

# WMPO: World Model-based Policy Optimization for Vision-Language-Action Models

Fangqi Zhu<sup>1,2</sup>, Zhengyang Yan<sup>1</sup>, Zicong Hong<sup>1</sup>, Quanxin Shou<sup>1</sup>, Xiao Ma<sup>2,\*</sup>, Song Guo<sup>1,\*</sup>  
<sup>1</sup>Hong Kong University of Science and Technology <sup>2</sup>ByteDance Seed

\*Corresponding Author

## Abstract

Vision-Language-Action (VLA) models have shown strong potential for general-purpose robotic manipulation, but their reliance on expert demonstrations limits their ability to learn from failures and perform self-corrections. Reinforcement learning (RL) addresses these through self-improving interactions with the physical environment, but suffers from high sample complexity on real robots. We introduce *World-Model-based Policy Optimization (WMPO)*, a principled framework for on-policy VLA RL without interacting with the real environment. In contrast to widely used latent world models, WMPO focuses on pixel-based predictions that align the “*imagined*” trajectories with the VLA features pretrained with web-scale images. Crucially, WMPO enables the policy to perform on-policy GRPO that provides stronger performance than the often-used off-policy methods. Extensive experiments in both simulation and real-robot settings demonstrate that WMPO (i) substantially improves sample efficiency, (ii) achieves stronger overall performance, (iii) exhibits emergent behaviors such as self-correction, and (iv) demonstrates robust generalization and lifelong learning capabilities.

**Date:** November 7, 2025

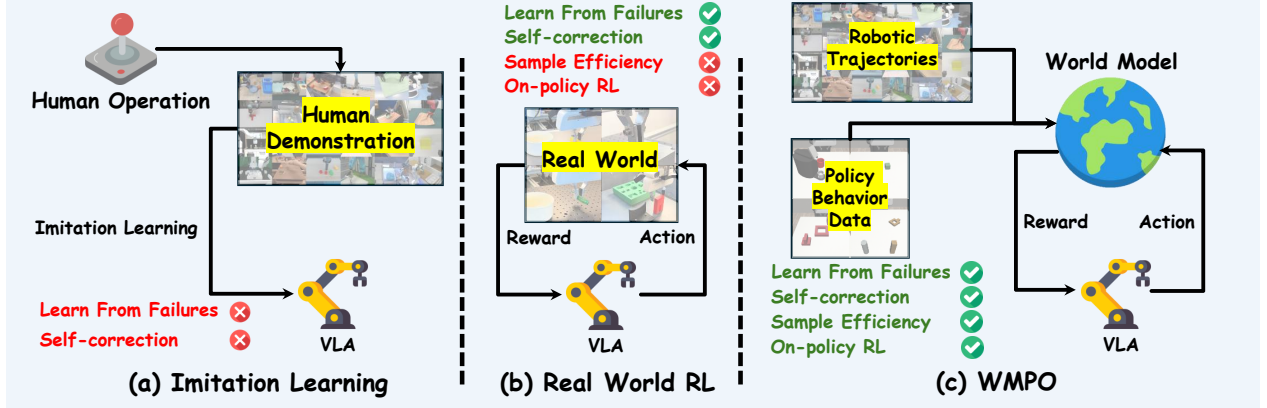
**Corresponding Email:** [songguo@ust.hk](mailto:songguo@ust.hk); [xiao.ma@bytedance.com](mailto:xiao.ma@bytedance.com)

**Project Page:** <https://wm-po.github.io/>

## 1 Introduction

Vision-Language-Action (VLA) models [1–3] have emerged as a promising paradigm for general-purpose robotic manipulation, enabling robots to follow natural language instructions in complex, unstructured environments. The dominant approach for training these models is imitation learning (IL) from large-scale human demonstrations [4]. While effective in mimicking demonstrated behaviors, IL-trained policies are often brittle. When encountering out-of-distribution states not seen during training, they can take suboptimal actions that lead to compounding errors [5], making task completion or recovery nearly impossible (Fig. 1a).

Reinforcement learning (RL) [6] offers a natural solution to this brittleness by allowing an agent to learn and improve through active interaction with its environment. This self-improvement process can lead to policies that are more robust and capable of recovering from failure. However, applying RL directly to real robots is notoriously sample-inefficient, requiring millions of interactions that are impractical, unsafe, and time-consuming to collect (Fig. 1b). Recent efforts to improve sample efficiency fall into two main strategies. The first leverages human intervention to guide learning [7–9], which reduces exploration cost but is labor-intensive and hard to scale. The second relies on simulation to reduce real-world interactions [10, 11],



**Figure 1** Three different VLA training paradigms: (a) Imitation learning learns from human demonstrations but lacks the ability for learning from failures and self-correction; (b) Real-world RL improves policy through direct interaction but suffers from high sampling costs and difficulty in achieving on-policy RL; (c) WMPO pretrains a world model on large-scale robotic trajectories and fine-tunes it with limited policy behavior data, enabling sample-efficient on-policy RL for VLA without real-world interaction.

but is limited by the difficulty of building accurate simulators for diverse scenarios.

The advent of large-scale generative models, particularly video-generative world models [12, 13], presents a compelling new frontier for leveraging model-based RL [14] to mitigate the sample inefficiency challenge in VLA RL. These models can learn the dynamics of the world from data and simulate future transitions, offering a path to scalable RL without costly real-world explorations. Nevertheless, integrating these models with existing VLAs remains a challenge. Many classical model-based RL approaches [15–20] operate in an abstract latent space, which creates a fundamental mismatch with powerful VLA foundation models that are pretrained on real-world images. The rich, pretrained visual understanding of VLAs cannot be directly applied within a mismatched latent dynamics model. We argue that leveraging a pixel-space video-generative world model is crucial, as it allows the VLA policy to operate on generated visual data that is consistent with its pretraining, effectively bridging the world model with the policy’s pre-trained knowledge.

To this end, we propose World Model-based Policy Optimization (WMPO), as illustrated in Fig. 1c, a principled framework that grounds VLA RL entirely in an action-conditioned video world model. By pretraining a high-fidelity, pixel-space video-generative model on millions of robotic trajectories[21], WMPO leverages realistic visual dynamics to create a scalable training environment for downstream tasks.

WMPO incorporates several key innovations. First, to mitigate the state-distribution mismatch between expert demonstrations and policy rollouts, we introduce policy behavior alignment, finetuning the world model with behavioral data collected by the policy itself. Second, short-horizon prediction makes it difficult to define accurate rewards and is prone to reward hacking. To address this, WMPO generates complete trials through clip-level autoregressive video generation, enabling more reliable outcome-based reward assignment. While this design supports long-horizon rollouts, it also introduces challenges such as visual distortion and action-frame misalignment. WMPO addresses these with noisy-frame conditioning and frame-level action control, ensuring robustness and accurate trajectory simulation. Third, we train a lightweight reward model that predicts task success or failure, providing a learned sparse reward signal and avoiding both complex reward shaping and reward hacking. Together, these components form a self-contained environment that enables on-policy RL entirely “in imagination”. Specifically, we adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024), whose robustness and scalability have been demonstrated in DeepSeek-AI et al. [22]. Notably, WMPO naturally supports repeated rollouts from the same initial state, which is difficult to realize in the physical world but crucial for large-scale GRPO training.

