RISE: Self-Improving Robot Policy with Compositional World Model

Jiazhi Yang[1,2*†] Kunyang Lin[2*] Jinwei Li[2,6*] Wencong Zhang[2*] Tianwei Lin[5] Longyan Wu[4]
Zhizhong Su[5] Hao Zhao[6] Ya-Qin Zhang[6] Li Chen[3] Ping Luo[3] Xiangyu Yue[1♮] Hongyang Li[3♮]
[1] The Chinese University of Hong Kong [2] Kinetix AI [3] The University of Hong Kong
[4] Shanghai Innovation Institute [5] Horizon Robotics [6] Tsinghua University
[*]Equal Contribution [†]Project lead [♮]Equal Advising

https://opendrivelab.com/kai0-rl

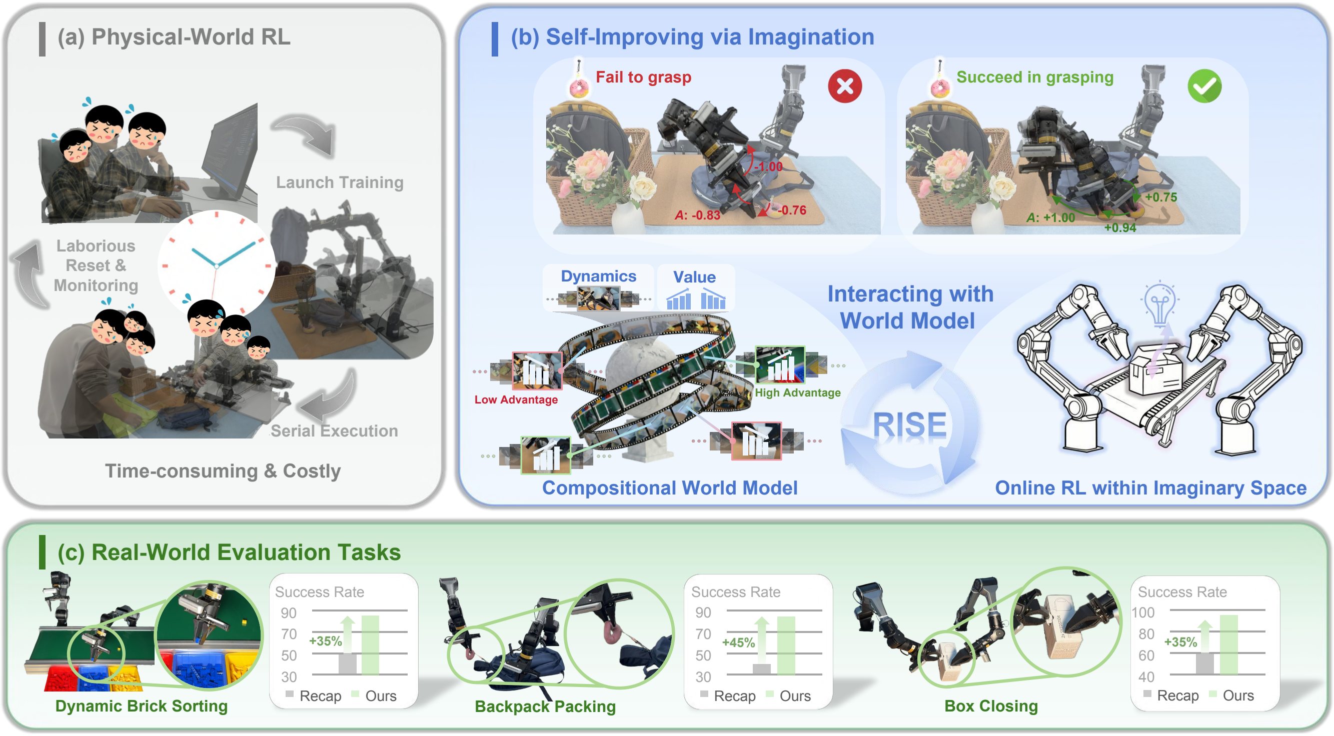


Fig. 1: We present **RISE**, a framework for **R**einforcement learning via **I**magination for **SE**lf-improving robots. (a) Conventional physical-world RL is bottlenecked by hardware cost, slow serial interaction, and the need for manual reset. (b) RISE shifts the learning environment to a **Compositional World Model**, which first emulates future observations for proposed actions, then evaluates imagined states to derive advantage for policy improvement. (c) Training on massive imaginative rollouts effectively bootstraps RISE's performance across a variety of complex, contact-rich tasks, surpassing prior art by a non-trivial margin.

***Abstract*—Despite the sustained scaling on model capacity and data acquisition, Vision–Language–Action (VLA) models remain brittle in contact-rich and dynamic manipulation tasks, where minor execution deviations can compound into failures. While reinforcement learning (RL) offers a principled path to robustness, on-policy RL in the physical world is constrained by safety risk, hardware cost, and environment reset. To bridge this gap, we present RISE, a scalable framework of robotic reinforcement learning via imagination. At its core is a Compositional World Model that (i) predicts multi-view future via a controllable dynamics model, and (ii) evaluates imagined outcomes with a progress value model, producing informative advantages for the policy improvement. Such compositional design allows state and value to be tailored by best-suited yet distinct architectures and objectives. These components are integrated into a closed-loop self-improving pipeline that continuously generates imaginary rollouts, estimates advantages, and updates the policy in imaginary space without costly physical interaction. Across three challenging real-world tasks, RISE yields significant improvement over prior art, with more than +35% absolute performance increase in dynamic brick sorting, +45% for backpack packing, and +35% for box closing, respectively.**

I. Introduction

The trajectory of embodied intelligence has been reshaped by the scaling of foundation models. Particularly, VLA models have emerged as the dominant paradigm for generalist robot control, leveraging massive pre-training on web-scale data to acquire broad semantic understanding and instruction-following capabilities [9, 7, 8, 46, 22, 76]. Despite the progress on high-level semantic competence, such VLAs still fall short of robust manipulation under complex physical dynamics, such as precise grasping of moving objects or effective bi-manual coordination [65, 37]. This discrepancy highlights the inherent limitation of Imitation Learning (IL), a core mechanism enabling VLAs to generate executable actions. Concretely, IL is inherently limited by the quality and coverage of the expert demonstrations while suffering from the exposure bias problem: once the robot drifts slightly off the expert's manifold, it lacks the recovery skills to correct its course, leading to compounding errors [73, 45, 37, 15]. Reinforcement Learning (RL), which improves agents through their own success and failure, offers a potential remedy.

In virtual simulators such as LIBERO [60], agents can play massive interactions in parallel, where both state and reward updates are controllable and accessible. Such properties of highly-crafted simulators have inspired successful RL adaptations upon recent VLAs [63, 56, 61]. Nonetheless, such controllability and parallelization do not hold in a real-world regime, where robot executions are serial, time-consuming, and labor-intensive due to manual monitoring and resets, as depicted in Fig. 1(a). These physical challenges largely confine previous methods of real-world RL to offline data with heavy distribution shift to current policy [85, 64, 65, 80]. Ultimately, the policy improvement could be bottlenecked without sufficient on-policy data stream [52, 90, 72].

The gap between the simulator and the physical world motivates the development of world models, which first learn from passive experience and then simulate future outcomes conditioned on different actions [78, 27, 29, 30, 31, 50]. Nevertheless, constructing a world model applicable to real-world robotics poses fundamental challenges. For control, world models must faithfully follow actions to represent the accurate consequences. Despite the improved visual realism by integrating high-capacity generative models [87, 26, 99], how to improve controllability over various actions remains an open problem [55]. Furthermore, learning from imagination necessitates informative learning signals for intermediate actions, rather than relying solely on a binary indicator. Otherwise, determining terminal success would require the world model to simulate the entire task execution, which is beyond the reliable horizon of most generative world models [55, 57].

To handle these issues, we present RISE, a holistic learning framework that **R**einforces robot foundation model via **I**magination to enable **SE**lf-improving, as shown in Fig. 1(b). At its core is an online learning environment achieved by a learned world model. Inspired by prior works that decompose world modeling into tractable sub-problems to flexibly leverage heterogeneous architectures and priors [5, 20, 97, 86], we build a Compositional World Model that factorizes the simulation problem into two objectives, dynamics prediction and value estimation, allowing each to be instantiated with architectures and training objectives best suited to its role.

Built on an efficient video diffusion model [59, 28], we pre-train our dynamics model on large-scale robot datasets with a Task-centric Batching strategy to improve action controllability, which contributes to effective fine-tuning on targeted tasks. The value model is initialized from a pre-trained VLA backbone [8] and adapted with both progress estimate [66, 92, 25] and Temporal-Difference learning [77] objectives, providing dense and failure-sensitive evaluation of imagined states. These components are combined to compute advantages for candidate actions, enabling stable policy improvement via advantage-conditioned training. As a result, RISE performs on-policy reinforcement learning effectively in imagination. As presented in Fig. 2, we rigorously evaluate RISE on a suite of real-world tasks that stress-test dynamic adaptation and precision. The results demonstrate that RISE outperforms previous RL methods by a non-trivial margin, while avoiding costly real-world trial-and-error.

Our contributions are threefold: **(1)** We propose RISE, a principled framework for robotic reinforcement learning, that enables autonomous self-improvement in a scalable and online manner. RISE overcomes the physical restrictions posed by prior art by shifting the robotic interactions from physical environment to imaginative space. **(2)** At the core of this system is an online learning environment achieved by a Compositional World Model that builds reliable dynamics and value estimates for real-world tasks. We unveil critical design choices to derive stable learning signals for policy improvement. **(3)** Through extensive experiments on dexterous tasks, we demonstrate that RISE exhibits significantly higher performance compared to existing RL methods.

We view our work as the first study on leveraging world models as an effective learning environment for challenging real-world manipulation, bootstrapping performance on tasks requiring high dynamics, dexterity, and precision. Code and models will be released publicly.

II. Preliminary

A. World Model Formulation

We aim to construct a world model consisting of a dynamics model for predicting future states and a value model for predicting rewards over different courses of action. Crucially, these predicted rewards are converted into advantages to guide RL training. Formally, let $o_t = [m_t^1, \dots, m_t^n]$ be the multi-view observation at time t with n camera views. We apply a history window of length N as $\mathbf{O}_t = \{o_{t-N}, \dots, o_{t-1}, o_t\}$ to capture temporal dependency. The conditional action $\mathbf{a}_t$ is drawn from a running policy π as $\mathbf{a}_t = [a_t, a_{t+1}, \dots, a_{t+H-1}] \sim \pi(\cdot|o_t, \ell)$, where $\mathbf{a}_t$ is commonly applied as a sequence of actions with chunk length H, *i.e.*, action chunk, and ℓ is a language instruction describing the task. The dynamics model $\mathcal{D}$ predicts future observations $\{\hat{o}_{t+1}, \dots, \hat{o}_{t+H}\}$ conditioned on both the historical context and the proposed action sequence:

$$\hat{o}_{t+1}, \dots, \hat{o}_{t+H} = \mathcal{D}(\mathbf{O}_t, \mathbf{a}_t). \tag{1}$$

To evaluate the utility of imagined trajectories, we further introduce a value model $\mathcal{V}$, which assigns a progress signal towards successful completion conditioned on observation and task instruction as $\mathcal{V}(\hat{o}_t, \ell)$. We define the advantage as

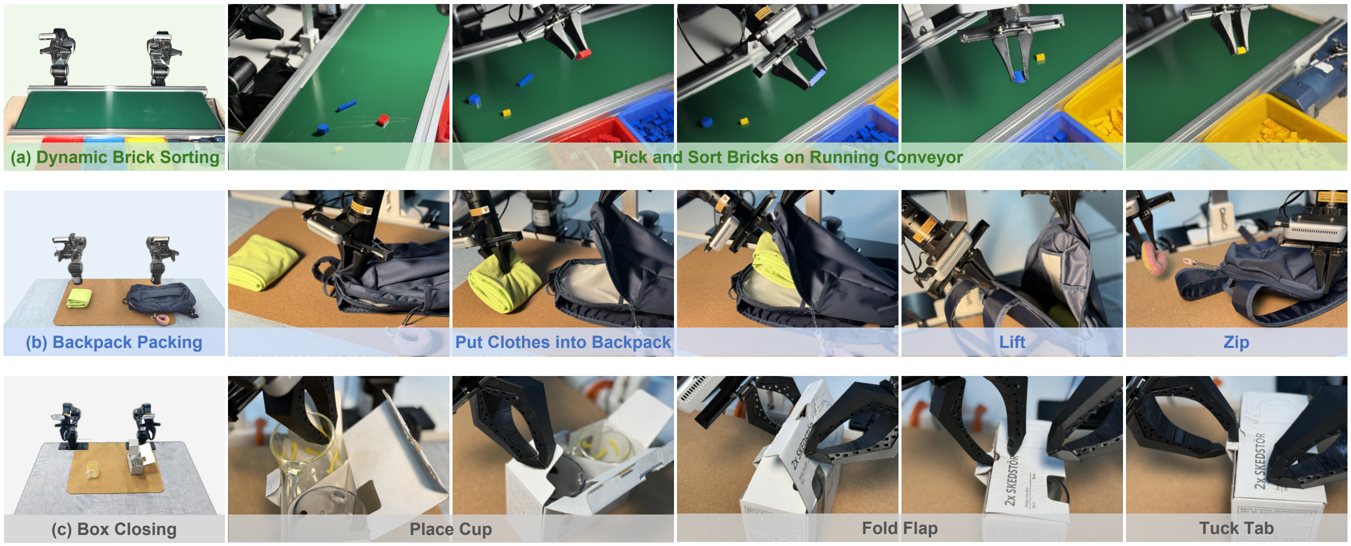

(a) Dynamic Brick Sorting
Pick and Sort Bricks on Running Conveyor
(b) Backpack Packing
Put Clothes into Backpack
Lift
Zip
(c) Box Closing
Place Cup
Fold Flap
Tuck Tab

