WISE: WORLD-MODEL-GUIDED IMAGINATION SCHEDULING FOR EFFICIENT POST-TRAINING OF VISION-LANGUAGE-ACTION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-training VLA policies typically rely on supervised fine-tuning with costly expert demonstrations or reinforcement learning with expensive and potentially unstable real-world exploration. World models offer a promising alternative by evaluating candidate behaviors through imagined futures, yet effective post-training requires more than accurate prediction: imagination must be scheduled where it is useful, bounded to limit prediction errors, and translated into reliable policy-learning signals. In robotic manipulation, the value of imagination varies substantially across execution stages, while extended rollouts can accumulate prediction errors and introduce unreliable learning signals. We introduce **WISE (World-model-guided Imagination Scheduling for Efficient Post-training of Vision-Language-Action Models)**, a unified framework that coordinates when and how world-model imagination is used during policy refinement. WISE selectively invokes imagination at interaction-relevant states, performs bounded multi-view rollouts, evaluates candidate futures using progress and completion signals, and uses their relative outcomes to refine actions generated from real interaction contexts. Extensive experiments with both π_0 and $\pi_{0.5}$ demonstrate consistent improvements across diverse manipulation tasks while reducing GPU computation time by approximately 80% compared with full imagination. Real-world evaluations further show substantial gains in robustness and generalization under diverse real-world distribution shifts.

1 INTRODUCTION

Vision-Language-Action (VLA) models have emerged as a promising foundation for general-purpose robotic manipulation by integrating visual perception, language understanding, and action generation (Brohan et al., 2023b;a; Kim et al., 2025b; Black et al., 2026). Leveraging large-scale vision-language pretraining, recent VLA models such as the π_0 family (Intelligence et al., 2025b;a) exhibit strong semantic understanding and generalization capabilities. However, adapting pretrained policies to specific environments and interaction patterns remains challenging (Ma et al., 2026). Imitation learning-based post-training (Ren et al., 2024) relies on large amounts of high-quality robot demonstrations, which are expensive and difficult to collect. Reinforcement learning provides an alternative through trial-and-error policy improvement (Shao et al., 2024; Wagenmaker et al., 2025; Lu et al., 2025; Luo et al., 2025). However, extensive real-world exploration incurs substantial interaction costs and may cause collisions, hardware wear, or damage during trial-and-error, as illustrated in Figure 1(a) (Hou et al., 2026). These limitations motivate post-training strategies that can improve pretrained VLAs without repeatedly executing exploratory behaviors in the physical world.

World models (Hafner et al., 2019; 2024; Bruce et al., 2024) offer a promising alternative by predicting the consequences of candidate actions before physical execution. Recent approaches have begun to exploit imagined rollouts for robot policy refinement (Jiang et al., 2026; Zhu et al., 2025; Li et al., 2025). However, effective world-model-guided post-training depends not only on the fidelity of predicted futures, but also on how imagination is allocated and translated into policy supervision (Hou et al., 2026). In our simulated manipulation experiments, the distribution of failures across task stages in Figure 1(c) shows that pretrained VLAs fail predominantly during grasp and operation,

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.

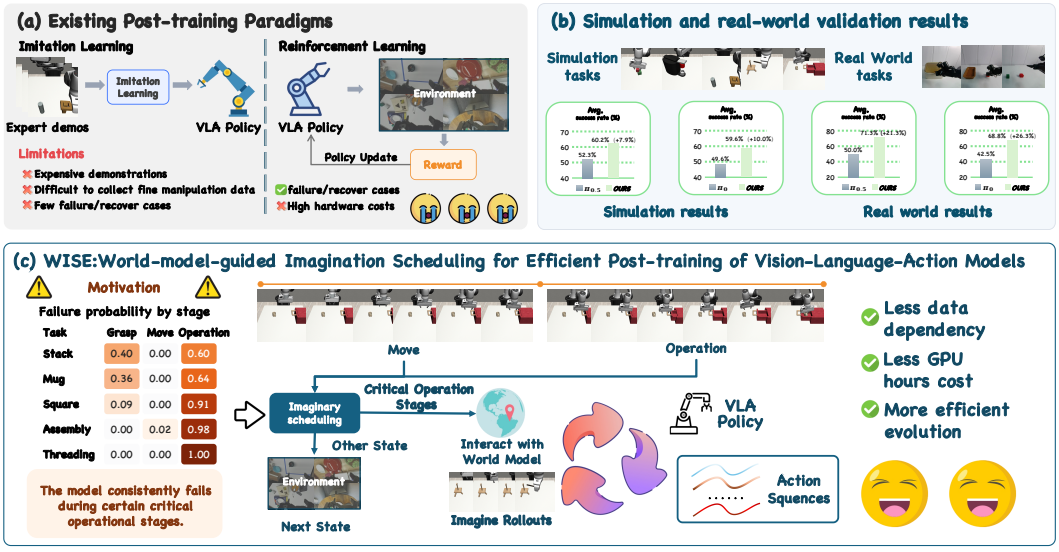

Figure 1: Motivation and key idea of WISE. Existing VLA post-training methods rely on costly demonstrations or extensive real-world exploration, whereas WISE coordinates world-model imagination to provide efficient and reliable policy-learning signals.

while coarse motion rarely accounts for failures, suggesting that full-trajectory imagination is often unnecessary. The utility of imagination is therefore highly non-uniform across execution stages. Meanwhile, long imagined rollouts accumulate prediction errors and can degrade policy supervision. Thus, more imagination is not necessarily more useful. Effective post-training requires focusing imagination on failure-prone interaction stages, bounding rollout horizons, and converting predicted outcomes into reliable learning signals.

To this end, we propose **WISE (World-model-guided Imagination Scheduling for Efficient Post-training of Vision-Language-Action Models)**, a unified framework that coordinates where and how world-model imagination supports VLA policy refinement (Figure 1(c)). At interaction-relevant states, WISE generates multiple candidate action chunks and uses an action-conditioned multi-view world model (Zheng et al., 2024) to perform bounded prediction of their future consequences. A reward model built upon visual representations (Oquab et al., 2024) assesses imagined trajectories through progress and completion signals, while their relative outcomes provide feedback for group-relative policy optimization (Shao et al., 2024). Importantly, multi-step imagination is used for counterfactual evaluation rather than direct supervision on synthetic state-action pairs.

We evaluate WISE on five MimicGen manipulation tasks (Mandlekar et al., 2023) and real-world robotic manipulation scenarios using both π_0 and $\pi_{0.5}$. WISE consistently improves both pretrained VLA backbones across diverse tasks and generalization conditions. Compared with full imagination, WISE reduces world-model inference calls by approximately 80% while achieving higher policy performance, demonstrating the effectiveness and efficiency of selective imagination scheduling.

Our contributions are summarized as follows: We propose **WISE**, a unified world-model-guided VLA post-training framework that coordinates interaction-relevant state selection, bounded future prediction, trajectory evaluation, and real-context policy refinement to enable more reliable and efficient use of imagination for policy improvement. Extensive experiments with π_0 and $\pi_{0.5}$ in simulation and the real world demonstrate consistent performance gains, substantially reduced world-model computation, and improved robustness under diverse real-world distribution shifts.

2 Related Work

2.1 Vision-Language-Action Models

Vision-Language-Action (VLA) models provide a promising paradigm for general-purpose robot policies by integrating multimodal understanding with action generation. Early works, including

RT-1 (Brohan et al., 2023b), RT-2 (Brohan et al., 2023a), RoboFlamingo (Li et al., 2024), and VIMA (Jiang et al., 2023), demonstrate the potential of large-scale multimodal pretraining for robotic manipulation. Open X-Embodiment and RT-X (Collaboration et al., 2025) further scale robot learning across diverse embodiments, while Octo and OpenVLA (Ghosh et al., 2024; Kim et al., 2025b) improve policy scalability and generalization through large-scale robot data. More recent models, such as the π_0 family (Black et al., 2026; Intelligence et al., 2025b;a), further strengthen action generation and general-purpose manipulation. Despite this progress, adapting pretrained VLAs to specific environments and interaction patterns remains an important challenge.

2.2 Post-training of Vision-Language-Action Models

Post-training enables pretrained VLA policies to adapt beyond their original training distributions. Imitation-learning-based approaches (Celemin et al., 2022; Ren et al., 2024; Torabi et al., 2018; Zhao et al., 2023) rely on task-specific demonstrations, but collecting diverse and high-quality robot data remains costly. Reinforcement learning (Tang et al., 2024; Shao et al., 2024; Schulman et al., 2017) instead improves policies through environment interaction. Recent VLA post-training methods (Luo et al., 2025; Xu et al., 2026), including VLA-RL (Lu et al., 2025), iRe-VLA (Guo et al., 2025), ConRFT (Chen et al., 2025), and other VLA-specific optimization approaches (Kim et al., 2025a), demonstrate substantial gains beyond supervised imitation. However, these methods often remain dependent on costly environment interaction and reliable reward signals, motivating more efficient sources of policy supervision.