We conduct extensive experiments in both Mimicgen simulation environments [23] and real-world environments to validate the effectiveness of WMPO. Our results show that WMPO achieves substantially higher sample

efficiency and consistently outperforms VLA RL methods that directly optimize with real trajectories. Crucially, we provide qualitative evidence of emergent behaviors, where the WMPO-trained policy demonstrates self-correction strategies not present in the demonstration data and often completes tasks faster and more smoothly, without noticeable stalls. We further demonstrate WMPO’s strong generalization compared to offline RL methods, as well as its capacity for lifelong learning through alternating updates between the VLA policy and the world model. Taken together, these findings highlight WMPO as a scalable and generalizable paradigm for advancing VLA RL.

## 2 Related Work

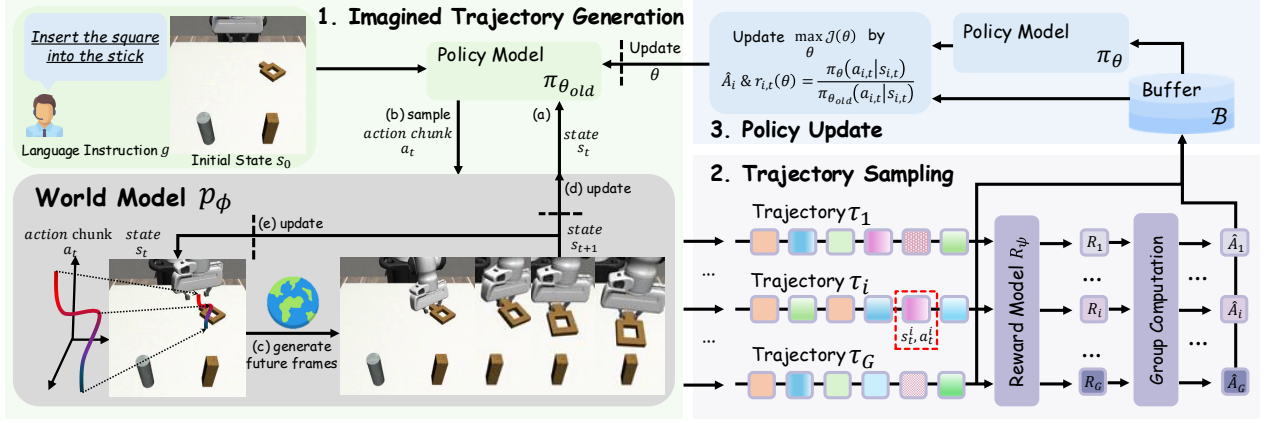
**Vision-Language-Action Models.** Vision-Language-Action (VLA) models aim to map visual inputs and natural language instructions into executable robotic actions, enabling general-purpose manipulation. Most VLAs build upon pretrained vision-language models (VLMs) and are further fine-tuned on robotic trajectories, thereby inheriting strong visual grounding and language understanding. This progress has been driven both by large-scale demonstration collections [21] and by advances in action parameterization, including discrete next-token prediction [2], parallel decoding [24], and continuous flow-based heads [25]. Despite these advances in data and model design, existing VLAs largely remain within the imitation learning (IL) paradigm, making them brittle when encountering out-of-distribution states and unable to leverage failed executions for improvement [5].

**Reinforcement Learning for VLA Models.** RL provides a natural complement to IL by enabling policies to learn from interaction, thereby improving robustness and recovery capabilities. However, applying on-policy RL to VLA remains challenging due to severe sample inefficiency and substantial system-level complexity. Prior works can be broadly divided into two strategies. The first introduces human intervention to guide exploration, e.g., supplying corrective signals when policies encounter unrecoverable states [8, 9, 26]. While effective at reducing exploration cost, this approach requires continuous human supervision, making it labor-intensive and difficult to scale. The second attempts to perform RL directly in simulation or in the real world. For instance, Zhang et al. [27] adopt trajectory-level DPO [28], while others apply PPO [29] or GRPO [30] to optimize VLA policies in simulation [10, 11]. These approaches avoid human supervision but still suffer from poor sample efficiency, and constructing accurate simulators for each real-world scenario introduces prohibitive engineering overhead. In contrast, WMPO enables policy optimization entirely within a learned world model, substantially improving sample efficiency and scalability.

**World Models.** World models aim to mitigate the need for costly real-world interactions by learning generative dynamics and enabling policies to learn from “imagine” trajectories. Early approaches [16–18] learned world model in the latent space of recurrent state-space models, which provided efficient but overly abstract rollouts. More recent work introduced diffusion-based world models, showing that retaining pixel-level fidelity is crucial for RL with Gaussian policies [31, 32]. Building on this trend, large-scale video world models [12, 13] have demonstrated impressive generality across diverse domains. However, when applied to robotics, they suffer from distribution mismatch—struggling to faithfully reproduce policy rollouts and fine-grained robot–object interactions. In contrast, WMPO addresses these challenges through policy behavior alignment and, for the first time, verifies the feasibility of leveraging high-fidelity world models for scalable VLA RL.

## 3 World Model-Based Policy Optimization

We introduce World Model-based Policy Optimization (WMPO), a novel framework for learning complex, vision-based robotic manipulation policies. The WMPO framework is grounded in the principles of model-based RL, where policy optimization is performed entirely within a learned world model, thereby eliminating the need for costly real-world interactions. WMPO operates directly in the pixel space, instead of latent space, which better bridges the pretrained VLA features from web-scale video-action pairs with the “imagined” trajectories. Specifically, the world model is trained to accurately reflect the policy’s behavior through a process we call Policy Behavior Alignment, where it is fine-tuned on a small set of real trajectories collected



**Figure 2** WMPO starts from an initial state  $s_0$ . The overall training procedure consists of three components: (1) *Imagined Trajectory Generation*, where policy model  $\pi_{\theta_{old}}$  and world model  $p_\phi$  interact alternately to generate a full imagined trajectory; (2) *Trajectory Sampling*, where multiple trajectories are sampled and evaluated by the reward model  $R_\psi$ ; and (3) *Policy Update*, where the policy parameters  $\theta$  are optimized via Eq. 4. This process is iteratively repeated throughout training.

from the policy itself. This ensures the model can faithfully simulate diverse scenarios, including failures. We also introduce noisy frame conditioning and frame-level action control techniques to overcome the problems of visual distortion and action-frame misalignment in long-horizon video prediction. With these modifications, we are able to perform strong on-policy Group Relative Policy Optimization (GRPO) using trajectories imagined by the learned world model, which significantly enhances sample efficiency compared to direct RL methods. Fig. 2 illustrates an overview of the training procedure of WMPO.

### 3.1 Problem Formulation

We formulate the VLA manipulation task as a decision-making problem in the form of an MDP  $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R)$ .