Fig. 2: **Evaluation task suite of RISE. Left**: Tabletop setting. **Right**: Zoomed-in details of each task procedure. **Dynamic Brick Sorting** involves precisely picking up colored bricks from a moving conveyor and placing them into the corresponding color-designated bins. **Backpack Packing** requires the robot to open, insert clothes, lift, and zip the backpack. **Box Closing** necessitates subtle controls to fold the flap and tuck the tab into the box precisely.

the average cumulative improvement across the entire chunk. Specifically, we compute the difference between the value of each predicted future observation $\hat{o}_{t+k}$ and the initial observation o_t as the reward of action a_{t+k}, then take the expectation over the horizon of the action chunk as the advantage:

$$A(o_t, \mathbf{a}_t, \ell) = \left(\frac{1}{H} \sum_{k=1}^{H} \mathcal{V}(\hat{o}_{t+k}, \ell) \right) - \mathcal{V}(o_t, \ell), \tag{2}$$

where A is associated with the action chunk proposed by the policy π, forming the learning signal for policy optimization. The interaction between $\mathcal{D}$ and $\mathcal{V}$ occurs in imagination space, and both modules are compatible with *multi-view* images.

B. Reinforcement Learning

We formulate the problem as a standard RL setting with decision-making process as a Markov Decision Process (MDP) characterized by the tuple $(\mathcal{O}, \ell, \mathcal{A}, H, r)$. At each timestep t, given an observation $o_t \in \mathcal{O}$ and task instruction ℓ, the policy π generates an action sequence $\mathbf{a}_t \in \mathcal{A}^H$ of horizon H, obtaining reward r for each step. The interaction between the policy and the environment induces a trajectory distribution $\rho_\pi(\tau)$, where $\tau = (o_0, \mathbf{a}_0, \ldots, o_T) \in \mathcal{O} \times \mathcal{A} \cdots \mathcal{O}$. The objective is to maximize the expected return $\mathcal{J}(\pi) = \mathbb{E}_{\tau \sim \rho_\pi}[\sum_{t=0}^{T} r(o_t, \mathbf{a}_t)]$. To quantify the quality of a specific action sequence relative to the average policy performance, we utilize the advantage function $A^\pi(o_t, \mathbf{a}_t, \ell)$, estimated via Eq. (2).

To ensure stable improvement over a reference policy π_{ref}, we adopt the probabilistic inference framework from $\pi^*_{0.6}$ [2]. Rather than maximizing a regularized objective directly, we construct a target distribution $\hat{\pi}$ by weighting π_{ref} with the probability of improvement I:

$$\hat{\pi}(\mathbf{a}_t | o_t, \ell) \propto \pi_{\text{ref}}(\mathbf{a}_t | o_t, \ell) \cdot p(I \mid A^{\pi_{\text{ref}}}(o_t, \mathbf{a}_t, \ell))^\beta. \tag{3}$$

Since improvement is fully determined by the advantage value, we have $p(I | A^{\pi_{\text{ref}}}) \equiv p(I | \mathbf{a}_t, o_t, \ell)$. Applying Bayes' rule allows us to express the improvement likelihood as a density ratio:

$$p(I \mid \mathbf{a}_t, o_t, \ell) \propto \frac{\pi_{\text{ref}}(\mathbf{a}_t \mid I, o_t, \ell) p(I | o_t, \ell)}{\pi_{\text{ref}}(\mathbf{a}_t \mid o_t, \ell)}. \tag{4}$$

Substituting Eq. (4) into the target distribution and setting $\beta = 1$ cancels the unconditional prior π_{ref}, yielding the simplified objective $\hat{\pi}(\mathbf{a}_t | o_t, \ell) = \pi_{\text{ref}}(\mathbf{a}_t \mid I, o_t, \ell)$. Practically, we implement this by conditioning the policy on discretized advantages, guiding generation toward high-return trajectories.

III. Methodology

Our approach is structured as follows: In Sec. III-A, we propose a Compositional World Model that composes dynamics prediction with value estimation, providing an interactive environment with informative learning signals. In Sec. III-B, we establish a Policy Warm-up stage on real-world experience to anchor the policy to practical behavioral distribution and equip it with advantage-conditioned capabilities. In Sec. III-C, we present a Self-Improving Loop that iteratively generates imaginary rollouts and optimizes the policy within the world model. Implementation details with compute allocation are covered in Sec. III-D.

A. Compositional World Model

Scalable RL necessitates precise environment modeling to map current states and policy actions to future dynamics and rewards. To this end, we introduce a Compositional World Model to disentangle dynamics prediction from value estimation, thereby enabling independent architectural optimization for each component. Starting from a context observation, the

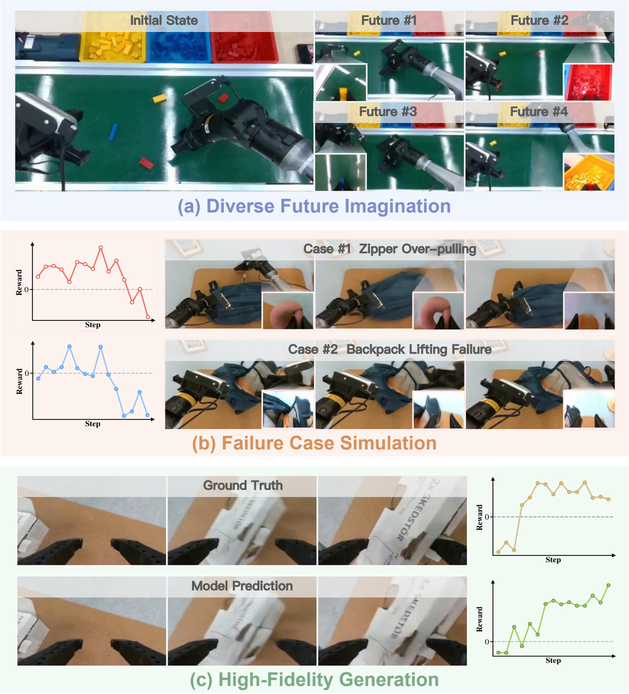

Initial State
Future #1
Future #2
Future #3
Future #4
(a) Diverse Future Imagination
Reward
Step
Case #1 Zipper Over-pulling
Case #2 Backpack Lifting Failure
(b) Failure Case Simulation
Ground Truth
Model Prediction
(c) High-Fidelity Generation

Fig. 3: **Qualitative imaginations produced by RISE.** Given initial multi-view context and candidate action chunks, RISE can (a) emulate a variety of future accordingly, (b) simulate failure cases with corresponding reward drops, and (c) maintain coherent predictions consistent with real executions.

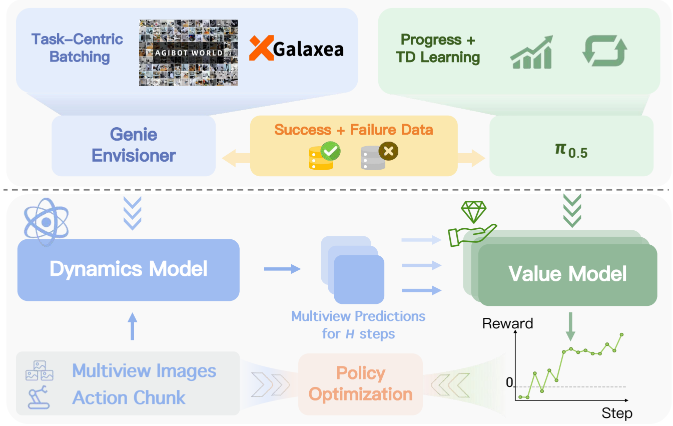


Fig. 4: **Workflow of compositional world model. Top**: Training recipe upon proper model initialization. **Bottom**: Inference pipeline that yields rewarded samples for policy optimization. Both modules are compatible with multi-view images. We omit text prompt for both policy and value model for brevity.

dynamics model emulates a faithful future under the candidate action chunk, which would be evaluated by the value model to derive an advantage for policy improvement. We show samples from imagination in Fig. 3 qualitatively. Crucially, the model is employed exclusively during training, imposing zero computational overhead at inference. The training recipe and inference pipeline of our world model are shown in Fig. 4.

Controllable Dynamics Model. Reliably simulating future states for RL yields two fundamental requirements: (i) The generation latency should not be prohibitively high, which would bottleneck the throughput of the RL system. (ii) The generated states should not only be plausible in visuals but also consistent with the conditional actions. Thereby, we initialize our dynamics model from pre-trained Genie Envisioner [59], *i.e.*, GE-base variant, which inherits the architectural advances in LTX-Video [28] and features a favorable trade-off between generation quality and inference speed. In comparison, advanced world models such as Cosmos [1] takes more than 10 minutes for synthesizing 25 multi-view observations, whereas GE only requires less than 2 seconds to achieve such a horizon, leading to 300x speedup. Such generation efficiency is a critical pillar for applicable RL training.

Despite its efficiency, GE-Base is originally conditioned on text rather than fine-grained robot actions. To endow the model with precise action controllability that could be further transferred into task-specific scenarios, we further optimize the model on large-scale action-labeled datasets, including Agibot World [11] and Galaxea [43], by incorporating an additional light-weight action encoder. Additionally, we impose stronger noise on context frames compared to the original GE-base training, to improve the generation robustness when encountering motion blurs and visual artifacts that might occur in both recorded and synthesized data. Nevertheless, fine-tuning a controllable world model on heterogeneous action data is prone to instability and slow convergence when diverse tasks and visual domains are included within the same batch for each optimization iteration. We mitigate this issue with a *Task-Centric Batching* strategy, where each batch is sampled from a small fraction of tasks while covering more samples of the same task correlated with different actions. Intuitively, this batching strategy prioritizes action diversity under the same scene over scenario diversity for batch optimization, thus contributing to improved action controllability. Empirically, applying this strategy improves both task-specific fine-tuning efficiency, as in Table V, and stronger policy improvement, as in Table IV. With these design choices, our dynamics model is capable of providing fast and faithful multi-view state prediction to support the self-improving loop.

Progress Value Model. Imagination-based policy improvement critically depends on a reward-related signal that is (i) dense over long horizons and (ii) sensitive to subtle failures in contact-rich manipulation. We therefore learn a value estimator $\mathcal{V}$ that maps sensory observations to a scalar value used to score imagined rollouts. $\mathcal{V}$ is parameterized from a pre-trained VLA policy $\pi_{0.5}$ [8], that brings in two advantages. First, $\pi_{0.5}$ has been trained on broad robot datasets and thus carries robot-centric understanding that transfers naturally to value estimation. Second, the policy backbone is compatible with multi-view inputs, whereas generic VLMs are mostly developed on single-view images without such adaptation.

As for training, we warm-start $\mathcal{V}$ with a simple temporal progress estimate as objective, which equips our value model with a coarse understanding of monotonic temporal structure.

$$\mathcal{L}_{\text{prog}} = \mathbb{E}_{(o_t,\ell)\sim\mathcal{D}_{\text{exp}}}\left[\left(\mathcal{V}(o_t,\ell) - {}^{t}\!/\!_{T}\right)^2\right], \tag{5}$$

where t indexes the current timestep within an episode of length T. While progress regression provides a dense signal, it is often overly smooth and can be insensitive to failures, especially in contact-rich settings where execution errors might be subtle in visuals. To conquer this, we augment the progress loss with Temporal-Difference (TD) learning [77], which uses both successful demonstrations and failure rollouts to establish a value function that distinguishes success from errors.

$$\begin{aligned} \mathcal{L}_{\text{TD}} &= \mathbb{E}_{(o_t,\ell,o_{t+1})\sim\mathcal{D}}\left[\left(\mathcal{V}(o_t,\ell) - y_t\right)^2\right], \\ y_t &= r_t + \gamma\mathcal{V}(o_{t+1},\ell), \end{aligned} \tag{6}$$

where γ is the temporal discount factor, and r_t is set to 0 in intermediate steps while being $+1/-1$ at the end of successful and failure episodes, respectively. Our final value learning objective simply combines both terms $\mathcal{L}_{\mathcal{V}} = \mathcal{L}_{\text{prog}} + \mathcal{L}_{\text{TD}}$ to leverage both the learning stability and error sensitivity provided by two terms, respectively.