2.3 World-Model-Guided Robot Learning

World models enable agents to predict environment dynamics and improve policies through imagined interactions (Hafner et al., 2019; 2024). They have been extended to physical robot learning for policy optimization, data generation, and reinforcement post-training (Wu et al., 2022; Yang et al., 2023; Jiang et al., 2026). More recently, WMPO, World4RL, and VLA-RFT (Zhu et al., 2025; Jiang et al., 2026; Li et al., 2025) exploit imagined rollouts to refine pretrained VLA policies, while AtomVLA (Sun et al., 2026) introduces predictive latent world models with task-oriented evaluation and RISE (Yang et al., 2026) explores offline imagined interactions with compositional world models. These approaches primarily focus on generating and exploiting imagined trajectories. In contrast, WISE focuses on how world-model imagination should be selectively allocated and reliably incorporated into VLA post-training, rather than treating imagined rollout generation as a uniformly applied component.

3 Method

Figure 2 illustrates the overall framework of WISE. Starting from a pretrained VLA policy, WISE coordinates world-model-guided post-training through an imagination–evaluation–optimization loop. During interaction, imagination is scheduled at interaction-relevant states, where multiple candidate actions are evaluated through bounded world-model rollouts. Their predicted outcomes are converted into relative feedback for policy refinement, while updates remain grounded in real interaction contexts. This design jointly controls when imagination is used, how imagined rollouts are bounded, and how it contributes to policy learning. Detailed training configurations and implementation settings are provided in Appendix A.

3.1 Problem Formulation

We consider the post-training of a pretrained Vision-Language-Action (VLA) policy π_{θ_0} for robotic manipulation. Let $h_t = (\mathbf{O}_t, \mathbf{s}_t, \ell)$ denote the real interaction context, including recent multi-view observations, robot state, and language instruction. The policy predicts an action chunk as

$$\mathbf{a}_t \sim \pi_\theta(\cdot \mid h_t).$$

Our objective is to maximize the expected task success:

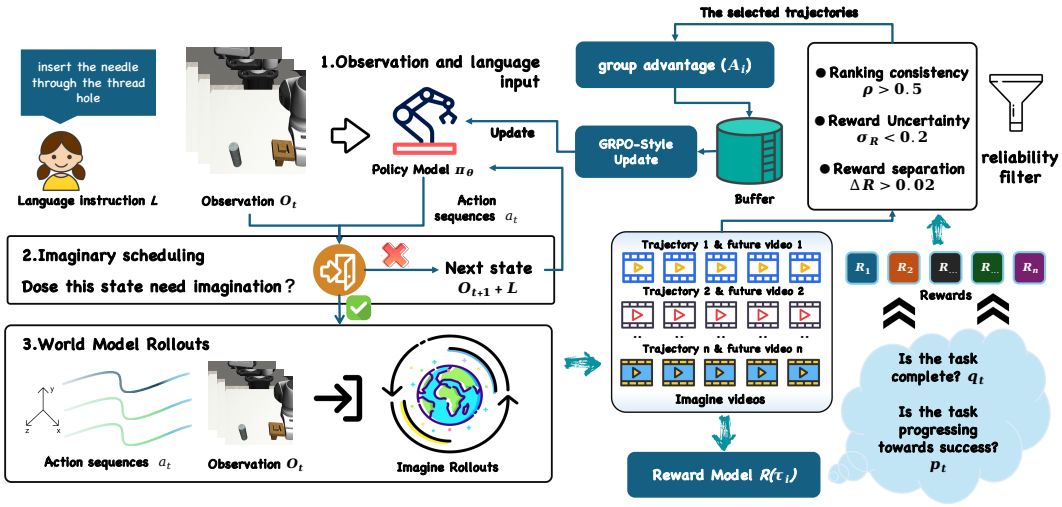

Figure 2: Overview of WISE. Scheduled world-model imagination evaluates candidate actions through bounded imagined futures and converts their relative outcomes into feedback for VLA policy refinement.

$$\theta^\star = \arg\max_\theta \mathbb{E}_{\ell\sim p(\ell),\, \tau\sim p(\tau|\pi_\theta,\ell)}[y(\tau,\ell)],$$

where $y(\tau,\ell) \in 0,1$ indicates whether trajectory τ successfully completes the instructed task.

Directly optimizing this objective in the real world is costly due to sparse feedback and expensive exploration. We therefore use a world model p_ψ to predict the counterfactual consequences of candidate actions:

$$\widehat{\mathbf{O}}_{t+1:t+H} \sim p_\psi(\cdot \mid \mathbf{O}_t, \ell, \mathbf{a}_t).$$

Such predictions allow candidate behaviors to be evaluated without physical execution. Effective world-model-guided post-training, however, requires deciding when imagination should be invoked, how imagined rollouts should be bounded, and how predicted outcomes should be translated into policy supervision. As illustrated in Figure 2, WISE addresses these decisions within a coordinated imagination–evaluation–refinement framework.

3.2 Bounded Counterfactual Imagination

At each imagination context selected by the scheduler (Section 3.3), the policy samples M candidate action chunks

$$\mathbf{a}_{t,i} \sim \pi_\theta(\cdot \mid \mathbf{O}_t, \mathbf{s}_t, \ell), \qquad i = 1, \ldots, M.$$

A frozen world model p_ψ, adapted from Open-Sora (Zheng et al., 2024), predicts action-conditioned multi-view futures from wrist and third-person observations.

The world model is trained on observed frames followed by H future frames. Let $\mathbf{z}_{\text{obs}}$ and $\mathbf{z}_{\text{fut}}$ denote their VAE latents. Following rectified-flow training (Lipman et al., 2023), with $\lambda \sim \mathcal{U}(0,1)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,

$$\mathbf{z}_{\text{fut},\lambda} = (1-\lambda)\boldsymbol{\epsilon} + \lambda \mathbf{z}_{\text{fut}},$$

and

$$\mathcal{L}_{\text{WM}} = \mathbb{E}\Big[\|v_\psi([\mathbf{z}_{\text{obs}}; \mathbf{z}_{\text{fut},\lambda}], \lambda, \ell, \mathbf{a})_{\text{fut}} - (\mathbf{z}_{\text{fut}} - \boldsymbol{\epsilon})\|_2^2\Big].$$

Only future latents receive rectified-flow supervision, while $\mathbf{z}_{\text{obs}}$ is used as conditioning context.

During post-training, each candidate forms a bounded closed-loop rollout. Starting from $\widehat{\mathbf{O}}_i^{(0)} = \mathbf{O}_t$, $\widehat{\mathbf{s}}_i^{(0)} = \mathbf{s}_t$, and $\widehat{\mathbf{a}}_i^{(0)} = \mathbf{a}_{t,i}$, we recursively apply

$$\widehat{\mathbf{O}}_i^{(r+1)} \sim p_\psi(\cdot \mid \widehat{\mathbf{O}}_i^{(r)}, \ell, \widehat{\mathbf{a}}_i^{(r)}),$$

$$\widehat{\mathbf{s}}_i^{(r+1)} = F_s(\widehat{\mathbf{s}}_i^{(r)}, \widehat{\mathbf{a}}_i^{(r)}),$$

and, for $r < L - 1$,

$$\widehat{\mathbf{a}}_i^{(r+1)} \sim \pi_\theta(\cdot \mid \widehat{\mathbf{O}}_i^{(r+1)}, \widehat{\mathbf{s}}_i^{(r+1)}, \ell),$$

where F_s deterministically updates proprioception from the action chunk. Repeating this process for L chunks yields

$$\widehat{\tau}_i = \left\{ \widehat{\mathbf{O}}_i^{(1:L)}, \widehat{\mathbf{a}}_i^{(0:L-1)} \right\}.$$

Since each action chunk predicts H future frames, an L-chunk rollout contains LH imagined frames in total, indexed by $j = 1, \ldots, LH$. Bounding the rollout limits accumulated prediction error. Additional configurations are provided in Appendix A, and qualitative comparisons with real futures are provided in Appendix B.