- **State space.**  $\mathcal{S} = \mathcal{I} \times \mathcal{G}$ , where  $\mathcal{I}$  denotes the image observation space, i.e., image sequences  $I_{0:K}$ , and  $\mathcal{G}$  denotes the language instruction space. Here, we make the assumption that robot states can be solely defined by their image observations. For more complicated setups, e.g., Partially Observable MDPs (POMDPs) [33, 34], we leave them for future studies.
- **Action space.**  $\mathcal{A}$  denotes the space of action chunks, i.e., sequences of robot actions. Each action in an action chunk of length  $K$ ,  $a_t \in \mathbb{R}^{K \times D}$ , represents a  $D$ -DoF (degree of freedom) control vector. For policy optimization, each dimension is discretized into 256 bins. Specifically, we learn to sample action chunks using a parameterized VLA policy,  $a \sim \pi_\theta(a | s)$ .
- **Transition function.**  $P : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$  is realized by a parameterized world model  $s_{t+1} \sim p_\phi(s_{t+1} | s_t, a_t)$ , which generates future observations conditioned on past observations and actions. In particular, we sample *imagined trajectories*  $\tau = \{s_0, \hat{s}_1, \dots, \hat{s}_T\}$  by iteratively sampling from the world model  $\hat{s}_{t+1} \sim p_\phi(s_{t+1} | \hat{s}_t, a_t)$  and the VLA policy  $a_t \sim p_\theta(a_t | s_t)$ , given an initial state  $s_0$  sampled in the real environment.
- **Reward function.** The reward is given by a learned model  $R_\psi$  that determines task success from the full trajectory, assigning a binary outcome  $R_\psi(\tau) \in \{0, 1\}$ .

Our objective is to train a policy  $\pi_\theta(a | s)$  such that the predicted cumulative return of the imagined trajectories will be maximized

$$\max_{\theta} \mathbb{E}_{\tau \sim \pi_\theta, p_\phi} [R_\psi(\tau)]. \quad (1)$$

This formulation highlights a general paradigm: RL for VLA can be decoupled from real-world interactions by leveraging a generative world model as the imaginary training ground.

### 3.2 Generative World Model

**Imagined Trajectory Generation.** Given  $c$  initial frames  $I_{0:c}$ , the policy  $\pi_\theta$  takes the most recent  $m$  frames and language instruction  $g$  as input and predicts an action chunk<sup>1</sup>, i.e.,  $a_{i:i+K} \sim \pi_\theta(I_{i-m:i}, g)$ . The world model  $p_\phi$  then conditions on the last  $c$  observed frames and the predicted action chunk to generate the next  $K$  frames:

$$I_{i:i+K} \sim p_\phi(I_{i-c:i}, a_{i:i+K}). \quad (2)$$

Repeating this process until a maximum length  $N$  yields a trajectory  $\tau = \{I_{0:N}, a_{0:N}\}$ . A reward model  $R_\psi$  evaluates the frames in  $\tau$  and outputs a binary label  $y = R_\psi(I_{0:N}) \in \{0, 1\}$ . Thus, each imagined trajectory in the world model is represented as a labeled pair  $(\tau, y)$ , which is then used for policy optimization.

**Model Architecture.** Our world model is based on a video diffusion backbone inherited from OpenSora [35], with modifications designed for simulating robot-object interactions. Specifically, we replace the 3D VAE in OpenSora with the 2D VAE from SDXL [36], which better preserves fine-grained motion details and avoids temporal distortions caused by excessive compression. Consequently, the diffusion process operates in the VAE’s latent space. When applying the imagined trajectory to VLA optimization, we decode the images back into pixel space to better leverage the pretrained knowledge, rather than retraining the VLA in a new latent space such as that from RSSM [18].

Since our world model generates full trajectories in an autoregressive manner—using previously generated frames as conditioning for future predictions—errors can accumulate, leading to severe degradation over long horizons, ultimately leading to failed prediction. To mitigate this issue, we introduce a noisy-frame conditioning technique: during training, conditional frames  $I_{i-m:i}$  are perturbed with diffusion noise at 50/1000 steps rather than kept clean, which improves robustness to imperfect conditioning. As a result, the world model achieves stable long-horizon generation, producing trajectories of hundreds of frames without noticeable quality loss.

To enable precise action conditioning, inspired by Zhu et al. [37], we incorporate a frame-level action control mechanism for better action-frame alignment. Specifically, we extend the AdaLN [38] block to inject both action signals and diffusion timestep embeddings at the frame level. For each action  $a_i$ , an MLP generates modulation coefficients: scale  $\gamma_1^i$  and shift  $\beta_1^i$  for the LayerNorm output, and scale  $\alpha_1^i$  for the residual connection of either the MHA or FFN block. Let  $\mathbf{x}^i$  denote the feature representation at frame  $i$ ; the update rule within each transformer block is given as:  $\mathbf{x}^i = \mathbf{x}^i + (1 + \alpha_1^i) \cdot \text{Block}(\gamma_1^i \cdot \text{LayerNorm}(\mathbf{x}^i) + \beta_1^i)$ .

**Policy Behavior Alignment.** We pretrain the world model on Open X-Embodiment (OXE) [21] trajectories, which offer diverse demonstrations of robot interactions and endow the model with broad knowledge of physical dynamics. However, because OXE trajectories primarily consist of successful executions, failure scenarios are underrepresented in the observation distribution. Likewise, training only on expert demonstrations from downstream tasks leaves the model unable to simulate failures, making the imagined trajectories unsuitable for training. To address this mismatch, we fine-tune the world model on real rollout trajectories collected from the policy itself, thereby adapting it to the downstream (state, action) distribution and capturing failure modes more faithfully. Without this adaptation, the model’s imagination of failure cases remains brittle and unfaithful.

### 3.3 Reward Model

A key requirement for scalable policy optimization in the world model is automatically judging whether an imagined trajectory indicates task success. We construct a lightweight reward model trained on real trajectories. Given a trajectory  $\tau = \{I_{0:N}\}$ , we define a clip of length  $L$  as  $c_i = I_{i-L:i}$ . The terminal clip  $c_N$  of a successful trajectory serves as a positive sample, whereas negatives are drawn from  $\{c_i : L \leq i \leq N - L\}$  of successful trajectories and from arbitrary clips of failed trajectories. To address class imbalance, we balance

<sup>1</sup>To avoid confusion in subscripts, we specify that  $i$  is used as the frame-level subscript and  $t$  as the state-level subscript, where a state often includes multiple frames.  $N$  and  $T$  denote the maximum length of an imagined trajectory at the frame level and state level, respectively.

the number of positive and negative samples within each training batch. The reward model, implemented as a VideoMAE [39] encoder with a linear head, is trained with binary cross-entropy loss. At inference, the model applies a sliding window with stride  $s$  over  $\tau$  to compute the success probability of each clip. A trajectory is classified as successful if any clip exceeds a threshold  $\tau_{\text{thr}}$ , which is selected via validation experiments.