B. Policy Warm-up on Real-world Experience

Before performing the on-policy improvement, we first warm-start the learning process with offline-collected data, which anchors the policy to a physically plausible behavior distribution on the targeted task, avoiding careless exploration in the later stage. Both data composition and training objective mainly follow RECAP [2]. For each task, we fine-tune the pre-trained policy, *i.e.*, $\pi_{0.5}$ [8], on offline collected data, comprising expert demonstrations, policy rollout with success and failure, and human-intervened correction. During training, the policy is conditioned on an advantage signal, labeled by our learned value model $\mathcal{V}$ as in Eq. (2), by treating $\hat{o}_{t+k}$ as later frames from an offline recorded video. Different from the practice in RECAP that labels advantage for offline data and policy rollout, in early experiments, we found that assigning advantages for both sources yields worse results than labeling for rollout only. Thereby, only rollout data is assigned the learned advantages whereas both expert and human correction data are directly paired with optimal advantages, denoted as $\mathbb{1}$, in our experiments. Consequently, this warm-up stage empowers the policy to absorb action data in different qualities, which is critical for the next self-improvement stage that learns from trial-and-error in an online manner.

C. Self Improving with World Model

With the advantage conditioning capability acquired from the warm-up stage on offline data, we then apply the compositional world model as an interactive simulator to improve the policy. The self-improving loop executes the Rollout stage and Training stage iteratively, as shown in Fig. 5.

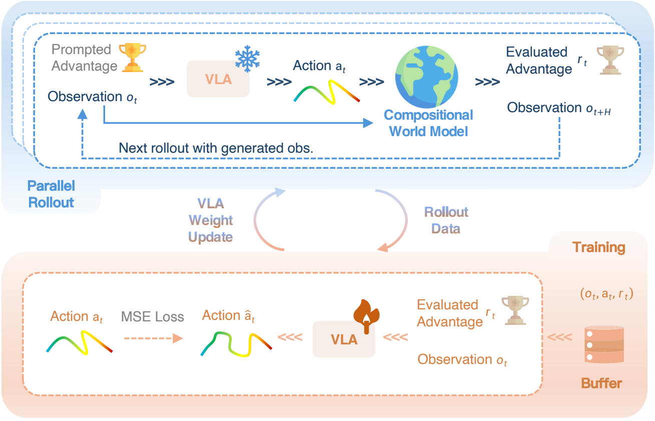


Fig. 5: **Self-improving loop of RISE.** Our learning pipeline encompasses two stages. **Top**: Rollout stage. Prompted with an optimal advantage, the rollout policy interacts with the world model to produce rollout data. **Bottom**: Training stage. The behavior policy is then trained to generate proper action under an advantage-conditioning scheme.

Rollout Stage. To start off, we sample an initial state from the warm-up offline dataset. Along with the observation, we additionally prompt the rollout policy π_{rollout} with an optimal advantage $\mathbb{1}$, to infer an action with positive intent.

$$\hat{\mathbf{a}}_t = \pi_{\text{rollout}}(\mathbb{1}, o_t, \ell). \tag{7}$$

Visual history and action proposal are fed into the dynamics model to synthesize the next H visual states. These imagined states are then evaluated by the value model to compute the actual advantage of the proposed action, denoted as $A^{\pi_{\text{rollout}}}(o_t, \hat{\mathbf{a}}_t, \ell)$. We define $\mathbb{1}$ as the prompted advantage for inferring optimal actions, whereas $A^{\pi_{\text{rollout}}}(o_t, \hat{\mathbf{a}}_t, \ell)$ denotes the evaluated advantage, reflecting the true utility of the generated action. This advantage is discretized into one of N uniform bins representing the practical advantage of the action in the current state. To broaden the state coverage during the online training process, imagined states would also serve as input for the subsequent rollout. From each offline state, such consecutive interaction would be conducted at most two times, considering the known error accumulation issue of generative video models [38]. The rollout policy parameters are updated via an Exponential Moving Average (EMA) [40], blended from behavior policy weights. One major difference between RISE and prior approaches that also leverage world model as learning environment [44, 93, 13] is that RISE avoids explicitly simulating terminal states to obtain rewards, yet produces chunk-wise advantage for proposed actions directly.

Training Stage. The on-policy rollout data $\langle o, \hat{a}, A\rangle$ form batch samples to optimize the policy. The VLA is trained to minimize the distance between its output and the proposed action $\hat{a}$, given the evaluated advantage A as a condition. This allows the policy to learn from both high-advantage successes and low-advantage failures discovered in imagination. To prevent catastrophic forgetting during exploration, we also mix offline labeled data into the batch data. Both offline and online

TABLE I: **Performance comparisons on real-world tasks.** We evaluate success rates and scores across three diverse tasks, ranging from dynamic sorting to precise packing. RISE exhibits superior performance compared to baselines in all scenarios.

Method	Dynamic Brick Sorting		Backpack Packing		Box Closing	
	Succ. (%)	Score	Succ. (%)	Score	Succ. (%)	Score
$\pi_{0.5}$ [8]	35.00	8.28	30.00	4.25	35.00	7.50
$\pi_{0.5}$+DAgger [73, 45]	15.00	6.10	50.00	7.00	40.00	7.50
$\pi_{0.5}$+PPO [75]	10.00	7.68	35.00	5.88	10.00	4.75
$\pi_{0.5}$+DSRL [80]	10.00	6.65	10.00	3.50	10.00	7.63
RECAP [2]	50.00	9.00	40.00	6.13	60.00	8.13
RISE (Ours)	**85.00**	**9.78**	**85.00**	**9.50**	**95.00**	**9.88**

experiences are leveraged under unified learning objective:

$$\pi(A^{\pi_{\text{rollout}}}(o, \hat{\mathbf{a}}_t, \ell), o_t, \ell) \to \hat{\mathbf{a}}. \tag{8}$$

which is optimized under generic flow-matching criteria [7, 8].

D. Implementation Details

World Model Training. The dynamics model goes through two phases. The pre-training stage on Galaxea [43] and Agibot World [11] is conducted on 16 NVIDIA H100 GPUs with a global batch size of 512, taking about seven days. Subsequently, for task-specific fine-tuning, we utilize 8 NVIDIA H100 GPUs with a global batch size of 64, which takes about three days to complete. Parameterized from a pre-trained VLA [8], the value model is directly fine-tuned on task-specific data, thanks to the robot-centric knowledge inherited from the policy backbone. We apply progress estimate loss only for the first 10k training steps and include TD learning loss additionally for the remaining 40k steps. With a total batch size of 64 on 8 GPUs, the model converges in about one day of training. Importantly, both modules of our world model are only applied during the policy learning phase, thus posing **zero** inference overhead during real-world policy execution.

Policy Training. The policy warm-up phase largely follows the training procedure of RECAP [2] on an offline collected dataset, where the policy is conditioned on advantage labeled by our learned value model. The following self-improving stage then goes around 10k steps. For both stages, global batch size is 64 on 8 GPUs.

Task-specific Data. Both our world model and policy share the same set of offline data for each task, including expert demonstrations and policy rollouts with success and failure, except that policy learning also consumes a fraction of DAgger data to enrich the recovery mode, similar to RECAP [2].

IV. Evaluations

We conduct a comprehensive evaluation to investigate the capabilities of RISE. In particular, we focus on the following questions: **Comparative Analysis:** Does RISE outperform existing mainstream RL and IL methods, particularly in real-world dexterous and long-horizon tasks? **Design Choices**: How can the world model be effectively integrated into the RL loop, and is each module design essential?

A. Real-world Experimental Setup

Our real-world experiments employ a dual 7-DoF AgileX robot with absolute joint control. We benchmark three dexterous, long-horizon tasks, including: **Dynamic Brick Sorting**: The robot is required to sort diverse bricks dynamically on an operating conveyor belt, shown in Fig. 2(a), **Backpack Packing**: This task presents challenges involving compliant and deformable object manipulation as in Fig. 2(b). **Box Closing** The task requires precise bi-manual coordination to package a cup, as in Fig. 2(c). Notably, ablations are conducted on the most challenging task in practice, *i.e.*, Dynamic Brick Sorting. Hyperparameters remain fixed across variants. Detailed robot setup and evaluation metrics are included in the Appendix.

B. Main Results

Baselines. We benchmark RISE against state-of-the-art imitation and reinforcement learning baselines. Each counterpart is developed with a close compute budget. Implementation and data composition for each variant are detailed in the Appendix.

- $\pi_{0.5}$ [8]: A state-of-the-art VLA pre-trained on web-scale multi-robot data and fine-tuned on task demonstrations.
- $\pi_{0.5}$ **+ DAgger** [73, 45]: An interactive baseline utilizing on-policy human corrections to mitigate exposure bias.
- $\pi_{0.5}$ **+ PPO** [75]: A standard online RL baseline fine-tuning VLA weights via PPO.
- $\pi_{0.5}$ **+ DSRL** [80]: A sample-efficient method steering frozen VLAs by optimizing diffusion latent noise via RL.
- **RECAP** [2]: An advantage-conditioned offline RL approach [23, 48] originally built off a proprietary pre-trained policy, *i.e.*, $\pi_{0.6}$ [2]. Due to the inaccessibility of $\pi_{0.6}$, we apply this approach to $\pi_{0.5}$ upon the same parameter-tuning and offline data corpora as ours.

Results. We present quantitative results in Table I, reporting both *Success Rate* and *Stage-wise Score*, with evaluation criteria provided in the Appendix. Although $\pi_{0.5}$ offers preliminary capability, we observe that online adaptation (PPO, DSRL) incurs severe instability. This leads to performance degradation, *e.g.*, a sharp drop (35%→ 10%) in the *Dynamic Brick Sorting* task. RECAP validates the benefit of advantage conditioning but falls short of RISE. Notably, our method yields a 40% margin in *Backpack Packing*, while increasing success rates to 85% and 95% on the brick and box tasks, respectively. Overall,

TABLE II: **Ablation on offline data ratio.** Overall performance peaks at 0.6, indicating that balanced offline data is crucial for complex generalization.

Ratio	Pick&Place Succ. (%)	Sort Acc. (%)	Complete Succ. (%)	 Score
0.1	15.00	83.33	5.00	1.35
0.3	78.75	80.95	25.00	7.03
0.6	**90.00**	**87.50**	**50.00**	**8.32**
0.9	90.00	80.56	30.00	7.90

TABLE III: **Ablation on online action and state integration.** Results demonstrate the necessity of incorporating both online action proposed by the rollout policy and the online state generated by the dynamics model.

Online Action	Online State	Pick&Place Succ. (%)	Sort Acc. (%)	Complete Succ. (%)	 Score
✗	✗	80.00	76.56	35.00	6.98
✓	✗	96.25	84.42	40.00	8.73
✓	✓	**98.75**	**92.41**	**70.00**	**9.43**

TABLE IV: **Ablations on the modular designs of dynamics and value models.** "w/o Progress" indicates that the value model is trained without the auxiliary progress loss. Our full architecture proves to be the most effective across all metrics.

Module Variants		Pick&Place Succ. (%)	Sort Acc. (%)	Complete Succ. (%)	 Score
Dynamics	w/o Pre-train	97.50	60.26	15.00	7.43
	w/o Task-Centric	93.75	89.33	40.00	8.78
Value	w/o Progress	95.00	86.84	50.00	8.78
	w/o TD Learning	98.75	72.15	35.00	8.38
RISE (Ours)	w/ all designs	**98.75**	**92.41**	**70.00**	**9.43**

RISE significantly outperforms all RL and IL baselines across all tasks, with consistently high success rate.

C. Ablation Study

What ratio of the offline data should be allocated during RL training? Relying solely on online experience often leads to performance collapse due to the distribution shift between offline demonstrations and online rollouts. To address this, we investigate the optimal mixing ratio of offline data to retain performance. As shown in Table II, we observe a distinct trade-off. When the offline data ratio is too low (*e.g.*, 0.1), the success rate plummets to 5%. This confirms our hypothesis that insufficient offline retention leads to catastrophic forgetting in the face of massive online data. Conversely, an excessive ratio (*e.g.*, 0.9) also degrades performance. We attribute this to over-regularization, where the policy becomes too constrained to the offline distribution, hindering its ability to explore and discover superior policies.

Can VLA models benefit from world-model generated online actions or states? To validate this, we evaluate three variants: a baseline without online signals, one with online actions only, and the full RISE with both. Our results confirm the necessity of online signals. As shown in Table III, introducing online actions increases the success rate from 35% to 40%. We attribute this improvement to the expanded action space exploration; unlike the static behavioral mode typically found in offline data, online rollouts allow the VLA to distinguish between high-advantage actions and suboptimal failures. Crucially, incorporating online states further raises the success rate to 70%. This suggests that dynamically generated online states provide a richer, virtually unbounded training distribution, overcoming the limitations of fixed offline datasets.

