WORLD ACTION VERIFIER: SELF-IMPROVING WORLD MODELS VIA FORWARD-INVERSE ASYMMETRY

Yuejiang Liu[1,*], **Fan Feng**[2,*], **Lingjing Kong**[3,*], **Weifeng Lu**, **Jinzhou Tang**[2]
Kun Zhang[3], **Kevin Murphy**[4], **Chelsea Finn**[1], **Yilun Du**[5]
[1]Stanford University [2]UC San Diego [3]Carnegie Mellon University
[4]Google DeepMind [5]Harvard University
[*]Equal contribution world-action-verifier.github.io

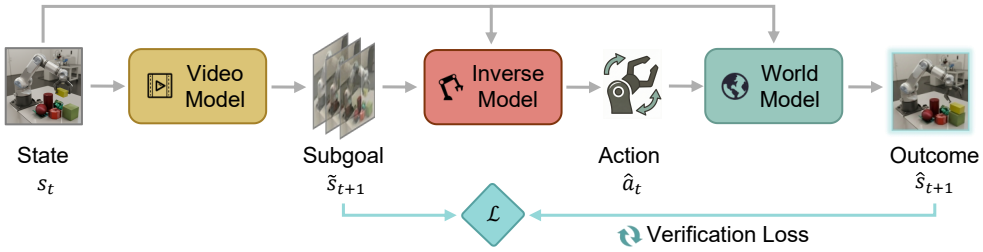

Figure 1: Overview of World Action Verifier, a framework that enables action-conditioned world models to self-improve from an asymmetric forward-inverse cycle: (i) a *diverse* subgoal generator proposes plausible future states, (ii) a *sparse* inverse model infers actions from a relevant subset of state features, and (iii) a world model rolls forward and verifies consistency between its predicted state and the proposed state.

ABSTRACT

General-purpose world models promise scalable policy evaluation, optimization, and planning, yet achieving the required level of robustness remains challenging. Unlike policy learning which primarily focuses on optimal actions, a world model needs to be reliable over a much broader range of suboptimal actions, which are often insufficiently covered by action-labeled interaction data. To address this challenge, we propose World Action Verifier (WAV), a framework that enables world models to identify their own prediction errors and self-improve. The key idea is to decompose action-conditioned state prediction into two factors—state plausibility and action reachability—and verify each separately. We show that these verification problems can be substantially easier than predicting future states due to two underlying asymmetries: the broader availability of action-free data and the lower dimensionality of action-relevant features. Leveraging these asymmetries, we augment a world model with (i) a diverse subgoal generator obtained from video corpora and (ii) a sparse inverse model that infers actions from a subset of state features. By enforcing cycle consistency among generated subgoals, inferred actions, and forward rollouts, WAV provides an effective verification mechanism in under-explored regimes, where existing methods typically fail. Across nine tasks spanning MiniGrid, RoboMimic, and ManiSkill, our method achieves 2× higher sample efficiency while improving downstream policy performance by 18%.

1 INTRODUCTION

World models—action-conditioned forward dynamics models that predict future states given specific actions or action chunks—have come to play an increasingly important role in robot learning (Ha & Schmidhuber, 2018; Wu et al., 2022; Assran et al., 2025; Team et al., 2025; Zheng et al., 2025; Huang et al., 2026). Recent works have shown that, when trained on action-labeled robot interactions

alongside action-free internet videos (Rigter et al., 2025; Huang et al., 2025; Ye et al., 2026b), world models have the potential to not only generate controllable future dynamics but also enable scalable policy evaluation (Quevedo et al., 2025; Team et al., 2026; Zhu et al., 2025b; Li et al., 2025b), policy optimization (Hafner et al., 2019b; Yang et al., 2023; 2024; Guo et al., 2025), and test-time planning (Hafner et al., 2019a; 2023; Jain et al., 2025; Zhou et al., 2025; Qi et al., 2025).

Despite remarkable progress, building a general-purpose world model that is robust enough for various downstream uses remains difficult. A central challenge is *action following*: predicting future states that faithfully reflect the effects of the given actions (Shang et al., 2026). Unlike policy learning from demonstrations, which primarily focuses on modeling optimal actions, a world model must be reliable across a much broader action distribution, including suboptimal, exploratory, and even random actions encountered during policy learning or evaluation (LeCun, 2022; Zhang et al., 2024; Jain et al., 2025). However, collecting such action-labeled robot interactions at scale is often slow, expensive, and sometimes even unsafe in the physical world. Given a limited budget of robot interactions, determining how to allocate them most effectively remains a pressing challenge.

Previous work has sought to address this through two main approaches. One line of work relies on on-policy exploration, *i.e.*, gathering data by rolling out the policies of interest (Jain et al., 2025; Guo et al., 2026; Liu et al., 2026). While effective for the considered policies, the learned model often degrades sharply beyond predefined policy sets, compromising its generality. Another line of work focuses on info-max exploration, actively seeking interactions that maximize information gain (Pathak et al., 2017; Sekar et al., 2020; Kim et al., 2020). A common proxy for information gain is the expected error of the world model, estimated prior to collecting the labeled transition—a process we refer to as *world model verification*. However, this process often suffers from a practical challenge: existing methods tend to be reliable in well-explored regions where additional data is less valuable, but unreliable in under-explored regions where verification matters most. After all, the data that offer the most new information for exploration are those where the least existing information is available for verification. This paradox raises a central question: *How can we reliably verify the predictions of a world model in under-explored regimes?*

To this end, we propose World Action Verifier (WAV), a framework that enables world models to verify their own prediction errors and self-improve through an asymmetric forward-inverse cycle. The core idea is to decompose the verification problem into more tractable subproblems. Specifically, we factorize action-conditioned state predictions into two complementary components: *state plausibility*, *i.e.*, whether a predicted state is visually and physically realistic, and *action reachability*, *i.e.*, whether the predicted transition is achievable under the given actions. This decomposition not only allows each factor to be verified separately, but also exposes two crucial asymmetries: (i) a broader availability of action-free data: state plausibility can be verified using internet videos without action labels, which are far more abundant than the action-labeled robot interactions used to train the world model, and (ii) lower dimensionality of action-relevant features: action reachability can be verified based on a subset of state features relevant to the actions, which are much lower-dimensional than the full state the world model must predict.

Motivated by these asymmetries, we augment a world model with two additional components: a diverse subgoal generator obtained from video corpora, and a sparse inverse model trained to infer actions from a learned subset of state features. Together, these components induce a goal-oriented self-improving cycle over proposed subgoals, inferred actions, and forward rollouts, where the consistency between proposed subgoals and predicted states provides an effective verification mechanism (Figure 1). Theoretically, we show that verification via a sparse inverse process is easier than dense forward generation, particularly in high-dimensional stochastic environments. Empirically, we evaluate WAV on nine tasks spanning MiniGrid (Chevalier-Boisvert et al., 2023), RoboMimic (Zhu et al., 2020), and ManiSkill (Mu et al., 2021). Compared to existing methods, WAV improves the sample efficiency of world models by $2\times$ and boosts downstream policy performance by more than 18%. Our results suggest that exploiting the asymmetries between forward and inverse dynamics can be a promising route toward self-improving world models.

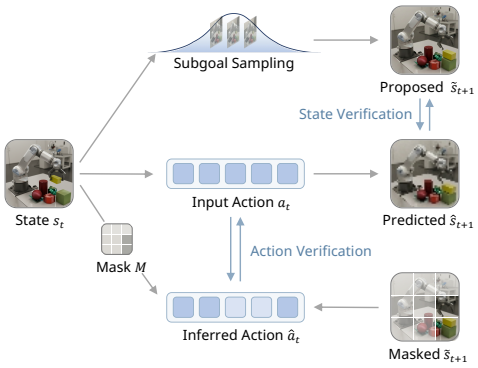

Figure 2: Decompose verification into state plausibility and action reachability.

Algorithm 1: WAV-Guided Exploration.

```
# s: current state, f: world model
# v: subgoal generator, h: inverse model
# D: current data, K: number of candidates

for each exploration iteration:
  s_g = v.sample(s, K)        # subgoals
  a = h.inverse(s, s_g)       # actions
  s_p = f.predict(s, a)       # outcomes
  scores = dist(s_g, s_p)     # disagreement
  idx = argmax(scores)        # max surprise
  s_n = env.step(a[idx])      # collect data
  D.append((s, a[idx], s_n))
  f.update(D), h.update(D)
```

2 Method: Self-Improving World Model with World Action Verifier

World models excel when grounded in action-labeled interaction data, yet collecting such data at scale is often prohibitively expensive. In this section, we present World Action Verifier (WAV), a self-improving framework that enables a world model to identify its own prediction errors and to prioritize the most informative interactions for exploration. We first formalize the verification problem in a semi-supervised setting (Sec. 2.1), then decompose it into two more tractable subproblems (Sec. 2.2), and finally couple them into a self-improvement cycle (Sec. 2.3).

2.1 Preliminary: Semi-Supervised Verification of World Models

We consider a world model f_θ as an action-conditioned forward dynamics model $\hat{s}^{t+1} = f_\theta(s^t, a^t)$, where s^t and a^t are the state and action (or action chunk) at time t, and $\hat{s}^{t+1}$ is the predicted successor state. Following recent training recipes (Huang et al., 2025; Gao et al., 2026), we study a semi-supervised setting with two data sources: a small action-labeled robot interaction dataset $\mathcal{D}_{\text{act}} = \{(s^t, a^t, s^{t+1})\}$ and a large action-free video dataset $\mathcal{D}_{\text{vid}} = \{(s^t, s^{t+1}, \ldots)\}$. Typically, $\mathcal{D}_{\text{vid}}$ covers a much broader range of state transitions than $\mathcal{D}_{\text{act}}$.

Our goal is to improve f_θ not only on the narrow action distribution represented in $\mathcal{D}_{\text{act}}$, but also on the broader transition support reflected in $\mathcal{D}_{\text{vid}}$. This is challenging due to the lack of action labels in video data. In practice, world models pre-trained on $\mathcal{D}_{\text{vid}}$ and then post-trained on $\mathcal{D}_{\text{act}}$ often struggle with *action following*, hallucinating future states misaligned with the given action (Shang et al., 2026; Mei et al., 2026). A natural remedy is to collect additional action-labeled interactions. However, since such annotations are typically expensive to obtain, a key question is *which new interactions should be queried so as to improve the world model most effectively*?

Intuitively, not all interactions are equally informative. Some are already well modeled and add little value. Ideally, we would spend the data budget on transitions where the current world model is most likely to make large prediction errors, since these are the cases where additional data can drive the *greatest improvement*. More formally, for a given state s^t and candidate action a^t, the prediction error of the world model is

$$\varepsilon(s^t, a^t) := \ell\big(\hat{s}^{t+1}, s^{t+1}\big) = \ell\big(f_\theta(s^t, a^t), s^{t+1}\big)\,, \tag{1}$$

where $\ell(\cdot, \cdot)$ is a discrepancy measure in the state space. Since the ground-truth s^{t+1} is unavailable prior to execution, we aim to construct a *verification mechanism* that produces a verifier $\hat{\varepsilon}$ for estimating this error, or at least preserving its relative ordering across candidate interactions. Concretely, given two candidate actions a_i^t and a_j^t, we would like the verifier $\hat{\varepsilon}$ to correctly rank their difficulties:

$$\varepsilon(s^t, a_i^t) < \varepsilon(s^t, a_j^t) \implies \hat{\varepsilon}(s^t, a_i^t, \hat{s}_i^{t+1}) < \hat{\varepsilon}(s^t, a_j^t, \hat{s}_j^{t+1}). \tag{2}$$

2.2 TWO COMPLEMENTARY FACTORS OF VERIFICATION

A common strategy to approximate such a verifier for prioritizing informative interactions is to estimate difficulty directly from the current world model, *e.g.*, through epistemic uncertainty (Pathak et al., 2017), ensemble disagreement (Sekar et al., 2020), or learning progress (Kim et al., 2020). However, these methods often inherit the weaknesses of the learned world model itself: they provide relatively reliable estimates in well-explored regimes, where the current world model is already accurate and additional exploration is less needed, but become much less reliable in under-explored regimes, where such estimates are most critical.

To overcome this issue, we take a different perspective: instead of estimating difficulty with prediction error directly, we build our approach around verifying two simpler conditions that any correct action-conditioned prediction should satisfy. Our starting point is Bayes' rule,

$$p(s^{t+1} \mid s^t, a^t) = \frac{p(a^t \mid s^t, s^{t+1})\, p(s^{t+1} \mid s^t)}{p(a^t \mid s^t)} \;\propto\; \underbrace{p(s^{t+1} \mid s^t)}_{\text{state}}\, \underbrace{p(a^t \mid s^t, s^{t+1})}_{\text{action}}, \tag{3}$$

which suggests that, instead of directly estimating model error, verification can be decomposed into two complementary factors:

- *State Plausibility: whether the predicted next state is plausible under the environment dynamics.*
- *Action Reachability: whether the transition from s^t to s^{t+1} is consistent with the given action.*

A correct prediction should satisfy both conditions: it should remain on the manifold of plausible futures and should be reachable under the intended action. More importantly, each factor admits a verification strategy that is easier than directly predicting the forward dynamics:

State verification via distribution asymmetry. A key asymmetry is that action-free video data are orders of magnitude more abundant than action-labeled interactions, providing a much broader prior over plausible transitions. A common failure mode in under-explored regimes is leaving this data manifold: rollouts become unrealistic or physically implausible. To detect such failures, we train a subgoal generator p_ϕ on $\mathcal{D}_{\text{vid}}$ and use it as a prior over plausible future states. Specifically, given the current state s^t, we sample K candidate subgoal states

$$\{\tilde{s}_k^{t+1}\}_{k=1}^{K} \sim p_\phi(\cdot \mid s^t), \tag{4}$$

which serve as on-manifold references that anchor verification to the much broader transition support available in action-free video data.

Action verification via dimensionality asymmetry. State plausibility alone does not imply correct forward predictions: for downstream policy uses, predicted transitions must be reachable under a specified action or action chunk. A second asymmetry is that, in many robotic tasks, the action is identifiable from a small subset of state features—such as end-effector pose or manipulated-object motion—making inverse verification far lower-dimensional than forward prediction. We instantiate this by learning a sparse inverse dynamics model h_ψ on $\mathcal{D}_{\text{act}}$, with a learned mask M that selects action-relevant state features:

$$\hat{a}^t = h_\psi\big(M \odot s^t,\; M \odot s^{t+1}\big)\,. \tag{5}$$

As shown in Figure 2, the subgoal generator p_ϕ and inverse model h_ψ provide two complementary components for verification: the former checks whether a candidate future is plausible, while the latter checks whether it is reachable through an inferred action.

2.3 VERIFICATION-GUIDED SELF-IMPROVING CYCLE

Given the two verification criteria above, we next connect them into a self-improvement cycle for exploration, where the current model sets (i.e., the world model and inverse model) autonomously generate verification signals to prioritize informative interactions, and the resulting data are fed back to update the world model under a fixed training recipe, without additional human intervention. A natural design is a forward-first cycle: sample actions, roll out f_θ, then apply h_ψ to recover actions

from generated state pairs (Ye et al., 2025). However, this design can be brittle in practice, since early errors in f_θ produce off-manifold rollouts on which the inverse model is unreliable.

We instead use a *reverse* cycle that anchors verification on the action-free state manifold. Given the current state s^t, we apply the subgoal prior, inverse model, and world model in sequence:

$$s^t \xrightarrow{p_\phi} \tilde{s}^{t+1} \xrightarrow{h_\psi} \hat{a}^t \xrightarrow{f_\theta} \hat{s}^{t+1}. \tag{6}$$

This ordering first proposes a plausible target state and then tests whether the action-conditioned world model can realize it. We measure their discrepancy as the estimated error $\hat{\varepsilon}(s^t, \hat{a}^t, \hat{s}^{t+1}) = \ell\big(\tilde{s}^{t+1}, \hat{s}^{t+1}\big)$ for prioritizing informative interactions, as summarized in Algorithm 1.