### 3.4 On-Policy Reinforcement Learning with WMPO

Reinforcement learning in VLA tasks faces two key bottlenecks. (1) Physical interaction bottleneck. Unlike LLM settings where feedback is cheap, VLA tasks require repeated rollouts in the physical world, which incur high hardware costs, safety concerns, and limited scalability. (2) Value estimation bias by off-policy RL. Given the aforementioned physical constraints, existing real-world RL methods often consider off-policy RL methods [8, 40–42]. However, off-policy methods naturally cause value estimation errors [43] and on-policy methods are often favorable for better performance.

To overcome these challenges, we optimize policies entirely within a world model: replacing costly real-world rollouts with model-generated trajectories alleviates reliance on physical interaction and enables scalable online learning. We adopt Group Relative Policy Optimization (GRPO) as the policy optimization algorithm, since it provides stable and scalable training in settings with sparse rewards. In our case, state transitions are simulated by the world model (Eq. 2), and rewards are binary with  $R_\psi(\tau) \in \{0, 1\}$  depending on task success.

**Trajectory Sampling.** From each initial frames  $I_{0:c}$  sample from real environment  $\mathcal{D}$ , we sample a group of imagined trajectories  $\{\tau_1, \dots, \tau_G\}$  from current policy  $\pi_{\theta_{\text{old}}}$  inside the world model. The reward model is then employed to predict the success or failure of each trajectory. To mitigate vanishing gradients, we adopt a *Dynamic Sampling* strategy [44]: if all trajectories in a group are predicted to be successful or all unsuccessful, the group is discarded and additional rollouts are sampled until the batch is fully populated. The log-probability of each action chunk under  $\pi_{\theta_{\text{old}}}$  is pre-computed as reference:

$$\log \pi_{\theta_{\text{old}}}(a_t | s_t) = \sum_{i=1}^K \sum_{j=1}^D \log \pi_{\theta_{\text{old}}}(a_t^{i,j} | s_t), \quad (3)$$

where  $a_t$  denotes the action chunk at time  $t$ , and  $a_t^{i,j}$  represents the  $i$ -th action in the  $j$ -th DoF.

**Policy Update.** Following DAPO [44], we remove the KL divergence regularization so that no reference model is required during training, thereby reducing memory consumption and encouraging the policy to explore novel behaviors. The final training objective is given by

$$\mathcal{J}(\theta) = \mathbb{E}_{s_0 \sim \mathcal{D}, \{\tau_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}} \left[ \frac{1}{G} \sum_{i=1}^G \frac{1}{T} \sum_{t=0}^T \min(r_{i,t}(\theta) \hat{A}_i, \text{clip}(r_{i,t}(\theta), 1 - \epsilon_{\text{low}}, 1 + \epsilon_{\text{high}}) \hat{A}_i) \right], \quad (4)$$

with

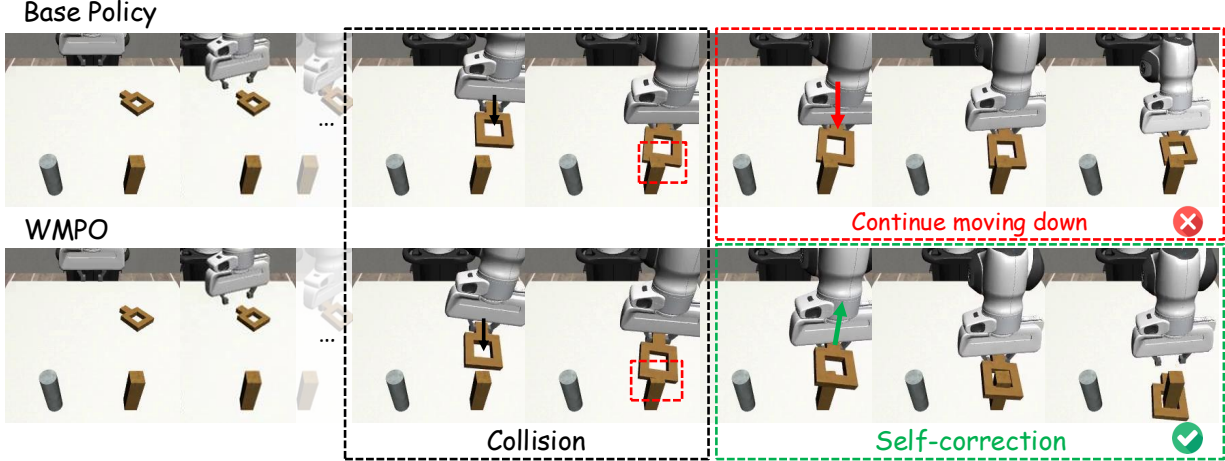
$$r_{i,t}(\theta) = \frac{\pi_\theta(a_{i,t} | s_{i,t})}{\pi_{\theta_{\text{old}}}(a_{i,t} | s_{i,t})}, \quad \hat{A}_i = \frac{R_i - \text{mean}(\{R_i\}_{i=1}^G)}{\text{std}(\{R_i\}_{i=1}^G)}. \quad (5)$$

Here  $r_{i,t}(\theta)$  is the probability ratio between new and old policies at step  $t$  of trajectory  $\tau_i$ ,  $R_i = R(\tau_i)$ , and  $\hat{A}_i$  is the normalized advantage of trajectory  $\tau_i$  over the horizon  $N$ . The overall training pipeline is detailed in Algorithm 1 in Appendix A.

## 4 Experiments

We conduct extensive experiments to evaluate the effectiveness of WMPO, focusing on the following questions: (1) can WMPO outperform online and offline RL in simulation environments; (2) how does the behavior of WMPO differ from imitation learning; (3) can WMPO generalize to unseen settings; (4) can WMPO achieve iterative improvement during deployment; and (5) can WMPO be applied on a real robot?





**Figure 3** Behavior analysis of the *Square* task (inserting the square into the stick) shows that, compared with the base policy, WMPO demonstrates the ability to self-correct.

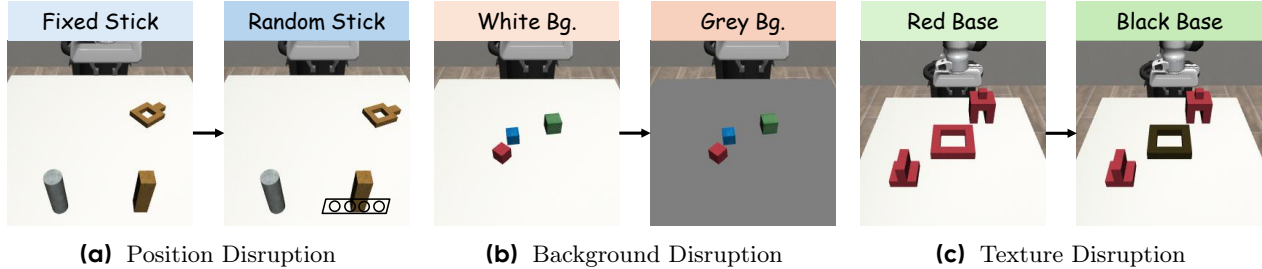
## 4.1 Experiment Settings

**Implementation Details** In this work, we fine-tune OpenVLA-OFT [24] via imitation learning on target manipulation tasks as our base policy. For simplicity, we omit the robot proprioceptive state and wrist camera inputs, and set the action chunk length  $K$  to 8. We collect  $P$  real trajectories using the base policy, and conduct experiments with  $P = 128$  and  $P = 1280$  to evaluate the scalability of our approach. These trajectories are further used to fine-tune a world model, which predicts the next  $K = 8$  frames given  $c = 4$  conditioning frames and one action chunk. In addition, we train reward models on the collected trajectories, using a video clip length of  $L = 8$  and a stride of 1 during evaluation. Further training details are provided in Appendix A.