How significant is the impact of the modules on RISE? Quantitative results in Table IV highlight the criticality of each component. In the dynamics model, removing visual pre-training drops sorting accuracy by 32.15% and completion

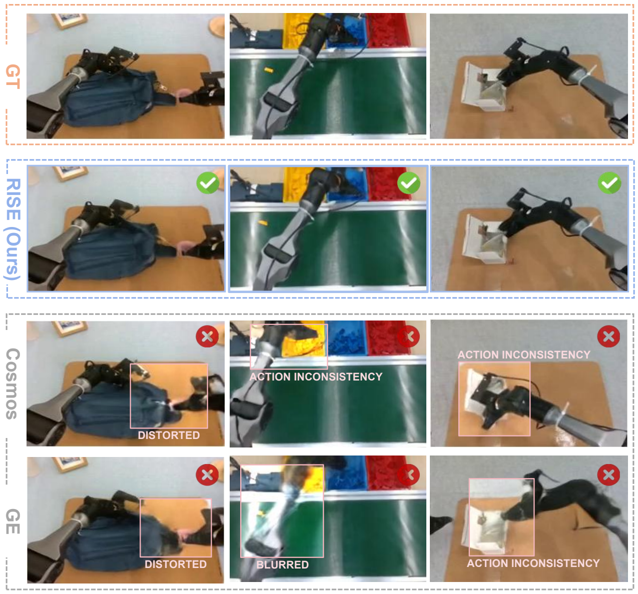


Fig. 6: **Qualitative Comparison on Dynamics Models.** Cosmos [1] and Genie Envisioner [59] suffer from geometric distortion, motion blurring, and physical inconsistency, whereas our method showcases temporally coherent and physically consistent results with Ground Truth (GT).

TABLE V: **Quantitative comparison of dynamics models.** ↑ (↓) denotes higher (lower) is better. Our method shows superior motion accuracy (EPE) and perceptual quality across both real-world tasks in Fig. 2 and the Bridge dataset [81].

Method	PSNR ↑	LPIPS ↓	SSIM ↑	FVD ↓	EPE ↓
Experiment #1: Fine-tuning on our real world tasks					
Cosmos	21.17	0.14	0.79	97.90	1.21
GE	21.16	0.11	0.79	85.72	1.05
RISE (w/o Task-Centric)	22.67	0.08	0.80	**61.22**	0.68
RISE (Ours)	**23.90**	**0.07**	**0.82**	66.84	**0.54**
Experiment #2: Fine-tuning on Bridge dataset [81]					
Cosmos	21.32	0.14	0.80	73.21	1.18
GE	21.47	0.12	0.79	64.55	0.96
RISE (w/o Task-Centric)	22.61	0.10	0.78	49.07	0.72
RISE (Ours)	**23.68**	**0.10**	**0.82**	**45.21**	**0.64**

to 15%, underscoring the need for visual priors. Absence of task-centric design reduces completion by 30%, validating the filtering of distractions. For the value model, ablating progress regression lowers success by 20%, confirming the importance of dense signals. Furthermore, omitting TD learning leads to a 35% decline, demonstrating its role in robust estimation.

How reliable is the dynamics model? We compare RISE with Cosmos [1] and Genie Envisioner (GE) [59] to investigate the reliability. We evaluate generation quality using PSNR, SSIM [82], LPIPS [95], and FVD [79], alongside optical flow end-point error (EPE) [93] for action controllability. Quantitatively, Table V underscores the superiority of RISE across all baselines under identical experimental settings. Notably, the significant reduction in EPE validates our task-centric pre-training, confirming that prioritizing action-conditioned dynamics effectively enhances motion awareness beyond standard pixel-level reconstruction. Qualitatively (Fig. 6), while baselines suffer from blurring and kinematic inconsistencies, RISE generates physically plausible dynamics with high fidelity. Additional comparisons are provided in the Appendix.

V. Related Work

A. World Models for Robot Learning

World models have been envisioned as a pathway to enable effective planning and learning through internal imagination [50, 27, 29, 78]. Early approaches in robotics and control focused on abstract state modeling in latent space with low-capacity dynamics model, which are limited in capturing the rich visual and contact dynamics required for real-world manipulation [29, 30, 31, 34, 35, 33, 49]. Recent advances in large-scale generative modeling renewed world modeling in high-fidelity observation space [10, 1, 32, 87, 88, 99]. However, adapting such models to serve as interactive environments for reinforcement learning remains challenging. Most approaches prioritize visual plausibility over action controllability, incurring prohibitive inference costs that prevent their use inside a reinforcement learning loop. Beyond dynamics prediction, reward and value shaping also introduce an additional bottleneck to apply these models to policy improvement. Prior efforts heavily rely on sparse terminal rewards or heuristic distance towards the goal state, which provide insufficient guidance for long-horizon manipulation and are brittle under long-term prediction errors [96, 93, 100, 68]. Importantly, prior works center around either simulated benchmarks [34, 35, 29, 30, 31, 32, 33, 24], low-level control problems [54, 84, 36, 74], or short-term tasks (*e.g.*, pick and place), with limited validation in real-world tasks under contact-rich and complex dynamics [93, 13, 44, 39, 49, 26, 41, 3, 98]. Motivated by prior efforts that carefully integrate heterogeneous modules to tackle the challenging world modeling problem [97, 5, 20], we seamlessly compose a dynamics model and a value function to achieve faithful simulation for various actions.

B. Reinforcement Learning for Foundation Policies

Reinforcement learning is increasingly used to strengthen VLA foundation policies on robustness and precision of manipulation. A large body of work adapts VLA post-training with RL within simulated environments [60, 69, 16, 67], where interactions are cheap, resettable, and parallelizable [63, 56, 14, 61, 17]. However, such scalability does not hold in the physical world, where interactions are serial, slow, and labor-intensive. Thereby, prior work on real-world RL is constrained to heavily reuse off-policy data while online interactions are performed on limited robot hardware only, which potentially bottlenecks the policy improvement and is hard to scale [65, 85, 4, 64]. Regarding learning stability, some work proposes to freeze the large-scale pre-trained policy while optimizing an additional residual policy [85] or input noise distribution only [80, 58]. With most parameters unchanged, such approaches sacrifice the adaptability of the policy to target tasks. In contrast, RECAP [2] enables finetuning the pre-trained policy via an advantage-conditioned formulation [23, 48], eliminating the complexity of adjusting the denoising chain for diffusion or flow-matching policy [51]. To derive reliable advantages for policy optimization, recent works resort to vision language models with a progress estimate formulation, which is numerically stable and free from laborious annotations [66, 94, 2, 92, 25]. However, such an objective is prone to the over-fitting problem and is less sensitive to subtle failures [2, 58]. Distinguished from prior approaches, we enable on-policy RL by shifting the learning environment from the physical world into an imaginative space via a learned world model. Furthermore, our value model benefits from both progress estimate and Temporal-Difference learning [77] in stability and failure sensitivity.

VI. Conclusion

We introduced RISE, a framework for on-policy reinforcement learning of robot foundation policies through imagination. RISE replaces the physical environment with imagination during training, enabling scalable online improvement without the prohibitive cost and risk of real-world exploration. Central to the system is a compositional world model that coherently orchestrates dynamics and value models, built from proper recipes, to efficiently emulate state and estimate advantage for policy improvement. Across real-world tasks spanning dynamic interaction, deformable-object handling, and bi-manual coordination, RISE consistently outperforms strong post-training baselines, proving that world models can be applied as an effective learning environment to improve policy performance on challenging manipulation tasks. We hope this work serves as a reference for the community in exploring scalable self-improving VLA models.

VII. Limitations and Future Work

The Gap between Imagination and Realism. The effectiveness of RISE is constrained by the accuracy and coverage of the learned world model. Although our compositional design improves controllability and consistency relative to prior generative simulators, the model can still produce physically implausible transitions in rare or underrepresented scenarios. Addressing this gap requires future work on uncertainty-aware imagination and principled integration of physical constraints that explicitly encode geometry properties.

The Simulated–Real Data Balance. Our results indicate that a non-trivial amount of real-world data remains essential to anchor the learning procedure. However, the optimal ratio between simulated rollouts and real-world experience requires further parameter tuning. Understanding the effectiveness and principles of these offline data represents an open problem.

From Physical Cost to Compute Cost. RISE shifts the primary bottleneck in robot learning from physical interaction to computation. While this trade-off releases the burden of physical interaction, training high-fidelity world models incurs a high computational cost. Improving the efficiency of world models will be critical for the compute-constrained regime.

Outlook. Taken together, these limitations suggest a promising pathway in integrating learned simulation into a broader data ecosystem, where model-based reinforcement learning complements scarce physical interaction. Discovering the right balance between these two key components points to a future of adaptive, robust, and sample-efficient robotic intelligence.

VIII. Acknowledgments

This study is supported by National Natural Science Foundation of China (62206172). This work is in part supported by the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025. 4, 7, 8, 19

[2] Ali Amin, Raichelle Aniceto, Ashwin Balakrishna, Kevin Black, Ken Conley, Grace Connors, James Darpinian, Karan Dhabalia, Jared DiCarlo, Danny Driess, et al. $\pi^*_{0.6}$: a vla that learns from experience. *arXiv preprint arXiv:2511.14759*, 2025. 3, 5, 6, 8, 14, 16, 17, 19

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 8

[4] Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. In *ICML*, 2023. 8

[5] Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to Manipulate: Compositional world models empowering robot imitation learning with imagination. *arXiv preprint arXiv:2412.14957*, 2024. 2, 8

[6] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 19

[7] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024. 1, 6, 19

[8] 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. *arXiv preprint arXiv:2504.16054*, 2025. 1, 2, 4, 5, 6, 16, 17, 19

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. In *RSS*, 2023. 1

[10] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *ICML*, 2024. 8

[11] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. AgiBot World Colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *IROS*, 2025. 4, 6, 19

[12] Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. In *RSS*, 2025. 19

[13] Akshay L Chandra, Iman Nematollahi, Chenguang Huang, Tim Welschehold, Wolfram Burgard, and Abhinav Valada. DiWA: Diffusion policy adaptation with world models. *arXiv preprint arXiv:2508.03645*, 2025. 5, 8

[14] Kang Chen, Zhihao Liu, Tonghe Zhang, Zhen Guo, Si Xu, Hao Lin, Hongzhi Zang, Quanlu Zhang, Zhaofei Yu, Guoliang Fan, et al. π_{RL}: Online rl fine-tuning for flow-based vision-language-action models. *arXiv preprint arXiv:2510.25889*, 2025. 8

[15] Li Chen, Chonghao Sima, Kashyap Chitta, Antonio Loquercio, Ping Luo, Yi Ma, and Hongyang Li. Intelligent robot manipulation requires self-directed learning. *OpenReview*, 2026. URL https://openreview.net/forum?id=Seb7rprW1Y. Accessed: 2026-01-02. 2

[16] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. 8

[17] Zengjue Chen, Runliang Niu, He Kong, Qi Wang, Qianli Xing, and Zipei Fan. TGRPO: Fine-tuning vision-language-action model via trajectory-wise group relative policy optimization. *arXiv preprint arXiv:2506.08440*, 2025. 8

[18] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor policy learning via action diffusion. In *RSS*, 2023. 19

[19] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal Manipulation Interface: In-the-wild robot teaching without in-the-wild robots. In *RSS*, 2024. 19

[20] Yilun Du, Sherry Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B. Tenenbaum, Leslie Pack Kaelbling, Andy Zeng, and Jonathan Tompson. Video Language Planning. In *ICLR*, 2024. 2, 8

[21] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 15

[22] Kuan Fang, Fangchen Liu, Pieter Abbeel, and Sergey Levine. MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting. In *RSS*, 2024. 1

[23] Kevin Frans, Seohong Park, Pieter Abbeel, and Sergey Levine. Diffusion guidance is a controllable policy improvement operator. *arXiv preprint arXiv:2505.23458*, 2025. 6, 8

[24] Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. *arXiv preprint arXiv:2503.18938*, 2025. 8

[25] Seyed Kamyar Seyed Ghasemipour, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, and Igor Mordatch. Self-improving embodied foundation models. *arXiv preprint arXiv:2509.15155*, 2025. 2, 8, 19

[26] Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025. 2, 8

[27] David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *NeurIPS*, 2018. 2, 8

[28] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. LTX-Video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 4

[29] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. *arXiv preprint arXiv:1912.01603*, 2019. 2, 8

[30] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering Atari with Discrete World Models. In *ICLR*, 2021. 2, 8

[31] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104*, 2023. 2, 8

[32] Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025. 8

[33] Nicklas Hansen, Yixin Lin, Hao Su, Xiaolong Wang, Vikash Kumar, and Aravind Rajeswaran. MoDem: Accelerating visual model-based reinforcement learning with demonstrations. *arXiv preprint arXiv:2212.05698*, 2022. 8

[34] Nicklas Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *ICML*, 2022. 8

[35] Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023. 8

[36] Nicklas Hansen, Jyothir SV, Vlad Sobal, Yann LeCun, Xiaolong Wang, and Hao Su. Hierarchical world models as visual whole-body humanoid controllers. *arXiv preprint arXiv:2405.18418*, 2024. 8

[37] Zheyuan Hu, Robyn Wu, Naveen Enock, Jasmine Li, Riya Kadakia, Zackory Erickson, and Aviral Kumar. RaC: Robot learning for long-horizon tasks by scaling recovery and correction. *arXiv preprint arXiv:2509.07953*, 2025. 2