3 Theory: Robustness and Efficiency of World Action Verifier

WAV (Sec. 2) exploits two asymmetries, broader coverage of action-free data and lower dimensionality of action-relevant features, to verify world-model predictions via a sparse inverse dynamics model. We now formalize the conditions under which these asymmetries make sparse inverse verification both more *robust* under distribution shift and more *sample-efficient* to learn than forward dynamics. We focus on two questions:

1. *Under what conditions* can the sparse inverse verifier generalize beyond the training distribution of labeled transitions? (Sec. 3.1)
2. *What factors* influence the asymmetry between forward and inverse dynamics models? (Sec. 3.2)

3.1 Distribution-Level Robustness

To formalize the first question, we model the observed state s^t as arising from a latent vector $\mathbf{z}^t = (\mathbf{z}_1^t, \ldots, \mathbf{z}_k^t)$. The learned mask M in (5) selects an action-relevant block $\mathcal{S}$ of this latent space; intuitively, $\mathcal{S}$ captures agent-centric variables (*e.g.*, proprioception or end-effector motion) and is largely insulated from the rest of the scene. The sparse inverse model h_ψ from Sec. 2 thus operates on $(\mathbf{z}_\mathcal{S}^t, \mathbf{z}_\mathcal{S}^{t+1})$; we write the verifier as $\hat{\mathbf{a}}^t = h_\psi(\hat{\mathbf{z}}_\mathcal{S}^t, \hat{\mathbf{z}}_\mathcal{S}^{t+1})$, where $\hat{\mathbf{z}}^t$ denotes the encoder's latent estimate of $\mathbf{z}^t$.

Let P_{seed} denote the distribution induced by $\mathcal{D}_{\text{act}}$; we call a state–action pair *out-of-support* (OOS) when $(\mathbf{z}^t, \mathbf{a}^t) \notin \text{supp}(P_{\text{seed}})$. The key structural condition is the presence of a *generation–verification gap*: the full pair $(\mathbf{z}^t, \mathbf{a}^t)$ may be OOS, while the restricted pair $(\mathbf{z}_\mathcal{S}^t, \mathbf{a}^t)$ remains on support. This captures the regime in which scene-level consequences are novel, but the agent-side motion that encodes the action is still familiar.

Proposition 3.1 (Informal)**.** *Assume there exists an identifiable verification subset $\mathcal{S}$ such that: (i) $\mathbf{z}_\mathcal{S}^{t+1}$ depends only on $(\mathbf{z}_\mathcal{S}^t, \mathbf{a}^t)$ and not on the rest of the scene; (ii) $(\mathbf{z}_\mathcal{S}^t, \mathbf{a}^t)$ stays on-support even when $(\mathbf{z}^t, \mathbf{a}^t)$ is OOS; and (iii) the action is identifiable from the subset transition $(\mathbf{z}_\mathcal{S}^t, \mathbf{z}_\mathcal{S}^{t+1})$. Then an inverse model trained on the seed data can recover the correct action from $(\hat{\mathbf{z}}_\mathcal{S}^t, \hat{\mathbf{z}}_\mathcal{S}^{t+1})$ on such compositional OOS transitions. Consequently, the forward–inverse mismatch used by WAV localizes forward-model error rather than action-label ambiguity.*

Interpretation. Proposition 3.1 guarantees that the sparse inverse model h_ψ produces correct pseudo-labels whenever the agent's own motion pattern (*e.g.*, joint-angle trajectories) was seen during training, even if the full scene transition is novel. This is a strictly weaker requirement than asking the *full* transition to be on-support, which is what a dense forward model or a full-observation inverse model would need. The direct consequence for the self-improving cycle in Sec. 2.3 is that the discrepancy $\ell(\tilde{s}^{t+1}, \hat{s}^{t+1})$ between the subgoal and the forward rollout reflects genuine world-model error rather than action-label noise, so each exploration round adds trustworthy data that expands the world model's effective coverage. Appendix F.1 formalizes this claim and shows how the verification subset $\mathcal{S}$ can be identified from observations.

3.2 Sample-Efficiency Advantage

Proposition 3.1 says *when* sparse verification transfers. We now characterize *what factors* determine the asymmetry between inverse verification and forward prediction. To isolate the key factors, we

idealize the learned element-wise mask M in (5) as a fixed rank-d_z linear projection $z := Ms \in \mathbb{R}^{d_z}$ and analyze a stylized linear–Gaussian model. This setting cleanly separates three sources of the forward–inverse asymmetry: *dimensionality*, *stochasticity*, and *sample size*.

For tractability, we work directly with the observed state $s \in \mathbb{R}^{d_s}$ and action $a \in \mathbb{R}^{d_a}$, noting that the results extend to the latent setting of Theorem 3.1 whenever s admits a factored latent representation (see Appendix F.2). Assume one-step dynamics

$$s' = As + Ba + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_s^2 \mathbf{I}_{d_s}), \tag{7}$$

where σ_s captures transition stochasticity. We further assume that the action is recoverable from a low-dimensional action-relevant slice $z := Ms \in \mathbb{R}^{d_z}$ with $d_z \mathbf{l} d_s$:

$$a = h(z, z') + \eta, \qquad h(z, z') := H \begin{bmatrix} z \\ z' \end{bmatrix}, \qquad \eta \sim \mathcal{N}(0, \sigma_a^2 \mathbf{I}_{d_a}), \tag{8}$$

where $z' := Ms'$ and σ_a measures irreducible ambiguity in recovering the action from (z, z').

We compare a dense forward model f_θ trained on $[s; a] \in \mathbb{R}^{d_s + d_a}$ against a sparse inverse model h_ψ trained on $[z; z'] \in \mathbb{R}^{2d_z}$, both fit by OLS on n transitions from $\mathcal{D}_{\text{act}}$. To compare them in the same units, we evaluate in state space:

$$\mathcal{E}_F := \frac{1}{d_s} \mathbb{E}\big[\|f_\theta(s, a) - f^\star(s, a)\|_2^2\big], \qquad \mathcal{E}_I := \frac{1}{d_s} \mathbb{E}\big[\|f^\star(s, h_\psi(z, z')) - f^\star(s, h(z, z'))\|_2^2\big], \tag{9}$$

where $f^\star(s, a)$ denotes the true dynamics and $\lambda := \|B\|_{\text{op}}$ converts action error into state-space error.

Proposition 3.2 (Informal)**.** *Under the stylized setup above, if both models are fit by OLS on n labeled transitions, then*

$$\frac{\mathbb{E}[\mathcal{E}_F]}{\mathbb{E}[\mathcal{E}_I]} \;\geq\; \underbrace{\Big(\frac{d_s + d_a}{2d_z} \cdot \frac{d_s}{d_a}\Big)}_{\textit{dimensionality}} \cdot \underbrace{\Big(\frac{\sigma_s}{\lambda\, \sigma_a}\Big)^2}_{\textit{stochasticity}} \cdot \underbrace{\Big(\frac{n - 2d_z - 1}{n - (d_s + d_a) - 1}\Big)}_{\textit{sample size}}, \tag{10}$$

provided $n > d_s + d_a + 1$ and $n > 2d_z + 1$. The exact statement and proof are in Appendix F.2.

Interpretation. The ratio in (10) factorizes into three terms. *Dimensionality*: the forward model must estimate a map from $d_s + d_a$ inputs, whereas the sparse inverse uses only $2d_z$. *Stochasticity*: forward prediction suffers from environment noise σ_s, while inverse verification suffers only from action-recovery ambiguity σ_a (scaled by λ). *Sample size*: when n is only modestly larger than $d_s + d_a$, the forward estimator is far less stable. In practice, WAV helps most when (i) the verifier needs only a small agent-centric subset while the world model predicts a large scene (*large* d_s/d_z); (ii) uncontrolled dynamics inflate σ_s while the action imprint stays clean (*large* σ_s/σ_a); and (iii) action-labeled data are limited (*small* n). We validate each factor empirically in Sec. 4.1.1: varying the data budget isolates the sample-size term, increasing the number of objects raises the effective state dimension d_s, and adding noisy floors inflates σ_s while leaving σ_a unchanged.

4 Experiments

In this section, we empirically evaluate the central claims of WAV. We begin by validating whether inverse verification, especially with sparsity, is more robust and easier to learn than action-conditioned forward prediction under limited data and distribution shift. We then examine whether WAV can effectively improve world model learning through self-improvement, and finally, whether the resulting gains translate into better downstream policy learning.

Concretely, we study the following research questions:

RQ1: Is learning the inverse dynamics model generally easier and more robust than learning an action-conditioned world model, and what factors drive the gap between them?

RQ2: Do sparse IDMs generalize better than vanilla IDMs to unseen objects and novel interactions?

RQ3: How effective is forward–inverse asymmetry for self-improving the world model?

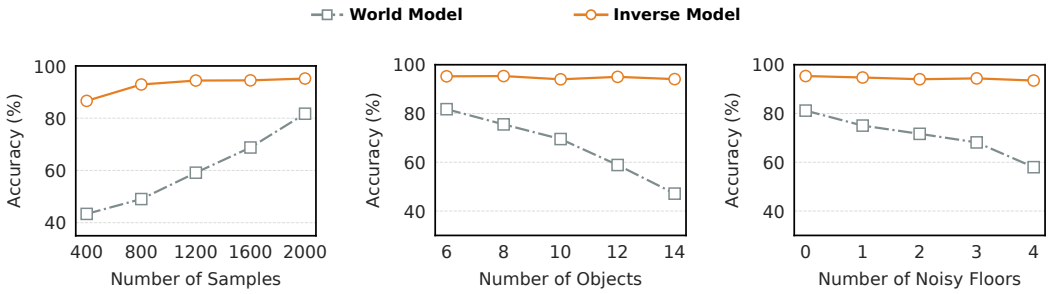

Figure 3: **Verification of robustness of WAV on MiniGrid.** (Left) Sample efficiency comparison between Sparse IDM and the World Model with six objects. (Mid) Robustness to increasing state complexity. (Right) Robustness to growing environment stochasticity.

RQ4: Does this self-improvement translate into improved downstream policy learning?

To address RQ1 and RQ2, we conduct controlled experiments in MiniGrid (Chevalier-Boisvert et al., 2023), with a particular emphasis on out-of-distribution generalization to unseen objects and novel interactions. Building on these findings, we then evaluate the proposed framework in more complex settings, assessing how it enables self-improvement of world models quality (*RQ3*) on both MiniGrid and simulated robotic manipulation tasks (RoboMimic (Zhu et al., 2020) and ManiSkill (Mu et al., 2021)), and how these enhance downstream policy learning performance (*RQ4*) on robotic manipulation tasks.

Baselines. We compare our method (Sec. 2) against *Random* sampling (lower bound), *Uncertainty*-based (Sekar et al., 2020) and *Progress*-based (Kim et al., 2020) acquisition, a *Vanilla IDM* without sparsity, and an *Oracle* that uses ground-truth action labels (upper bound). Detailed descriptions are given in Appendix C.

4.1 Experiments in Synthetic MiniGrid

Dataset. We evaluate on three MiniGrid tasks (`Key Delivery`, `Ball Delivery`, `Object Matching`), splitting 50k sequences into an action-free set for pretraining the subgoal generator and an exploration pool with a 200-sequence labeled seed set. We additionally construct datasets with varying object counts and noisy floor tiles for controlled robustness evaluation. Full details are in Appendix D.

4.1.1 Robustness of World Action Verification

To verify the robustness of the world action verifier, we first compare the robustness of forward world models and inverse dynamics models under controlled distribution shifts (***RQ1*** & ***RQ2***) - to evaluate the feasibility of using WAV. We then evaluate whether the estimated errors with WAV preserve the same ordering as the oracle errors, in comparison with the baseline data selection strategies.

Setup. We vary labeled data budgets, scene complexity (object counts), and observation noise; the IDM is converted to next-state predictions for direct comparison. Full setup and metric details are in Appendix D.4 and D.3.

Results. For ***RQ1***, Figure 3 empirically validates the three factors identified in Proposition 3.2, each isolated by a separate controlled variable. Figure 3 (Left) isolates the *sample-size* factor: across data regimes, action inference using IDMs consistently outperforms learning action-conditioned world models, with the performance gap being most pronounced in the low-data regime, consistent with the diverging finite-sample term in (10) when n is only modestly larger than $d_s + d_a$. Figure 3 (Mid) isolates the *dimensionality* factor: increasing the number of objects raises the effective state dimension d_s while leaving the action-relevant subset d_z unchanged. The WM's performance degrades rapidly as state complexity increases, whereas the IDM remains stable, consistent with the growing ratio $(d_s+d_a)/2d_z$ in the dimensionality term. Figure 3 (Right) isolates the *stochasticity* factor: noisy floor tiles inflate observation noise σ_s while the action imprint on the agent-centric features remains clean (σ_a unchanged). The WM exhibits clear sensitivity to the induced observation noise, whereas the

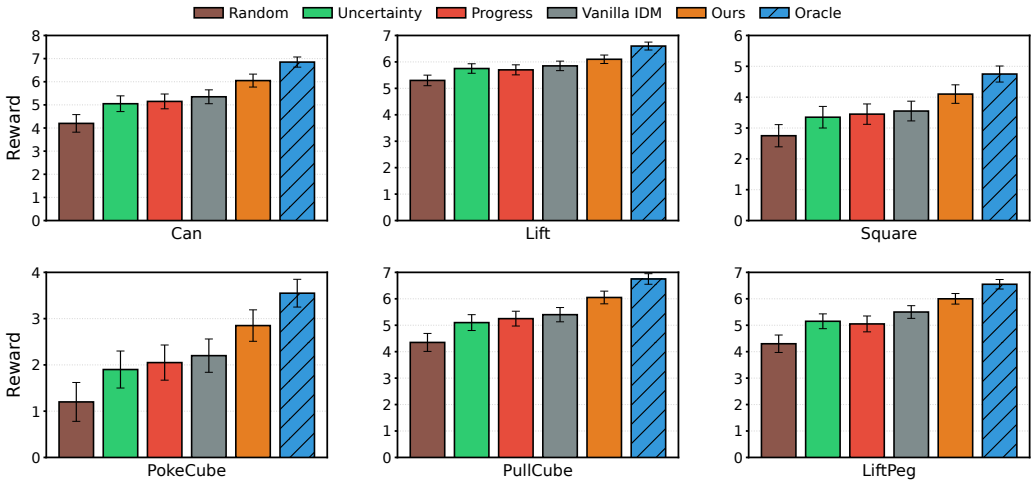

Figure 4: **Downstream policy performance on RoboMimic and ManiSkill using learned world models.** On average, our method achieves 18% higher average reward than the strongest baseline. Error bars denote the standard error over 3 seeds.

IDM maintains largely invariant performance, consistent with the $(\sigma_s/\lambda\sigma_a)^2$ ratio in the stochasticity term. Together, these results confirm the predicted advantage of sparse inverse verification across all three factors, providing empirical justification for *using the IDM as a reliable verifier of the world model*.

For ***RQ2***, we additionally evaluate sparse vs. vanilla IDMs and the correlation of verification scores with Oracle rankings. We further assess the effectiveness of WAV for self-improving world model quality (***RQ3***) in MiniGrid. These results, along with the corresponding evaluation figures, are presented in Appendix D.7.