3.3 Imagination Scheduling

World-model imagination is computationally expensive and need not be invoked uniformly throughout an execution. WISE therefore schedules imagination using the current interaction context, focusing predictive computation on interaction-relevant states where counterfactual evaluation is likely to be useful. Given synchronized wrist and third-person observations $\mathbf{O}_t = I_t^w, I_t^a$, a view-shared DINOv2 encoder (Oquab et al., 2024) E_{ϕ_s} and lightweight head g_η predict an interaction-relevance score

$$s_t = \sigma(g_\eta \left[E_{\phi_s}(I_t^w); E_{\phi_s}(I_t^a)\right]), \qquad s_t \in [0, 1].$$

The scheduler is trained with weak interaction-relevance supervision derived from robot execution signals, which serves as a visual proxy for identifying states where counterfactual evaluation is likely to be useful. Only visual observations are required for scheduling after training. Imagination is triggered by

$$m_t = \mathbb{I}[s_t > \kappa],$$

where κ is the scheduling threshold. When $m_t = 1$, the corresponding interaction context is selected for candidate action generation and bounded world-model rollout; otherwise, imagination is skipped. In this way, WISE reduces unnecessary world-model inference while retaining interaction-relevant contexts for subsequent policy refinement. The scheduler architecture, weak-supervision construction, and training procedure are detailed in Appendix C.1.

3.4 Counterfactual Trajectory Evaluation and Policy Refinement

For each imagined trajectory $\widehat{\tau}_i$, a reward model f_ω evaluates the imagined wrist and agent views $\widehat{\mathbf{O}}_{i,j} = \{\widehat{I}_{i,j}^w, \widehat{I}_{i,j}^a\}$ through a separately trained DINOv2 encoder E_{ϕ_r}:

$$\mathbf{e}_{i,j} = [E_{\phi_r}(\widehat{I}_{i,j}^w); E_{\phi_r}(\widehat{I}_{i,j}^a)],$$

the model predicts progress, confidence, and completion:

$$p_{i,j} = \sigma(h_p(\mathbf{e}_{i,j})), \quad c_{i,j} = \sigma(h_c(\mathbf{e}_{i,j})), \quad q_{i,j} = \tanh(h_q(\mathbf{e}_{i,j})).$$

With $p_{i,0} = p_t$ evaluated from the real context, we define

$$G_i = \sum_{j=1}^{LH} c_{i,j}[p_{i,j} - p_{i,j-1}]_0^\delta, \quad B_i = \sum_{j=1}^{LH} [p_{i,j-1} - p_{i,j}]_0^\delta,$$

$$T_i = \max_j q_{i,j}, \qquad R_i = G_i - \alpha B_i + \beta T_i,$$

where $[x]_0^\delta = \min(\max(x, 0), \delta)$. Here, G_i rewards confidence-weighted forward progress, B_i penalizes regression, and T_i captures completion, while clipping limits abrupt score changes. Reward-model training is detailed in Appendix C.2.

Each candidate is imagined twice, yielding $\bar{R}_i = (R_i^{(1)} + R_i^{(2)})/2$. WISE retains a group only if

$$\Delta R = \max_i \bar{R}_i - \min_i \bar{R}_i \geq 0.02, \qquad \rho_{\text{rank}} \geq 0.5, \qquad \sigma_{\text{WM}} \leq 0.2,$$

where ΔR ensures reward separation, while ρ_{rank} and σ_{WM} measure ranking consistency and reward variation across repeated imaginations, rejecting ambiguous or unstable groups.

For each accepted group,

$$A_i = \frac{\bar{R}_i - \mu_{\bar{R}}}{\max(\sigma_{\bar{R}}, \epsilon)},$$

which measures each candidate relative to alternatives from the same real context. The signed advantages define the update direction

$$\mathbf{g}_{\text{WISE}} = \frac{1}{M} \sum_{i=1}^{M} \text{sg}(A_i) \nabla_\theta \mathcal{L}_{\text{FM}}(\theta; \mathbf{a}_{t,i}, h_t) + \lambda_{\text{ref}} \nabla_\theta \mathcal{L}_{\text{ref}}.$$

The policy is updated along $-\mathbf{g}_{\text{WISE}}$; positive and negative advantages respectively reinforce and suppress candidate actions, while sg blocks reward gradients. Detailed policy optimization configurations and trajectory reliability filtering procedures are provided in Appendix A. Only the first action chunk $\mathbf{a}_{t,i}$ from the real context is supervised; imagined rollouts are used solely for candidate evaluation.

4 Experiments

We conduct extensive experiments to evaluate WISE in simulated and real-world robotic manipulation environments. We focus on three questions: **Policy Improvement.** Does WISE consistently improve different pretrained VLA policies over existing post-training methods? **Design Choices.** How do imagination scheduling and reward model contribute to effective world-model-guided policy refinement? **Efficiency and Generalization.** Can WISE reduce unnecessary world-model computation while maintaining effective policy improvement, and does it generalize to real-world manipulation under varying conditions?

4.1 Experimental Setup

Tasks and Metrics. We evaluate WISE on five MimicGen (Mandlekar et al., 2023) manipulation tasks: *Stack*, *Coffee*, *Square*, *Threading*, and *Mug Cleanup*. These tasks span object placement, articulated interaction, contact-rich insertion, and long-horizon manipulation. We report episode-level success rates and the average performance across all tasks. Detailed simulation environments and evaluation protocols are provided in Appendices E.1 and E.3.

Baselines. We evaluate two pretrained VLA backbones, π_0 and $\pi_{0.5}$. For π_0 Black et al. (2026), we compare WISE against representative post-training methods, including GRPO (Shao et al., 2024), DSRL (Wagenmaker et al., 2025), and DPO (Rafailov et al., 2024). We further apply WISE to the stronger $\pi_{0.5}$ (Intelligence et al., 2025b) backbone to evaluate whether the proposed framework transfers across pretrained VLA policies. All methods use the same task environments, evaluation protocols, and policy initialization within each backbone. Training data, compute, and post-training configurations are summarized in Appendix A.

4.2 Main Results

Table 1 summarizes the simulation results. WISE produces consistent improvements on both pretrained VLA backbones and improves every evaluated task relative to its corresponding base policy. For π_0, WISE increases the average success rate from 49.6% to 59.6%, a gain of 10.0 percentage points, and outperforms the strongest post-training baseline, $\pi_0 + \text{DPO}$, by 5.8 points. Applying WISE to $\pi_{0.5}$ similarly improves the average success rate from 52.3% to 60.2% ($+7.9$ points). Across the two backbones, WISE variants achieve the best performance on all five tasks, with particularly clear gains on *Stack*, *Square*, *Threading*, and *Mug Cleanup*. These results indicate that the benefits of WISE are not specific to a single pretrained policy and that coordinated world-model imagination provides more consistent post-training signals than direct policy optimization alone.

Table 1: Success rates (%) of different post-training methods on five robotic manipulation tasks. **Avg.** denotes the average success rate across tasks.

	Success rate (%)					
Method	**Stack**	**Coffee**	**Square**	**Threading**	**Mug Cleanup**	**Avg.**
π_0	71.9	46.9	28.1	61.5	39.6	49.6
π_0 + GRPO	74.0	50.0	30.2	60.4	42.7	51.5
π_0 + DSRL	78.1	50.0	20.8	61.5	38.5	49.8
π_0 + DPO	78.1	52.1	27.1	64.6	46.9	53.8
$\pi_{0.5}$	81.3	40.6	39.6	45.8	54.2	52.3
π_0 + WISE	84.4	**56.3**	36.5	**68.8**	52.1	59.6
Δ vs. π_0	**+12.5**	**+9.4**	**+8.4**	**+7.3**	**+12.5**	**+10.0**
$\pi_{0.5}$ + WISE	**88.5**	46.9	**50.0**	53.1	**62.5**	**60.2**
Δ vs. $\pi_{0.5}$	**+7.2**	**+6.3**	**+10.4**	**+7.3**	**+8.3**	**+7.9**

4.3 Ablation Studies

We conduct ablation studies on *Stack*, *Threading*, and *Mug Cleanup* to examine two key components of WISE: the reward model and imagination scheduling.

Reward Model. Table 2 studies the contribution of different reward components. The full formulation achieves an average success rate of 68.4%, substantially outperforming all reduced variants. Using progress or completion alone yields only 57.3%, indicating that dense progress feedback and terminal completion provide complementary supervision. Removing confidence weighting or the backward-progress penalty also degrades performance, demonstrating the importance of suppressing uncertain progress estimates and penalizing trajectory regression.

Table 2: Ablation of counterfactual trajectory evaluation components.

	Success Rate (%)			
Variant	**Stack**	**Threading**	**Mug Cleanup**	**Avg.**
Progress only	75.0	58.3	38.5	57.3
Completion only	78.1	56.3	37.5	57.3
w/o Confidence	74.0	61.5	38.5	58.0
w/o Backward Penalty	79.2	61.5	37.5	59.4
Full Reward	**84.4**	**68.8**	**52.1**	**68.4**

Imagination Scheduling. Table 3 compares different strategies for allocating world-model imagination. WISE achieves the highest average success rate of 68.4% with only 2.61 GPU hours. Compared with full imagination, WISE reduces selected states by approximately 80% and GPU time by 77%, while improving success by 8.0 percentage points. Under comparable computation budgets, WISE also outperforms uniform and random scheduling by 7.3 and 9.0 points, respectively. These results demonstrate that selectively allocating imagination to interaction-relevant states provides a better performance–efficiency trade-off than uniformly applying world-model prediction. Further analysis of these ablation results is provided in Appendix D.