[38] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 5

[39] Chia-Yu Hung, Navonil Majumder, Haoyuan Deng, Liu Renhang, Yankang Ang, Amir Zadeh, Chuan Li, Dorien Herremans, Ziwei Wang, and Soujanya Poria. NORA-1.5: A vision-language-action model trained using world model-and action-based preference rewards. *arXiv preprint arXiv:2511.14659*, 2025. 8

[40] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *UAI*, 2018. 5

[41] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. DreamGen: Unlocking generalization in robot learning through video world models. In *CoRL*, 2025. 8, 19

[42] Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards unified latent vla for whole-body loco-manipulation control. In *ICLR*, 2026. 19

[43] Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025. 4, 6, 19

[44] 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. *arXiv preprint arXiv:2509.19080*, 2025. 5, 8

[45] Michael Kelly, Chelsea Sidrane, Katherine Driggs-Campbell, and Mykel J Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. In *ICRA*, 2019. 2, 6

[46] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Paul 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 *CoRL*, 2024. 1, 19

[47] Moo Jin Kim, Chelsea Finn, and Percy Liang. Fine-tuning vision-language-action models: Optimizing speed and success. *arXiv preprint arXiv:2502.19645*, 2025. 19

[48] Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-conditioned policies. *arXiv preprint arXiv:1912.13465*, 2019. 6, 8

[49] Patrick Lancaster, Nicklas Hansen, Aravind Rajeswaran, and Vikash Kumar. MoDem-V2: Visuo-motor world models for real-world robot manipulation. In *ICRA*, 2024. 8

[50] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 2022. 2, 8

[51] Kun Lei, Huanyu Li, Dongjie Yu, Zhenyu Wei, Lingxiao Guo, Zhennan Jiang, Ziyu Wang, Shiyu Liang, and Huazhe Xu. RL-100: Performant robotic manipulation with real-world reinforcement learning. *arXiv preprint arXiv:2510.14830*, 2025. 8

[52] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020. 2

[53] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. BEHAVIOR-1K: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv preprint arXiv:2403.09227*, 2024. 19

[54] Chenhao Li, Andreas Krause, and Marco Hutter. Robotic World Model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*, 2025. 8

[55] Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E Gonzalez, et al. WorldModelBench: Judging video generation models as world models. *arXiv preprint arXiv:2502.20694*, 2025. 2

[56] Haozhan Li, Yuxin Zuo, Jiale Yu, Yuhao Zhang, Zhaohui Yang, Kaiyan Zhang, Xuekai Zhu, Yuchen Zhang, Tianxing Chen, Ganqu Cui, et al. SimpleVLA-RL: Scaling vla training via reinforcement learning. *arXiv preprint arXiv:2509.09674*, 2025. 2, 8

[57] Xinqing Li, Xin He, Le Zhang, Min Wu, Xiaoli Li, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025. 2

[58] Yunfei Li, Xiao Ma, Jiafeng Xu, Yu Cui, Zhongren Cui, Zhigang Han, Liqun Huang, Tao Kong, Yuxiao Liu, Hao Niu, et al. GR-RL: Going dexterous and precise for long-horizon robotic manipulation. *arXiv preprint arXiv:2512.01801*, 2025. 8

[59] Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, et al. Genie Envisioner: A unified world foundation platform for robotic manipulation. *arXiv preprint arXiv:2508.05635*, 2025. 2, 4, 7, 8, 16, 19

[60] Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *NeurIPS*, 2023. 2, 8

[61] Jijia Liu, Feng Gao, Bingwen Wei, Xinlei Chen, Qingmin Liao, Yi Wu, Chao Yu, and Yu Wang. What can rl bring to vla generalization? an empirical study. In *NeurIPS*, 2025. 2, 8

[62] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. RDT-1B: A diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 19

[63] 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. *arXiv preprint arXiv:2505.18719*, 2025. 2, 8

[64] 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 *ICRA*, 2024. 2, 8

[65] Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. *Science Robotics*, 2025. 2, 8, 17

[66] Yecheng Jason Ma, Joey Hejna, Chuyuan Fu, Dhruv

Shah, Jacky Liang, Zhuo Xu, Sean Kirmani, Peng Xu, Danny Driess, Ted Xiao, et al. Vision language models are in-context value learners. In *ICLR*, 2024. 2, 8

[67] Oier Mees, Lukas Hermann, Erick Rosete-Beas, and Wolfram Burgard. CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks. *RA-L*, 2022. 8

[68] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. In *CoRL*, 2023. 8

[69] Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. RoboTwin: Dual-arm robot benchmark with generative digital twins. In *CVPR*, 2025. 8

[70] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. RoboCasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024. 19

[71] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open X-Embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration. In *ICRA*, 2024. 19

[72] Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019. 2

[73] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011. 2, 6

[74] Pascal Roth, Jonas Frey, Cesar Cadena, and Marco Hutter. Learned perceptive forward dynamics model for safe and platform-aware robotic navigation. In *RSS*, 2025. 8

[75] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 6

[76] Modi Shi, Li Chen, Jin Chen, Yuxiang Lu, Chiming Liu, Guanghui Ren, Ping Luo, Di Huang, Maoqing Yao, and Hongyang Li. Is diversity all you need for scalable robotic manipulation? *arXiv preprint arXiv:2507.06219*, 2025. 1

[77] Richard S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 1988. 2, 5, 8

[78] Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 1991. 2, 8

[79] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019. 8

[80] 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 *CoRL*, 2025. 2, 6, 8, 14, 16, 17

[81] Homer Rich Walke, Kevin Black, Tony Z Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, et al. BridgeData v2: A dataset for robot learning at scale. In *CoRL*, 2023. 7, 19

[82] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE transactions on image processing*, 2004. 8

[83] Longyan Wu, Checheng Yu, Jieji Ren, Li Chen, Ran Huang, Guoying Gu, and Hongyang Li. FreeTacMan: Robot-free visuo-tactile data collection system for contact-rich manipulation. In *ICRA*, 2026. 19

[84] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. DayDreamer: World models for physical robot learning. In *CoRL*, 2023. 8

[85] Wenli Xiao, Haotian Lin, Andy Peng, Haoru Xue, Tairan He, Yuqi Xie, Fengyuan Hu, Jimmy Wu, Zhengyi Luo, Linxi Fan, et al. Self-improving vision-language-action models with data generation via residual rl. *arXiv preprint arXiv:2511.00091*, 2025. 2, 8, 17, 19

[86] Jiazhi Yang, Kashyap Chitta, Shenyuan Gao, Long Chen, Yuqian Shao, Xiaosong Jia, Hongyang Li, Andreas Geiger, Xiangyu Yue, and Li Chen. ReSim: Reliable World Simulation for Autonomous Driving. In *NeurIPS*, 2025. 2

[87] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Leslie Pack Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024. 2, 8

[88] Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. In *ICML*, 2024. 8

[89] Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024. 19

[90] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. In *NeurIPS*, 2020. 2

[91] Hongzhi Zang, Mingjie Wei, Si Xu, Yongji Wu, Zhen Guo, Yuanqing Wang, Hao Lin, Liangzhi Shi, Yuqing Xie, Zhexuan Xu, et al. RLinf-VLA: A unified and efficient framework for vla+ rl training. *arXiv preprint arXiv:2510.06710*, 2025. 17

[92] Shaopeng Zhai, Qi Zhang, Tianyi Zhang, Fuxian Huang, Haoran Zhang, Ming Zhou, Shengzhe Zhang, Litao Liu, Sixu Lin, and Jiangmiao Pang. A vision-language-

action-critic model for robotic real-world reinforcement learning. *arXiv preprint arXiv:2509.15937*, 2025. 2, 8

[93] Jiahui Zhang, Ze Huang, Chun Gu, Zipei Ma, and Li Zhang. Reinforcing action policies by prophesying. *arXiv preprint arXiv:2511.20633*, 2025. 5, 8

[94] Jiahui Zhang, Yusen Luo, Abrar Anwar, Sumedh Anand Sontakke, Joseph J Lim, Jesse Thomason, Erdem Biyik, and Jesse Zhang. ReWiND: Language-guided rewards teach robot policies without new demonstrations. *arXiv preprint arXiv:2505.10911*, 2025. 8

[95] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 8

[96] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024. 8

[97] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024. 2, 8

[98] Chuning Zhu, Raymond Yu, Siyuan Feng, Benjamin Burchfiel, Paarth Shah, and Abhishek Gupta. Unified world models: Coupling video and action diffusion for pretraining on large robotic datasets. In *RSS*, 2025. 8

[99] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024. 2, 8

[100] Fangqi Zhu, Zhengyang Yan, Zicong Hong, Quanxin Shou, Xiao Ma, and Song Guo. WMPO: World model-based policy optimization for vision-language-action models. *arXiv preprint arXiv:2511.09515*, 2025. 8

[101] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. In *NeurIPS*, 2023. 19

[102] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong T. Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag R. Sanketi, Grecia Salazar, Michael S. Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Lisa Lee, Tsang-Wei Edward Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil J. Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Kumar Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Marcin Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, 2023. 19

Appendix

The appendix is organized as follows:

- In **Appendix Sec. IX**, we present additional experimental results, including some ablation studies on minor components and visualization of the learned representations.
- In **Appendix Sec. X**, we list the comprehensive experimental settings, including hyperparameter configurations and hardware environments.
- In **Appendix Sec. XI**, we elaborate on the implementation details of modules of RISE.
- In **Appendix Sec. XII**, we provide a conceptual comparison between our method and several highly related works.
- In **Appendix Sec. XIII**, we provide more qualitative visualizations.
- In **Appendix Sec. XIV**, we analyze failure mode of RISE.
- In **Appendix Sec. XV**, we include additional related work on vision-language-action models.
- In **Appendix Sec. XVI**, we list the license for each asset used in this paper, *i.e.*, data and pre-trained weights.
- In **Appendix Sec. XVII**, we envision the broader impact of the proposed method.

IX. Additional Results

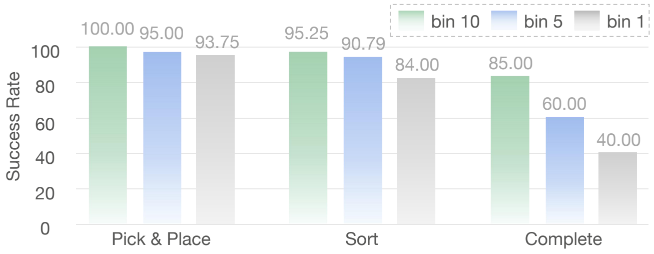


Fig. 7: **Task success rate across advantage bins.** A clear performance drop is observed from high to low advantage levels, especially in Sorting. This confirms that our policy effectively captures behavior diversity through advantage conditioning.

Can bins with different advantages reveal different performance of the policy? RISE utilizes advantage-based bins to guide RL training. We investigate whether the policy yields diverse task performance when conditioned on different bins. To this end, we evaluate the policy conditioned on high (Bin 10), neutral (Bin 5), and low (Bin 1) advantage bins. Results in Fig. 7 show a performance drop from bin 10 to bin 1, which supports our hypothesis. This performance drop is primarily attributed to sorting errors, as the success rate for sort deteriorates more significantly than for pick-and-place. Furthermore, the agent displays increased instability with lower bin indices. These findings demonstrate that our learned advantages are convincing and that the policy effectively captures the diversity of behaviors through our conditioning RL.

Can extended training of RL baselines match RISE's performance? To verify that our gains are not simply due to more training, we extended the RECAP and DSRL baselines with an extra 50k steps under the same batch size of our method. As shown in Fig. 8, RECAP saturates at a 30% to 50% success rate, while DSRL saturates at 5% to 10%. In contrast, RISE yields a +35% improvement (boosting success rate from 50% to 85%) with only $9k$ additional steps. We attribute this efficiency to online world model interaction, providing diverse samples to mitigate overfitting.

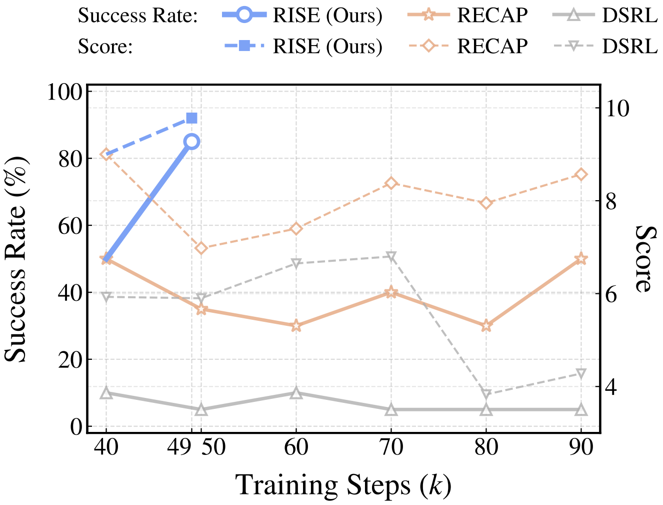


Fig. 8: **Learning dynamics of RL alternatives.** Compared to RECAP [2] and DSRL [80], RISE yields significantly higher results, which cannot be attained by the competing methods even with extended training [2] and increased real-world interactions [80].