4.2 Experiments on Simulated Robot Manipulations

We evaluate on six robotic manipulation tasks from Roboverse (Geng et al., 2025): RoboMimic (Zhu et al., 2020) (`Lift`, `Can`, `Square`) and ManiSkill (Mu et al., 2021) (`PullCube`, `PokeCube`, `LiftPeg`), using Dreamer-v3 (Hafner et al., 2023) as the world model and a sparsity-regularized variant of CLAM (Liang et al., 2025) as the inverse model. Full dataset construction and model details are given in Appendix E and E.4.

We additionally verify that WAV's verification robustness and its positive impact on world model quality, as established in MiniGrid, extend to the robotic manipulation setting; these results, including verification rank correlation analysis and world model prediction comparisons, are reported in Appendix E.6. Here, we focus on the most practically important downstream question (***RQ4***): whether improved world models translate into better policy learning.

4.2.1 Effectiveness for Policy Learning

To address ***RQ4***, we evaluate downstream policy learning by using the world model for imagination-based planning following the SAILOR protocol (Jain et al., 2025) with a data budget of 1,000 trajectories (setup details in Appendix E.2).

Results. As shown in Figure 4, across all RoboMimic and ManiSkill tasks, policies trained with our world model consistently achieve higher rewards than those using baseline world models, and are second only to the oracle with access to ground-truth actions. On average, our method achieves 18% higher reward than the strongest baseline, demonstrating that the improved dynamics representations translate directly into more effective imagination-based policy learning. A detailed per-task analysis is provided in Appendix E.6.3.

ACKNOWLEDGEMENT

We thank the members of the IRIS Lab for valuable feedback and discussions. We also thank Xiangcheng Zhang for help with the simulation setup. This work was supported in part by the Robotics and AI Institute, DARPA, ONR, CIFAR, SNSF, Schmidt Science, NSF, NIH, and the AI Institute for Societal Decision Making.

REFERENCES

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning, 2017. URL https://arxiv.org/abs/1703.01732.

Christopher Agia, Rohan Sinha, Jingyun Yang, Rika Antonova, Marco Pavone, Haruki Nishimura, Masha Itkina, and Jeannette Bohg. Cupid: Curating data your robot loves with influence functions. *arXiv preprint arXiv:2506.19121*, 2025.

Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Advances in neural information processing systems*, 29, 2016.

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.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation, 2016. URL https://arxiv.org/abs/1606.01868.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning*, volume 2, pp. 4, 2021.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Rich Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pre-trained image-editing diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation, 2018. URL https://arxiv.org/abs/1810.12894.

Junhao Cai, Zetao Cai, Jiafei Cao, Yilun Chen, Zeyu He, Lei Jiang, Hang Li, Hengjie Li, Yang Li, Yufei Liu, et al. Internvla-a1: Unifying understanding, generation and action for robotic manipulation. *arXiv preprint arXiv:2601.02456*, 2026.

Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

Annie S Chen, Alec M Lessing, Yuejiang Liu, and Chelsea Finn. Curating demonstrations using online experience. *arXiv preprint arXiv:2503.03707*, 2025a.

Boyuan Chen, Tianyuan Zhang, Haoran Geng, Kiwhan Song, Caiyi Zhang, Peihao Li, William T Freeman, Jitendra Malik, Pieter Abbeel, Russ Tedrake, et al. Large video planner enables generalizable robot control. *arXiv preprint arXiv:2512.15840*, 2025b.

Xiaoyu Chen, Junliang Guo, Tianyu He, Chuheng Zhang, Pushi Zhang, Derek Cathera Yang, Li Zhao, and Jiang Bian. IGOR: Image-GOal representations are the atomic building blocks for next-level generalization in embodied AI, 2025c. URL https://openreview.net/forum?id=bpdIZTIVq8.

Maxime Chevalier-Boisvert, Bolun Dai, Mark Towers, Rodrigo Perez-Vicente, Lucas Willems, Salem Lahlou, Suman Pal, Pablo Samuel Castro, and Jordan Terry. Minigrid & miniworld: Modular & customizable reinforcement learning environments for goal-oriented tasks. *Advances in Neural Information Processing Systems*, 36:73383–73394, 2023.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025a.

Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025b.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Yinpei Dai, Hongze Fu, Jayjun Lee, Yuejiang Liu, Haoran Zhang, Jianing Yang, Chelsea Finn, Nima Fazeli, and Joyce Chai. Robomme: Benchmarking and understanding memory for robotic generalist policies. *arXiv preprint arXiv:2603.04639*, 2026.

Marc Peter Deisenroth and Carl Edward Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 465–472. Omnipress, 2011. URL https://icml.cc/2011/papers/323_icmlpaper.pdf.

Yilun Du, Chuang Gan, and Phillip Isola. Curious representation learning for embodied intelligence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10408–10417, 2021.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Advances in neural information processing systems*, 36:9156–9172, 2023.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

Haoran Geng, Feishi Wang, Songlin Wei, Yuyang Li, Bangjun Wang, Boshi An, Charlie Tianyue Cheng, Haozhe Lou, Peihao Li, Yen-Jen Wang, et al. Roboverse: Towards a unified platform, dataset and benchmark for scalable and generalizable robot learning. *arXiv preprint arXiv:2504.18904*, 2025.

Alex Graves, Marc G. Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks, 2017. URL https://arxiv.org/abs/1704.03003.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation, 2025. URL https://arxiv.org/abs/2510.10125.

Yanjiang Guo, Tony Lee, Lucy Xiaoyang Shi, Jianyu Chen, Percy Liang, and Chelsea Finn. Vlaw: Iterative co-improvement of vision-language-action policy and world model. *arXiv preprint arXiv:2602.12063*, 2026.

Anthony GX-Chen, Kenneth Marino, and Rob Fergus. Efficient exploration and discriminative world model learning with an object-centric abstraction, 2025. URL https://arxiv.org/abs/2408.11816.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

Nick Haber, Damian Mrowca, Li Fei-Fei, and Daniel L. K. Yamins. Learning to play with intrinsically-motivated self-aware agents, 2018. URL https://arxiv.org/abs/1802.07442.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019a.

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

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

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

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, pp. 8387–8406. PMLR, 2022.

Joey Hejna, Suvir Mirchandani, Ashwin Balakrishna, Annie Xie, Ayzaan Wahid, Jonathan Tompson, Pannag Sanketi, Dhruv Shah, Coline Devin, and Dorsa Sadigh. Robot data curation with mutual information estimators. *arXiv preprint arXiv:2502.08623*, 2025.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration, 2017. URL https://arxiv.org/abs/1605.09674.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14(3):569–600, 2014.

Edward S Hu, Richard Chang, Oleh Rybkin, and Dinesh Jayaraman. Planning goals for exploration. *arXiv preprint arXiv:2303.13002*, 2023.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *International Conference on Machine Learning*, pp. 24328–24346. PMLR, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation, 2026. URL https://arxiv.org/abs/2601.03782.

Arnav Kumar Jain, Vibhakar Mohta, Subin Kim, Atiksh Bhardwaj, Juntao Ren, Yunhai Feng, Sanjiban Choudhury, and Gokul Swamy. A smooth sea never made a skilled sailor: Robust imitation via learning to search, 2025. URL https://arxiv.org/abs/2506.05294.

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. *arXiv preprint arXiv:2505.12705*, 2025.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems*, 32, 2019.

Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Kuno Kim, Megumi Sano, Julian De Freitas, Nick Haber, and Daniel Yamins. Active world model learning with progress curiosity, 2020. URL https://arxiv.org/abs/2007.07853.

Jacky Kwok, Christopher Agia, Rohan Sinha, Matt Foutter, Shulu Li, Ion Stoica, Azalia Mirhoseini, and Marco Pavone. Robomonkey: Scaling test-time sampling and verification for vision-language-action models. *arXiv preprint arXiv:2506.17811*, 2025.

Jacky Kwok, Xilun Zhang, Mengdi Xu, Yuejiang Liu, Azalia Mirhoseini, Chelsea Finn, and Marco Pavone. Scaling verification can be more effective than scaling policy learning for vision-language-action alignment. *arXiv preprint arXiv:2602.12281*, 2026.

Sébastien Lachapelle. On the identifiability of latent action policies. *arXiv preprint arXiv:2510.01337*, 2025.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric partial disentanglement via mechanism sparsity: Sparse actions, interventions and sparse temporal dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025a.

Yaxuan Li, Yichen Zhu, Junjie Wen, Chaomin Shen, and Yi Xu. Worldeval: World model as real-world robot policies evaluator, 2025b. URL https://arxiv.org/abs/2505.19017.

Anthony Liang, Pavel Czempin, Matthew Hong, Yutai Zhou, Erdem Biyik, and Stephen Tu. Clam: Continuous latent action models for robot learning from unlabeled demonstrations. *arXiv preprint arXiv:2505.04999*, 2025.

Shalev Lifshitz, Sheila A McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*, 2025.

Junhong Lin, Xinyue Zeng, Jie Zhu, Song Wang, Julian Shun, Jun Wu, and Dawei Zhou. Plan and budget: Effective and efficient test-time scaling on reasoning large language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ctspw4CqbS.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning, 2023. URL https://arxiv.org/abs/2306.03310.

Grace Liu, Michael Tang, and Benjamin Eysenbach. A single goal is all you need: Skills and exploration emerge from contrastive rl without rewards, demonstrations, or subgoals. *arXiv preprint arXiv:2408.05804*, 2024.

Xiaokang Liu, Zechen Bai, Hai Ci, Kevin Yuchen Ma, and Mike Zheng Shou. World-vla-loop: Closed-loop learning of video world model and vla policy. *arXiv preprint arXiv:2602.06508*, 2026.

Yuejiang Liu, Jubayer Ibn Hamid, Annie Xie, Yoonho Lee, Max Du, and Chelsea Finn. Bidirectional decoding: Improving action chunking via guided test-time sampling. *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2408.17355.

Calvin Luo, Zilai Zeng, Mingxi Jia, Yilun Du, and Chen Sun. Self-adapting improvement loops for robotic learning. *arXiv preprint arXiv:2506.06658*, 2025.

Qi Lv, Weijie Kong, Hao Li, Jia Zeng, Zherui Qiu, Delin Qu, Haoming Song, Qizhi Chen, Xiang Deng, and Jiangmiao Pang. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951*, 2025.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Zhiting Mei, Tenny Yin, Ola Shorinwa, Apurva Badithela, Zhonghe Zheng, Joseph Bruno, Madison Bland, Lihan Zha, Asher Hancock, Jaime Fernández Fisac, et al. Video generation models in robotics-applications, research challenges, future directions. *arXiv preprint arXiv:2601.07823*, 2026.

Russell Mendonca, Oleh Rybkin, Kostas Daniilidis, Danijar Hafner, and Deepak Pathak. Discovering and achieving goals via world models. *Advances in Neural Information Processing Systems*, 34: 24379–24391, 2021.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Cathera Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. Maniskill: Generalizable manipulation skill benchmark with large-scale demonstrations. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Mitsuhiko Nakamoto, Oier Mees, Aviral Kumar, and Sergey Levine. Steering your generalists: Improving robotic foundation models via value guidance. *arXiv preprint arXiv:2410.13816*, 2024.

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.

Georg Ostrovski, Marc G. Bellemare, Aaron van den Oord, and Remi Munos. Count-based exploration with neural density models, 2017. URL https://arxiv.org/abs/1703.01310.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction, 2017. URL https://arxiv.org/abs/1705.05363.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. URL https://arxiv.org/abs/1709.07871.

Han Qi, Haocheng Yin, Aris Zhu, Yilun Du, and Heng Yang. Strengthening generative robot policies through predictive world modeling. *arXiv preprint arXiv:2502.00622*, 2025.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. Worldgym: World model as an environment for policy evaluation, 2025. URL https://arxiv.org/abs/2506.00613.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. In *Reinforcement Learning Conference*, 2025.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *International conference on machine learning*, pp. 8583–8592. PMLR, 2020.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

Pranav Shyam, Wojciech Jaśkowski, and Faustino Gomez. Model-based active exploration, 2019. URL https://arxiv.org/abs/1810.12162.

Charles Spearman. The proof and measurement of association between two things. 1961.

Bradly C. Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models, 2015. URL https://arxiv.org/abs/1507.00814.

Richard S. Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In Bruce W. Porter and Raymond J. Mooney (eds.), *Machine Learning, Proceedings of the Seventh International Conference on Machine Learning, Austin, Texas, USA, June 21-23, 1990*, pp. 216–224. Morgan Kaufmann, 1990. doi: 10.1016/B978-1-55860-141-3.50030-4. URL https://doi.org/10.1016/b978-1-55860-141-3.50030-4.

Gemini Robotics Team, Krzysztof Choromanski, Coline Devin, Yilun Du, Debidatta Dwibedi, Ruiqi Gao, Abhishek Jindal, Thomas Kipf, Sean Kirmani, Isabel Leal, Fangchen Liu, Anirudha Majumdar, Andrew Marmon, Carolina Parada, Yulia Rubanova, Dhruv Shah, Vikas Sindhwani, Jie Tan, Fei Xia, Ted Xiao, Sherry Yang, Wenhao Yu, and Allan Zhou. Evaluating gemini robotics policies in a veo world simulator, 2026. URL https://arxiv.org/abs/2512.10675.

PAN Team, Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, Davit Abrahamyan, Arif Ahmad, Ganesh Bannur, Junrong Chen, Kimi Chen, Mingkai Deng, Ruobing Han, Xinqi Huang, Haoqiang Kang, Zheqi Liu, Enze Ma, Hector Ren, Yashowardhan Shinde, Rohan Shingre, Ramsundar Tanikella, Kaiming Tao, Dequan Yang, Xinle Yu, Cong Zeng, Binglin Zhou, Zhengzhong Liu, Zhiting Hu, and Eric P. Xing. Pan: A world model for general, interactable, and long-horizon world simulation, 2025. URL https://arxiv.org/abs/2511.09057.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. *arXiv preprint arXiv:2412.15109*, 2024.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=meRCKuUpmc.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, pp. 4950–4957. International Joint Conferences on Artificial Intelligence Organization, 2018.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018. URL https://arxiv.org/abs/1711.00937.

An Dinh Vuong, Tuan Van Vo, Abdullah Sohail, Haoran Ding, Liang Ma, Xiaodan Liang, Anqing Duan, Ivan Laptev, and Ian Reid. World2act: Latent action post-training via skill-compositional world models. *arXiv preprint arXiv:2603.10422*, 2026.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

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.

Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *Advances in neural information processing systems*, 32, 2019.

Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen, Zheng Lin, and Weiping Wang. Dynamic early exit in reasoning models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=NpU7ZXafRi.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators, 2024. URL https://arxiv.org/abs/2310.06114.

Angen Ye, Boyuan Wang, Chaojun Ni, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, Jie Li, Jindi Lv, Jingyu Liu, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026a.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026b.

Yang Ye, Tianyu He, Shuo Yang, and Jiang Bian. Reinforcement learning with inverse rewards for world model post-training. *arXiv preprint arXiv:2509.23958*, 2025.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Xiangcheng Zhang, Haowei Lin, Haotian Ye, James Zou, Jianzhu Ma, Yitao Liang, and Yilun Du. Inference-time scaling of diffusion models through classical search. *arXiv preprint arXiv:2505.23614*, 2025.

Zhilong Zhang, Ruifeng Chen, Junyin Ye, Yihao Sun, Pengyuan Wang, Jingcheng Pang, Kaiyuan Li, Tianshuo Liu, Haoxin Lin, Yang Yu, et al. Whale: Towards generalizable and scalable world models for embodied decision-making. *arXiv preprint arXiv:2411.05619*, 2024.