Table 3: Ablation of imagination scheduling strategies and computation efficiency.

	Success Rate (%)				Efficiency	
Method	**Stack**	**Threading**	**Mug Cleanup**	**Avg.**	**selected states**	**GPU Hours**
Full Imagination	76.0	63.5	41.7	60.4	138	11.45
Uniform	72.9	62.5	47.9	61.1	28	2.90
Random	77.1	60.4	40.6	59.4	28	2.89
WISE	**84.4**	**68.8**	**52.1**	**68.4**	28	2.61

4.4 Real-World Evaluation

We further evaluate WISE on the Galaxea R1 Lite robot across four real-world manipulation tasks (Figure 3): *Pick-and-Place*, *Two-Cube Stacking*, *Open*, and *Open-and-Close*. Here, *Open* is a subtask of *Open-and-Close*, requiring only the opening phase. These tasks span object transportation, contact-sensitive placement, articulated interaction, and multi-stage manipulation. We compare π_0 (Black et al., 2026), its DPO (Rafailov et al., 2024) variant, and the stronger $\pi_{0.5}$ (Intelligence et al., 2025b), together with their WISE-post-trained counterparts. Details of the robotic platform, sensing configuration, and task setup are provided in Appendix E.2.

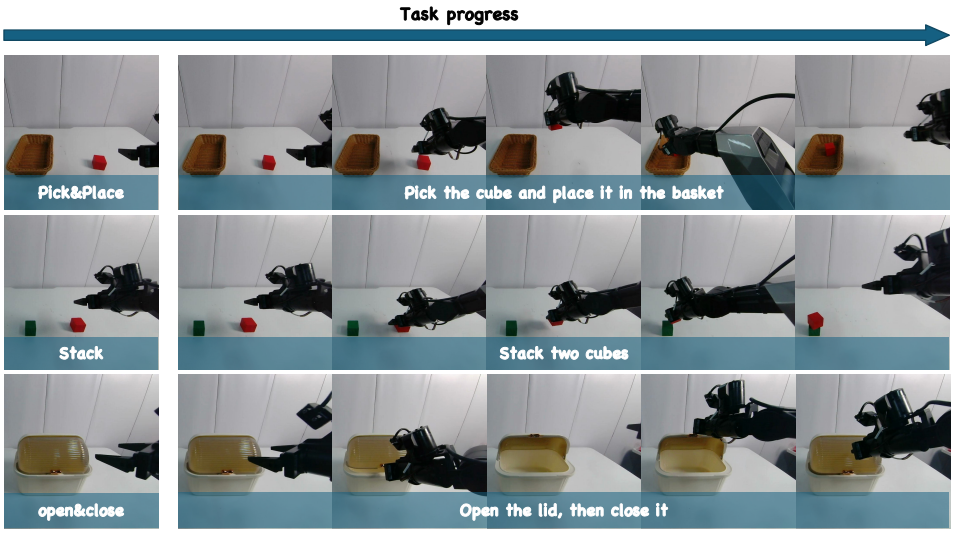

Figure 3: Real-world manipulation tasks evaluated on the Galaxea R1 Lite robot: Pick-and-Place, Two-Cube Stacking, Open, and Open-and-Close.

We report episode-level success rates under the standard setting (**Std.**) and a generalization setting (**Gen.**). The latter averages four controlled variations summarized in Figure 4, with detailed configurations provided in Appendix E.4. For composite tasks, success requires completion of the full instructed sequence. All methods use identical task configurations, reset protocols, and success criteria; the complete evaluation protocol is provided in Appendix E.3.

Table 4 shows that WISE consistently improves both VLA backbones in the real world. For π_0, WISE increases the average success rate from 60.0% to 77.5% under the standard setting and from 42.5%

Table 4: Real-world manipulation success rates (%) under standard (**Std.**) and generalization (**Gen.**) settings. **Avg.** denotes the average success rate across four tasks.

	Pick & Place		Two-Cube Stacking		Open		Open & Close		Avg.	
Method	**Std.**	**Gen.**	**Std.**	**Gen.**	**Std.**	**Gen.**	**Std.**	**Gen.**	**Std.**	**Gen.**
π_0	100.0	70.0	50.0	45.0	60.0	40.0	30.0	15.0	60.0	42.5
π_0 + DPO	100.0	85.0	55.0	40.0	45.0	40.0	20.0	15.0	55.0	45.0
π_0 + WISE	**100.0**	**95.0**	**80.0**	**70.0**	**80.0**	**70.0**	**50.0**	**40.0**	**77.5**	**68.8**
Δ vs. π_0 (pp)	**+0.0**	**+25.0**	**+30.0**	**+25.0**	**+20.0**	**+30.0**	**+20.0**	**+25.0**	**+17.5**	**+26.3**
$\pi_{0.5}$	100.0	85.0	55.0	40.0	60.0	50.0	30.0	25.0	61.3	50.0
$\pi_{0.5}$ + WISE	**100.0**	**95.0**	**80.0**	**65.0**	**80.0**	**75.0**	**50.0**	**50.0**	**77.5**	**71.3**
Δ vs. $\pi_{0.5}$ (pp)	**+0.0**	**+10.0**	**+25.0**	**+25.0**	**+20.0**	**+25.0**	**+20.0**	**+25.0**	**+16.3**	**+21.3**

to 68.8% under generalization, corresponding to gains of 17.5 and 26.3 percentage points. Similar improvements are observed for $\pi_{0.5}$, from 61.3% to 77.5% and from 50.0% to 71.3%, respectively. The larger gains under distribution shifts suggest that world-model-guided post-training particularly improves robustness beyond the training conditions.

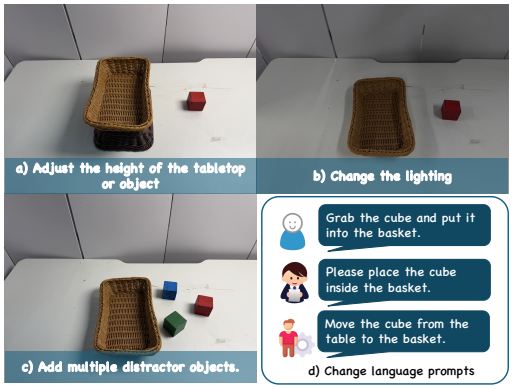

Figure 4: Generalization settings for real-world evaluation.

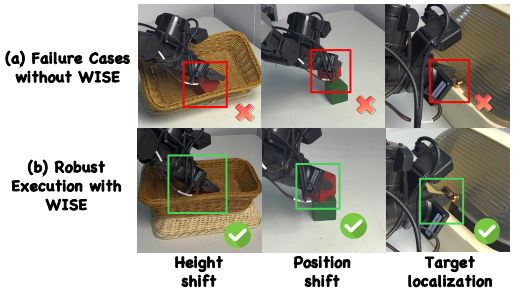

Figure 5: Representative failure cases and recovery behaviors during real-world manipulation.

We further examine representative behaviors under the generalization settings. As shown in Figure 5, WISE better adapts to height changes, relative object positions, and handle localization across the evaluated tasks. Together with the controlled variations in Figure 4, these cases provide qualitative evidence of improved robustness beyond the standard evaluation setting.

5 Conclusion

We presented **WISE**, a world-model-guided post-training framework that coordinates imagination scheduling, bounded counterfactual prediction, trajectory evaluation, and policy refinement for pretrained VLA models. Rather than invoking world-model reasoning uniformly throughout an execution, WISE allocates imagination to interaction-relevant states and converts the relative quality of predicted futures into supervision grounded in real interaction contexts. Experiments with both π_0 and $\pi_{0.5}$ demonstrate consistent improvements across simulated and real-world manipulation tasks, while substantially reducing unnecessary world-model computation. WISE also shows strong robustness under diverse real-world generalization conditions.

Several extensions remain promising. The current framework uses a fixed imagination horizon and coarse stage supervision for scheduling; future work could explore adaptive imagination horizons and finer-grained scheduling based on prediction uncertainty and policy confidence. Extending WISE to a broader range of embodiments, longer-horizon tasks, and more diverse pretrained VLA backbones would further investigate the scalability of imagination scheduling for general-purpose robot post-training.

AI USE STATEMENT

Generative AI tools were used for LaTeX formatting and limited copy-editing assistance. The authors reviewed all AI-assisted changes, verified the citations and technical content, and take responsibility for the final content of this paper.

ETHICS STATEMENT

This work does not involve human subjects or private user data. All real-world robot experiments are conducted in controlled laboratory environments following standard safety procedures. The proposed method is intended for research on robot learning and does not target safety-critical autonomous deployment.