**Simulation Environment Details** We conduct experiments in the Mimicgen simulation [23]. We select four fine-grained manipulation tasks: *Coffee\_D0*, *StackThree\_D0*, *ThreePieceAssembly\_D0*, and *Square\_D0*, and fine-tune the OpenVLA-OFT model with 300 expert trajectories per task as the base policy. For evaluation, we test 128 different initial states for each task and report the average success rate.

**Table 1** Comparison of policy optimization methods across four manipulation tasks in the Mimicgen simulation benchmark.  $P$  denotes the rollout budget, i.e., the number of full real trajectories available for policy optimization. Results show that WMPO consistently outperforms both GRPO and DPO baselines under different budgets. As the rollout budget increases from 128 to 1280, WMPO continues to exhibit substantial improvements, highlighting both its data efficiency and scalability. Performance is reported as the task success rate (%).

Rollout budget $P$	Methods	Coffee	StackThree	ThreePieceAssembly	Square	Mean (%)
–	<i>Base policy</i>	43.8	46.9	19.5	24.2	33.6
128	GRPO	38.3	52.3	17.2	25.0	33.2
	DPO	43.8	53.9	23.4	28.1	37.3
	Ours	<b>61.7</b>	<b>56.3</b>	<b>37.5</b>	<b>32.8</b>	<b>47.1</b>
1280	GRPO	47.7	54.7	20.3	25.8	37.1
	DPO	52.3	57.0	26.7	33.6	42.4
	Ours	<b>75.0</b>	<b>64.1</b>	<b>46.1</b>	<b>45.3</b>	<b>57.6</b>



**Figure 4** (a) For the *Square* task, we vary the stick’s position from fixed to a random position inside a rectangle. (b) For the *StackThree* task, we substitute the tabletop background with a gray background. (c) For the *ThreePieceAssembly* task, we substitute the red base with a dark wooden base.

## 4.2 Comparison Experiments

We compare WMPO with two established RL algorithms, GRPO [30] and DPO [28], both widely used for optimizing large language models. To ensure fairness, all methods are allocated the same real rollout budget  $P$ . We consider both online and offline baselines: GRPO is implemented in an online setting, where the policy is updated directly from trajectories collected in the environment; DPO is implemented in an offline setting, where the base policy serves as the reference and trajectory pairs (success vs. failure) are constructed for optimization using the standard DPO loss. Unlike GRPO, which discards trajectories after each update, DPO can repeatedly reuse collected data, but it lacks the ability to update the policy online as WMPO does. More implementation details are provided in Appendix B.

As shown in Tab. 1, WMPO consistently outperforms all baselines across all tasks. With a small rollout budget of  $P=128$ , it already surpasses the strongest baseline by +9.8 points, demonstrating strong data efficiency. When the budget increases to  $P=1280$ , the margin further expands to +15.2 points on average, indicating that WMPO leverages additional trajectories more effectively than existing methods and scales robustly with more rollouts. In contrast, GRPO often underperforms with limited updates, and DPO plateaus due to static data reuse, whereas WMPO continues to improve steadily as the rollout budget grows. Furthermore, we evaluate the reward model and find that it achieves an F1 score above 0.95 across all tasks, reliably distinguishing success from failure and effectively mitigating reward hacking. These results highlight the effectiveness and scalability of WMPO for policy optimization in robotic manipulation.

## 4.3 Emergent Behavior of WMPO

To better understand the source of WMPO’s strong performance, we conduct a visual comparison of its test-time behavior against the base policy. We identify two emergent behaviors unique to WMPO: (1) the WMPO policy learns to self-correct, recovering from nearly failure states; and (2) the WMPO policy executes tasks more efficiently, as it rarely becomes “stuck” in suboptimal states.

First, taking the *Square* task as an illustrative example (see Fig. 3), we observe that when both the base policy and WMPO deviate from the correct trajectory due to error accumulation and encounter a collision, their behaviors diverge. The baseline policy, trained only on expert demonstrations, has never observed collisions during training; it continues to push the square against the stick until the maximum time horizon is reached, resulting in failure. In contrast, WMPO benefits from large-scale imagined trajectories generated by the world model, enabling it to learn the self-correction behaviors, which is challenging to obtain through imitation learning alone. Specifically, the policy autonomously learns to lift the square, realign it, and then insert it correctly, ultimately succeeding in the task. Second, we analyzed the lengths of successful trajectories generated by different policies, as shown in Fig. 5. We found that the trajectories of policies trained with WMPO are significantly shorter. This is because WMPO discourages stuck behaviors, which often result in failures due to timeouts. As a side benefit, WMPO also makes the policy’s behavior faster and smoother.

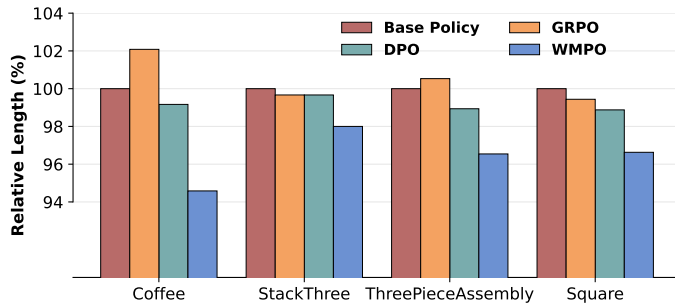
## 4.4 Generalization to Novel Tasks



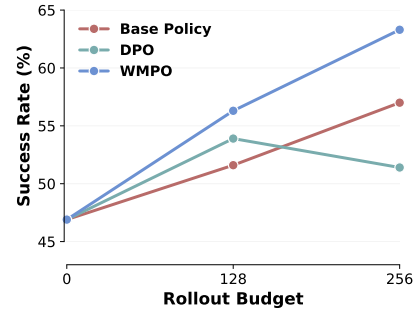
In this section, we evaluate the generalization ability of WMPO across three novel disruption scenarios (Fig. 4), which systematically assess generalization under spatial, background, and texture shifts. As shown in Tab. 2, WMPO consistently achieves the best performance across all disruption types. DPO attains modest improvements in the in-distribution setting compared to the base policy, but its performance degrades significantly under background and texture changes, suggesting reliance on spurious visual cues rather than transferable manipulation skills. GRPO exhibits performance similar to the base policy, and both are worse than WMPO across all disruption scenarios. In contrast, WMPO, trained entirely in the world model, captures more generalizable strategies and maintains reliable performance across spatial, background, and texture variations.