Chongyi Zheng, Jens Tuyls, Joanne Peng, and Benjamin Eysenbach. Can a misl fly? analysis and ingredients for mutual information skill learning. *arXiv preprint arXiv:2412.08021*, 2024.

Ruijie Zheng, Jing Wang, Scott Reed, Johan Bjorck, Yu Fang, Fengyuan Hu, Joel Jang, Kaushil Kundalia, Zongyu Lin, Loic Magne, Avnish Narayan, You Liang Tan, Guanzhi Wang, Qi Wang, Jiannan Xiang, Yinzhen Xu, Seonghyeon Ye, Jan Kautz, Furong Huang, Yuke Zhu, and Linxi Fan. Flare: Robot learning with implicit world modeling, 2025. URL https://arxiv.org/abs/2505.15659.

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.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning, 2025. URL https://arxiv.org/abs/2411.04983.

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. *arXiv preprint arXiv:2504.02792*, 2025a.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation, 2025b. URL https://arxiv.org/abs/2406.14540.

Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. *arXiv preprint arXiv:2009.12293*, 2020.

Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłos, Błażej Osiński, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=S1xCPJHtDB`.

A Related Work

World Models for Robotics. Model-based reinforcement learning (MBRL) learns predictive environment dynamics for planning and policy improvement. Early approaches focused on probabilistic, data-efficient control (Sutton, 1990; Deisenroth & Rasmussen, 2011), while modern deep MBRL combines expressive dynamics models with planning and imagined rollouts in off-policy learning (Chua et al., 2018; Janner et al., 2019). By reasoning over predicted futures, these methods can be highly sample-efficient and effective for control. However, they are often tied to specific tasks or policy distributions, which limits their scalability to high-dimensional perception, diverse interaction regimes, and broad generalization. More recently, general-purpose world models that learn predictive representations from large and diverse sequential data have been developed. One line of work focuses on *latent world models*, which learn compact action-conditioned dynamics from high-dimensional observations and perform prediction, imagination, and control in a learned latent space (Ha & Schmidhuber, 2018; Hafner et al., 2019b; Łukasz Kaiser et al., 2020; Hafner et al., 2019a; 2023; Zhou et al., 2024; Garrido et al., 2024; Maes et al., 2026). A second line is more *planning- and control-centric*, with objectives that are directly for decision making and policy improvement (Schrittwieser et al., 2020; Hansen et al., 2022; 2023; Zhou et al., 2024). A third line studies *pixel-based* world models trained on large-scale robotics or even internet video data (Wu et al., 2024; Rigter et al., 2025; Chen et al., 2025b), either by combining video generation with inverse dynamics models (IDMs) (Chi et al., 2025b) or by directly learning action-conditioned dynamics (Guo et al., 2025; Hafner et al., 2025; Gao et al., 2026), with the goal of producing visually realistic yet controllable futures. A key advantage of these models is that they can leverage internet-scale data and increasingly scalable model sizes to acquire broad predictive priors. Despite their strong generative capacity, these models still face important challenges in physical consistency and action following (Shang et al., 2026; Mei et al., 2026), especially when deployed for robotics control. Our work addresses this through targeted exploration, which actively collects informative interactions with verification to improve the robustness of action-conditioned world models.

Exploration for World Models. A central challenge in exploration for world models is determining which actions yield the most informative data for updating the world model. Early count-based methods encourage coverage of the state or transition space through visitation statistics or density surrogates (Bellemare et al., 2016; Ostrovski et al., 2017; Burda et al., 2018; GX-Chen et al., 2025), but they are largely agnostic to transition learnability and therefore can be inefficient for reducing model error. More recent approaches estimate informativeness through epistemic uncertainty or disagreement (Houthooft et al., 2017; Shyam et al., 2019; Sekar et al., 2020), or through learning progress (Graves et al., 2017; Achiam & Sastry, 2017; Kim et al., 2020). Related prediction-error and curiosity methods directly use model mismatch as an intrinsic signal (Pathak et al., 2017; Haber et al., 2018; Stadie et al., 2015; Du et al., 2021), while goal-discovery approaches use world models to construct intrinsic objectives for exploration and skill learning (Mendonca et al., 2021; Hu et al., 2023; Zheng et al., 2024; Liu et al., 2024). A common issue across these lines of work is that they rely on signals derived from the current action-conditioned world model itself, and are therefore often most reliable in well-explored regimes and least reliable in the under-explored regimes where verification matters most. Our method addresses this by verifying two simpler factors of informative transitions, *plausible future states* and *sparse inverse-dynamics reachability*, instead of employing dense forward-model error directly.

Inverse Dynamics on Videos. A common strategy for leveraging action-free observations is to infer actions from state transitions. Existing work uses inverse dynamics in different roles in world model and policy learning. As *policies*, IDMs convert predicted or planned future observations into executable actions, as in visual-planning and foresight-based systems (Du et al., 2023; Black et al., 2023; Chen et al., 2025c; Tian et al., 2025; Hu et al., 2025; Lv et al., 2025; Luo et al., 2025; Cai et al., 2026; Ye et al., 2026b). As *regularizers*, inverse objectives have long been used to shape self-supervised representations and dynamics features (Agrawal et al., 2016; Pathak et al., 2017). As *labelers*, inverse models can impute missing actions from state-only or video demonstrations (Torabi et al., 2018; Yang et al., 2019; Baker et al., 2022; Jang et al., 2025). As *verifiers*, they can test whether a transition is action-consistent; conceptually, our approach is closest to this line and to RLIR (Ye et al., 2025). However, our verifier differs in two key ways: it uses a reverse cycle anchored on

plausible future states, and it checks reachability with a sparse action-relevant inverse model rather than dense full-state generation.

World Action Models. World action models (WAMs) have recently made substantial progress in policy learning by jointly leveraging action-free internet video and action-labeled robot data. One line jointly predicts future video and actions within a unified model, allowing large-scale action-free data to improve representations without changing the downstream policy interface (Cheang et al., 2024; Rigter et al., 2025; Huang et al., 2025; Zhu et al., 2025a; Li et al., 2025a; Bi et al., 2025; Wu et al., 2024). A second line performs visual planning or foresight generation and conditions action prediction on future frames, shifting more of the planning burden into semantic image or video space while keeping low-level control at the action level (Du et al., 2023; Black et al., 2023; Chen et al., 2025b; Jang et al., 2025; Vuong et al., 2026; Hu et al., 2025; Lv et al., 2025; Cai et al., 2026; Ye et al., 2026b;a; Li et al., 2026; Yuan et al., 2026). Our method is formally closer to the second family: like these approaches, it combines future-state generation with inverse action prediction. The difference is that we use this structure not as a policy model, but to expose an accuracy advantage of inverse verification over world models and to repurpose a WAM-like pipeline as a verifier for action-conditioned world models.

B Conclusion and Discussion

In this work, we identified an important asymmetry between forward and inverse dynamics: inferring which action caused a plausible transition can be substantially easier than predicting its full outcome. Building on this insight, we introduced World Action Verifier, a self-improving framework that exploits cycle consistency among a diverse subgoal generator, a sparse inverse model, and a forward world model to gather informative interactions. Across MiniGrid, RoboMimic, and ManiSkill, our method enables $2\times$ greater sample efficiency in exploration and improves downstream policy performance by 18%.

Although our primary focus is verification-guided exploration for acquiring new interactions, the proposed WAV could be potentially useful in other settings that benefit from robust error estimation, such as test-time scaling (Nakamoto et al., 2024; Liu et al., 2025; Kwok et al., 2025) and offline data curation Hejna et al. (2025); Chen et al. (2025a); Agia et al. (2025). Nevertheless, the current instantiation of our method requires three inference passes, making it computationally more expensive than prior exploration methods. Improving its efficiency through shared intermediate representations (Zhu et al., 2025a; Li et al., 2025a) or adaptive computation mechanisms (Yang et al., 2026; Zhang et al., 2025; Lin et al., 2026) is an important direction for enabling more affordable real-time deployment.

More broadly, our results suggest that the forward–inverse asymmetry may be especially pronounced in high-dimensional, uncertain environments. However, fully self-improving world models in such complex settings remain elusive. Unlike language models, which can already improve from purely synthetic data on some reasoning tasks, our method still relies critically on additional environment feedback to correct action-conditioned prediction errors. Reducing this reliance will likely require substantially stronger verification mechanisms (Lifshitz et al., 2025; Kwok et al., 2026), likely scaffolding on more capable pretrained models. Extending World Action Verifier to incorporate richer generative priors (Chen et al., 2025b; Gao et al., 2026), more expressive inverse models (Tian et al., 2024; Ye et al., 2026b), and longer-horizon embodied tasks (Liu et al., 2023; Nasiriany et al., 2024; Bu et al., 2025; Dai et al., 2026) can be promising directions for future work.

C Baseline Descriptions

We compare our method (Sec. 2) against the following exploration strategies:

- *Random:* uniformly sample candidates from unlabeled examples (lower bound).
- *Uncertainty:* select candidates with the highest predictive uncertainty (Sekar et al., 2020).
- *Progress:* select candidates with the largest learning progress, measured as the change in model loss on the candidate set between two rounds (Kim et al., 2020).
- *Vanilla IDM*: our method without the sparsity constraint.

- *Oracle:* select candidates with the largest prediction loss of the world model, computed using ground-truth action labels (upper bound).

D MiniGrid Setting

We conduct experiments in the MiniGrid simulation environment, using `EmptyEnv` with three object types: key, ball, and box, each of which can be either red or blue. The agent has seven discrete actions: *turn left, turn right, move forward, pick up, drop, toggle, and swap.* The behavior of the toggle action is object-dependent: for keys and balls, it switches the object's color; for boxes, it acts as an exchange mechanism, swapping the item currently held by the agent with the item inside the box (or placing the held item inside if the box is empty). The swap action is not part of the original EmptyEnv. We define it as exchanging the object in front of the agent with the object it is carrying. Based on this setup, we designed three tasks in EmptyEnv, the details of which can be found in Sec. D.1.

D.1 Task Definitions

To evaluate the agent's ability to handle long-horizon dependencies and compositional logic, we design three complex tasks in the MiniGrid environment. Each task requires the agent to manipulate objects (Key, Ball, Box) based on their attributes (Red, Blue).

- **Task 1: Key Delivery.** The agent must: (1) locate a key, (2) change its color to match the target box, (3) place the key inside the box, (4) swap the box with a ball, (5) adjust the ball's color to match the box, and (6) reach the goal.
- **Task 2: Ball Delivery.** This is a structural mirror of Task 1 but swaps the roles of the key and the ball. The agent must place the ball inside the box before manipulating the key and reaching the goal.
- **Task 3: Object Matching.** The agent must: (1) identify the reference color of the box, (2) locate the key and ball, (3) synchronize the key's color with the box, (4) synchronize the ball's color with the box, (5) place both the key and the ball around the box, and (6) reach the goal.

D.2 Dataset Composition.

Random Play Dataset. We construct random play datasets based on the `EmptyEnv`, where objects can be freely placed. To study state complexity, we vary the number of objects $\{6, 8, 10, 12, 14\}$ and collect trajectories using random policies, resulting in environments with increasingly complex object configurations. In this setting, only the environment with 6 objects is used to construct both the training and test sets, while environments with higher object counts are used exclusively for testing, enabling controlled evaluation of generalization to more complex scenes.

To study environmental stochasticity, we vary the number of noisy floor tiles $\{0, 1, 2, 3, 4\}$, whose colors change randomly at every step. For each noise level, we collect trajectories with random policies and construct both training and test sets, allowing us to evaluate robustness under different levels of environmental noise.

Exploration Pool. We collect a total of 56,273 transitions across the three tasks defined in Sec. D.1, covering diverse state–action–next-state tuples. More than half of the collected data (28,000 transitions) is used as an unlabeled pre-training set to train the video model without action annotations. The remaining 28,273 transitions form the exploration pool, from which different acquisition strategies iteratively select informative samples. We additionally construct an action-balanced test set of 10,368 transitions for evaluation.

Compositional OOD Generalization. To evaluate compositional out-of-distribution (OOD; equivalently, out-of-support in our theoretical framework, Sec. 3) generalization, we partition action–object–color combinations as summarized in Table 2. During training, models are exposed only to a restricted subset of combinations (e.g., *pick up blue keys*), while evaluation is conducted on held-out compositions involving unseen combinations (e.g., *pick up blue balls*). This setup tests whether the model can generalize compositionally beyond observed training distributions.

Table 1: Statistics of the collected dataset for exploration. The data is split into an unlabeled pre-training set for learning the video model, an exploration pool for sample acquisition, and an action-balanced test set for evaluation.

Category	**Count (Transitions)**
Total Collected	66,641
Unlabeled Pre-training Set	28,000
Exploration Pool	28,273
Test Set (Action-balanced)	10,368

Table 2: Action–object composition coverage in the training and OOS test sets. ✓ denotes compositions seen during training, ⋆ denotes OOS-only compositions evaluated at test time, and × denotes combinations absent from both sets.

Action \ Object	red key	blue key	red ball	blue ball
Pick up	×	✓	×	⋆
Drop	⋆	×	✓	×
Toggle	⋆	✓	✓	⋆
Swap (box for . . .)	⋆	×	✓	×

D.3 Evaluation Metrics.

We first report **Dynamics Accuracy**, which measures prediction accuracy only over elements that undergo temporal changes, including both visual grid cells and internal agent attributes (e.g., carried status), while masking out invariant background regions. By focusing on these dynamic components, Dynamics Accuracy mitigates metric inflation caused by static background dominance and better reflects the model's ability to capture action-driven dynamics.

In exploration experiments, we further evaluate all methods using the world model's next-state prediction loss on the held-out test set. We utilize prediction loss in this context because it provides a more sensitive signal of training stability and convergence behavior during exploration, whereas Dynamics Accuracy is better suited for interpreting final predictive performance.

To assess the quality of data selection, we compute the rank correlation between method-assigned scores and Oracle scores using **Spearman's rank correlation coefficient** (ρ) Spearman (1961) and **Kendall's rank correlation coefficient** (τ) Kendall (1938). Given a set of samples with scores $\{s_i\}$ and corresponding Oracle scores $\{o_i\}$, the Spearman correlation is defined as

$$\rho = 1 - \frac{6\sum_i (r_i - q_i)^2}{n(n^2-1)}, \tag{11}$$

where r_i and q_i denote the ranks of s_i and o_i, respectively.

Kendall's τ measures the consistency of pairwise orderings:

$$\tau = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}, \tag{12}$$

where N_c and N_d are the numbers of concordant and discordant pairs. Higher values indicate stronger agreement with the Oracle ranking.

D.4 Robustness Evaluation Setup

We vary the amount of labeled training data $\{400, 800, 1200, 1600, 2000\}$ collected in environments with 6 objects, and evaluate on test transitions in environments with $\{6, 8, 10, 12, 14\}$ objects. In addition, to examine robustness against observation noise, we construct training datasets in 6-object environments with $\{0, 1, 2, 3, 4\}$ noisy floors. For direct comparison, we convert inverse model predictions into next state predictions: for each test pair, the IDM predicts an action, which we then execute in the simulator to obtain the induced next state. We report the same dynamics accuracy defined in Sec. D.3 for both the world model and the IDM-induced transition. Additionally, we

compute the Spearman's rank correlation coefficient Spearman (1961) and Kendall's rank correlation coefficient Kendall (1938) between the data selection scores of each method and those of the Oracle method. Details of the two rank metrics are provided in Sec. D.3.

D.5 MODELS IN MINIGRID

World Model. We employ a physics-aligned architecture that preserves spatial structure via coordinate-aware convolutions. The model incorporates a **supervised vector-quantized bottleneck** van den Oord et al. (2018), which explicitly maps latent codes to discrete actions and conditions a residual dynamics engine through Feature-wise Linear Modulation (FiLM) Perez et al. (2017), promoting object persistence and physical consistency.