REPRODUCIBILITY STATEMENT

We provide detailed model architectures, training objectives, hyperparameters, and evaluation protocols in the main paper and supplementary material. Source code, training and evaluation configurations, and implementation details are provided in an anonymous repository https://anonymous.4open.science/r/WISE-72F0 and will be made publicly available upon publication.

REFERENCES

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael 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, James Tanner, Quan Vuong, Anna Walling, Haohuan Wang, and Ury Zhilinsky. π_0: A vision-language-action flow model for general robot control, 2026. URL https://arxiv.org/abs/2410.24164.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-2: Vision-language-action models transfer web knowledge to robotic control, 2023a. URL https://arxiv.org/abs/2307.15818.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale, 2023b. URL https://arxiv.org/abs/2212.06817.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments, 2024. URL https://arxiv.org/abs/2402.15391.

Carlos Celemin, Rodrigo Pérez-Dattari, Eugenio Chisari, Giovanni Franzese, Leandro de Souza Rosa, Ravi Prakash, Zlatan Ajanović, Marta Ferraz, Abhinav Valada, and Jens Kober. Interactive imitation learning in robotics: A survey, 2022. URL https://arxiv.org/abs/2211.00600.

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, 2025. URL https://arxiv.org/abs/2502.05450.

Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga

Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Muhammad Zubair Irshad, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Martín-Martín, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Vitor Guizilini, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open x-embodiment: Robotic learning datasets and rt-x models, 2025. URL https://arxiv.org/abs/2310.08864.

Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, Jianlan Luo, You Liang Tan, Lawrence Yunliang Chen, Pannag Sanketi, Quan Vuong, Ted Xiao, Dorsa Sadigh, Chelsea Finn, and Sergey Levine. Octo: An open-source generalist robot policy, 2024. URL https://arxiv.org/abs/2405.12213.

Yanjiang Guo, Jianke Zhang, Xiaoyu Chen, Xiang Ji, Yen-Jen Wang, Yucheng Hu, and Jianyu Chen. Improving vision-language-action model with online reinforcement learning, 2025. URL https://arxiv.org/abs/2501.16664.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. URL https://arxiv.org/abs/2301.04104.

Bohan Hou, Gen Li, Jindou Jia, Tuo An, Xinying Guo, Sicong Leng, Haoran Geng, Yanjie Ze, Tatsuya Harada, Philip Torr, Oier Mees, Marc Pollefeys, Zhuang Liu, Jiajun Wu, Pieter Abbeel, Jitendra Malik, Yilun Du, and Jianfei Yang. World model for robot learning: A comprehensive survey, 2026. URL https://arxiv.org/abs/2605.00080.

Physical Intelligence, Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, Michael Equi,

Adnan Esmail, Yunhao Fang, Chelsea Finn, Catherine Glossop, Thomas Godden, Ivan Goryachev, Lachy Groom, Hunter Hancock, Karol Hausman, Gashon Hussein, Brian Ichter, Szymon Jakubczak, Rowan Jen, Tim Jones, Ben Katz, Liyiming Ke, Chandra Kuchi, Marinda Lamb, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Yao Lu, Vishnu Mano, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Charvi Sharma, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, Will Stoeckle, Alex Swerdlow, James Tanner, Marcel Torne, Quan Vuong, Anna Walling, Haohuan Wang, Blake Williams, Sukwon Yoo, Lili Yu, Ury Zhilinsky, and Zhiyuan Zhou. $\pi^*_{0.6}$: a vla that learns from experience, 2025a. URL https://arxiv.org/abs/2511.14759.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Manuel Y. Galliker, Dibya Ghosh, Lachy Groom, Karol Hausman, Brian Ichter, Szymon Jakubczak, Tim Jones, Liyiming Ke, Devin LeBlanc, Sergey Levine, Adrian Li-Bell, Mohith Mothukuri, Suraj Nair, Karl Pertsch, Allen Z. Ren, Lucy Xiaoyang Shi, Laura Smith, Jost Tobias Springenberg, Kyle Stachowicz, James Tanner, Quan Vuong, Homer Walke, Anna Walling, Haohuan Wang, Lili Yu, and Ury Zhilinsky. $\pi_{0.5}$: a vision-language-action model with open-world generalization, 2025b. URL https://arxiv.org/abs/2504.16054.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/jiang23b.html.

Zhennan 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, 2026. URL https://arxiv.org/abs/2509.19080.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, Vitor Guizilini, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Muhammad Zubair Irshad, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset, 2025. URL https://arxiv.org/abs/2403.12945.

Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success, 2025a. URL https://arxiv.org/abs/2502.19645.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 06–09 Nov 2025b. URL https://proceedings.mlr.press/v270/kim25c.html.

Hengtao Li, Pengxiang Ding, Runze Suo, Yihao Wang, Zirui Ge, Dongyuan Zang, Kexian Yu, Mingyang Sun, Hongyin Zhang, Donglin Wang, and Weihua Su. Vla-rft: Vision-language-action reinforcement fine-tuning with verified rewards in world simulators, 2025. URL https://arxiv.org/abs/2510.00406.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 26703–26721, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/71639c317fb0bf398835627b4418693e-Paper-Conference.pdf.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

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.

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning, 2025. URL https://arxiv.org/abs/2410.21845.

Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 37 (7):3031–3051, 2026. doi: 10.1109/TNNLS.2025.3650584.

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 Jie Tan, Marc Toussaint, and Kourosh Darvish (eds.), *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1820–1864. PMLR, 06–09 Nov 2023. URL https://proceedings.mlr.press/v229/mandlekar23a.html.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Allen Z. Ren, Justin Lidard, Lars L. Ankile, Anthony Simeonov, Pulkit Agrawal, Anirudha Majumdar, Benjamin Burchfiel, Hongkai Dai, and Max Simchowitz. Diffusion policy policy optimization, 2024. URL https://arxiv.org/abs/2409.00588.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

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.

Xiaoquan Sun, Zetian Xu, Chen Cao, Zonghe Liu, Yihan Sun, Jingrui Pang, Ruijian Zhang, Zhen Yang, Kang Pang, Dingxin He, Mingqi Yuan, and Jiayu Chen. Atomvla: Scalable post-training for robotic manipulation via predictive latent world models, 2026. URL https://arxiv.org/abs/2603.08519.

Chen Tang, Ben Abbatematteo, Jiaheng Hu, Rohan Chandra, Roberto Martín-Martín, and Peter Stone. Deep reinforcement learning for robotics: A survey of real-world successes, 2024. URL https://arxiv.org/abs/2408.03539.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pp. 4950–4957. AAAI Press, 2018. ISBN 9780999241127.

Andrew Wagenmaker, Yunchu Zhang, Mitsuhiko Nakamoto, Seohong Park, Waleed Yagoub, Anusha Nagabandi, Abhishek Gupta, and Sergey Levine. Steering your diffusion policy with latent space reinforcement learning. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 258–282. PMLR, 27–30 Sep 2025. URL https://proceedings.mlr.press/v305/wagenmaker25a.html.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. Daydreamer: World models for physical robot learning, 2022. URL https://arxiv.org/abs/2206.14176.

Charles Xu, Jost Tobias Springenberg, Michael Equi, Ali Amin, Adnan Esmail, Sergey Levine, and Liyiming Ke. Rl token: Bootstrapping online rl with vision-language-action models, 2026. URL https://arxiv.org/abs/2604.23073.

Jiazhi Yang, Kunyang Lin, Jinwei Li, Wencong Zhang, Tianwei Lin, Longyan Wu, Zhizhong Su, Hao Zhao, Ya-Qin Zhang, Li Chen, Ping Luo, Xiangyu Yue, and Hongyang Li. Rise: Self-improving robot policy with compositional world model, 2026. URL https://arxiv.org/abs/2602.11075.

Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1389–1399, 2023. doi: 10.1109/CVPR52729.2023.00140.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware, 2023. URL https://arxiv.org/abs/2304.13705.

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.

Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. Wmpo: World model-based policy optimization for vision-language-action models, 2025. URL https://arxiv.org/abs/2511.09515.

Appendix

A Implementation Details

The world model is first pretrained on the full DROID (Khazatsky et al., 2025) dataset and subsequently adapted to each target task using the same task-specific demonstrations as the corresponding VLA policy. Both π_0 and $\pi_{0.5}$ (Black et al., 2026; Intelligence et al., 2025b) are independently fine-tuned for each task before WISE post-training.

During WISE post-training, counterfactual rollouts are generated only from scheduler-selected real interaction contexts. For each selected context, the policy samples $M = 256$ candidate first action chunks. Each candidate is independently evaluated twice using different random seeds, and the resulting two trajectory rewards are averaged before computing group-relative advantages. The world model and the main VLA backbone remain frozen, while only parameters in the action head are optimized. Detailed training configurations are summarized in Table 5.

