Ping Wang, Junhao Cai, Jia Zeng, Hao Dong, and Jiangmiao Pang. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 976–985, June 2026.

Wei-Cheng Tseng, Gashon Hussein, Yuzhu Dong, Allen Z. Ren, Lucy X. Shi, XuDong Wang, Sergey Levine, Zhaoshuo Li, Jinwei Gu, Florian Shkurti, Ming-Yu Liu, and Quan Vuong. Sc3-eval: Evaluating robot foundation models via self-consistent video generation, 2026.

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

Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *Computer Vision – ECCV 2024*, pp. 55–72, Cham, 2025. ISBN 978-3-031-73195-2.

Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive World Simulator for Robot Policy Training and Evaluation. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.018.

Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. RoboGen: Towards unleashing infinite data for automated robot learning via generative simulation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51936–51983, 21–27 Jul 2024.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Adam Wei, Abhinav Agarwal, Boyuan Chen, Rohan Bosworth, Nicholas Pfaff, and Russ Tedrake. Empirical analysis of sim-and-real cotraining of diffusion policies for planar pushing from pixels. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5625–5632, 2025a. doi: 10.1109/IROS60139.2025.11246304.

Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Jian Tong, Haodong Duan, Qipeng Guo, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. VideoroPE: What makes for good video rotary position embedding? In *Forty-second International Conference on Machine Learning*, 2025b.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14662–14672, June 2024a.

Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. Resim: Reliable world simulation for autonomous driving. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167710–167741, 2025. doi: 10.52202/085713-5591.

Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *The Twelfth International Conference on Learning Representations*, 2024b.

Junjie Ye, Rong Xue, Basile Van Hoorick, Runhao Li, Harshitha Rajaprakash, Pavel Tokmakov, Muhammad Zubair Irshad, Vitor Guizilini, and Yue Wang. Robodream: Compositional world models for scalable robot data synthesis. *arXiv preprint arXiv:2606.02577*, 2026a.

Junjie Ye, Rong Xue, Basile Van Hoorick, Pavel Tokmakov, Muhammad Zubair Irshad, Yue Wang, and Vitor Guizilini. Anchordream: Repurposing video diffusion for embodiment-aware robot data synthesis. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026b.

Tenny Yin, Zhiting Mei, Zhonghe Zheng, Miyu Yamane, David Wang, Jade Sceats, Samuel M. Bateman, Lihan Zha, Apurva Badithela, Ola Shorinwa, and Anirudha Majumdar. Playworld: Learning robot world models from autonomous play. In *Proceedings of The 10th Conference on Robot Learning*, 2026.

Mingtong Zhang and Dhruv Shah. Visual Verification Enables Inference-time Steering and Autonomous Policy Improvement. In *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.079.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

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

Fangqi Zhu, YAN Zhengyang, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World model-based policy optimization for vision-language-action models. In *The Fourteenth International Conference on Learning Representations*, 2026.

A Appendix

A.1 Extended Related Work

World Models for Robotics The development of generative modeling (Ho et al., 2020; Song et al., 2021; Lipman et al., 2023; Rombach et al., 2022; Blattmann et al., 2023) has enabled the advance of high-fidelity action-conditioned world models (Guo et al., 2026c; Yang et al., 2024b; Zhu et al., 2025; NVIDIA et al., 2025; Hafner et al., 2025; Wang et al., 2026; Huang et al., 2026; Guo et al., 2026a; Ye et al., 2026b;a) that can be applied for diverse downstream applications, such as policy evaluation (Team et al., 2025; Zhu et al., 2025; Guo et al., 2026c; Tseng et al., 2026; Wang et al., 2026; Quevedo et al., 2026; Li et al., 2026; Yin et al., 2026; Jain et al., 2026; Lu et al., 2026), inference-time planning (Zhu et al., 2025; Assran et al., 2025; Qi et al., 2026; Jain et al., 2026), and synthetic data generation for policy improvement (Yang et al., 2024b; Chandra et al., 2025; Guo et al., 2026c; Yin et al., 2026; Jain et al., 2026; Guo et al., 2026b; Jiang et al., 2026; Zhu et al., 2026). However, world models (Guo et al., 2026c; Jain et al., 2026; Quevedo et al., 2026) that are trained on large-scale offline robot data (Khazatsky et al., 2024; Walke et al., 2023) show limited capabilities when predicting future frames of non-optimal or non-successful action trajectories, because the datasets bias to successful demonstrations by human teleoperators and rarely include counterfactual dynamics (Pitis et al., 2020). Although techniques such as diffusion forcing (Chen et al., 2024) and rotary position embedding (Su et al., 2024; Wei et al., 2025b) have been applied to improve the fidelity of world models (Jain et al., 2026), they do not address the fundamental problem of narrow data distribution. To broaden the data coverage of interaction dynamics, especially failure cases such as missed grasps and slips, recent works (Yin et al., 2026; Guo et al., 2026b) execute vision-language-action (VLA) policies (Black et al., 2025a) to iteratively collect online interactions between robots and the real world. Although they can obtain high-quality real-world counterfactual data to finetune the world model, collecting data on real robots is intrinsically expensive and time-consuming considering the robot time, hardware resources, and unpredictable failure modes that