Inverse Dynamics Models (IDM). To verify the robustness of sparse IDM, we compare two IDMs on the OOD test set:

- **Vanilla IDM:** Takes the entire observation frame and the agent's proprioceptive state as input.
- **Sparse IDM:** Built upon the vanilla IDM by applying a learnable feature mask to the input, which selectively filters out irrelevant information and yields a sparse representation. The mask is learned automatically during training, encouraging the model to focus on the most informative local features.

D.6 EXPLORATION METHODS IN MINIGRID

The **Oracle** strategy used in Sec. 4 selects samples with the highest prediction loss, corresponding to the hardest transitions under the current world model. To better understand the role of sample difficulty during exploration, we further introduce two oracle variants:

- **Oracle-Easy:** selects samples with the lowest prediction loss, corresponding to the easiest transitions.
- **Oracle-Uniform:** partitions samples into disjoint prediction-loss intervals and selects high-loss samples within each interval, ensuring balanced coverage across different difficulty levels.

To analyze the behavior of different exploration strategies, we visualize the distribution of prediction errors—measured by the world model's prediction loss—over the samples selected by each method (Figure 5). This analysis reveals how different selection criteria bias the collected data toward specific difficulty regimes and provides insight into their impact on world model learning.

Qualitative Results. Figure 10 presents a qualitative comparison of world model predictions across interactive actions. While most methods perform similarly on simple motion-dominated transitions, clear differences arise for interaction-centric actions such as *Toggle* and *Swap*. In these cases, models trained with actively selected data more accurately capture interaction-induced state changes, whereas **Random** exhibit a strong bias toward predicting frequent but uninformative movement actions (e.g., *Turn*), highlighting the benefit of informative data selection under distribution shift.

D.7 ADDITIONAL MINIGRID EVALUATION RESULTS

In this section, we present the evaluation of sparse vs. vanilla IDMs (*RQ2*), the correlation analysis of verification scores (*RQ1–RQ2*), and the world model quality improvement results (*RQ3*) on MiniGrid, complementing the robustness analysis in the main text (Sec. 4.1.1).

For ***RQ2***, Figure 6 (Left) reveals a pronounced divergence in out-of-distribution generalization when data is limited. The vanilla IDM fails on interaction-heavy actions, whereas the sparse IDM maintains strong performance, showing that enforcing sparsity promotes more robust action inference.

Further, we assess how faithfully each method's verification scores rank sample difficulty. The correlation analysis in Figure 6 (Mid) shows that our method achieves the highest Spearman and Kendall scores, indicating strong alignment with the Oracle's data ranking. Together, these results validate the robustness of the proposed world action verification mechanism.

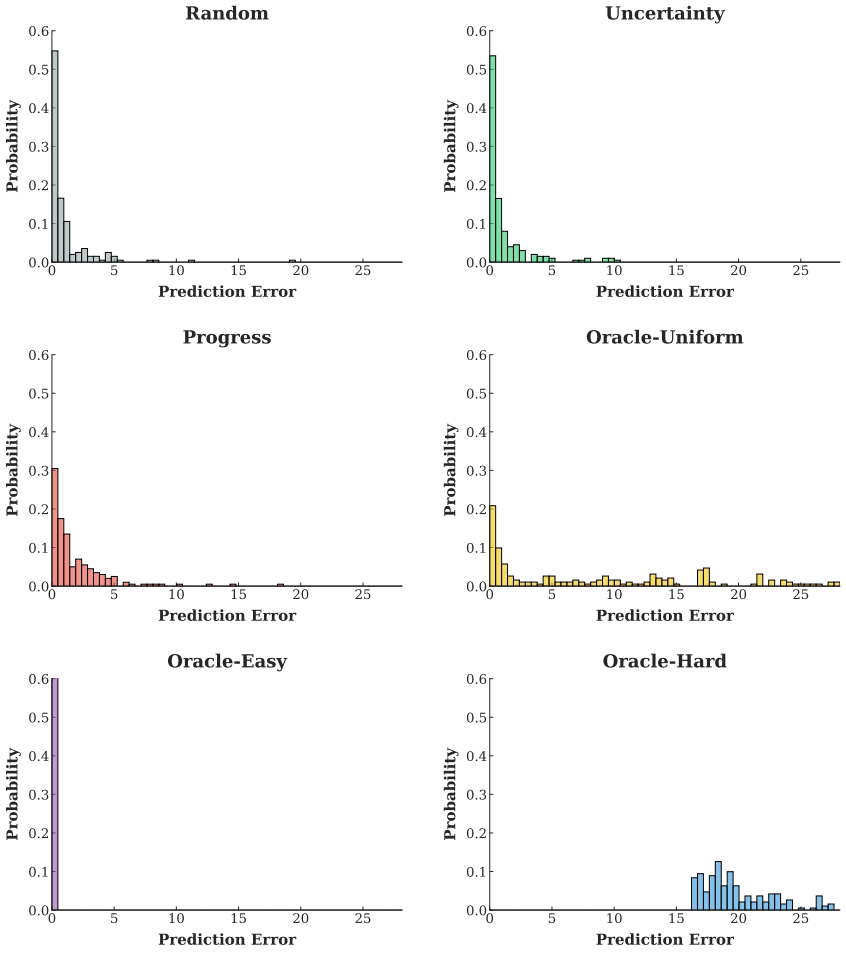

Figure 5: The distribution of world model's prediction error on the data selected by different exploration methods in MiniGrid.

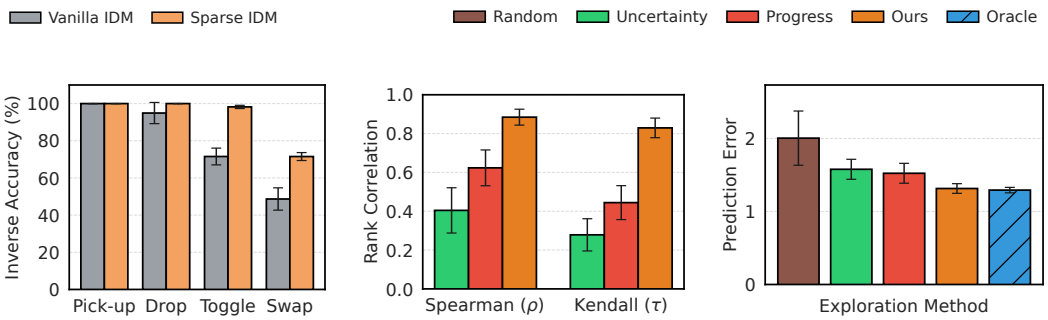

Figure 6: **Evaluation of world model learning with WAV on MiniGrid.** (Left) Action prediction accuracy of Sparse IDM and Vanilla IDM. Sparse IDM achieves better out-of-distribution generalization under limited data. (Mid) Correlation with Oracle ranking. We measure how well each method ranks informative samples using Spearman and Kendall correlations between method-assigned scores and Oracle scores. (Right) Comparison of acquisition strategies. Our proposed WAV outperforms standard baselines and approaches Oracle performance by prioritizing interaction-rich transitions.

D.7.1 Improvements on World Model Quality

Building on the above justifications, we now evaluate ***RQ3*** by examining whether the proposed framework improves world model learning quality.

Setup. We first train a base model on 200 uniformly sampled labeled transitions, followed by three exploration rounds where each strategy acquires a budget of 100 transitions. We report the prediction error averaged over five random seeds, focusing on the second round where differences are most pronounced for clearer comparison.

Results. Figure 6 (Right) demonstrates that both the Oracle and our method significantly outperform baseline methods. We attribute this advantage to the structural imbalance of the dataset, where complex interaction actions are sparse relative to simple movement. The *Random* strategy fails to sample these critical sparse events with sufficient frequency. The *Progress* method struggles with sample redundancy; it tends to over-prioritize transitions where the model is already competent, yielding negligible marginal information gain. Similarly, the *Uncertainty* baseline suffers from a slow warm-up, leading to suboptimal performance in the early stages of exploration. In contrast, by explicitly filtering for interaction-rich transitions, our method closely matches the Oracle's selection efficacy, achieving superior generalization under the same data budget. Qualitative results are given in Sec. D.6.

E Robotic Doamins Setting

The experiments are conducted on simulated robotic manipulation tasks from Robomimic (Zhu et al., 2020) (`Lift`, `Can`, `Square`) and ManiSkill (Mu et al., 2021) (`PullCube`, `PokeCube`, `LiftPeg`). Fig. 7 provides a visualization of these tasks. For each task, the agent observes RGB images from both a wrist-mounted camera and a front-facing camera. In addition, proprioceptive observations are provided, including the end-effector position and orientation, as well as the gripper position. The action space is a 7-dimensional continuous vector in $[-1, 1]$, including the control changes in the end-effector position and orientation, and the opening and closing states of the gripper.

E.1 Dataset Construction and Model Choices.

Datasets & Setups. We consider a set of challenging robotic manipulation tasks from two evaluation suites in Roboverse (Geng et al., 2025): RoboMimic (Zhu et al., 2020) (`Lift`, `Can`, `Square`) and ManiSkill (Mu et al., 2021) (`PullCube`, `PokeCube`, `LiftPeg`). For both suites, we curate training data using expert demonstrations in a two-stage process. We first pretrain diffusion policies (Chi et al., 2025a) for different numbers of training steps, yielding a diverse collection of behavior trajectories with varying levels of optimality. Based on these trajectories, we partition the data into *two subsets*:

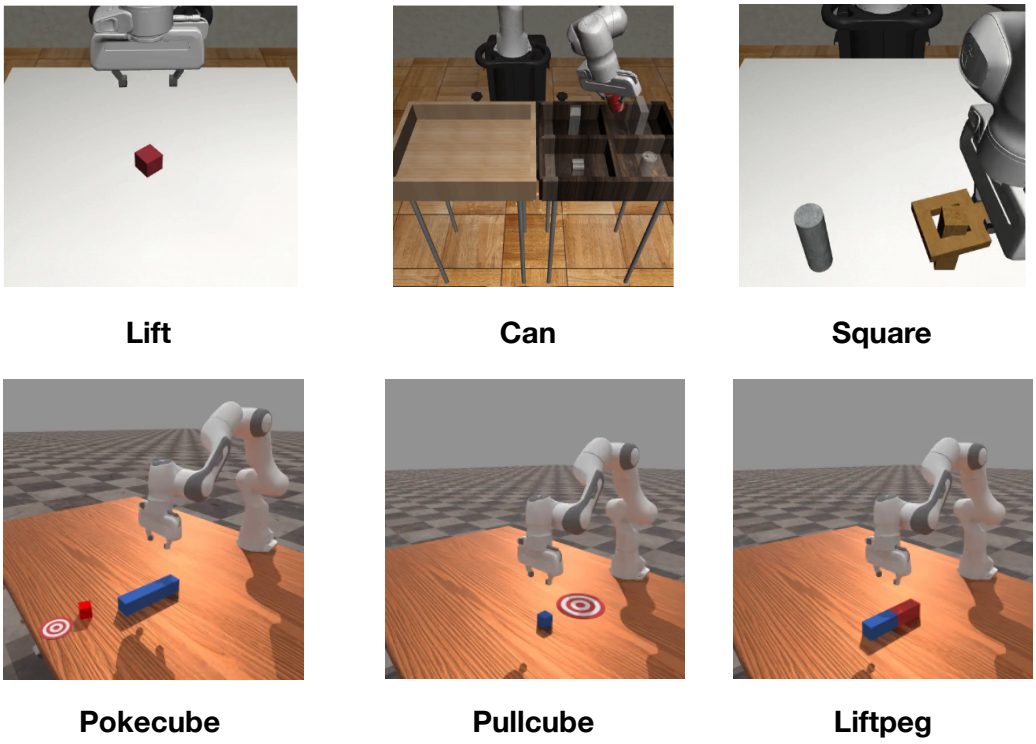

Figure 7: Visual description of all tasks used from RoboMimic (Row 1) and Maniskill (Row 2).

(1) *the world-model warm-up dataset*, which includes the expert demonstrations together with on-policy trajectories collected from the best-performing diffusion policy checkpoint trained on those demonstrations; (2) *the exploration dataset*, which consists of trajectories generated by imperfect diffusion policy checkpoints, capturing diverse exploratory behaviors.

Model Choices. For the world model, we adopt Dreamer-v3 (Hafner et al., 2023), which learns a latent recurrent state-space model (RSSM). For sparse IDM, we employ the model backbones from CLAM (Liang et al., 2025) and further impose sparsity on this latent action space. Details are given in Sec. E.4.

E.2 Additional Details on Setups.

Dataset Collection. For both benchmarks, we train 10 diffusion policy models to collect a diverse set of trajectories, resulting in a total of 1500 samples per task. We follow the default horizon settings of each environment: 100 steps for `Lift`, 200 for `Can` and `Square`, 50 for `PullCube`, 100 for `PokeCube`, and 150 for `LiftPeg`. We use nine of these checkpoints to construct the exploration dataset, and reserve the remaining checkpoint as a validation set for evaluating world model learning quality.

Reward Evaluation. We follow the reward formulation of SAILOR (Jain et al., 2025). Specifically, we train a reward model (using the same architecture and hyperparameters as in SAILOR) to score the latent states of our world model based on how expert-like they are. The reward model is trained as a discriminator between latent embeddings from expert rollouts and learner rollouts, using a moment-matching objective with a gradient penalty.

Policy Learning Setup. We evaluate whether the learned world model leads to improved policy learning through imagination. Following SAILOR (Jain et al., 2025), we use the world model to perform online search over imagined latent trajectories for policy refinement of base diffusion policies. Specifically, we adopt their reward modeling setup, where a learned reward model scores latent states based on how closely they resemble expert behavior. We then evaluate policy performance after

Table 3: World model training hyperparameters.

Hyperparameter	Value
Replay capacity	1×10^5
Batch size	16
Batch length	32
Optimizer	Adam
Reconstruction loss scale	1.0
Learning rate	1×10^{-4}

fine-tuning the world model using a data budget of 1,000 trajectories under different world model variants.

E.3 Details on World Models.

We adopt an action-conditioned world model based on Dreamer-V3 (Hafner et al., 2023). Visual observations are encoded using a convolutional encoder with stride-2 convolutions, and proprioceptive states are embedded with a 5-layer MLP. Based on empirical performance, image and state inputs are processed by separate encoders. For decoding, image observations are reconstructed using a transposed-convolutional decoder with stride-2 upsampling, and proprioceptive states are reconstructed via a 5-layer MLP.

The latent state z_t comprises a deterministic recurrent component h_t and a stochastic component s_t. The deterministic state is modeled by a GRU with a 512-dimensional hidden state and is updated using the previous latent state z_{t-1} and action a_{t-1}. The resulting recurrent state is then used by the dynamics model to parameterize the distribution of the stochastic latent variable s_t. The number of dimensions of stochastic representation is 1024. The number of the rollout horizon is 32. Other training hyperparameters are given in Table 3.

E.4 Details on Inverse Dynamic Models.

We adopt the inverse dynamics modeling framework from CLAM (Liang et al., 2025) as our base IDM architecture. Given consecutive observations (o_t, o_{t+1}), a latent inverse dynamics model infers a continuous latent action $z_t = f_\phi(o_t, o_{t+1})$ that captures the underlying transition. This latent action is then used to condition a latent forward dynamics model, $g_\psi(o_{t+1} \mid o_t, z_t)$, which predicts the next observation $\hat{o}_{t+1}$. An action decoder $p_\omega(a_t \mid z_t)$ maps the latent action back to the environment action space. The IDM, FDM, and action decoder are jointly trained using a combination of reconstruction and action prediction losses over both labeled and unlabeled trajectories. Compared to CLAM, we additionally We encourage sparsity in the latent action space by applying an ℓ_1 regularization term to the inferred latent actions z_t. encouraging the model to discover structured and task-relevant factors.