Table 5: Main training and post-training configurations.

Stage	Configuration	Setting
WM Pre-training	Dataset	Full DROID
	Compute	16× NVIDIA A800
	Training time	∼7 days
Task-specific WM	Demonstrations	Sim.: 300 / Real: 100
	Compute	8× NVIDIA A800
	Training time	∼10 h / task
Task-specific VLA	Backbone	π_0, $\pi_{0.5}$
	Demonstrations	Sim.: 300 / Real: 100
	Training time	Sim.: ∼12 h / Real: ∼8 h
WISE Post-training	Candidates per selected state M	256
	Repeated imaginations	2
	Future frames H	8
	Rollout chunks	4
	Optimizer	AdamW
	Learning rate	1×10^{-8}
	Gradient clipping	0.25
	Reference coefficient λ_{ref}	0.1
	Advantage normalization ϵ	10^{-6}
	Trainable parameters	Action head
	Frozen modules	WM and VLA backbone

A.1 Policy Optimization Details

Flow-matching objective. WISE optimizes the trainable action head using the flow-matching objective inherited from the pretrained VLA policy. Let $\mathbf{a}$ denote an action chunk after applying the same action transformation and normalization as the pretrained policy, and let $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. For $\tau \sim \mathcal{U}(0, 1)$, the noisy action and target velocity are constructed as

$$\mathbf{x}_\tau = (1 - \tau)\mathbf{a} + \tau\boldsymbol{\epsilon}, \qquad \mathbf{u}_\tau = \boldsymbol{\epsilon} - \mathbf{a}.$$

Let $v_\theta(\mathbf{x}_\tau, \tau, h_t)$ denote the action-flow prediction of the VLA policy conditioned on the real interaction context h_t. The flow-matching loss used during WISE optimization is

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}} \left[\frac{1}{8 \times 7} \sum_{h=1}^{8} \sum_{d=1}^{7} \left(v_\theta(\mathbf{x}_\tau, \tau, h_t)_{h,d} - u_{\tau,h,d} \right)^2 \right].$$

The loss is averaged over the first eight action steps and the seven valid action dimensions. Only the first action chunk sampled from the real interaction context is used as a policy supervision target. Actions generated at subsequent imagined states are used only to extend counterfactual rollouts and are never used as direct flow-matching supervision.

Reference regularization. To limit policy drift during WISE post-training, the trainable parameters are anchored to the task-specific SFT initialization. Let θ^{SFT} denote the trainable action-head parameters saved immediately before WISE optimization. We define

$$\mathcal{L}_{\mathrm{ref}} = \frac{1}{N_\theta} \sum_{k=1}^{N_\theta} \left(\theta_k - \theta_k^{\mathrm{SFT}}\right)^2,$$

where N_θ is the number of trainable parameters. The reference parameters θ^{SFT} remain fixed throughout WISE optimization and correspond to the task-specific SFT policy rather than the original pretrained π_0 initialization. Frozen VLM parameters are excluded from the regularization. We use $\lambda_{\mathrm{ref}} = 0.1$.

Optimization. The WISE action head is optimized using AdamW with a learning rate of 1×10^{-8}. The gradient norm is clipped to 0.25. Group-relative advantages use $\epsilon = 10^{-6}$ for numerical stability during normalization.

A.2 Reliability Filtering

For each candidate i, let $R_i^{(1)}$ and $R_i^{(2)}$ denote the trajectory rewards obtained from two independently seeded imagined evaluations. Their average reward is

$$\bar{R}_i = \frac{R_i^{(1)} + R_i^{(2)}}{2}.$$

We measure the consistency of candidate rankings between the two repeated imaginations using pairwise ranking agreement,

$$\rho_{\mathrm{rank}} = \frac{1}{\binom{M}{2}} \sum_{i<j} \mathbb{I}\left[\mathrm{sign}\left(R_i^{(1)} - R_j^{(1)}\right) = \mathrm{sign}\left(R_i^{(2)} - R_j^{(2)}\right)\right].$$

A candidate pair is counted as consistent when the two repeated imaginations induce the same pairwise ordering. A tie is counted as consistent only when both repeats produce a tie.

We additionally measure the maximum reward variation across the repeated imaginations as

$$\sigma_{\mathrm{WM}} = \max_i \sqrt{\frac{1}{2} \sum_{r=1}^{2} \left(R_i^{(r)} - \bar{R}_i\right)^2} = \max_i \frac{\left|R_i^{(1)} - R_i^{(2)}\right|}{2}.$$

For multi-chunk rollouts, this variation can reflect both stochastic world-model prediction and differences in subsequent policy samples, since later action chunks are independently re-sampled within each repeated imagination.

A candidate group is retained only when all three reliability conditions are satisfied:

$$\Delta R = \max_i \bar{R}_i - \min_i \bar{R}_i \geq 0.02,$$

$$\rho_{\mathrm{rank}} \geq 0.5, \qquad \sigma_{\mathrm{WM}} \leq 0.2.$$

The reward-separation constraint rejects groups with insufficient discrimination among candidate actions, the ranking-consistency constraint requires stable relative ordering across repeated imaginations, and the variation constraint rejects groups containing candidates whose rewards are excessively sensitive to stochastic rollout variation.

Group-relative advantage normalization. For each retained candidate group, the mean and standard deviation of the averaged rewards are

$$\mu_{\bar{R}} = \frac{1}{M} \sum_{i=1}^{M} \bar{R}_i,$$

and

$$\sigma_{\bar{R}} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} \left(\bar{R}_i - \mu_{\bar{R}}\right)^2}.$$

The normalized group-relative advantage is then

$$A_i = \frac{\bar{R}_i - \mu_{\bar{R}}}{\max(\sigma_{\bar{R}}, 10^{-6})}.$$

The advantage is treated as a stop-gradient weight during policy optimization.

B World-Model Prediction Quality

Figure 6 qualitatively compares world-model predictions with the corresponding real future observations. The predicted multi-view futures preserve the main scene structure and task-relevant interaction evolution over the bounded imagination horizon used by WISE.

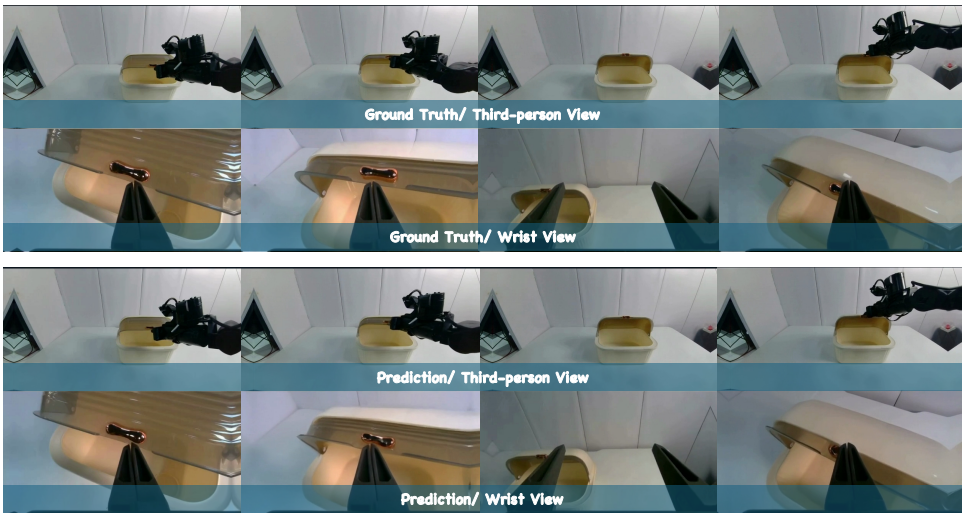

Figure 6: Qualitative comparison between real and world-model-predicted futures. The upper two rows show real third-person and wrist observations, while the lower two rows show the corresponding predictions.

C Training Details of WISE Components

WISE contains two lightweight visual modules that are trained offline and subsequently frozen during VLA post-training: an imagination scheduler that determines when world-model rollouts should be invoked and a trajectory evaluator that scores the imagined consequences of candidate action chunks. Both modules consume the same two camera views used by the policy, namely the third-person agent view and the wrist-mounted eye-in-hand view.

C.1 Imagination Scheduler Training

Architecture. The imagination scheduler predicts the interaction relevance of the current visual context for world-model imagination. Wrist and agent views are encoded by a view-shared pretrained DINOv2 ViT-L/14 encoder E_{ϕ_s} (Oquab et al., 2024). The two view-level features are concatenated and processed by an MLP with two 256-dimensional hidden layers, LayerNorm, GELU activations, and dropout of 0.1. The scheduler outputs

$$s_t = \sigma(g_\eta\left[E_{\phi_s}(I_t^w); E_{\phi_s}(I_t^a)\right]),$$

where I_t^w and I_t^a denote the wrist and agent views, respectively. During training, the classifier head and the final transformer block of E_{ϕ_s} are optimized, while the remaining backbone parameters are frozen. The complete scheduler is frozen during WISE post-training.