**Table 2** We evaluate each policy in its corresponding disruption scenario and report the success rate (%).

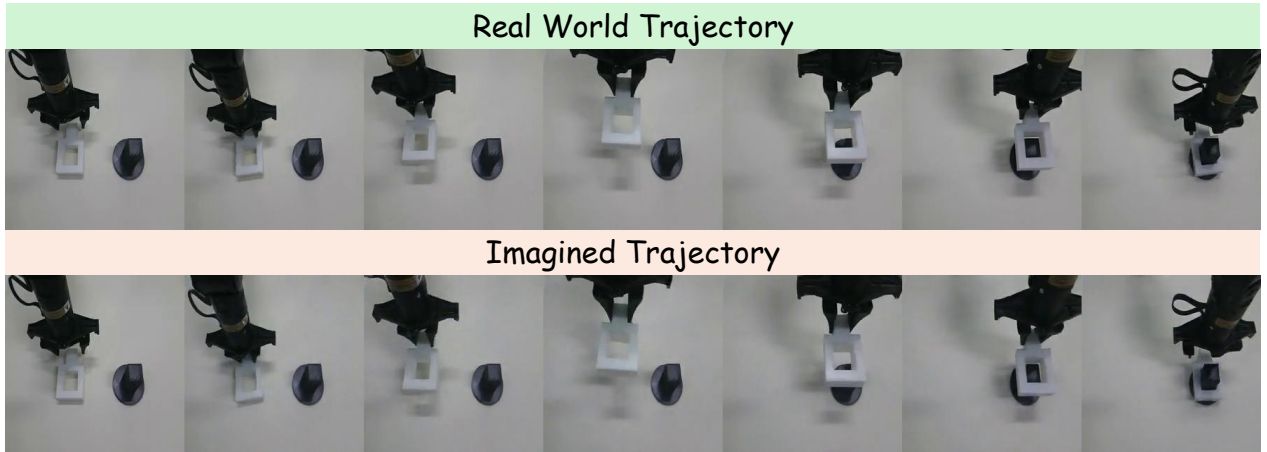
Methods	Pos. Dis.	Bg. Dis.	Tex. Dis.	Mean
Base policy	14.1	46.1	10.9	23.7
GRPO	15.6	47.7	10.9	24.7
DPO	16.4	34.4	7.8	19.5
<b>Ours</b>	<b>22.3</b>	<b>50.0</b>	<b>16.4</b>	<b>29.6</b>



**Figure 5** Relative average trajectory length of successful trials across different policies (Base Policy = 100%).



**Figure 6** Lifelong learning results of WMPO and baselines.



**Figure 7** Real-world experiments on the fine-grained manipulation task “Insert the square into the stick” where the clearance between the square and the stick is only 5 mm. The top row shows the real-world trajectory of the base policy executed in the real world, while the bottom row depicts the corresponding imagined trajectory starting from the same initial state within our learned world model. Despite never observing this trajectory during training, the world model accurately predicts the future evolution, demonstrating its ability to capture precise task dynamics.

## 4.5 Lifelong Learning

In this section, we demonstrate that WMPO can continuously improve the performance of VLA by iteratively collecting real trajectories from the environment. Specifically, we iteratively collect  $P = 128$  real trajectories, perform WMPO to optimize the policy, and then use the updated policy to collect another  $P$  real trajectories.

We apply the same setting to the DPO baseline. To compare WMPO with an imitation learning-based policy using more expert demonstrations, we leverage 300, 428, and 556 expert trajectories to train the base policy as a reference. It is important to note that the base policy requires human-collected trajectories, whereas WMPO only relies on trajectories collected by the policy itself, making it more scalable. The results on the *StackThree* task, shown in Fig. 6, demonstrate that WMPO achieves stable and substantial improvements over both baselines, whereas DPO fails to improve iteratively due to unstable training.

## 4.6 Real-world Experiments

In this section, we evaluate the challenging real-world manipulation task, "Insert the square into the stick" (see Fig. 7, more cases including failure could be found in Appendix C), to validate the effectiveness of WMPO. Using the Cobot Mobile ALOHA platform, we collect 200 high-quality expert demonstrations to fine-tune the OpenVLA-OFT model as the base policy. We then deploy this policy to collect an additional 128 trajectories, which are used to further fine-tune the world model and optimize the policy within it. For comparison, we also train an offline DPO policy using the same dataset. All models are evaluated under identical experimental conditions, and we report the average success rate over 30 trials. The results show that the base policy, DPO, and WMPO achieve success rates of 53%, 60%, and 70%, respectively, demonstrating the effectiveness of WMPO on real robots.

## 5 Conclusion

In this work, we introduced WMPO, a novel framework for on-policy RL of VLA models. By grounding policy optimization entirely in a video-generative world model, WMPO eliminates the need for costly real-world interactions while maintaining consistency with pretrained VLA representations. Through policy behavior alignment, robust autoregressive video generation, and a lightweight reward model, WMPO enables scalable training with strong sample efficiency. Extensive experiments in both simulation and real-world settings demonstrated that WMPO (i) consistently outperforms state-of-the-art model-free baselines, (ii) exhibits emergent self-correction behavior, (iii) generalizes reliably to unseen scenarios, and (iv) supports iterative lifelong learning. Together, these findings highlight WMPO as a scalable and generalizable paradigm for advancing VLA RL.

## References

- [1] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.
- [2] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.
- [3] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al.  $\pi_{0.5}$ : a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.
- [4] Fanqi Lin, Yingdong Hu, Pingyue Sheng, Chuan Wen, Jiacheng You, and Yang Gao. Data scaling laws in imitation learning for robotic manipulation, 2024. URL <https://arxiv.org/abs/2410.18647>.
- [5] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, pages 627–635, 2011.
- [6] Qiyang Li, Zhiyuan Zhou, and Sergey Levine. Reinforcement learning with action chunking, 2025. URL <https://arxiv.org/abs/2507.07969>.
- [7] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *arXiv preprint arXiv:2410.21845*, 2024.
- [8] Yuhui Chen, Shuai Tian, Shugao Liu, Yingting Zhou, Haoran Li, and Dongbin Zhao. Conrft: A reinforced fine-tuning method for vla models via consistency policy. *arXiv preprint arXiv:2502.05450*, 2025.
- [9] Wenke Xia, Yichu Yang, Hongtao Wu, Xiao Ma, Tao Kong, and Di Hu. Robotic policy learning via human-assisted action preference optimization, 2025. URL <https://arxiv.org/abs/2506.07127>.
- [10] Guanxing Lu, Wenkai Guo, Chubin Zhang, Yuheng Zhou, Haonan Jiang, Zifeng Gao, Yansong Tang, and Ziwei Wang. Vla-rl: Towards masterful and general robotic manipulation with scalable reinforcement learning, 2025. URL <https://arxiv.org/abs/2505.18719>.
- [11] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, Dehui Wang, Dingxiang Luo, Yuchen Fan, Youbang Sun, Jia Zeng, Jiangmiao Pang, Shanghang Zhang, Yu Wang, Yao Mu, Bowen Zhou, and Ning Ding. Simplevla-rl: Scaling vla training via reinforcement learning, 2025. URL <https://arxiv.org/abs/2509.09674>.
- [12] NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezani, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchampi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL <https://arxiv.org/abs/2501.03575>.
- [13] Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttmore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu,