For visual observations, following CLAM, we adopt a ST Transformer (Bertasius et al., 2021) for encoders. Each $64 \times 64 \times 3$ RGB image is first partitioned into non-overlapping 16×16 patches, yielding 16 visual tokens per frame, which are projected into a shared hidden space via a linear embedding layer. The encoder is composed of the stacked ST attention layers. In the decoder, each ST block performs cross-attention between visual tokens and the latent action representations produced by the encoder. Other hyperparameters are given in Table 4.

E.5 Visualization

Fig. 11–12 visualize open-loop rollouts on Robomimic-Lift and Robomimic-Square. Overall, the base model exhibits poor visual predictions, with noticeable degradation in both rendering quality and dynamical consistency. Incorporating uncertainty- and progress-aware exploration substantially improves visual fidelity with more samples, producing sharper and more coherent renderings over time. In contrast, the vanilla IDM improves the accuracy of the underlying dynamics, indicating that inverse dynamics supervision helps recover action-relevant transitions. However, over long horizons, particularly in the final one to two frames, noticeable discrepancies remain, such as misaligned gripper orientations and inaccurate object occlusions. Our sparse IDMs further mitigate these long-horizon

Table 4: Hyperparameters for inverse dynamics and action decoding.

Hyperparameter	Value
Num updates	500,000
Train action decoder every	2
Action decoder batch size	128
Action decoder loss weight	1
Action decoder hidden dim	[1024, 1024, 1024]
Action decoder embedding dim	512
Reconstruction loss weight	1
Sparsity loss weight	0.1
Latent action dim	16
Context len	2
Embedding dim	128

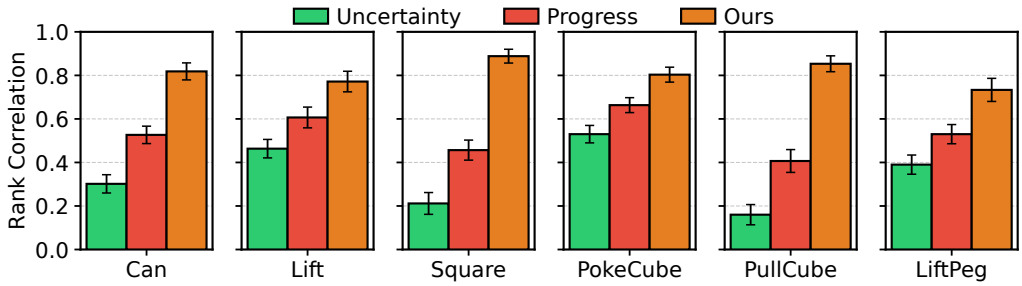

Figure 8: **Verification robustness of WAV on RoboMimic and ManiSkill.** Correlation with Oracle ranking. We evaluate how well each method orders informative samples by computing Spearman rank correlations between the method-assigned scores and Oracle scores on RoboMimic and ManiSkill environments. Higher correlation indicates closer agreement with the Oracle ranking.

errors, yielding more stable dynamics and better-preserved fine-grained details in the predicted rollouts.

E.6 Additional Robustness and World Model Quality Results

In this section, we present additional experimental results on the robustness of WAV (*RQ1–RQ2*) and world model quality improvement (*RQ3*) in the robotic manipulation domain, complementing the MiniGrid evaluation in the main text (Secs. 4.1.1 and D.7.1).

E.6.1 Robustness of World Action Verification

Setup. To evaluate the robustness of WAV, similar to the evaluation in MiniGrid, we compute the Spearman's rank correlation coefficient Spearman (1961) between the data selection scores of each method and those of the Oracle method. We use 100 samples that are held out from the world model training data.

Results. As shown in Figure 8, our verification scores more faithfully reflect the true (oracle) difficulty ranking of samples, verifying the robustness of WAV in the robotic setting.

E.6.2 Improvement of World Model Quality

To address ***RQ3***, we assess world-model learning quality by using the prediction loss of the learned world models. Concretely, we measure the mean squared error (MSE) of next-observation prediction on a held-out test set.

Setup. We first warm-start the world model with training on $\{200, 400, 600, 800, 1000\}$ trajectories to tune the world model for another 200 epochs with our self-improving loop. We measure the

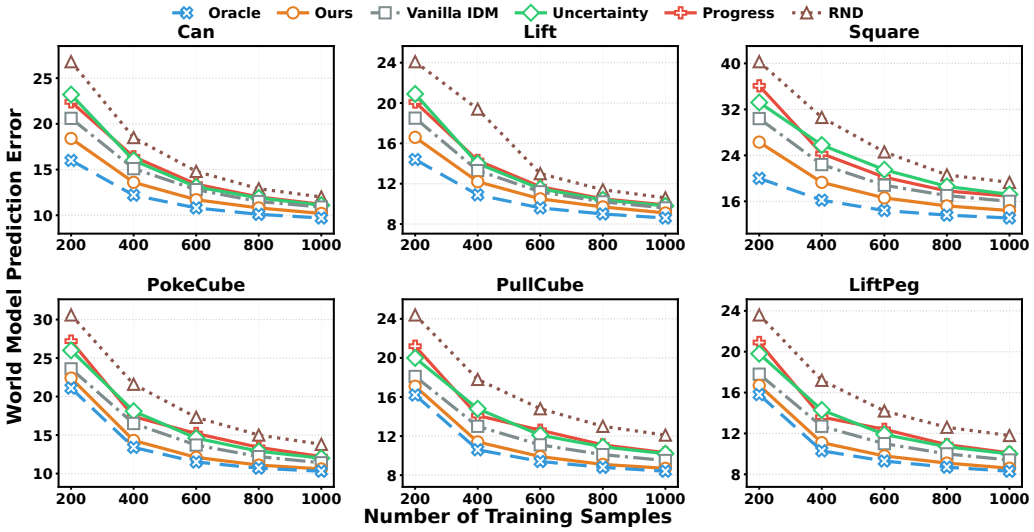

Figure 9: **Evaluation of world model learning with WAV on RoboMimic and ManiSkill.** We evaluate observation prediction as the number of training samples increases. Compared to existing methods, our approach yields more accurate world models with fewer samples. Error bars denote the standard error over 3 seeds.

error as the average observation MSE for 32 predicted frames after 2 exploration rounds. To ensure statistical reliability, we test using 3 seeds and report the mean error for each.

Results. Figure 9 shows the world model prediction results under different data budgets. Our method consistently outperforms all baselines, with particularly large gains in the low-data regime. In addition, sparse inverse dynamics models consistently outperform the dense ones on object manipulation tasks.

E.6.3 Detailed Policy Learning Results

Figure 4 evaluates whether improved world model learning translates into stronger downstream policy learning when planning in imagination. The performance gap is particularly pronounced on tasks with higher ambiguity or contact complexity, such as Can, Square, and PokeCube, where accurate latent dynamics are critical for effective imagination-based planning. In contrast, Lift exhibits smaller performance gaps across methods, indicating that simpler dynamics reduce sensitivity to world model quality. Overall, these results demonstrate that our approach learns more task-relevant and accurate dynamics representations, leading to more effective policy tuning in imagination and improved sample efficiency compared to both dense and heuristic-based baselines.

F Missing Details for Theory

F.1 Detailed Derivation for Distribution-Level Robustness

This appendix provides a compact, self-contained proof of the self-improvement guarantee stated in Sec. 3.1. The key idea is that, under a *generation–verification gap*, the full transition $(\mathbf{z}^t, \mathbf{a}^t) \mapsto \mathbf{z}^{t+1}$ may be out-of-support, while a small *verification subset* of latent variables remains on-support and is sufficient to recover the action. WAV uses this subset to *verify* (infer) missing action labels on OOS transitions, thereby expanding its action-labeled support.

F.1.1 Time-Lagged Latent Causal Model (TLCM)

We consider a *time-lagged latent causal model* (TLCM) with k latent blocks $\mathbf{z}^t := (\mathbf{z}_1^t, \ldots, \mathbf{z}_k^t)$, actions $\mathbf{a}^t$, and observations $\mathbf{x}^t$. For clarity, we present the (deterministic) Markovian case used in

our analysis:

$$\mathbf{z}^{t+1} = g(\mathbf{z}^t, \mathbf{a}^t), \tag{13}$$

$$\mathbf{x}^t = \varphi(\mathbf{z}^t), \tag{14}$$

where φ is a diffeomorphism onto its image (so φ^{-1} is well-defined on observed states). The induced causal graph over $(\mathbf{z}^t, \mathbf{a}^t, \mathbf{z}^{t+1})$ factorizes as

$$p(\mathbf{z}^{t+1} \mid \mathbf{z}^t, \mathbf{a}^t) \;=\; \prod_{i=1}^{k} p(\mathbf{z}_i^{t+1} \mid \mathrm{Pa}(\mathbf{z}_i^{t+1})). \tag{15}$$

F.1.2 Support, compositional OOS, and the verification subset

Let $\mathcal{D}_{\mathrm{act}} \subset \{(\mathbf{x}^t, \mathbf{a}^t, \mathbf{x}^{t+1})\}$ be the action-labeled seed dataset inducing a seed distribution $P_{\text{seed}}(\mathbf{z}^t, \mathbf{a}^t, \mathbf{z}^{t+1})$. Write

$$S_{\text{seed}} \;:=\; \mathrm{supp}\left(P_{\text{seed}}(\mathbf{z}^t, \mathbf{a}^t)\right)$$

for the on-support set of state–action pairs.

Definition F.1 (On-support vs. out-of-support (OOS))**.** A state–action pair $(\mathbf{z}^t, \mathbf{a}^t)$ is *on-support* if it lies in S_{seed} and is *out-of-support* (OOS) otherwise. For any index set $\mathcal{U} \subseteq \{1, \ldots, k\}$ we also define the marginal support

$$S_{\text{seed}}^{\mathcal{U}} \;:=\; \mathrm{supp}\left(P_{\text{seed}}(\mathbf{z}_{\mathcal{U}}^t, \mathbf{a}^t)\right),$$

and say $(\mathbf{z}_{\mathcal{U}}^t, \mathbf{a}^t)$ is on-support if it lies in $S_{\text{seed}}^{\mathcal{U}}$.

Definition F.2 (Compositional OOS transition)**.** A transition $(\mathbf{z}^t, \mathbf{a}^t, \mathbf{z}^{t+1})$ is *compositional OOS* (w.r.t. P_{seed}) if $(\mathbf{z}^t, \mathbf{a}^t) \notin S_{\text{seed}}$ but there exists a subset of variables $\mathcal{U}$ such that $(\mathbf{z}_{\mathcal{U}}^t, \mathbf{a}^t) \in S_{\text{seed}}^{\mathcal{U}}$. Intuitively, novelty comes from an unseen *combination* of factors, while at least one subset transition remains within the training support.

We now formalize which subset can serve as a verifier.

Definition F.3 (Source (insulated) set)**.** Let $\mathcal{S}_{\mathrm{src}} \subseteq \{1, \ldots, k\}$ be a set of latent blocks that is *causally insulated* from its complement in the TLCM graph:

$$\forall i \in \mathcal{S}_{\mathrm{src}}, \qquad \mathrm{Pa}(\mathbf{z}_i^{t+1}) \;\subseteq\; \{\mathbf{z}_j^t : j \in \mathcal{S}_{\mathrm{src}}\} \cup \{\mathbf{a}^t\}.$$

Equivalently, there are no directed edges from $\mathbf{z}^t_{\setminus \mathcal{S}_{\mathrm{src}}}$ into $\mathbf{z}^{t+1}_{\mathcal{S}_{\mathrm{src}}}$.

Definition F.4 (Verification subset)**.** Let $\mathcal{S}_{\mathrm{act}} := \{i : \mathbf{a}^t \in \mathrm{Pa}(\mathbf{z}_i^{t+1})\}$ denote the *action-influenced* latent blocks. We define the *verification subset* as

$$\mathcal{S} \;:=\; \mathcal{S}_{\mathrm{src}} \cap \mathcal{S}_{\mathrm{act}}.$$

The subset we require to remain on-support for verification is the *intersection* of (i) source/insulated variables and (ii) action-influenced variables.

Assumption F.5 (Generation–verification gap (information asymmetry))**.** *There exists a verification subset $\mathcal{S}$ such that for every (potentially OOS) transition $(\mathbf{z}^t, \mathbf{a}^t, \mathbf{z}^{t+1})$ we wish to label, the restricted state–action pair remains on-support:*

$$(\mathbf{z}_{\mathcal{S}}^t, \mathbf{a}^t) \in S_{\mathit{seed}}^{\mathcal{S}},$$

even though $(\mathbf{z}^t, \mathbf{a}^t)$ may lie outside S_{seed}.

F.1.3 Two identifiability ingredients from prior work

Our proof uses (i) identifiability of the latent blocks (up to permutation / element-wise transforms) and (ii) identifiability of the action from on-support subset transitions.

Condition F.6 (Identifiable latent blocks via mechanism sparsity)**.** *This condition is adapted from Proposition 7 (together with Assumption 5) of Lachapelle et al. (2024). Consider a TLCM whose observation model is a diffeomorphism (so φ is invertible on its image) and whose transition model is Markov with respect to a sparse dependency graph between $(\mathbf{z}^t, \mathbf{a}^t)$ and $\mathbf{z}^{t+1}$ (as in Equation* (15)*). Assume we learn a second TLCM $(\hat{\varphi}, \hat{g}, \hat{G})$ that is $(\mathbf{z}, \mathbf{a})$-consistent with the ground-truth model in the sense of Lachapelle et al. (2024) and that the ground-truth graph satisfies their* graphical criterion *(Assumption 5). Then the learned latent variables are identifiable up to a permutation and element-wise invertible transformations ("complete disentanglement"), so we can treat the learned blocks as the true blocks up to a fixed relabeling.*

Condition F.7 (Identifiable action from subset transitions)**.** *This condition is adapted from Theorem 1 of Lachapelle (2025) by substituting* $\mathbf{x} \equiv \mathbf{z}_{\mathcal{S}}^t$ *and* $\mathbf{x}' \equiv \mathbf{z}_{\mathcal{S}}^{t+1}$*. Let* $\mathcal{S}$ *be a fixed verification subset and define the subset dynamics* $g_{\mathcal{S}}(\mathbf{z}_{\mathcal{S}}^t, \mathbf{a}^t) := [g(\mathbf{z}^t, \mathbf{a}^t)]_{\mathcal{S}}$ *(well-defined whenever* $\mathcal{S} \subseteq \mathcal{S}_{\mathrm{src}}$*). Assume the following hold on* $S_{seed}^{\mathcal{S}}$*:*

1. ***Continuity:*** *for each action value* $\mathbf{a}$*, the map* $\mathbf{z}_{\mathcal{S}}^t \mapsto g_{\mathcal{S}}(\mathbf{z}_{\mathcal{S}}^t, \mathbf{a})$ *is continuous;*
2. ***Injectivity:*** *for every* $\mathbf{z}_{\mathcal{S}}^t$*,* $g_{\mathcal{S}}(\mathbf{z}_{\mathcal{S}}^t, \mathbf{a}_1) = g_{\mathcal{S}}(\mathbf{z}_{\mathcal{S}}^t, \mathbf{a}_2)$ *implies* $\mathbf{a}_1 = \mathbf{a}_2$*;*
3. ***Connected conditional support:*** *for each* $\mathbf{a}$ *in the action support,* $\mathrm{supp}\left(P_{seed}(\mathbf{z}_{\mathcal{S}}^t \mid \mathbf{a})\right)$ *is connected;*
4. ***Support overlap:*** *for any* $\mathbf{a}_1, \mathbf{a}_2$ *in the action support,* $\mathrm{supp}\left(P_{seed}(\mathbf{z}_{\mathcal{S}}^t \mid \mathbf{a}_1)\right) \cap \mathrm{supp}\left(P_{seed}(\mathbf{z}_{\mathcal{S}}^t \mid \mathbf{a}_2)\right) \neq \varnothing$*.*

Then the action is identifiable from the subset transition $(\mathbf{z}_{\mathcal{S}}^t, \mathbf{z}_{\mathcal{S}}^{t+1})$ *up to a fixed relabeling: there exists an injective map* v *(independent of* $\mathbf{z}_{\mathcal{S}}^t$*) such that any solution of the latent-action reconstruction problem in Lachapelle (2025) recovers* $v(\mathbf{a})$ *deterministically from* $(\mathbf{z}_{\mathcal{S}}^t, \mathbf{z}_{\mathcal{S}}^{t+1})$*. In particular, when the learned action alphabet matches the true discrete action set, this corresponds to a permutation of action labels.*

F.1.4 VERIFIED SELF-IMPROVEMENT

We now state and prove the main appendix result.