Weak supervision. The imagination scheduler is trained with weak interaction-relevance labels automatically derived from gripper and grasp states in the offline trajectories. These execution signals indicate whether the robot has entered an interaction-relevant manipulation state without requiring manual task-phase annotation. We assign $y_n = 1$ to samples associated with grasp or active manipulation and $y_n = 0$ to the remaining interaction contexts. Class-dependent weights are used to alleviate label imbalance. These execution signals are used only for offline label construction; during WISE post-training, the scheduler predicts interaction relevance solely from the current multi-view visual observations.

Objective and optimization. Let $s_n \in [0, 1]$ denote the scheduler prediction for the n-th sample and $y_n \in 0, 1$ its weak interaction-relevance label. The scheduler is optimized using weighted binary cross-entropy:

$$\mathcal{L}_{\text{sched}} = -\frac{1}{N}\sum_{n=1}^{N} w_n\left[y_n \log s_n + (1 - y_n)\log(1 - s_n)\right],$$

where w_n compensates for the imbalance between interaction-relevant and non-relevant samples. After training, the scheduler is frozen and used to select visual contexts for world-model imagination during WISE post-training.

C.2 Reward Model Training

Architecture. The reward model uses a separately trained, view-shared DINOv2 ViT-L/14 encoder E_{ϕ_r} with three independent MLP heads for task progress, prediction confidence, and task completion. Given wrist and agent observations, the two-view feature is

$$\mathbf{e}_t = \left[E_{\phi_r}(I_t^w); E_{\phi_r}(I_t^a)\right],$$

from which the model predicts

$$p_t = \sigma(h_p(\mathbf{e}_t)), \qquad c_t = \sigma(h_c(\mathbf{e}_t)), \qquad q_t = \tanh(h_q(\mathbf{e}_t)).$$

Here, p_t, c_t, and q_t denote task progress, prediction confidence, and task completion, respectively. These signals are subsequently combined to score imagined trajectories during WISE post-training.

Training objective. The reward model is trained offline using recorded task trajectories. Progress supervision follows temporal ordering, encouraging later states in successful trajectories to receive higher scores than earlier states:

$$\mathcal{L}_{\text{rank}} = \text{softplus}\left(m - (z^p_{t_l} - z^p_{t_e})\right), \qquad t_e < t_l,$$

where z^p denotes the pre-sigmoid progress logit. Hard-negative pairs further compare successful states with time-aligned states from failed trajectories. The confidence head is trained with binary cross-entropy, while the completion head uses signed success/failure supervision. Offline execution signals provide coarse task-progress and completion labels using the same supervision construction for both simulation and real-world tasks. These signals are used only for reward-model training and are not required during WISE post-training.

The overall training objective is

$$\mathcal{L}_{\text{RM}} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{abs}} + 0.5\mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{comp}} + 0.1\mathcal{L}_{\text{aug}},$$

where the additional terms supervise absolute progress, confidence, task completion, and augmentation consistency. The reward model is trained for 15 epochs using AdamW, jointly optimizing the prediction heads and the final four transformer blocks of E_{ϕ_r}.

Reward computation. The reward model remains frozen throughout WISE policy optimization. Its progress, confidence, and completion predictions are combined using the trajectory reward defined in the main paper,

$$R = G - \alpha B + \beta T, \qquad \alpha = 0.2, \quad \beta = 2.0, \quad \delta = 0.2,$$

where δ is the clipping threshold for per-step progress changes. The resulting rewards are used only to rank candidate actions and compute relative policy advantages; imagined observations are never used as direct state–action supervision.

Qualitative Visualization of WISE Components Figure 7 evaluates the reward models on *Square*, *Coffee*, *Stack*, *Threading*, and *Mug Cleanup*, arranged from top to bottom. Each task includes 20 successful and 20 failed videos, with 64 frames sampled per video. Trajectories are aligned by normalized time (0–100%). Progress estimates task advancement, completion indicates whether the task is finished, and confidence reflects the model's confidence rather than the probability of success. Cumulative reward combines progress changes, confidence, and completion.

Successful videos generally achieve higher late-stage progress, completion, and cumulative reward. Completion scores rise particularly late in *Threading* and *Mug Cleanup*, whereas confidence can remain high even for failed videos. Final-reward AUROC ranges from 0.938 to 1.000, indicating strong outcome discrimination on the evaluated videos; 0.5 denotes chance-level discrimination and 1.0 denotes perfect discrimination.

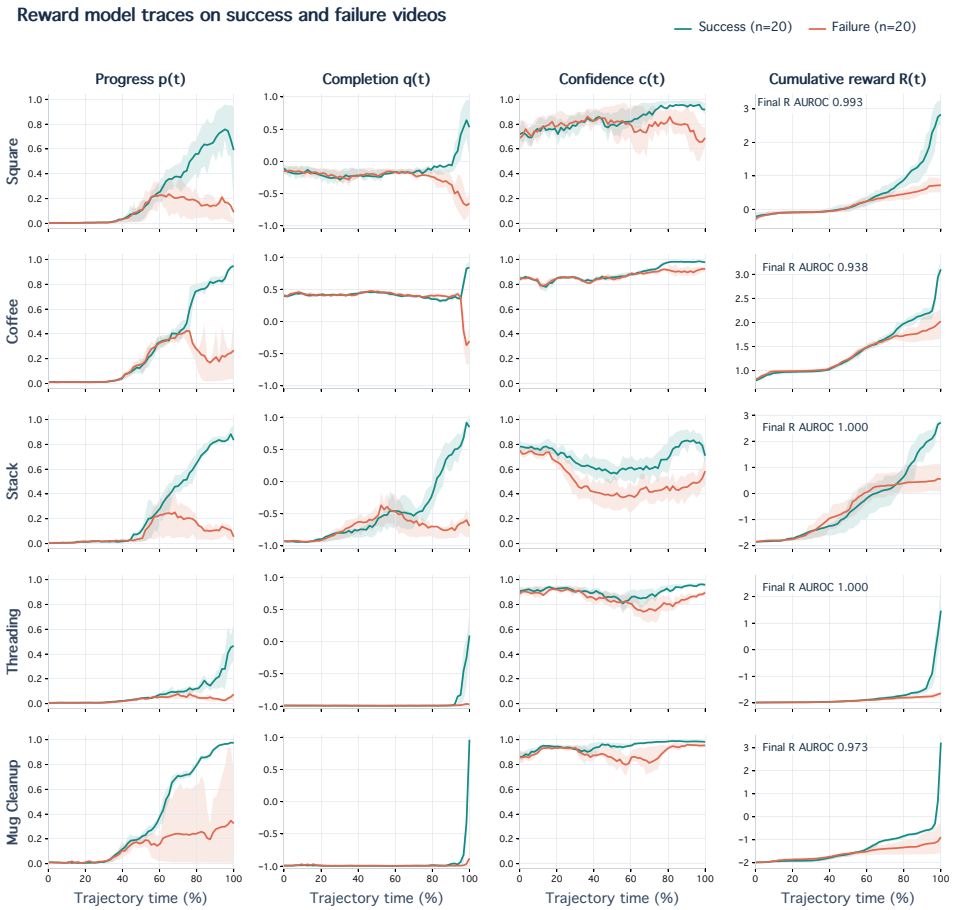

Figure 7: Reward-model scores on five simulated tasks. Green and red indicate successful and failed videos, respectively. Columns show progress $p(t)$, completion $q(t)$, confidence $c(t)$, and cumulative reward $R(t)$. Solid lines denote means; shaded regions indicate interquartile ranges.

D Further Analysis of Ablation Results

D.1 Counterfactual Trajectory Evaluation

The ablation results in Table 2 show that progress and terminal completion provide complementary supervision. Progress offers dense intermediate feedback but may overvalue trajectories that temporarily improve without completing the task, whereas terminal evaluation captures final outcomes but provides limited discrimination among intermediate behaviors. Confidence weighting suppresses uncertain progress estimates, while the backward-progress penalty discourages trajectories that regress after making temporary progress. Combining these signals therefore yields more reliable relative rankings of counterfactual action candidates.

D.2 Imagination Scheduling

Table 3 shows that increasing the amount of imagination does not necessarily improve policy learning. Full imagination uses substantially more world-model computation than WISE while achieving lower success, suggesting that many imagined states provide redundant or weakly informative supervision. Uniform and random scheduling further confirm that computation budget alone is insufficient. WISE instead predicts relevance from the current visual context, enabling finer-grained allocation of imagination to states where candidate actions are more likely to produce discriminative future outcomes.