- Demis Hassabis, Zoubin Ghahramani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025.
- [14] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In 2017 IEEE International Conference on Robotics and Automation (ICRA), page 2786–2793. IEEE Press, 2017. doi: 10.1109/ICRA.2017.7989324. URL <https://doi.org/10.1109/ICRA.2017.7989324>.
  - [15] 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, pages 2555–2565. PMLR, 2019.
  - [16] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. ICLR 2020, 2020.
  - [17] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193, 2020.
  - [18] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104, 2023.
  - [19] Xiao Ma, David Hsu, and Wee Sun Lee. Learning latent graph dynamics for visual manipulation of deformable objects. In 2022 International Conference on Robotics and Automation (ICRA), pages 8266–8273. IEEE, 2022.
  - [20] Bernd Frauenknecht, Devdutt Subhaskar, Friedrich Solowjow, and Sebastian Trimpe. On rollouts in model-based reinforcement learning. In The Thirteenth International Conference on Learning Representations, 2025. URL <https://openreview.net/forum?id=Uh5GRmLlvt>.
  - [21] Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. <https://arxiv.org/abs/2310.08864>, 2023.
  - [22] DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>.
  - [23] 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. arXiv preprint arXiv:2310.17596, 2023.
  - [24] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025. URL <https://arxiv.org/abs/2502.19645>.
  - [25] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al.  $\pi_0$ : A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164, 2024.
  - [26] Jianlan Luo, Zheyuan Hu, Charles Xu, You Liang Tan, Jacob Berg, Archit Sharma, Stefan Schaal, Chelsea Finn, Abhishek Gupta, and Sergey Levine. Serl: A software suite for sample-efficient robotic reinforcement learning. In IEEE International Conference on Robotics and Automation (ICRA), pages 16961–16969. IEEE, 2024.
  - [27] Zijian Zhang, Kaiyuan Zheng, Zhaorun Chen, Joel Jang, Yi Li, Siwei Han, Chaoqi Wang, Mingyu Ding, Dieter Fox, and Huaxiu Yao. Grape: Generalizing robot policy via preference alignment. arXiv preprint arXiv:2411.19309, 2024.
  - [28] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. Proceedings of Advances in Neural Information Processing Systems (NeurIPS), 2023.
  - [29] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.
  - [30] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL <https://arxiv.org/abs/2402.03300>.
  - [31] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024. URL <https://openreview.net/forum?id=NadTwTODgC>.

- [32] Zhenan Jiang, Kai Liu, Yuxin Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Chao Yu, Haoran Li, and Dongbin Zhao. World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation, 2025. URL <https://arxiv.org/abs/2509.19080>.
- [33] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International conference on machine learning*, pages 2117–2126. PMLR, 2018.
- [34] Xiao Ma, Siwei Chen, David Hsu, and Wee Sun Lee. Contrastive variational reinforcement learning for complex observations. In *Conference on robot learning*, pages 959–972. PMLR, 2021.
- [35] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. URL <https://arxiv.org/abs/2412.20404>.
- [36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=di52zR8xgf>.
- [37] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation, 2025. URL <https://arxiv.org/abs/2406.14540>.
- [38] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization, 2019. URL <https://arxiv.org/abs/1911.07013>.
- [39] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.
- [40] Stephen James, Kentaro Wada, Tristan Laidlow, and Andrew J Davison. Coarse-to-fine q-attention: Efficient learning for visual robotic manipulation via discretisation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13739–13748, 2022.
- [41] Younggyo Seo, Jafar Uruc, and Stephen James. Continuous control with coarse-to-fine reinforcement learning. arXiv preprint arXiv:2407.07787, 2024.
- [42] Andrew Wagenmaker, Mitsuhiko Nakamoto, Yunchu Zhang, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. arXiv preprint arXiv:2506.15799, 2025.
- [43] Seohong Park, Kevin Frans, Deepinder Mann, Benjamin Eysenbach, Aviral Kumar, and Sergey Levine. Horizon reduction makes rl scalable. arXiv preprint arXiv:2506.04168, 2025.
- [44] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Weinan Dai, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm reinforcement learning system at scale, 2025. URL <https://arxiv.org/abs/2503.14476>.
- [45] Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl, 2025. URL <https://arxiv.org/abs/2505.05470>.



# Appendix

## A Training Details

In this section, we provide the training details of WMPO. The supervised finetuning of OpenVLA-OFT is conducted on 8 H100 GPUs. We exclude robot proprioceptive states and wrist camera inputs for simplicity. Instead of applying L1 regression, we adopt an output head that predicts discrete action tokens in parallel. The subsequent world model training and policy optimization are performed on 32 H100 GPUs.

The world model is first pretrained on the Open X-Embodiment (OXE) dataset [21] and then fine-tuned on downstream policy behavior data. Table 3 summarizes the hyperparameters for training the world model, while the remaining settings follow OpenSora [35]. The hyperparameters for policy optimization with GRPO are provided in Tab. 4, with other settings inherited from Li et al. [11].

---

### Algorithm 1: World Model Policy Optimization

---

**Input** : World model  $p_\phi$ , reward model  $R_\psi$ , VLA policy  $\pi_\theta$ ; group size  $G$ ; batch size  $B$  (trajectories); mini-batch size  $M$  ( $M \mid B$ ); epochs  $E$ ; horizon  $T$

**Output**: Updated policy parameters  $\theta$

$\theta_{\text{old}} \leftarrow \theta$

**while** not converged **do**

$\mathcal{B} \leftarrow \emptyset$

**while**  $|\mathcal{B}| < B$  **do**

        Sample initial state  $s_0 = (I_{0:c}, g) \sim \mathcal{D}$

**for**  $i = 1$  **to**  $G$  **do**

            Imagine  $\tau_i = \{(s_t^i, a_t^i)\}_{t=0}^T \sim p_\theta(\tau_i \mid s_0, \pi_{\theta_{\text{old}}})$

$R_i \leftarrow R(\tau_i) \in \{0, 1\}$

**if** all( $R_1, \dots, R_G$ ) **or** none( $R_1, \dots, R_G$ ) **then**

**continue**

$\mu \leftarrow \frac{1}{G} \sum_{i=1}^G R_i, \quad \sigma \leftarrow \sqrt{\frac{1}{G} \sum_{i=1}^G (R_i - \mu)^2}$

**for**  $i = 1$  **to**  $G$  **do**

$\hat{A}_i \leftarrow (R_i - \mu)/\sigma$

            Precompute  $\{\log \pi_{\theta_{\text{old}}}(a_t^i \mid s_t^i)\}_{t=0}^T$

            Append  $(\tau_i, \hat{A}_i)$  to  $\mathcal{B}$

**if**  $|\mathcal{B}| \geq B$  **then**

**break**

**for**  $e = 1$  **to**  $E$  **do**

**for**  $j = 1$  **to**  $B/M$  **do**

$\mathcal{M} \leftarrow j$ -th contiguous block of  $M$  trajectories from  $\mathcal{B}$