require human intervention. In contrast, we broaden transition dynamics by sourcing large-scale simulation data that contain counterfactual cases as well as a large variety of camera viewpoints surrounding the robot workspace.

Simulation for Policy Learning Complementary to the effort in crowd-sourcing real-world robot data (Khazatsky et al., 2024; O'Neill et al., 2024), simulation (Kolve et al., 2017; Deitke et al., 2020; 2022; Raistrick et al., 2024; Kim et al., 2026) has been another primary source for automated generation of large-scale robot datasets (Mandlekar et al., 2023; Wang et al., 2024; Deng et al., 2025; Deshpande et al., 2026; Tian et al., 2026). Recent work has demonstrated that policies trained on massive amounts of simulation data can achieve impressive zero-shot transfer to the real world (Deng et al., 2025; Deshpande et al., 2026; Tian et al., 2026). An alternative line of work sidesteps explicit domain adaptation and extensive domain randomization by directly co-training policies on a mixture of simulated and real-world robot data (Bousmalis et al., 2018; Nasiriany et al., 2024; Ankile et al., 2025; Maddukuri et al., 2025). Notably, even straightforward sim-and-real co-training without specialized architectures for domain integration has been shown to outperform training on real-world data alone (Wei et al., 2025a; Lei et al., 2026). Whether the benefits observed in policy learning extend to action-conditioned world models for robotic manipulation, which must capture dynamics across domains, remains unclear. In this work, we investigate whether and how simulation and real-world robot data can be effectively leveraged to improve world model's faithfulness.

Simulation for World Models Similar to the challenges faced by action-conditioned world models for robotic manipulation, world models for autonomous driving are constrained by the coverage of real-world driving data (Caesar et al., 2020; Sun et al., 2020). While large-scale real-world driving datasets have enabled increasingly capable generative world models (Hu et al., 2023; Wang et al., 2025; Russell et al., 2025; Gao et al., 2025), they inevitably underrepresent rare and safety-critical scenarios. ReSim (Yang et al., 2025) enriches real-world human driving data (Dauner et al., 2024; Yang et al., 2024a) with non-expert data from driving simulator, i.e., CARLA (Dosovitskiy et al., 2017), and proposes multi-stage sim-and-real co-training to improve models' controllability. ReSim shares the most related philosophical idea to this paper, but we focus on co-training action-conditioned world models for robotic manipulation where there are richer and more complex contact dynamics and broader distribution of camera viewpoints. The most related work on training action-conditioned world model for robotic manipulation with simulation data is SimDist (Levy et al., 2026), which pretrains world models on online interaction in simulation and perturbs expert action sequences to gain counterfactual experience. But their models are task-oriented and trained on setup-specific data. In contrast, our models are co-trained on large-scale real-world general robot data (Khazatsky et al., 2024) and simulation data with a large variety of objects and scenes (Deshpande et al., 2026).