TABLE VI: **Quantitative ablation on the pre-training of our dynamics model.**

Method	PSNR ↑	LPIPS ↓	SSIM ↑	FVD ↓	EPE ↓
RISE (w/o pre-train)	20.95	0.11	0.78	83.36	1.09
RISE (Ours)	**23.90**	**0.07**	**0.82**	**66.84**	**0.54**

What is the impact of pre-training and task-centric strategies on the generation quality of future dynamics? We investigate the impact of strategies on the generation quality of future dynamics. As shown in Table VI, pre-training significantly enhances video generation fidelity. Moreover, Fig. 10 provides visual comparisons, revealing that ablated variants (specifically *w/o task-centric* and *w/o pre-train*) suffer from action misalignment and severe blurring, whereas our method maintains high consistency with ground truth dynamics. Additionally, a sample-wise optical flow analysis in Fig. 9 isolates the role of the task-centric mechanism. The results demonstrate that this objective effectively enhances motion sensitivity, yielding sharper and more physically coherent predictions.

X. Real-world Experimental Details

A. Task Evaluation Standard

To provide a fine-grained analysis of policy performance beyond binary success, we define a quantitative evaluation

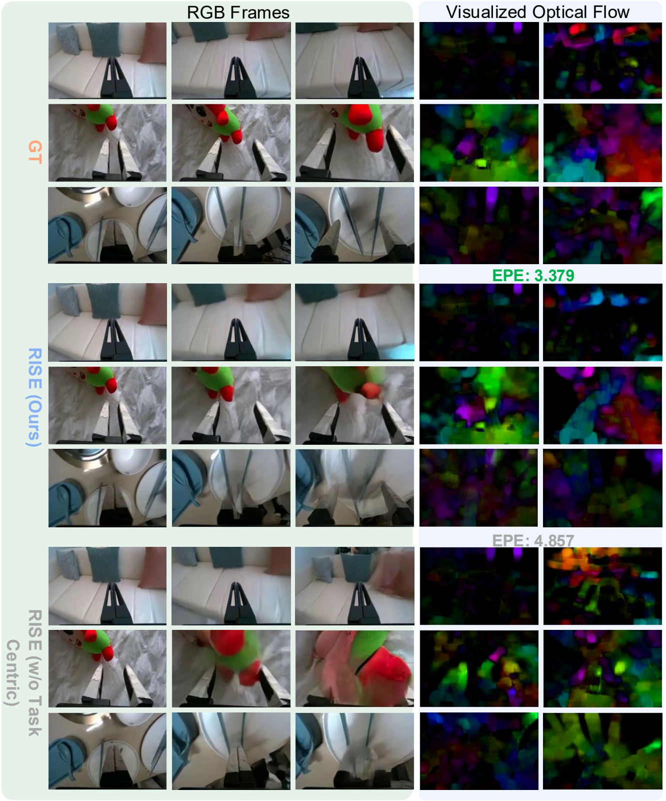


Fig. 9: **Task-centric versus non-task-centric during pre-train stage.** The optical flow maps demonstrate that our method captures action adherence more effectively during the initial stages of pre-training.

rubric detailed in Table VII. Given that our tasks involve multi-stage and long-horizon planning (as qualitatively illustrated in Figure 16), a simple success/failure metric fails to capture the incremental progress of the agent. Therefore, we decompose each task into distinct sub-goals, with a total score of 10 per episode to ensure consistency across different tasks. Throughout the paper, each of our evaluation results is based on an average score of 20 autonomous trials.

For the **Dynamic Brick Sorting** task, the scoring mechanism focuses on both manipulation robustness and classification accuracy. As the task involves processing a stream of objects, points are accumulated dynamically: successful grasping rewards the robot's low-level control, while correct placement into color-matched bins rewards semantic understanding. The score is capped at 10 to represent perfect clearing of the workspace.

For the **Backpack Packing** and **Box Closing** tasks, which are strictly sequential, we adopt a milestone-based scoring system. As shown in Table VII, these tasks are divided into four logical phases, with intermediate rewards assigned upon the completion of each sub-goal. This stepwise evaluation allows us to pinpoint exactly where a policy might degrade—whether during the initial interaction with deformable objects or during precision-critical phases like zipping or tab insertion.

TABLE VII: **Task evaluation standard.**

Task	Sub-goals	Total	Score
Conveyor	Grasp brick Place in matched bin Workspace cleared	10	1.0 each 1.5 each 10.0 max
Backpack	Open bag & Insert items Lift to settle contents Zip halfway Zip fully	10	2.5 5.0 7.5 10.0 max
Box	Load cup Fold side flaps Fold rear flap Tuck locking tab	10	2.5 5.0 7.5 10.0 max

B. Real-World Deployment

To bridge the gap between the discrete, low-frequency inference of the VLA model and the continuous, high-frequency requirements of real-world robotic control, we implemented an asynchronous control framework operating directly in joint space. Specifically, the VLA policy predicts action chunks with a horizon of $H = 50$ steps at a relatively low inference frequency, while the robot controller executes joint commands at a 30 Hz frequency. Instead of executing these chunks sequentially, which would cause motion freezing during inference, we adopt a Temporal Ensembling strategy that continuously integrates newly predicted action chunks into a running execution plan.

This integration is governed by a linear weighting scheme designed to smooth out transitions and suppress high-frequency jitter. When a new action chunk $\mathbf{a}^{\text{new}}$ is received from the inference thread, it overlaps with the unexecuted portion of the existing action sequence $\mathbf{a}^{\text{old}}$ in the buffer. For any time step t within this overlapping window, the final executed action command $\mathbf{a}_t \in \mathbb{R}^{14}$ (corresponding to the bi-manual setup in Figure 11) is derived via a time-varying linear interpolation between the previous plan $\mathbf{a}^{\text{old}}$ and the new prediction $\mathbf{a}^{\text{new}}$. This ensures that the robot's trajectory is primarily guided by the established plan at the beginning of the update to maintain continuity, while gradually shifting priority to the latest sensory observations towards the end.

XI. Implementation Details

A. Task-specific Data Composition

Dynamic Brick Sorting includes 3063 human demonstration data and 610 policy rollout data. Backpack Packing covers 2478 human demonstrations and 507 policy rollout data. Box Closing features 2286 human demonstrations, 524 policy rollouts, and 540 human corrections (DAgger) data.

B. Dynamics Model

Our dynamics model operates on multi-view RGB observations (192 × 256) captured from top-down and bilateral wrist cameras, conditioned on future actions. We employ a Flow Matching objective for training. For timestep scheduling, we adopt the Logit-Normal distribution following SD3 [21], defined as $\text{logit}(t) \sim \mathcal{N}(m, s^2)$, with $m = 0.2$ and $s = 1.0$.

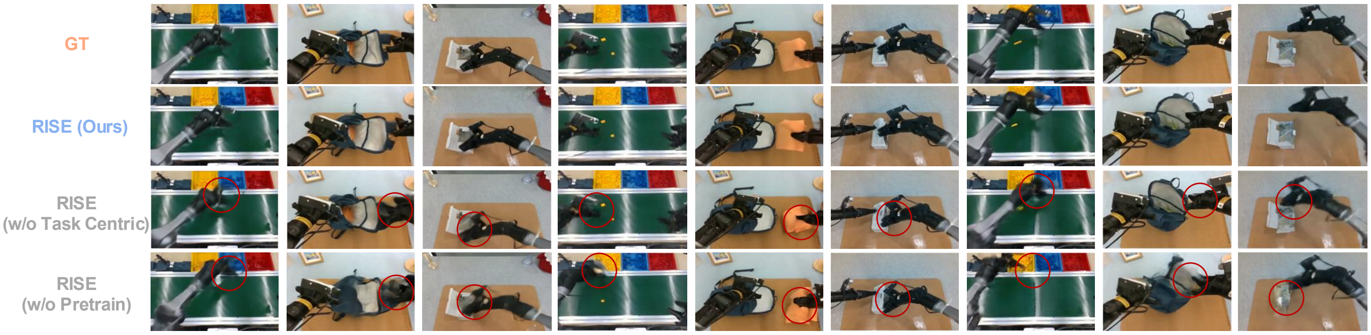


Fig. 10: **Visual ablation study on training strategies.** Compared to the other baselines, which exhibit significant degradation in image quality and motion coherence, our proposed method generates sharper, physically consistent predictions that strictly adhere to control actions.

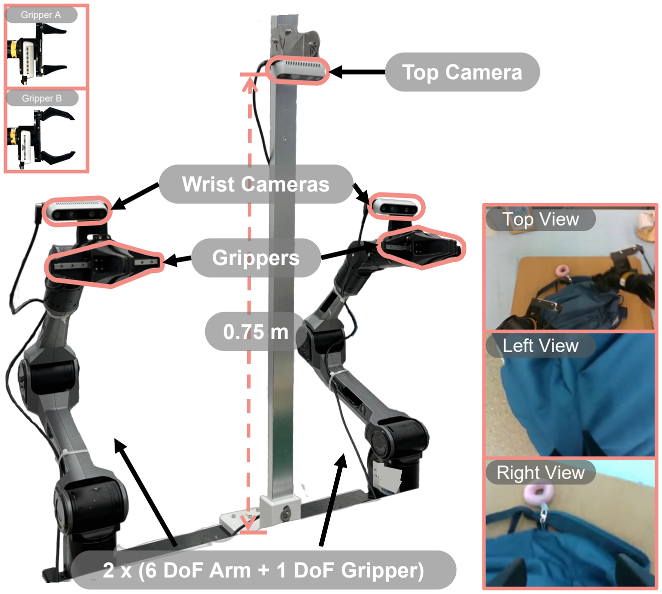


Fig. 11: **Experimental setup.** We utilize a bi-manual platform for our tasks. Each arm possesses 6 DoF along with a 1-DoF gripper, equipped with a wrist-mounted camera. To provide a global view, a top-down camera is positioned centrally between the arms at a height of approximately 0.75 m. The control frequency is set to 30 Hz. **Top Left:** We apply Gripper A for brick sorting and backpack packing, while applying Gripper B for box closing for the higher precision requirement.

Optimization is performed using AdamW with a constant learning rate of 1×10^{-4} after a linear warmup for 2k steps. During inference, we solve the flow ODE using the Euler discrete formulation, with 50 denoising steps. See Table VIII for more configurations.

TABLE VIII: **Hyper-parameters of dynamics model.**

Hyperparameter	**Value**
Basics	
Model initialization	GE-Base [59]
Input / Prediction frames	4 / 25
Number of views	3
Sampling frequency (pre-train / Fine-tune)	30 / 15 Hz
Optimization	
Training steps (pre-train / Fine-tune)	120k / 50k
Batch size (pre-train / Fine-tune)	512 / 64
Optimizer	AdamW
Learning rate	1×10^{-4}
Conditioned noise level σ	0.2

C. Value Model

The training configurations of the value model are listed in Table IX. For each task, the total training takes about 50k steps. For the first 10k steps, we apply progress estimate loss only, whereas for the remaining steps, we apply both progress estimate and Temporal Difference learning loss jointly. Notably, both dynamics model and value model are kept frozen during the self-improving loop for policy optimization.

D. Policy Optimization

The policy first gets warmed up mainly following the offline RL approach [2] with two differences. RECAP discretizes the labeled advantages into binary bins, yet we find that discretizing advantages into 10 bins with uniform intervals yields higher results. Moreover, directly assigning human demonstrations to the highest bins while labeling only the policy rollout data stabilizes learned behavior. These two discrepancies might emerge from the fact that our model initialization $\pi_{0.5}$ is not pre-trained with advantage conditioning, contrary to the offline RL pre-training as in $\pi^*_{0.6}$, where RECAP is instantiated. Subsequently, we start the self-improving loop with configurations listed in Table X.

E. Baseline Implementation

Throughout this paper, all policy variants, including baseline and our policy, are instantiated on pre-trained $\pi_{0.5}$ to fairly evaluate the effectiveness of various post-training strategies.

a) $\pi_{0.5}$: This variant is fine-tuned on our human demonstration corpus only via imitation learning, without using policy rollout or human correction data.

b) DSRL: The overall training configurations follow the official implementation of DSRL [80]. We utilize the $\pi_{0.5}$ model [8] as the base policy. To adapt the policy, we initialize

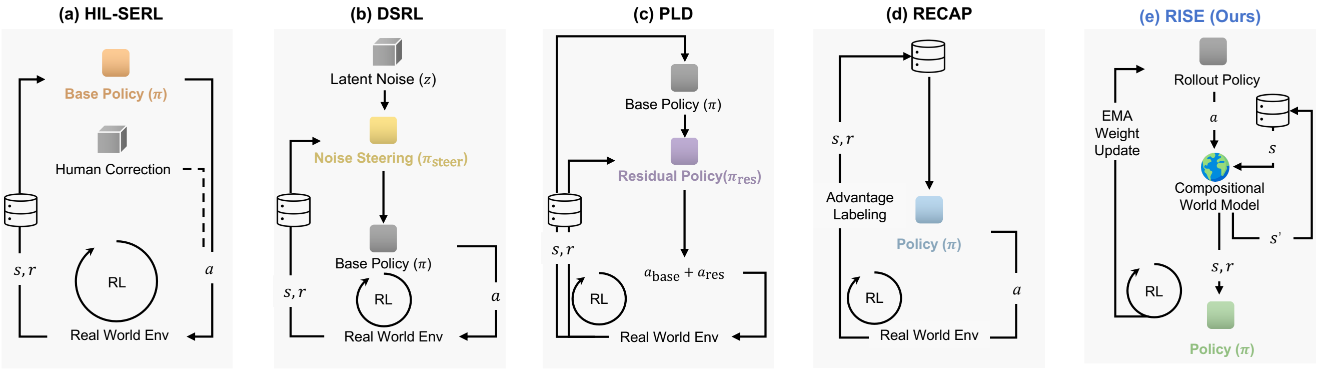


Fig. 12: **Conceptual comparisons with highly-related work.** Different from prior works that heavily rely on off-policy samples from real-world interactions for policy optimization [65, 80, 85, 2], RISE enables on-policy RL by building a world model as an interactive environment.