            Update  $\theta$  according to Eq. 4, where  $r_{i,t}(\theta) = \exp(\log \pi_\theta(a_{i,t} \mid s_{i,t}) - \log \pi_{\theta_{\text{old}}}(a_{i,t} \mid s_{i,t}))$

$\theta_{\text{old}} \leftarrow \theta$

---

## B Baseline Details

We provide the implementation details of our baselines, including online GRPO and offline DPO. For the GRPO baseline, the main hyperparameters are inherited from Tab. 4. We observe that the batch size has a significant impact on performance. Larger batch sizes (e.g., 64) yield more stable improvements; however, they require a substantial number of real trajectories. Specifically, a single model update requires at least  $64 \times 8 = 512$  real trajectories. Moreover, dynamic sampling further filters out groups with a success rate of 0 or 1. As a result, when the rollout budget is  $P = 128$ , such large batch sizes are infeasible, and even with  $P = 1280$ , the model can only be updated once or twice. Therefore, we additionally experimented with a

smaller batch size of 8, scaling the learning rate down proportionally (by a factor of 8). We report the best results obtained across both configurations (batch size = 8 and = 64).

For the DPO baseline, we follow the standard preference-based offline training paradigm. Specifically, we construct a preference dataset from trajectories collected by the supervised fine-tuned OpenVLA-OFT model and use it to optimize the policy with the DPO objective. The model architecture and optimizer settings are kept consistent with those used in GRPO for fair comparison. All baselines were trained under the same rollout budgets and evaluation protocols as WMPO to ensure fairness.

## C Real World Cases

In this section, we provide additional examples of trajectory predictions made by the world model in real-world settings. Fig 8 illustrates cases where the world model successfully predicts failure trajectories: it has learned that when the square and the stick are misaligned, the square cannot be inserted into the stick. In contrast, Fig 9 shows a failure case where the model does not correctly predict a failed trajectory. Although predictions remain accurate until the final frame, subtle perturbations prevent the model from faithfully capturing the moment when the square gets stuck in the stick. Empirically, such failures are relatively rare on the validation set, indicating that the world model can reliably predict both successful and failed outcomes in the vast majority of scenarios, which is crucial for stable policy optimization.

## D Limitation

While the WMPO framework can in principle support flow-based policies, this work focuses on discretized action representations. As future work, we plan to extend WMPO to more expressive policy classes, such as flow-matching based policies [25], and explore policy optimization with FlowGRPO [45], thereby broadening its applicability across diverse action spaces.

## E LLM Usage Statement

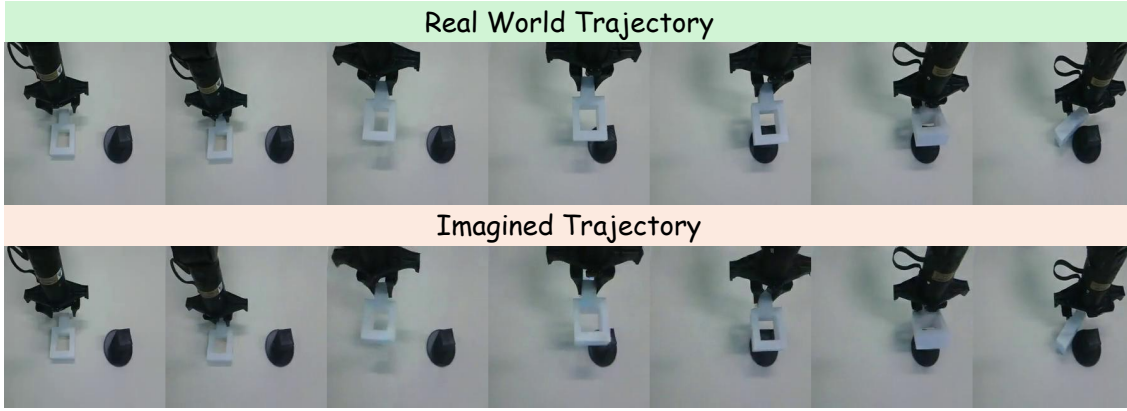
GPT-5 was used solely to assist with language refinement and stylistic polishing of the manuscript. The authors confirm that all scientific ideas, study design, data analyses, and conclusions presented in this work are entirely their own. The LLM did not contribute to the generation of research concepts, execution of experiments, or interpretation of findings.

Hyperparameter	Value
Optimizer	AdamW( $\beta = 0.9, \beta = 0.999$ )
Learning rate	0.0001
Batch size	128
Gradient clip	0.1
Pretrain training steps	12,000,000
Fine-tune training steps	3,000,000
EMA	0.9999
Weight decay	0.0
Prediction target	$\epsilon$

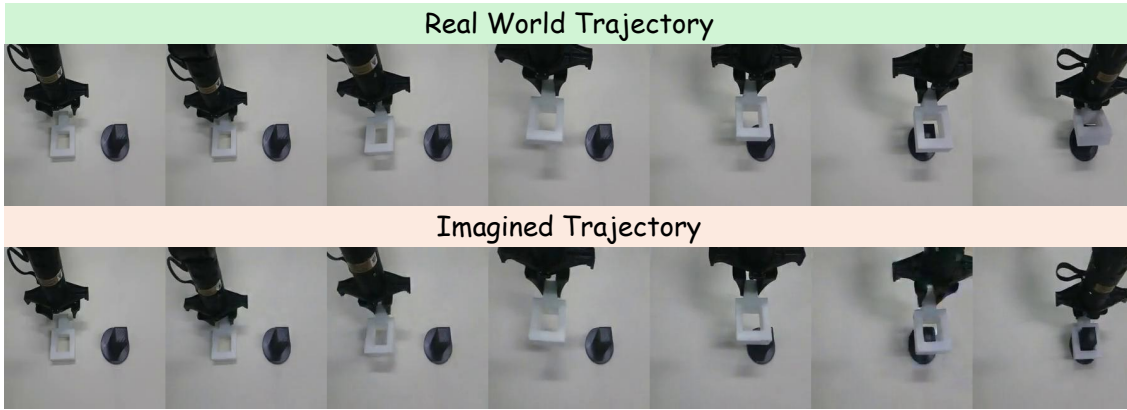
**Table 3** Hyperparameters for training the world model.

Hyperparameter	Value
Optimizer	AdamW( $\beta = 0.9, \beta = 0.999$ )
Learning rate	$5 \times 10^{-6}$
Training batch size	64
Group size $G$	8
Mini-batch size	128
Clip ratio $\epsilon_{low}$	0.20
Clip ratio $\epsilon_{high}$	0.28
Temperature	1.6

**Table 4** Hyperparameters for the GRPO algorithm.



**Figure 8** Real-world experiments on the fine-grained manipulation task “Insert the square into the stick”. The top row shows the rollout trajectory of the base policy executed in the real world, while the bottom row depicts the corresponding imagined trajectory starting from the same initial state within our learned world model. The world model successfully predicted failure cases.



**Figure 9** Example of a failure case. Although the predicted trajectory remains accurate until the final frame, the model fails to capture the square getting stuck in the stick due to subtle perturbations.