Theorem F.8 (Identifiability of Self-Improvement)**.** *Let* $\mathcal{D}_{\mathrm{act}} \subset \{(\mathbf{x}^t, \mathbf{a}^t, \mathbf{x}^{t+1})\}$ *be action-labeled transitions sampled from* P_{seed}*, and let* $(\mathbf{x}^{*,t}, \mathbf{x}^{*,t+1})$ *be an additional unlabeled transition sampled from some test distribution* p_{test}*. Let* $(\mathbf{z}^{*,t}, \mathbf{a}^*, \mathbf{z}^{*,t+1})$ *denote the corresponding latent transition in the TLCM, i.e.* $\mathbf{z}^{*,t+1} = g(\mathbf{z}^{*,t}, \mathbf{a}^*)$ *and* $\mathbf{x}^{*,t+1} = \varphi(\mathbf{z}^{*,t+1})$*. Assume:*

1. ***Latent blocks are identified*** *up to a fixed permutation / element-wise transform (Condition F.6);*
2. ***Generation–verification gap*** *holds for the verification subset* $\mathcal{S}$ *(Assumption F.5);*
3. ***Action is identifiable from subset transitions*** *on* $S_{seed}^{\mathcal{S}}$ *(Condition F.7).*

Let $h_\psi : (\mathbf{z}_{\mathcal{S}}^t, \mathbf{z}_{\mathcal{S}}^{t+1}) \mapsto \mathbf{a}^t$ *be an inverse dynamics model trained on the on-support subset transitions in* $\mathcal{D}_{\mathrm{act}}$*. Then the missing action label for the unlabeled transition is uniquely determined by* $(\mathbf{z}_{\mathcal{S}}^{*,t}, \mathbf{z}_{\mathcal{S}}^{*,t+1})$*, and*

$$\hat{\mathbf{a}}^* \;:=\; h_\psi(\mathbf{z}_{\mathcal{S}}^{*,t}, \mathbf{z}_{\mathcal{S}}^{*,t+1})$$

recovers the true action (up to the fixed label relabeling in Condition F.7; with labeled data this relabeling is resolved so that $\hat{\mathbf{a}}^* = \mathbf{a}^*$*). Consequently, the tuple* $(\mathbf{x}^{*,t}, \hat{\mathbf{a}}^*, \mathbf{x}^{*,t+1})$ *is correctly labeled while it may satisfy* $(\mathbf{z}^{*,t}, \mathbf{a}^*) \notin S_{seed}$*, i.e. it can expand the action-labeled support.*

Proof. By Condition F.6 (a restatement of Lachapelle et al. (2024, Prop. 7) in our notation), the learned representation can be aligned with the ground-truth latent blocks up to a fixed permutation and element-wise invertible transforms. This alignment preserves the block structure and (up to a fixed relabeling) the parent/child relations in the TLCM graph, so the verification subset $\mathcal{S} = \mathcal{S}_{\mathrm{src}} \cap \mathcal{S}_{\mathrm{act}}$ is well-defined and accessible from observations via the learned encoder.

By Assumption F.5, the subset state–action pair $(\mathbf{z}_{\mathcal{S}}^{*,t}, \mathbf{a}^*)$ lies in the on-support set $S_{\mathrm{seed}}^{\mathcal{S}}$, even if the full pair $(\mathbf{z}^{*,t}, \mathbf{a}^*)$ is OOS. Therefore, the on-support subset transitions contained in $\mathcal{D}_{\mathrm{act}}$ are sufficient to train an inverse model h_ψ for $g_{\mathcal{S}}$.

Finally, by Condition F.7 (adapted from Lachapelle (2025, Thm. 1) with $\mathbf{x} \equiv \mathbf{z}_{\mathcal{S}}^t$ and $\mathbf{x}' \equiv \mathbf{z}_{\mathcal{S}}^{t+1}$), the action is identifiable from the subset transition $(\mathbf{z}_{\mathcal{S}}^t, \mathbf{z}_{\mathcal{S}}^{t+1})$ up to a fixed relabeling. Hence applying h_ψ to the on-support subset transition $(\mathbf{z}_{\mathcal{S}}^{*,t}, \mathbf{z}_{\mathcal{S}}^{*,t+1})$ recovers the correct action label (after resolving the fixed relabeling using the labeled actions in $\mathcal{D}_{\mathrm{act}}$). This yields the correctly labeled tuple $(\mathbf{x}^{*,t}, \hat{\mathbf{a}}^*, \mathbf{x}^{*,t+1})$, which can lie outside the original support S_{seed} and thus expands the action-labeled coverage. □

F.1.5 IMPLICATIONS OF THE GENERATION–VERIFICATION GAP

We now interpret Theorem F.8 and Assumption F.5 through a practical lens, identifying when WAV is most beneficial, when it degrades, and when it fails.

How large is the gap? WAV separates *verification* (recovering the missing action) from *generation* (predicting the full next state). Verification only uses the subset transition on $\mathbf{z}_{\mathcal{S}}$, while generation must model the full $\mathbf{z}$. As a rule of thumb, WAV becomes most attractive when $\dim(\mathbf{z}_{\mathcal{S}}) \ll \dim(\mathbf{z})$: the inverse model stays simple and stable even as the world model faces increasingly many OOS compositions.

Condition 1: fixed verifier, growing scene (maximum benefit). If $\mathbf{z}_{\mathcal{S}}$ is agent-centric and fixed-dimensional (e.g., proprioception) while the rest of the scene grows in complexity (more objects, tools, contacts), then action recovery continues to rely on the same low-dimensional signal while the world model must extrapolate over a much larger state space. In this regime, the benefit of WAV grows with scene complexity. *Example (warehouse robot).* A 7-DoF manipulator arm has $\dim(\mathbf{z}_{\mathcal{S}}) = 7$ (joint angles/velocities). In a warehouse with many objects, the world model must predict the state of each object (and their interactions), so $\dim(\mathbf{z})$ grows with scene complexity. As scene complexity grows, predicting the full next state requires accounting for many interacting factors beyond the agent's direct control, which increases the difficulty of accurate forward prediction. In many control settings, however, the most useful signal is the action-imprinted change. This motivates focusing the inverse model on an agent-centric subset $\mathbf{z}_{\mathcal{S}}$, where action recovery can remain stable as the rest of the scene grows, yielding a practical forward–inverse gap that we exploit for self-improvement.

Condition 2: compositional OOS with preserved source-set (strong benefit). WAV succeeds when OOS novelty is concentrated in $\mathbf{z}_{\setminus\mathcal{S}}$ while the verifying subset behaves as it did in training. Concretely, even when the full pair $(\mathbf{z}^t, \mathbf{a}^t) \notin S_{\text{seed}}$, the subset transition on $\mathbf{z}_{\mathcal{S}}$ remains on-support, so an inverse model trained on $\mathcal{D}_{\text{act}}$ can still recover $\mathbf{a}^t$ reliably. *Example (tool–object contact).* Training contains "move knife in free space" and "touch apple with hand," but not "slice apple with knife." At test time, the full transition is OOS (contact dynamics are novel), yet the arm motion $\mathbf{z}_{\mathcal{S}}$ follows familiar trajectories. The inverse model can still recover the action, enabling the world model to learn the novel contact outcome.

Stochastic extension. For the remaining discussion we consider a stochastic generalization of the TLCM where $\mathbf{z}^{t+1} = g(\mathbf{z}^t, \mathbf{a}^t, \boldsymbol{\epsilon}^t)$ with exogenous noise $\boldsymbol{\epsilon}^t$; Theorem F.8 holds in the deterministic special case $\boldsymbol{\epsilon}^t = \mathbf{0}$.

Condition 3: weak injectivity / action aliasing (degradation). Condition F.7 (2) requires the mapping $\mathbf{a} \mapsto g_{\mathcal{S}}(\mathbf{z}^t_{\mathcal{S}}, \mathbf{a}, \boldsymbol{\epsilon})$ to be injective. When different actions produce indistinguishable (or nearly indistinguishable) subset transitions, recovered actions become ambiguous:

$$\exists\, \mathbf{a} \neq \mathbf{a}' \text{ s.t. } g_{\mathcal{S}}(\mathbf{z}^t_{\mathcal{S}}, \mathbf{a}, \boldsymbol{\epsilon}) \approx g_{\mathcal{S}}(\mathbf{z}^t_{\mathcal{S}}, \mathbf{a}', \boldsymbol{\epsilon}). \tag{16}$$

Example (underactuation / latency). In a soft gripper, different motor commands may produce nearly identical proprioceptive changes due to compliance. Similarly, unmodeled communication delays can smear the action's effect across timesteps, violating injectivity. In such cases, pseudo-labels drift and self-improvement degrades.

Condition 4: back-action from OOS into the verifier (failure). The verifying subset must remain insulated from OOS components (the *source set* property of Definition 3). WAV fails when OOS variables feed back into the verifier so that $\mathbf{z}_{\mathcal{S}}$ itself goes out-of-support. One way to view the failure mode is that the verifier dynamics no longer depend only on $(\mathbf{z}^t_{\mathcal{S}}, \mathbf{a}^t)$, but also on $\mathbf{z}^t_{\setminus\mathcal{S}}$:

$$\mathbf{z}^{t+1}_{\mathcal{S}} = g_{\mathcal{S}}(\mathbf{z}^t_{\mathcal{S}}, \mathbf{a}^t, \mathbf{z}^t_{\setminus\mathcal{S}}, \boldsymbol{\epsilon}^t). \tag{17}$$

Example (compliant contact). When a robot arm makes stiff contact with a deformable object, contact forces feed back into joint-level torques and velocities. The "verifying" proprioceptive dynamics now depend on the OOS object state, breaking the source-set assumption and causing action recovery to fail.

Summary. WAV is most effective when (i) the verifying subset is small and fixed-dimensional relative to the full state, (ii) OOS novelty is confined to non-verifying blocks while $\mathbf{z}_{\mathcal{S}}$ stays on-support, and (iii) the action imprint on $\mathbf{z}_{\mathcal{S}}$ is strong (high injectivity). It degrades under action aliasing and fails when OOS dynamics causally influence the verifier.

F.2 Detailed Derivation for Sample-Efficiency Advantage

This appendix provides the exact derivation behind Sec. 3.2. The purpose of the analysis is to isolate the statistical asymmetry exploited by WAV: predicting the full next state can be substantially harder than verifying the action from a low-dimensional action-relevant slice.

Setup. Let $s \in \mathbb{R}^{d_s}$ be the state, $a \in \mathbb{R}^{d_a}$ the action, and suppose the one-step dynamics are linear with additive Gaussian noise:

$$s' = As + Ba + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma_s^2 \mathbf{I}_{d_s}). \tag{18}$$

Assume also that there exists an action-relevant slice $z = Ms \in \mathbb{R}^{d_z}$, with $d_z \mathbf{l} d_s$, from which the action can be linearly recovered up to irreducible ambiguity:

$$a = H \begin{bmatrix} z \\ z' \end{bmatrix} + \eta, \qquad z' = Ms', \qquad \eta \sim \mathcal{N}(0, \sigma_a^2 \mathbf{I}_{d_a}). \tag{19}$$

We compare a dense forward regressor $\hat{f}$ trained on

$$x_F := \begin{bmatrix} s \\ a \end{bmatrix} \in \mathbb{R}^{d_s + d_a}$$

and a sparse inverse regressor $\hat{h}$ trained on

$$x_I := \begin{bmatrix} z \\ z' \end{bmatrix} \in \mathbb{R}^{2d_z}.$$

For analytic tractability, we assume both feature vectors have been whitened:

$$x_F \sim \mathcal{N}(0, \mathbf{I}_{d_s + d_a}), \qquad x_I \sim \mathcal{N}(0, \mathbf{I}_{2d_z}), \tag{20}$$

and both models are fit by ordinary least squares on n i.i.d. labeled transitions. (The whitening assumption simplifies the algebra; for general covariance Σ the excess risk scales with $\mathrm{tr}(\Sigma^{-1})$—see, *e.g.*, Hsu et al. (2014)—and the qualitative three-factor decomposition is preserved.)

As in the main text, we compare both models in the state space:

$$\mathcal{E}_F := \frac{1}{d_s} \mathbb{E}\Big[\|\hat{f}(s,a) - f^\star(s,a)\|_2^2 \Big], \tag{21}$$

$$\mathcal{E}_I := \frac{1}{d_s} \mathbb{E}\Big[\|f^\star(s, \hat{h}(z,z')) - f^\star(s, h(z,z'))\|_2^2 \Big]. \tag{22}$$

Lemma F.9 (OLS excess risk under isotropic Gaussian covariates)**.** *Consider scalar regression*

$$y = \langle w^\star, x \rangle + \epsilon, \qquad x \sim \mathcal{N}(0, \mathbf{I}_D), \qquad \epsilon \sim \mathcal{N}(0, \nu^2),$$

with $n > D + 1$. If $\hat{w}$ is the OLS estimator fit on n i.i.d. samples, then its expected excess risk is

$$\mathbb{E}\big[(\langle \hat{w} - w^\star, x \rangle)^2\big] = \nu^2 \, \frac{D}{n - D - 1}. \tag{23}$$

This is a classical exact expression for well-specified linear regression with isotropic Gaussian design; see, e.g., Hsu et al. (2014); Mourtada (2022).

Proposition F.10 (Exact forward–inverse gap in the linear–Gaussian model)**.** *Under the setup above, let $\lambda := \|B\|_{\mathrm{op}}$. If $n > d_s + d_a + 1$ and $n > 2d_z + 1$, then*

$$\mathbb{E}[\mathcal{E}_F] = \sigma_s^2 \, \frac{d_s + d_a}{n - (d_s + d_a) - 1}, \tag{24}$$

$$\mathbb{E}[\mathcal{E}_I] \le \lambda^2 \, \frac{d_a}{d_s} \, \sigma_a^2 \, \frac{2d_z}{n - 2d_z - 1}. \tag{25}$$

Consequently, the error ratio satisfies

$$\Gamma(n) := \frac{\mathbb{E}[\mathcal{E}_F]}{\mathbb{E}[\mathcal{E}_I]} \;\ge\; \Big(\frac{d_s + d_a}{2d_z} \cdot \frac{d_s}{d_a} \Big) \cdot \Big(\frac{\sigma_s}{\lambda \, \sigma_a} \Big)^2 \cdot \Big(\frac{n - 2d_z - 1}{n - (d_s + d_a) - 1} \Big). \tag{26}$$

Proof. We apply Theorem F.9 separately to the forward and inverse regressions.