A.2 Dataset Visualization

A.3 Robot Platform

Our robot platform follows the DROID setup (Khazatsky et al., 2024) and include a Franka arm equipped with 2F-85 Robotiq Gripper, 2 $\times$ ZED 2i exocentric cameras and a ZED Mini wrist camera, as shown in Fig. 5

A.4 Per-Dataset WorldArena Evaluation

Tables 3, 4, and 5 show the WorldArena benchmark scores on each of the 3 evaluation datasets.

A.5 Human Preference Evaluation.

Human Preference Results. As shown in Table 6, both co-trained variants achieve higher preference scores than Ctrl-World across all 3 evaluation sets. On Eval Sets 1 and 2, where there are regular shoulder camera viewpoints and no obvious counterfactual dynamics, the improvements are modest but consistent. The separation is substantially larger on our evaluation set consisting of transition dynamics and camera viewpoints that are underrepresented in real-robot data: Ctrl-World obtains 872.5, compared with 1073.1 for naïve mix and 1054.4 for LoRA with DCL. Thus, human

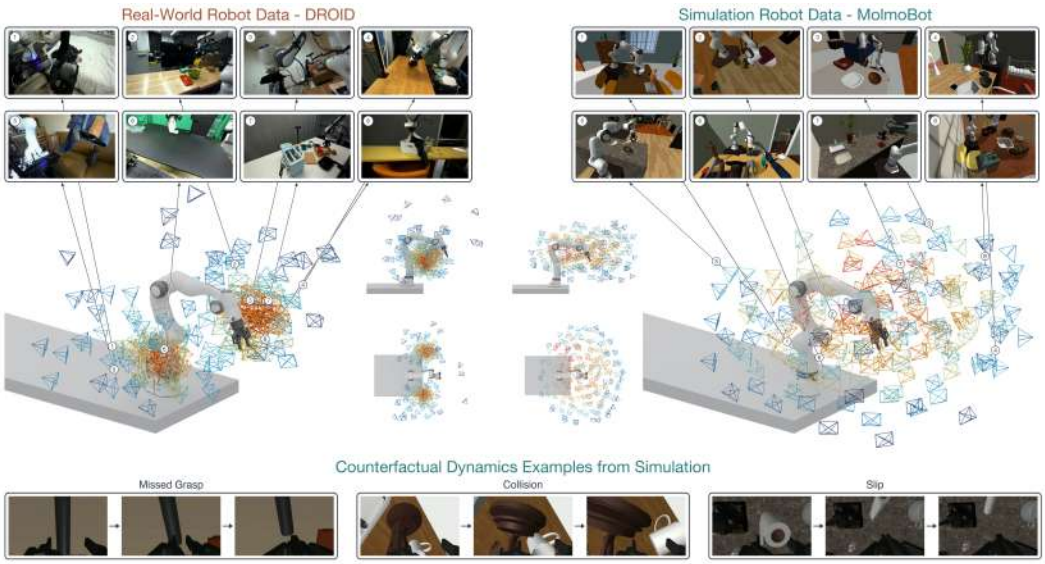


Figure 4: Comparison between real-world robot data and simulation robot data. DROID grounds real-world visual appearance and dynamics, and contains deformable object manipulation trajectories that simulation data lacks. MolmoBot-Data contains rigid object manipulation data that includes missed grasps, slips, inter-object collisions, with broader range of exocentric camera poses.

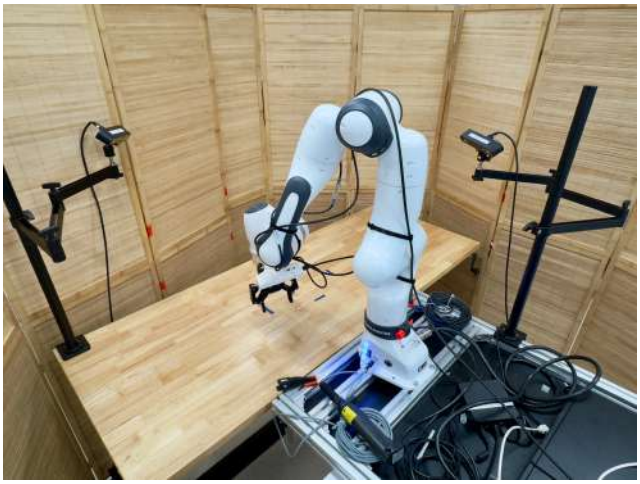


Figure 5: Hardware setup.

preferences consistently favor sim-and-real co-training, with the largest separation occurring on our independently collected evaluation trajectories.

