COMPETENCE WITHOUT PERFORMANCE: MODELS REPRESENT ARC-AGI-3 BUT CANNOT PLAY IT

Anonymous authors
Paper under double-blind review

ABSTRACT

ARC-AGI-3 tests whether a model can enter an unfamiliar interactive world, infer its hidden rules through play, and act successfully on what it learns. We build the first Gymnasium- and DreamerV3-compatible environment wrapper for ARC-AGI-3 and use it to establish the first model-based reinforcement-learning baseline on the benchmark with a learned latent world model. Run on six public games, the agent scores zero on five of them under the benchmark's metric, Relative Human Action Efficiency (RHAE). To find out why, we ask whether these zeros mean that the agent has learned nothing about the games, or that it has learned useful structure but cannot turn it into play. Because DreamerV3 trains its world model on real frames and its policy only on imagined trajectories, what the model has learned can be measured separately from what the policy does. **We find competence without performance.** The world model fits every game at the single-step level, and linear probes on its frozen latent state decode which level is being played, above raw-pixel, elapsed-time and label-permutation controls, in many games on which the agent never scores. Yet the policy stays at its uniform initialization on five of six games. Ablations that pretrain the world model across games and that replace the reinforcement learner with a supervised imitation model confirm that the gap lies in exploration and policy learning rather than in representation. Learning a world model and learning to act from it therefore fail separately on ARC-AGI-3, and the benchmark score alone cannot tell the two failures apart.

1 INTRODUCTION

ARC-AGI-3 is a benchmark of 25 public interactive abstract-reasoning games in which an agent enters an unfamiliar environment without being told its rules or goals, and must infer both through play (Foundation, 2026). Performance is measured by Relative Human Action Efficiency (RHAE), which rewards completing levels in fewer actions than human players and gives no credit for uncompleted levels. An agent that completes no level therefore scores zero, and the score says nothing about *why* it failed. A zero is compatible with at least two different failures: the agent may have learned no useful structure about the environment, or it may have learned such structure and be unable to convert it into successful play. Following the distinction drawn in developmental psychology (Firestone, 2020; Mitchell, 2026; Rane et al., 2025), we call the first a failure of *competence* and the second a failure of *performance*.

Recent work argues that games of this kind require an agent to build a model of each new environment as it plays. The ARC-AGI living survey (Vahdati et al., 2026) identifies learning such world models as an important next step across the ARC-AGI benchmark family, and Ying et al. (2025) argue that novel interactive games test the ability to construct task-specific models online from more abstract prior knowledge. Consistent with this view, Lee et al. (2024) show that DreamerV3 outperforms PPO on multi-step ARC-AGI-1 tasks. Yet it remains unclear whether a learned world model is enough to play ARC-AGI-3 and, when it is not, whether the agent fails to learn the environment or fails to act on what it has learned.

We study this question with DreamerV3 (Hafner et al., 2025), establishing a model-based reinforcement-learning baseline on six public ARC-AGI-3 games. DreamerV3 suits the question because its world model is trained on real frames, through a reconstruction loss and a dynamics loss, while its policy is trained only on trajectories that the model imagines, so what the model has learned can be measured separately from what the policy does.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