TABLE IX: **Hyper-parameters of value model.**

Hyperparameter	Value
Basics	
Model initialization	$\pi_{0.5}$ [8]
Input frames	1
Number of views	3
Optimization	
Training steps	50k
Batch size	64
Optimizer	AdamW
Learning rate	2.5×10^{-5}
Value discount factor	0.995

TABLE X: **Hyper-parameters of policy self-improving.**

Hyperparameter	Value
Batch size	64
Optimizer	cosine
Learning rate	1×10^{-4}
Minimum learning rate ratio	0.1
Rollout ema decay rate	0.995
Action chunk size	50
Action dimension	14

the replay buffer with 10 trajectories collected from the base policy sampled with standard Gaussian noise $w \sim \mathcal{N}(0, I)$, followed by 70 online steering episodes to fine-tune the behavior.

c) PPO: We initialize the PPO policy via a pre-trained $\pi_{0.5}$ model. At the rollout stage, we sample real-world trajectories, preserving the inference noise and log probabilities calculated according to RLinf [91]. During training, we use this stored inference noise to generate on-policy actions with gradient. We then compute the PPO loss by combining these actions with the new and old log probabilities and advantages. The PPO policy is updated by the PPO loss.

d) DAgger: Due to hardware constraints that preclude high-frequency mode switching, we adopt a single-intervention protocol where the human supervisor takes over upon imminent failure and completes the episode. This variant is trained on both expert demonstrations and additional human correction data via imitation learning.

e) RECAP: This variant follows the recipe of the policy warm-up stage, detailed in Sec. XI-D.

XII. Conceptual Comparisons with Highly-Related Work

We conceptually compare our method with highly-related work in Fig. 12. Contrary to prior methods that learn from off-policy data through costly real-world interactions, RISE enables on-policy reinforcement learning with a learned world model that generates new states and assigns advantage for each action chunk.

XIII. Qualitative Visualizations

Compositional World Model. We visualize rollout trajectories conditioned on distinct action sequences. As shown in Fig. 15, the dual-arm robot starts with the left arm grasping a blue brick. The expert trajectory executes a smooth pick-and-place operation into the target (blue) box, accompanied by an increasing reward curve. Similarly, the rollout driven by optimized actions exhibits a comparable trend. Notably, the generated video maintains high fidelity, accurately capturing complex environmental dynamics such as the operating conveyor belt. The corresponding reward curve shows improvement but remains slightly below the expert baseline, likely due to minor deviations in the optimized actions or subtle visual artifacts in the imagination. In contrast, the suboptimal trajectory clearly depicts the arm misplacing the brick into the wrong (yellow) box. Consequently, the reward rises during the picking phase but drops significantly once the arm moves toward the incorrect target. These results demonstrate the reliability of our world model in capturing both visual realism and logical consistency.

Dynamics Model. We provide a comprehensive visual assessment to benchmark our dynamics model against state-of-the-art alternatives. As shown in Fig. 19, our method distinctly outperforms other approaches, particularly in maintaining high image quality and precise action alignment. Extending this

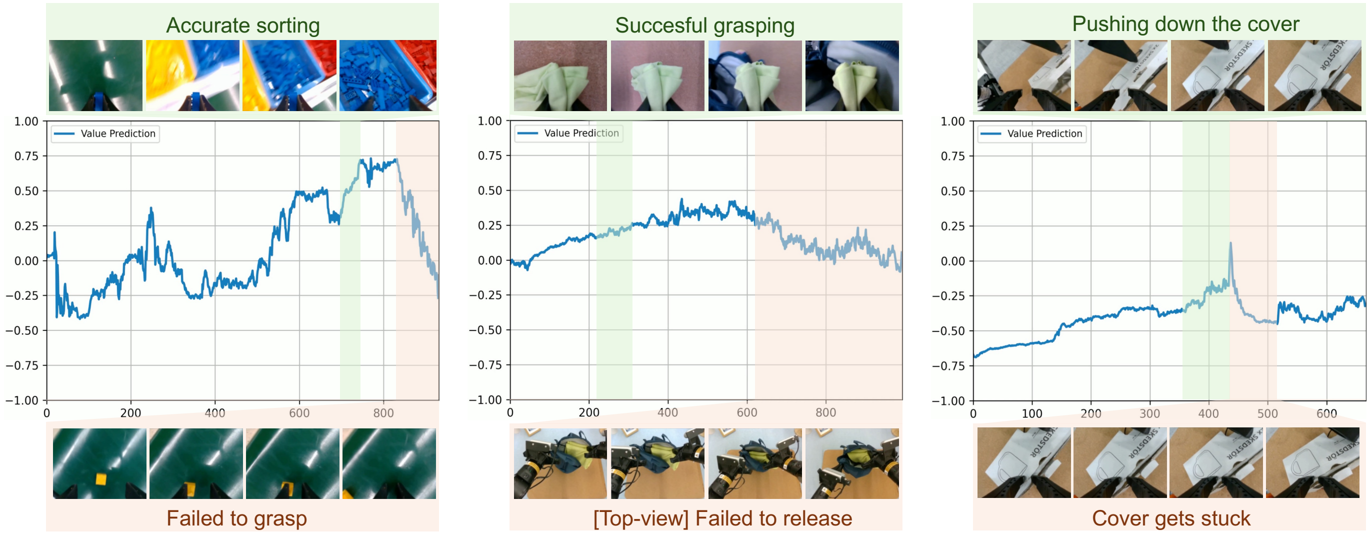


Fig. 13: **Qualitative visualizations of value prediction on real-world data.** Our value model is capable of distinguishing success and failure, highlighted in green and red, respectively.

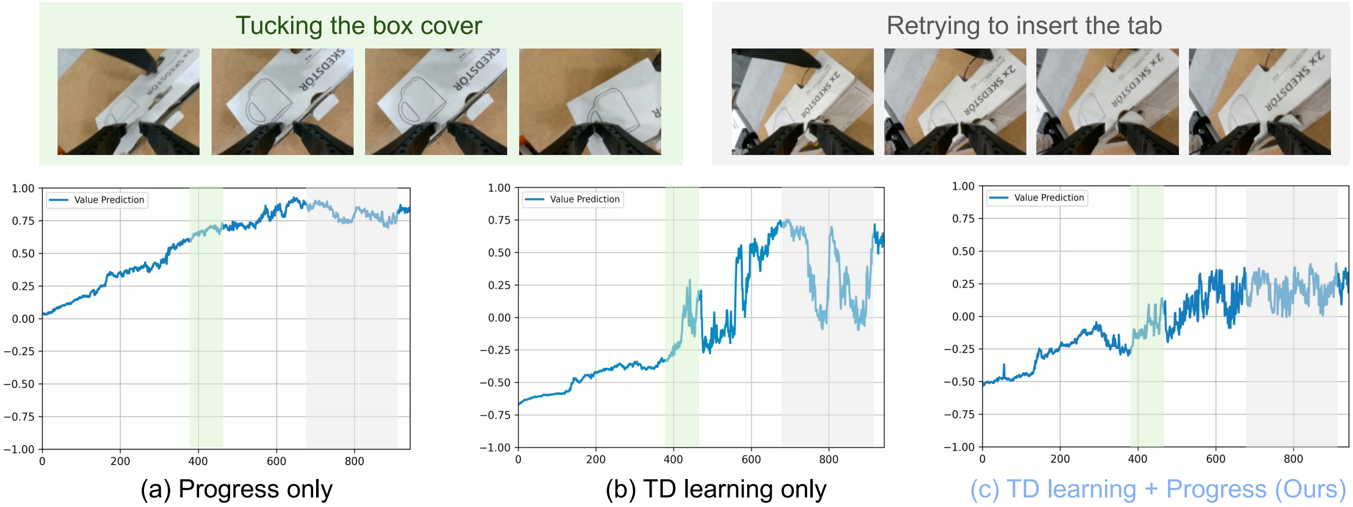


Fig. 14: **Qualitative ablation of value model.** This visualization ablates the effectiveness of imposing each loss during the training of the value model. Green and gray regions highlight the favorable and retrying behaviors, respectively. In the green region, (b) exhibits a stronger capability in detecting critical steps, compared to (a) progress only variant, where the result is simply monotonic. However, (b) is less numerically stable compared to (a), as depicted in the gray region. We jointly apply two losses to feature both visual sensitivity and numerical stability.

analysis, Fig. 18 presents additional rollout results on the Galaxea and Agibot World environments, confirming our model's consistency in complex domains.

Value Model. We showcase the predicted value trajectory over time alongside corresponding visual observations in Fig. 13. Green regions indicate successful execution, while red regions highlight inferior or suboptimal actions. The value model assigns increasing scores during successful executions (e.g., accurate sorting, stable grasping, and successful cover closure), while degrading when subtle failures occur, such as missed grasps, failure to release, or the cover getting stuck. Moreover, we visualize the impact of each loss for training value model in Fig. 14.

XIV. Failure Modes

We depict representative failure behaviors of the RISE policy in Fig. 17. In *Dynamic Brick Sorting*, failures stem from temporal inconsistency, manifesting as tracking failure or grasp slippage, alongside classification noise. In *Backpack Packing*, high deformability induces stowing failure and lifting instability, while surface compliance leads to zipper stuck or miss. In *Box Closing*, tight geometric tolerances cause incomplete loading, whereas bi-manual synchronization errors result in flap misalignment or tab deformation.

XV. Additional Related Work on VLA Models

One recent breakthrough in robot learning is the VLA framework that integrates general-purpose vision-language models with low-level robotic control. Building off pre-trained vision-language models, RT2 [102] and OpenVLA [46] represent actions as discretized bins following the training procedure of language models. OpenVLA-OFT [47] parallelizes the decoding process of chunked actions to improve inference latency. To overcome the multi-modality issue of robot actions where a variety of actions are correlated with the same state, GR00T [6], π-series [7, 8, 2], and RDT [62] further incorporate action generation with diffusion or flow matching-based architecture inspired by diffusion policy [18]. The massive training of these models is primarily supported by teleoperated robot datasets [71, 11, 43]. Other data corpora derived from simulators [70, 53], wearable devices [19, 83], neural synthesis [41], and generic internet [101] are also considered for the lack of costly real-world robot data. Effective approaches are proposed to incorporate heterogeneous data sources uniformly, even without sufficient action labels [12, 89, 42]. Despite advanced architecture and data scaling, VLAs still struggle with complex manipulation that requires high dexterity and precision [85, 25, 2], where our self-improving approach excels.

XVI. License of Assets

Our dynamics model is built on pre-trained Genie Envisioner [59] under the Apache License 2.0. The pre-training of our dynamics model leverages two large-scale public datasets, where Agibot World [11] is under CC BY-NC-SA 4.0 and Galaxea [43] is under Apache-2.0 license. Some comparisons of the dynamics model are conducted on the Bridge dataset [81] under Creative Commons Attribution 4.0 International License. Additionally, Cosmos-Predict2.5 [1] is applied as a baseline under the Apache License 2.0. Both our policy and value model are initialized from the pre-trained $\pi_{0.5}$ [8] under the Apache License 2.0.

XVII. Broader Impact

Overall, this work contributes to a growing vision of robots that learn continuously and efficiently by reasoning about the consequences of their actions via imagination. By improving robustness without excessive physical data collection costs, this work may contribute to safer and more reliable robotic systems that assist humans in physically demanding or hazardous tasks.