A.6 Policy Evaluation

In Sec. 5.3 we evaluate 5 policies on 3 tasks in the real world and measure rank-order correlation with our world model. Here we delineate the methodological details for each task.

PnP Banana. In this task, a robot must pick up a banana from a variety of initial configurations and place it on a plate, while ignoring a similarly colored distractor object. The task is deemed successful if and only if the banana is fully on the plate and the distractor object is not. The task prompt is "Put the yellow banana on the plate".

PnP Block. In this task, a robot must pick up a green block and place it on the plate, while ignoring other similarly-shaped but differently-colored blocks, from a variety of initial configurations. This

Table 3: WorldArena benchmark scores on evaluation data from Ctrl-World. Best scores are shown in **bold** and second-best results are underlined.

Model	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Background Consistency	Photometric Consistency
Ctrl-World	48.4	**34.2**	90.9	33.4	30.5	66.4	74.7	77.0	**41.2**
Naïve Mix (**Ours**)	48.9	34.0	90.7	34.1	30.6	66.5	75.8	78.3	41.1
LoRA w/ DCL (**Ours**)	**49.2**	34.1	**91.4**	**34.4**	**30.8**	**67.2**	**76.1**	**78.8**	40.3

Model	Physics Adherence		3D Accuracy		Controllability		EWM Score
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	
Ctrl-World	61.1	35.1	97.4	81.5	54.4	88.5	61.0
Naïve Mix (**Ours**)	**63.4**	**44.7**	**97.4**	82.1	57.6	**88.7**	62.3
LoRA w/ DCL (**Ours**)	63.2	44.6	97.4	**82.2**	**58.4**	88.1	**62.4**

Table 4: WorldArena benchmark scores on evaluation data from WEAVER. Best scores are shown in **bold** and second-best results are underlined.

Model	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Background Consistency	Photometric Consistency
Ctrl-World	45.3	**37.2**	88.9	44.7	36.3	70.9	82.7	86.0	49.6
Naïve Mix (**Ours**)	**47.1**	37.1	85.9	**47.0**	37.0	70.5	**85.6**	**89.1**	**50.6**
LoRA w/ DCL (**Ours**)	46.3	37.1	**90.9**	46.8	**37.1**	**71.4**	84.5	88.0	49.2

Model	Physics Adherence		3D Accuracy		Controllability		EWM Score
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	
Ctrl-World	62.0	36.2	**96.2**	81.4	55.3	87.7	64.0
Naïve Mix (**Ours**)	62.5	**50.1**	96.2	**82.2**	56.0	87.6	**65.6**
LoRA w/ DCL (**Ours**)	**62.8**	46.4	96.1	81.8	**57.0**	**88.1**	65.6

requires a policy to attend to specific color details in the language instruction. The task is deemed successful if and only if only the green block is fully on the plate. The task prompt is "Put the green block on the plate".

Fold Towel. In this task, a robot must fold a towel in half. The task is deemed successful if a significant portion of the towel has been folded onto the rest of the towel. The task prompt is "Fold the towel in half".

All policies are run at 15Hz with an action chunk length of 15. A rollout is terminated upon task success, if the robot triggers safety stops or performs dangerous movements, or after 45 seconds. Note that since policy inference often exceeds $1/15$hz ≈ 0.066s, we terminate after $45/0.066\text{s} = 682$ steps instead of 45 seconds of walltime for fairness.

A.7 Inference-Time Search

In Sec.5.4, we instantiate inference-time search using *Best-of-N* selection over action-conditioned world model predictions to improve policy performance. We illustrate a single iteration of *Best-of-N* in Fig.6. Given the current observation and robot proprioceptive state, we sample four candidate action chunks from $\pi_{0.5}$. Each candidate action chunk, together with the current observation and proprioceptive state, is provided to the world model, which predicts the corresponding future frames. We then rank the predicted futures using Robometer (Liang et al., 2026) and select the action chunk associated with the highest-ranked prediction. The robot executes the first 12 steps of the selected 15-step action chunk before proceeding to the next iteration of inference-time search.