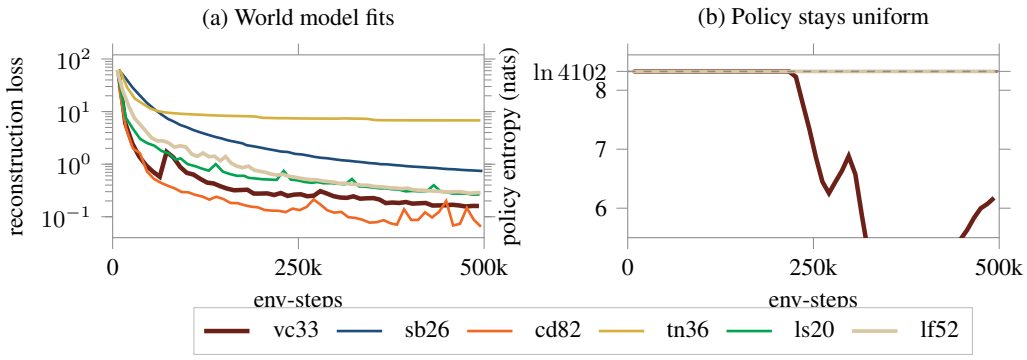


Figure 1: Reconstruction loss (a, log scale) and actor policy entropy (b) over 500k env-steps, warm arm, seed 0, one color per game. The dashed line in (b) is the uniform maximum $\ln 4102 = 8.32$ nats.

Figure 1 shows the main result. The world model learns to predict the frames well, but this competence does not lead to the right actions or to a better score. Across the six games, reconstruction loss falls sharply, yet five games remain at zero RHAE and their policy entropy stays at its maximum, the value of a uniformly random policy. Linear probes of the trained world model show that task-relevant information is nevertheless present in some of these failed runs: on `cd82` and `tn36`, the current level can be decoded linearly from the learned latent state above a label-permutation control, while probes on the raw pixels and on elapsed time do not pass the same test. The world model's multi-step predictions remain poor, so these probes do not show that it simulates the games faithfully. They show instead that task-relevant information can be present in the latent state without being used by the policy. Ablations support this reading (Section 3.3). Together, these results indicate that on ARC-AGI-3, learning task-relevant structure and turning that structure into successful play are distinct problems, and that the benchmark score alone cannot tell which one has failed.

Our contributions are summarized as follows:

- We build a Gymnasium- and DreamerV3-compatible environment wrapper for ARC-AGI-3 and use it to establish the first model-based reinforcement-learning baseline with a learned latent world model on the benchmark.
- We measure competence and performance separately in this setting: RHAE and policy entropy for what the agent does, and single-step fit, controlled linear probes and open-loop rollout accuracy for what the world model has learned.
- Across six public games we find competence without performance: the world model fits every game and encodes level identity in two games on which the policy never leaves its uniform initialization and never scores.

2 Measuring Competence and Performance

Our experiments are a set of measurements designed to separate two questions: whether an agent *learns useful structure* about a game, and, if it does, whether it *can convert that structure into play*. We run them on DreamerV3, which has previously been applied to ARC-AGI-1 (Lee et al., 2024) and whose world model and policy are trained separately (Appendix A gives the component table and a schematic). Section 2.1 defines how we measure performance and what the policy does, and Section 2.2 how we measure competence.

2.1 Measuring Performance

We measure performance with RHAE, the benchmark's metric, which scores level i by the squared ratio of the human baseline action count h_i to the agent's count n_i and weights levels by their index,

so that later levels count more and an uncompleted level earns nothing:

$$s_i \;=\; \min\Big(\big(h_i/n_i\big)^2,\; 1.15\Big), \qquad \text{RHAE} \;=\; \frac{\sum_{i\in\mathcal{C}} i\, s_i}{\sum_{i\in\mathcal{C}} i}, \tag{1}$$

where $s_i = 0$ for a level the agent does not complete and $\mathcal{C}$ is the set of levels with a human baseline. The agent trains on the environment's own reward, which is given only when a level is completed, so both the training signal and the score are zero until a level is completed. RHAE is therefore zero for every agent that completes no level, whatever that agent has learned, so we measure competence separately.

How the policy affects performance. Because RHAE cannot distinguish a policy that has learned nothing from one that has learned a little, we also record whether the policy has moved from its initialization. DreamerV3's actor starts as a uniform distribution over the 4102 actions, whose entropy is $\ln 4102 = 8.32$ nats. We report the actor's entropy, normalized by this maximum so that 1.000 means exactly uniform, together with the number of levels completed during training and during evaluation. A policy whose normalized entropy stays at 1.000 has received no usable learning signal, whatever the world model has learned.