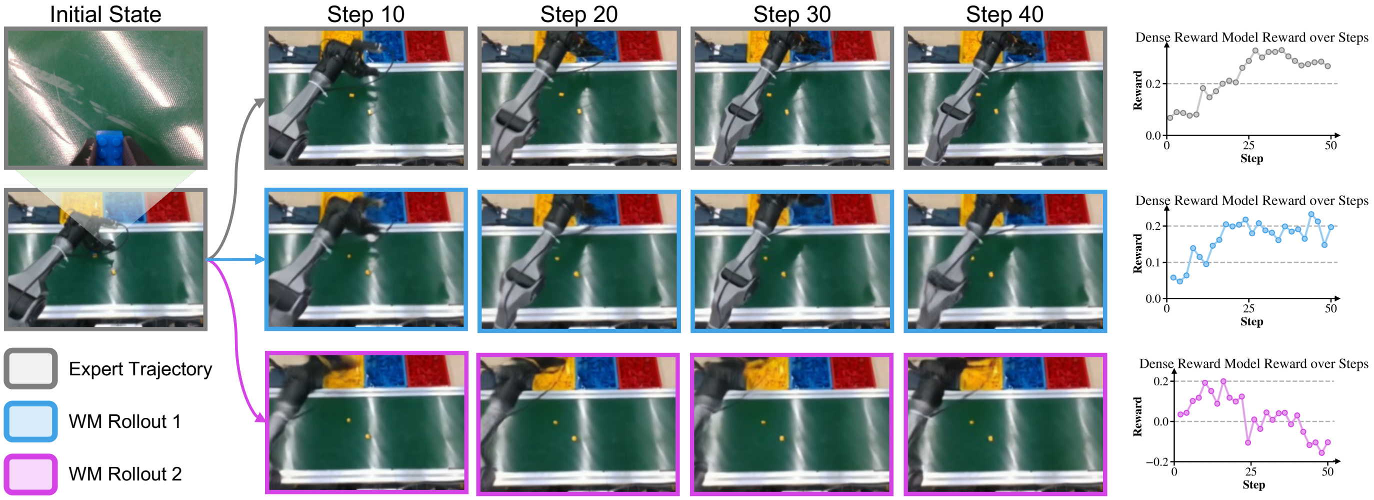


Fig. 15: **Multiple rollouts from the same initial state. Left**: Starting from the same state where the gripper grasps a blue brick, our world model can synthesize outcomes that accurately follow different actions. **Top Row**: Expert demonstration for reference. **Middle Row**: Imagined rollout of successful action that correctly put the blue brick into the blue basket, where the rewards go positive. **Bottom Row**: Imagined rollout of failed action that mistakenly put the blue brick into the yellow basket, where the rewards become negative.

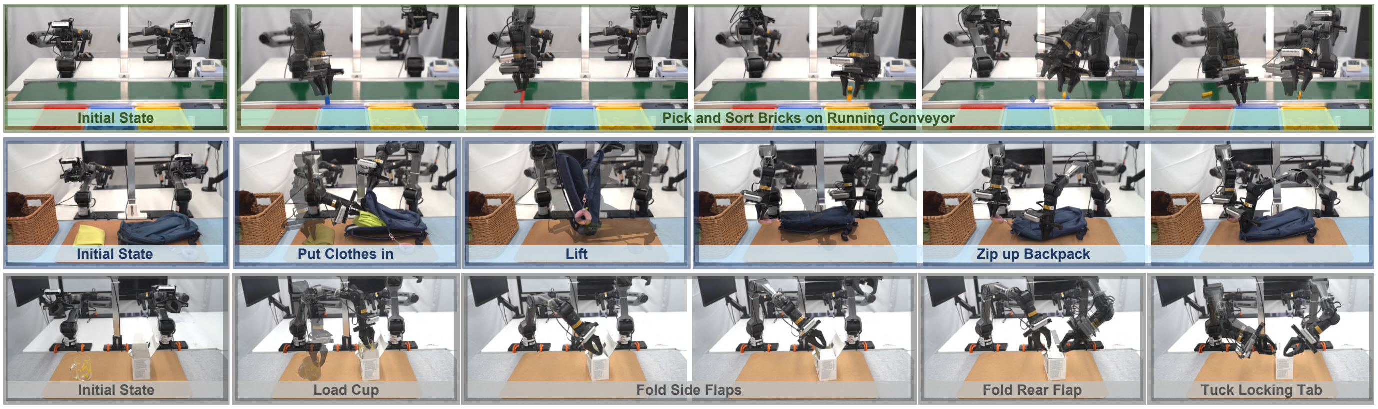


Fig. 16: **Policy rollout.** RISE demonstrates robust performance across diverse manipulation regimes. **Top:** Handling dynamic scenes by sorting bricks on a moving conveyor. **Middle:** Manipulating deformable objects in the Backpack Packing task. **Bottom:** Achieving high-precision bi-manual control in Box Closing.

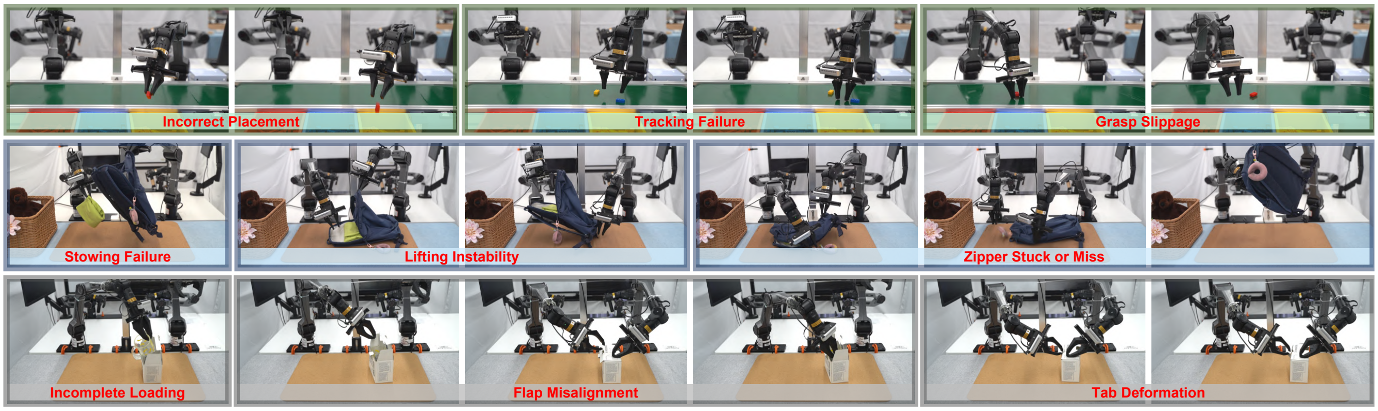


Fig. 17: **Failure modes during inference. Top:** Failures typically involve temporal inconsistency in tracking moving objects or precise grasping errors. **Middle:** The high deformability can lead to incomplete cloth insertion or slippage during the lifting and zipping stages. **Bottom:** Slight misalignments during bi-manual coordination can cause the cup to tip over during loading or result in unsuccessful folding and tucking.

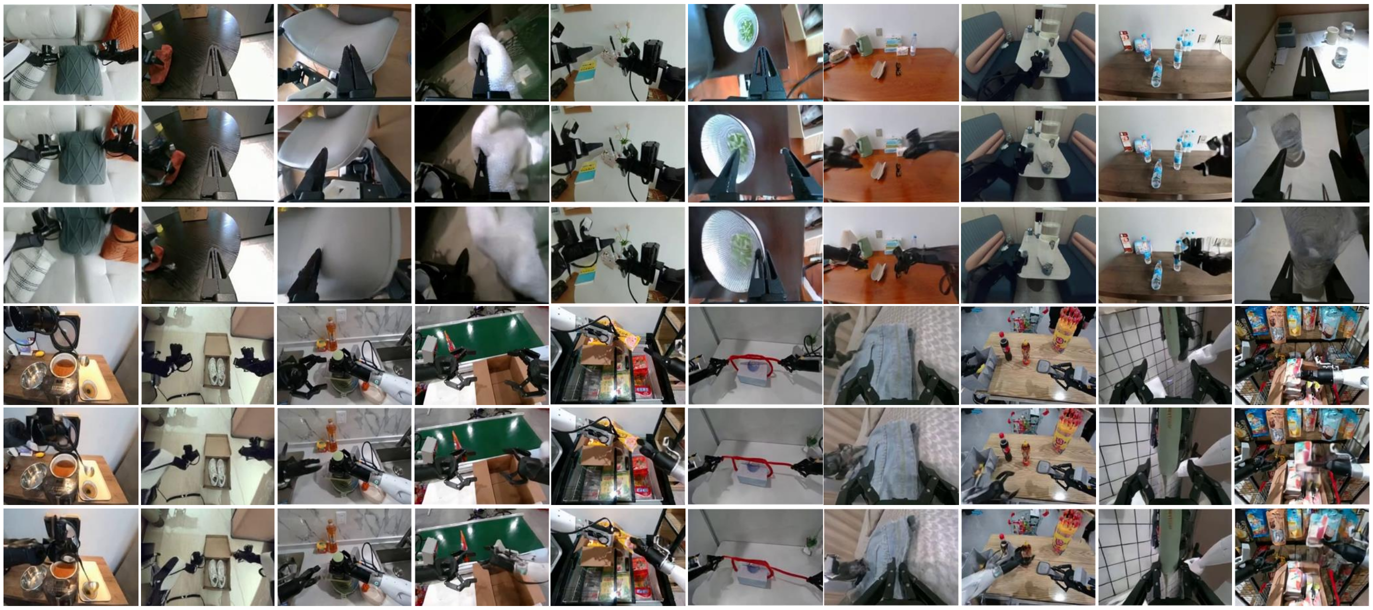
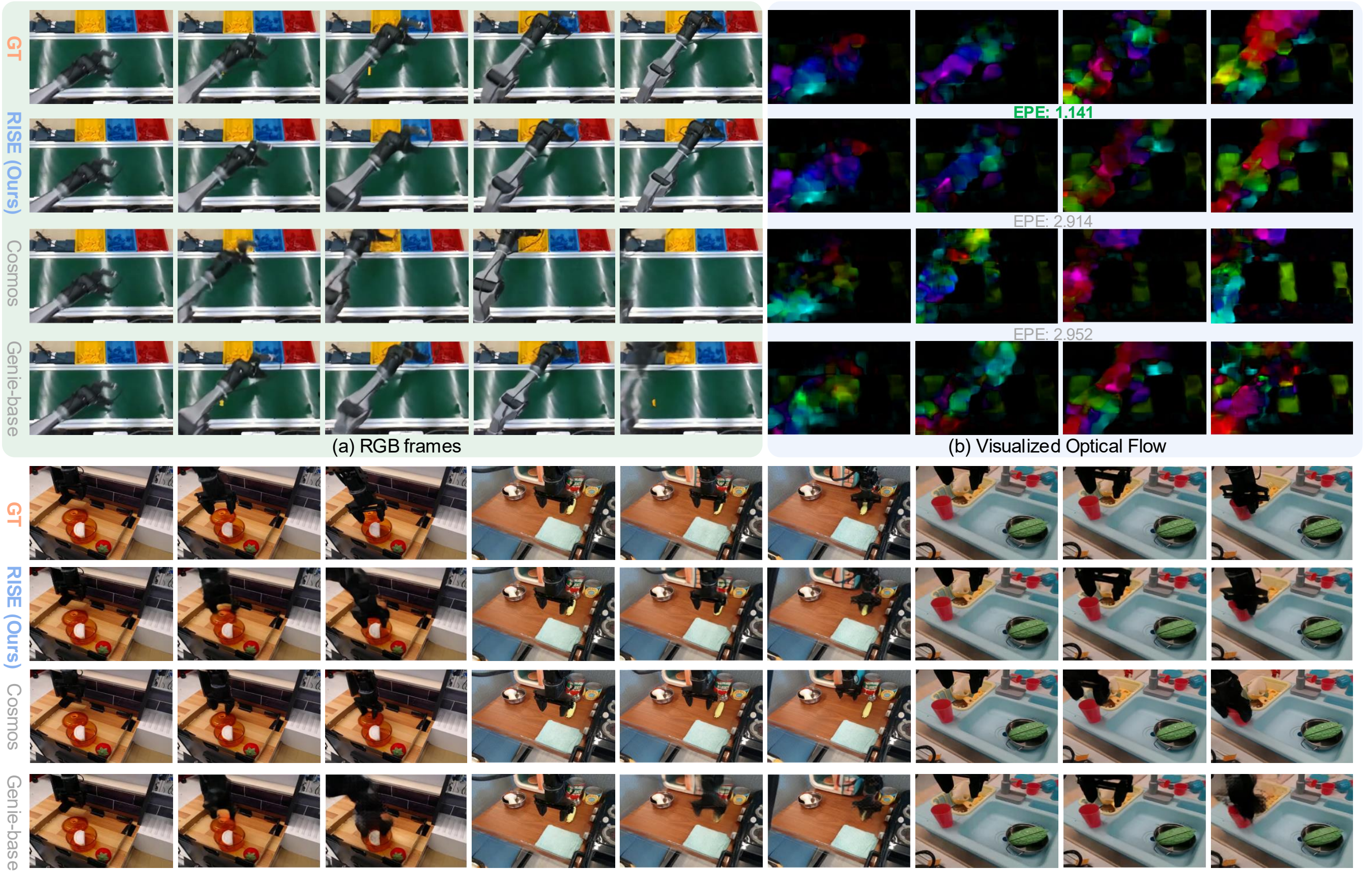


Fig. 18: **Dynamics model rollouts.** Each video clip is ordered top to bottom.



(a) RGB frames

(b) Visualized Optical Flow

(c) Comparison on Bridge Dataset

Fig. 19: **Comparisons with other generative counterparts.**