We report the average number of states selected for world-model imagination across tasks. Each selected state triggers the same candidate-generation and world-model evaluation procedure, so the number of selected states provides a direct measure of the imagination budget under a fixed per-state configuration.

E Detailed Experimental Configurations

E.1 Simulation Setup

We evaluate WISE on five MimicGen D0 environments: `Square_D0`, `Coffee_D0`, `Stack_D0`, `Threading_D0`, and `MugCleanup_D0`, as illustrated in Figure 8.

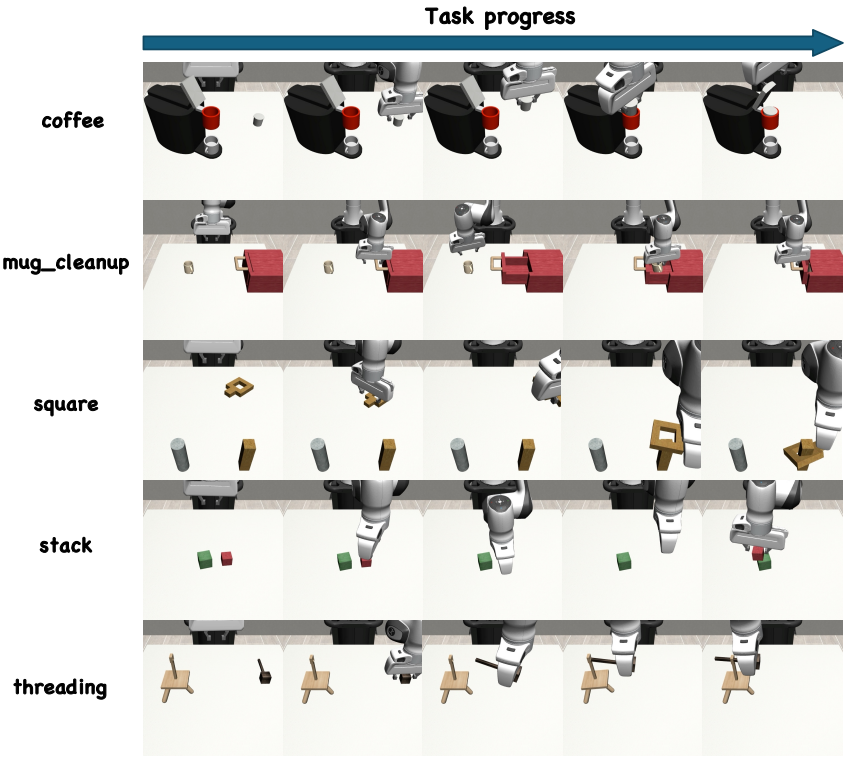

Figure 8: Simulation tasks used for evaluation: Coffee, Mug Cleanup, Square, Stack and Threading

E.2 Real-World Robot Setup

All real-world experiments are conducted on the Galaxea R1 Lite robotic platform. The policy takes synchronized RGB observations from a wrist-mounted camera and a fixed third-person camera, together with the robot proprioceptive state and the language instruction. We use the same robot configuration, camera placement, workspace, task objects, and nominal initialization distribution across all compared methods. Variations in object and scene configurations are introduced only for the generalization experiments described in Appendix E.4.

As illustrated in Figure 9, our real-world evaluation covers tasks with progressively increasing temporal horizon and manipulation complexity. Pick Cube represents a relatively simple short-horizon task, Stack Two Cubes requires medium-horizon multi-stage manipulation, and Open-and-Close constitutes a long-horizon task involving sustained interaction and precise end-effector control. In particular, the lid task requires the robot to accurately complete both opening and closing while maintaining stable contact with the articulated object. Despite this increased difficulty, our method consistently improves performance across all three settings, including stable gains on the challenging long-horizon lid task.

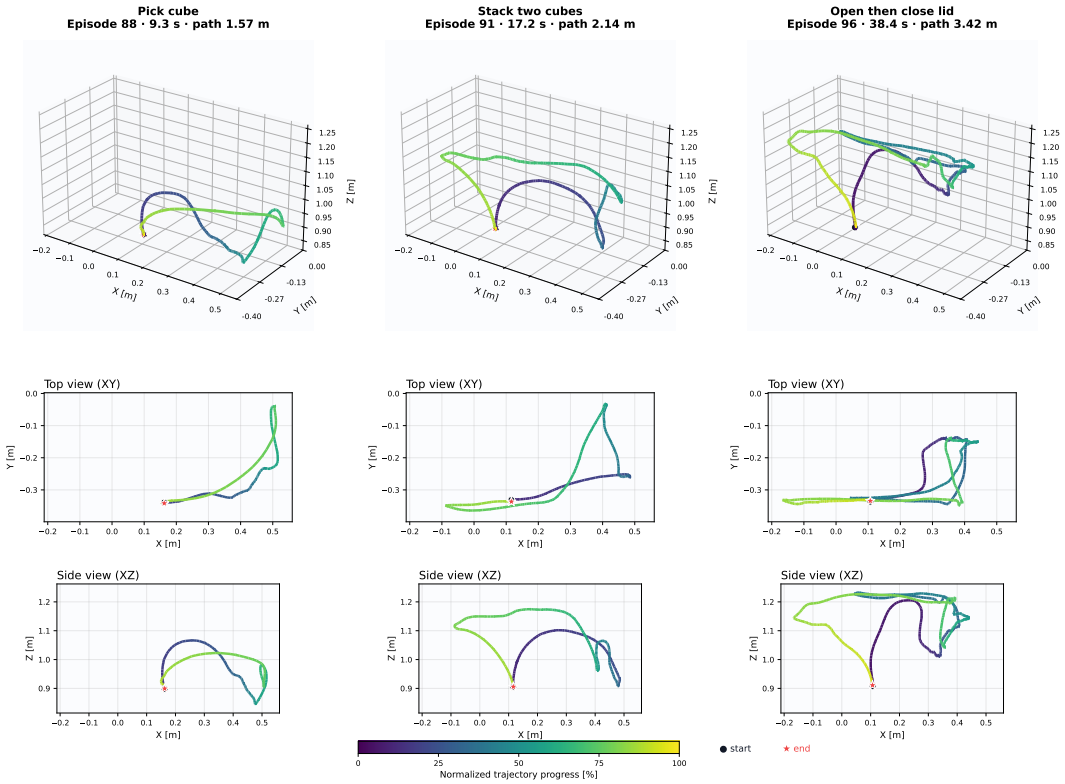

Figure 9: Representative right end-effector trajectories for the three real-world tasks on the Galaxea R1 Lite platform. Each column shows the 3D trajectory together with its top-view (XY) and side-view (XZ) projections. Colors indicate normalized task progress from start to completion, while the black circle and red star denote the start and end points, respectively.

E.3 Evaluation Protocol

We use episode-level task success rate as the primary evaluation metric in both simulation and real-world experiments. A trial is counted as successful only when the complete task-specific success criterion is satisfied within the maximum execution horizon. Partial completion is counted as failure. For multi-stage tasks, such as *Open-and-Close*, all required stages must be completed successfully within a single execution.

In simulation, success is determined using the native environment success criteria. As described in Appendix E.1, each task and policy variant is evaluated on 96 episodes using identical evaluation seeds for the base and post-trained policies, enabling paired comparisons under the same initial conditions.

For real-world evaluation, each task and policy variant is tested over 20 independent trials. Before each trial, the robot and task objects are reset according to the corresponding evaluation configuration. Failed executions are not retried, and no manual intervention is allowed after policy execution begins. Success is recorded according to predefined binary task-completion criteria that are kept identical across all compared methods.

The standard-setting success rate is computed over trials conducted under the nominal task configuration. The generalization score summarizes performance under the controlled distribution shifts described in Appendix E.4. Task-level averages reported in the main paper are computed as the arithmetic mean across the four real-world manipulation tasks.

E.4 Generalization Settings

We evaluate real-world generalization under four controlled distribution shifts: interaction height, lighting conditions, distractor objects, and language instructions. Representative configurations are shown in Figure 10.



Figure 10: Representative real-world generalization settings, including variations in interaction height, lighting conditions, and distractor objects.

Height variation. For *Pick-and-Place*, we use the height variation illustrated in the main paper. For the remaining tasks, the interaction height is changed by adjusting the height of the workstation while keeping the task configuration otherwise unchanged.

Lighting variation. Lighting conditions are varied by turning off the workstation illumination used during task-specific training.

Distractor variation. Task-irrelevant objects that are not observed during training are introduced into the workspace while leaving the target objects unchanged.

Language variation. The original task instruction is replaced with semantically equivalent paraphrases using different wording or sentence structures while preserving the same manipulation objective.

Each factor is varied independently, and identical configurations are used across all compared policies.