2.2 Measuring Competence

We measure competence on the trained world model, with its weights fixed, through three tests of increasing strength, each of which addresses what the previous one leaves open.

Single-step fit of the world model. The weakest requirement is that the world model predicts the game one step at a time, which we read from the reconstruction and dynamics losses during training. When these losses converge while RHAE stays at zero, single-step prediction is not what limits the agent. A low loss nevertheless shows only that frames are predictable from the model state. It does not show that the state encodes anything the task depends on, since boards that change little between steps can be reconstructed well without task-level information.

Task structure in the latent state. We therefore ask whether two task variables that the agent is never given as input, which level is being played and whether a level has just been completed, can be predicted from the latent state by a linear classifier (a linear probe). A successful probe is evidence of competence only if simpler explanations fail, so each probe is compared with three controls. A label-permutation control, a probe trained on shuffled labels, bounds what the probe achieves by chance. A raw-pixel probe tests whether the variable is already visible in the frame, in which case the latent state adds nothing. An elapsed-time probe tests whether the probe merely tracks how long the episode has run, since later levels occur later. We call a variable decodable only when the lower bound of the probe's bootstrap confidence interval (CI) exceeds the upper bound of the permutation control's, and we attribute it to the latent state only when the pixel and elapsed-time probes do not pass the same test (Table 2).

Multi-step simulation accuracy. Representing a task variable is weaker than simulating the game, and simulation is what DreamerV3's policy learning relies on. We therefore let the model predict several steps ahead without seeing the real frames (an open-loop rollout), under the real action sequence of a human replay, and measure the fraction of grid cells whose color it predicts correctly. We compare this accuracy with a baseline that repeats the last observed frame (copy-last-frame), which a model must beat to show that it predicts how the board changes.

Together with RHAE and policy entropy (Section 2.1), these three tests form the columns of the per-game summary in Table 3.

3 Experiments

We report the experimental setup, the main results on competence and performance, and ablations that vary the initialization of the world model and the learner. We observe that a better-fitting world model does not lead to better task performance, and we explain why.

3.1 EXPERIMENTAL SETUP

Games and protocol. We evaluate six public ARC-AGI-3 games: `vc33`, `sb26`, `cd82`, `tn36`, `ls20`, and `lf52`. The first three form the pre-specified baseline pilot set, with a progression criterion of RHAE > 0 on at least two games; the remaining three extend the sweep. Each game is run under the two training arms described below, using two random seeds (s0, s1) and 500k environment steps per run.

Training configuration. All main-sweep experiments use DreamerV3 (`size12m`) with the default hyperparameter configuration and the full unmasked 4102-way action interface. The same model size, optimization settings, interaction budget, and action representation are used across games, arms, and seeds; no game-specific hyperparameter tuning is performed. Each game is run in two arms that differ only in how the world model is initialized. In the **warm** arm, the world model is initialized from a single model pretrained on all 340 public human replays across games, using the world-model losses only, without actor or critic; each per-game run then starts with a newly initialized actor and critic and with that game's human replays in its replay buffer. In the **cold** arm, the world model is not pretrained. The main results below report the warm arm at seed 0, whose world model fits best; the full 24-run sweep is given in Appendix B, and the paired comparison between the two arms is treated as an ablation (Section 3.3).

Evaluation. We report RHAE computed post hoc from evaluation episodes collected at 500k environment steps. RHAE $= 0$ corresponds to clearing no level, while RHAE $= 1.0$ denotes human parity by construction. Per-cell evaluation episode counts are reported in Table 1. To calibrate the random-policy floor on the only game with nonzero trained performance, we additionally evaluate a uniformly random policy on `vc33` over 1,132 episodes; it scores RHAE 0.