A.8 Inference Time Search

Table 5: WorldArena benchmark scores on our trajectories. Best scores are shown in **bold** and second-best results are underlined.

Model	Visual Quality			Motion Quality			Content Consistency		
	Image Quality	Aesthetic Quality	JEPA Similarity	Dynamic Degree	Flow Score	Motion Smoothness	Subject Consistency	Background Consistency	Photometric Consistency
Ctrl-World	47.4	39.1	70.4	30.7	33.3	69.7	47.7	49.1	26.8
Naïve Mix (**Ours**)	**48.8**	39.7	82.9	**35.2**	33.8	69.7	**65.9**	**67.1**	**39.4**
LoRA w/ DCL (**Ours**)	48.7	**39.8**	**84.6**	34.3	**34.1**	**70.4**	62.3	63.6	35.3

Model	Physics Adherence		3D Accuracy		Controllability		EWM Score
	Interaction Quality	Trajectory Accuracy	Depth Accuracy	Perspectivity	Instruction Following	Semantic Alignment	
Ctrl-World	57.3	17.8	91.6	82.8	52.9	86.9	53.6
Naïve Mix (**Ours**)	**60.1**	23.3	**92.5**	**85.5**	**57.5**	**87.2**	**59.2**
LoRA w/ DCL (**Ours**)	59.3	**23.3**	91.7	84.8	56.3	87.0	58.4

Table 6: Human preference results across three sources of data.

Model	**Eval Set 1** (Guo et al., 2026c)	**Eval Set 2** (Jain et al., 2026)	**Our Eval Set**
Ctrl-World	989.9	985.6	872.5
Naïve Mix (**Ours**)	1004.3	**1017.2**	**1073.1**
LoRA w/ DCL (**Ours**)	**1005.7**	997.4	1054.4

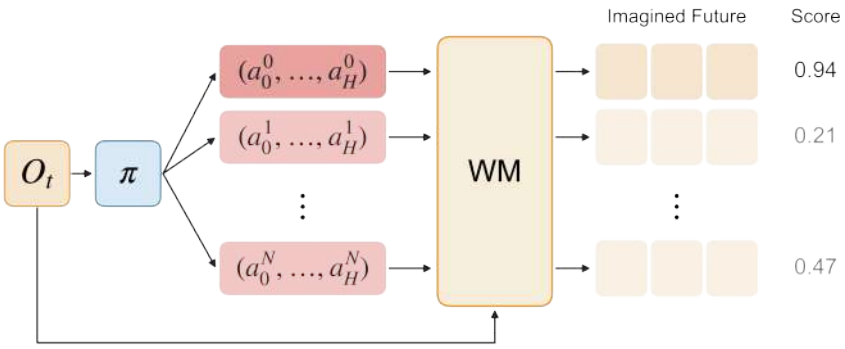


Figure 6: Illustration of one iteration of *Best-of-N* selection.

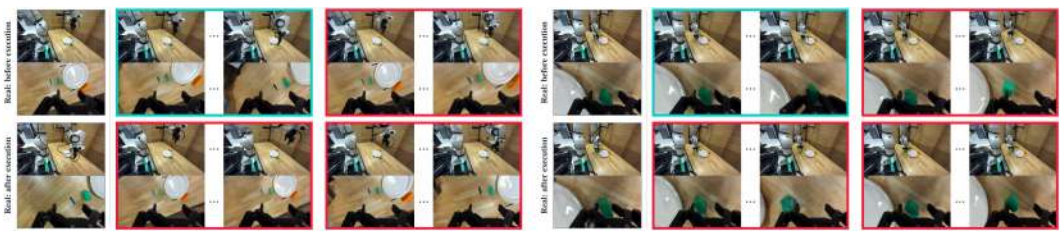


Figure 7: **Inference-time search examples.** We show examples during approach (left) and grasp (right), for placing a green block on a plate. Each example includes real observations before and after executing the selected action chunk, alongside four candidate world-model rollouts. Green borders denote selected candidates and red borders denote non-selected candidates