Forward model. Each coordinate of s' in (18) is a scalar linear regression on the feature vector $x_F \in \mathbb{R}^{d_s+d_a}$ with noise variance σ_s^2. Therefore,

$$\mathbb{E}\Big[(\hat{f}_j(s,a) - f_j^\star(s,a))^2\Big] = \sigma_s^2 \, \frac{d_s + d_a}{n - (d_s + d_a) - 1}$$

for each state coordinate $j \in \{1, \ldots, d_s\}$. Averaging over the d_s coordinates yields

$$\mathbb{E}[\mathcal{E}_F] = \sigma_s^2 \, \frac{d_s + d_a}{n - (d_s + d_a) - 1},$$

which is exactly (24).

Inverse model in action space. Similarly, each coordinate of a in (19) is a scalar linear regression on $x_I \in \mathbb{R}^{2d_z}$ with noise variance σ_a^2. Hence

$$\mathbb{E}\Big[(\hat{h}_k(z,z') - h_k(z,z'))^2\Big] = \sigma_a^2 \, \frac{2d_z}{n - 2d_z - 1}$$

for each action coordinate $k \in \{1, \ldots, d_a\}$. Summing across the d_a coordinates gives

$$\mathbb{E}\Big[\|\hat{h}(z,z') - h(z,z')\|_2^2\Big] = d_a \, \sigma_a^2 \, \frac{2d_z}{n - 2d_z - 1}. \tag{27}$$

Mapping inverse error back to state space. Because the true dynamics $f^\star$ are linear in the action,

$$f^\star(s, \hat{h}(z,z')) - f^\star(s, h(z,z')) = B\big(\hat{h}(z,z') - h(z,z')\big).$$

Therefore,

$$\mathcal{E}_I = \frac{1}{d_s} \, \mathbb{E}\left[\left\| B\big(\hat{h}(z,z') - h(z,z')\big)\right\|_2^2\right] \tag{28}$$

$$\leq \frac{\|B\|_{\mathrm{op}}^2}{d_s} \, \mathbb{E}\Big[\|\hat{h}(z,z') - h(z,z')\|_2^2\Big] \tag{29}$$

$$= \lambda^2 \, \frac{d_a}{d_s} \, \sigma_a^2 \, \frac{2d_z}{n - 2d_z - 1}, \tag{30}$$

which proves (25) after taking expectations and using (27).

Finally, dividing (24) by (25) yields (26). □

Reading the bound. The ratio in (26) cleanly factorizes into three interpretable pieces. The first term is a *dimensionality advantage*: the forward model must estimate a map from $d_s + d_a$ inputs, whereas the sparse inverse model only uses $2d_z$ inputs. The second term is a *stochasticity advantage*: forward prediction suffers from environment noise σ_s, while inverse verification only suffers from the ambiguity of recovering the action from the selected slice, measured by σ_a, after accounting for the state-space gain λ. The third term is a *sample-size advantage*: when n is only modestly larger than the dense forward dimension, the forward estimator is statistically much less stable.

Scope of the stylized model. This analysis is intentionally minimal. It does not claim that real robotic dynamics are linear or globally Gaussian. Instead, it isolates a statistical regime that matches the intuition behind WAV: if a low-dimensional subset preserves the action imprint while the full scene is high-dimensional, noisy, and sparsely labeled, then sparse inverse verification can be substantially more reliable than dense forward prediction.

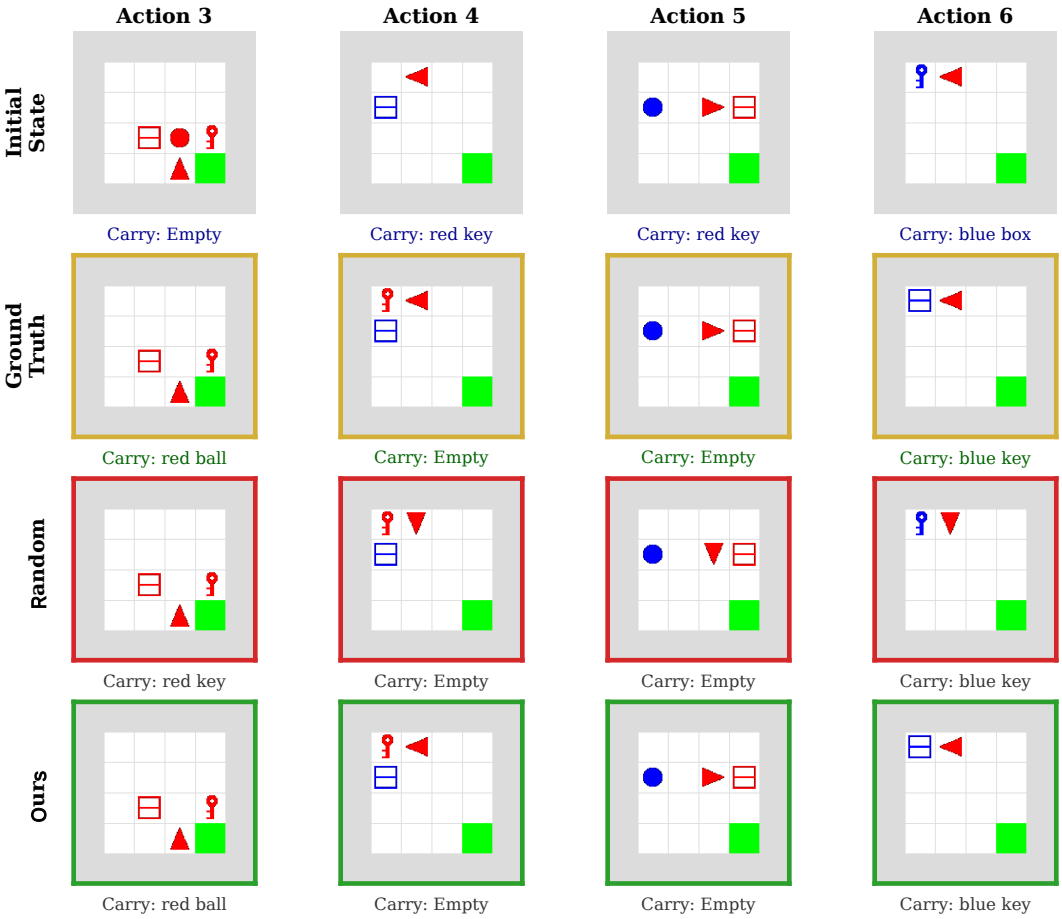

Figure 10: **Qualitative comparison of world-model rollouts under diverse interaction actions (Part I–II). Gold** borders denote ground-truth next observations; green and red borders indicate correct and incorrect predictions, respectively. Across both task sets, our method better preserves action-dependent state changes—notably for structured interactions such as *Toggle* and *Swap*—than exploration-based baselines.

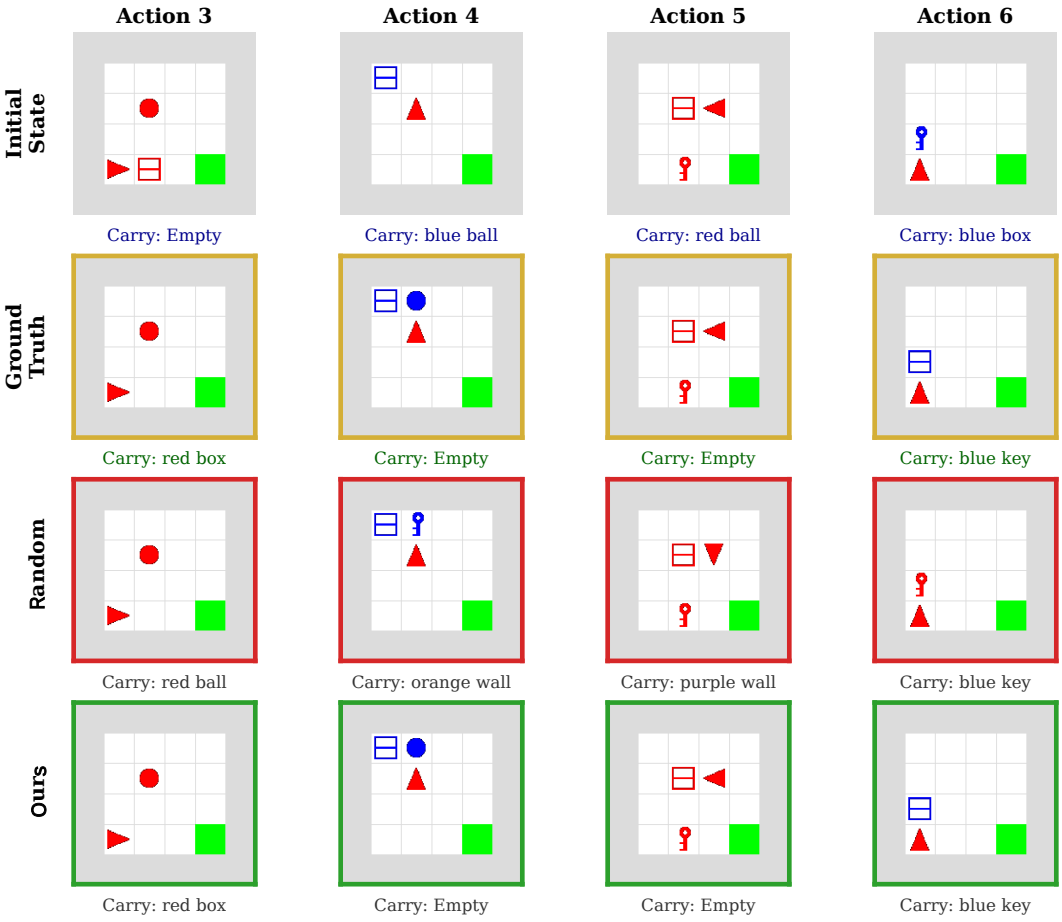

Figure 10: (Continued.) **Task Set B.** The **Random** baseline frequently collapses to predicting the most common primitive motions (e.g., *Turn*), failing to model interaction-induced state changes (e.g., *Toggle*). In contrast, models trained with data selected by our method capture these state transitions.

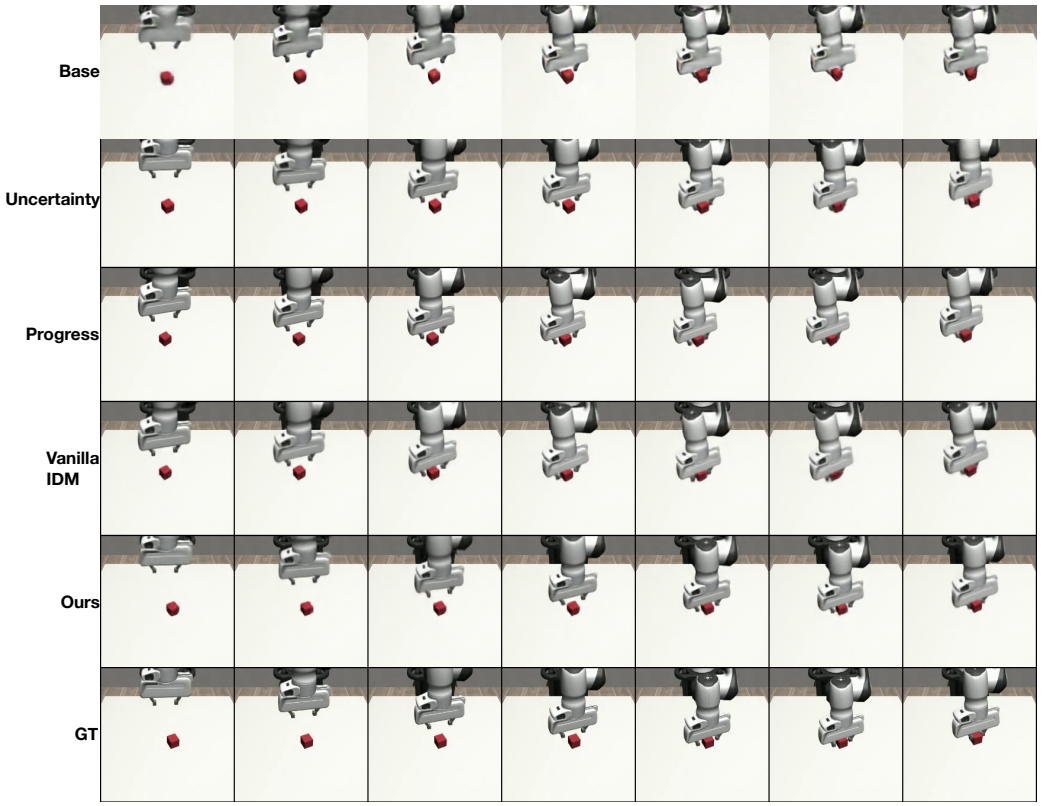

Figure 11: Qualitative comparison of world model predictions across different methods on Robomimic Lift.

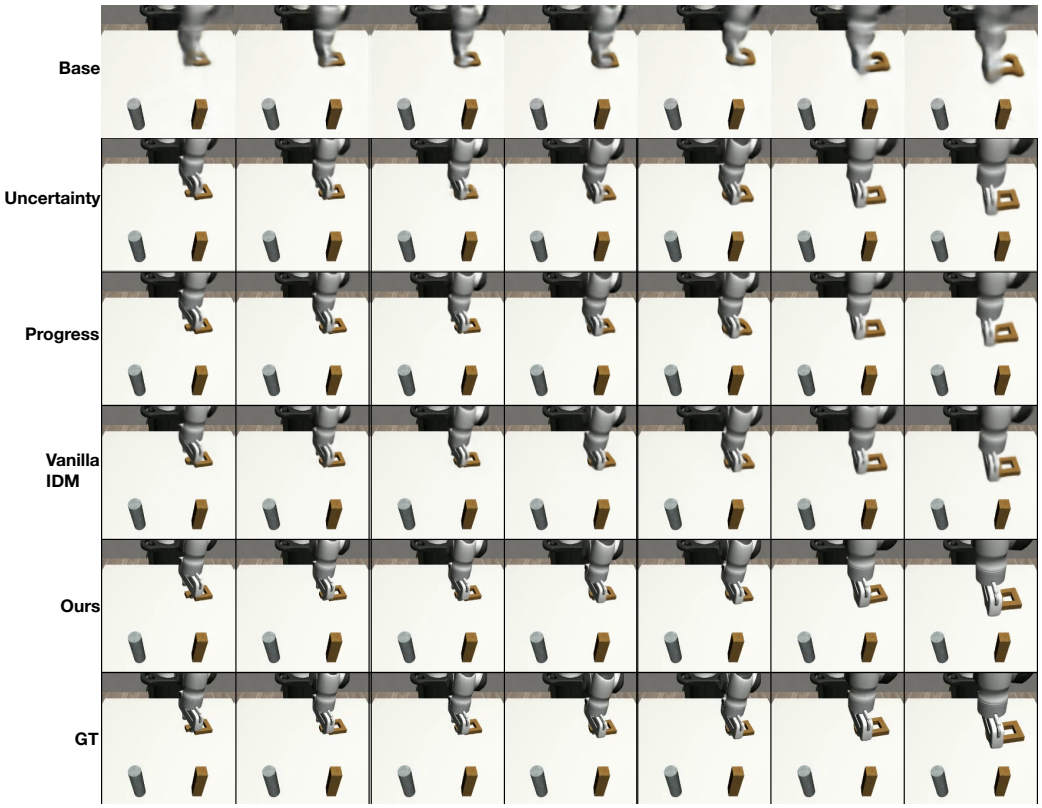

Figure 12: Qualitative comparison of world model predictions across different methods on Robomimic Square.