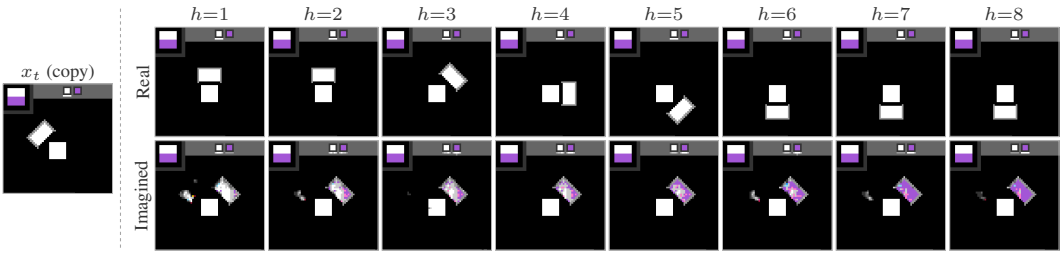
3.2 MAIN RESULTS

Figure 1 shows that the reconstruction loss of every game falls sharply, while the entropy of most actors stays at its uniform maximum. The world model can thus fit the environment well and, as we show below, encode task-relevant structure, while the agent as a whole still performs poorly. Table 1 gives the numbers at 500k steps. Image-reconstruction loss falls by $\geq 98\%$ on five of six games (`sb26` is the narrowest at 98.8%), and the dynamics KL converges to a $\sim$1.0–1.6 floor everywhere. However, only `vc33` ever departs from the uniform policy or clears a level. To understand this gap, we conduct a series of diagnostic analyses that separately examine policy behavior, game-specific constraints, latent task structure, and multi-step simulation.

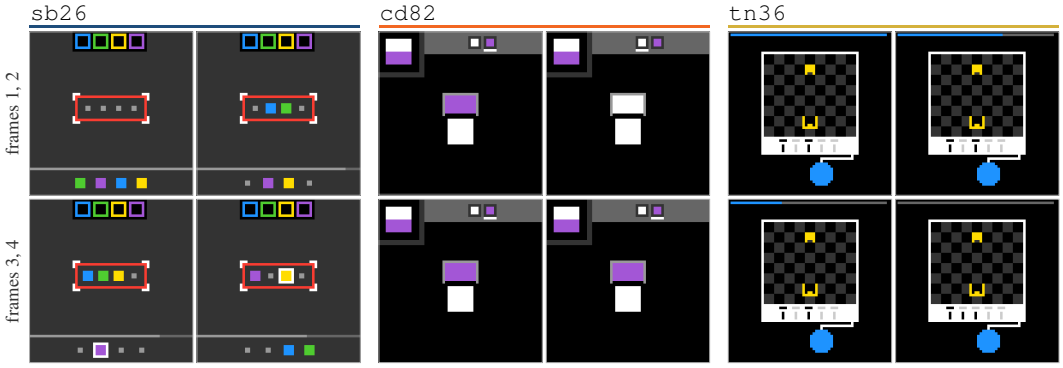
Table 1: World-model fit and policy activity at 500k env-steps, seed-0 warm cells. Recon. loss is the image-reconstruction loss at the first and last logged training window; Dyn. KL is the dynamics prior loss at the last window; Norm. entropy is the actor's entropy divided by the uniform maximum $\ln 4102$, so 1.000 is exactly uniform; clears are level completions out of episodes.

	World model fits		Policy acts			
Game	Recon. loss (first→last)	Dyn. KL (last)	Norm. entropy	Train clears	Eval clears	RHAE
vc33	$58.3 \rightarrow 0.16$	1.24	0.754	17/9751	4/23	**0.0548**
sb26	$62.8 \rightarrow 0.74$	1.55	1.000	0/502	0/24	0.0000
cd82	$31.0 \rightarrow \mathbf{0.06}$	1.03	1.000	0/4950	0/24	0.0000
tn36	$63.8 \rightarrow 6.80$	1.01	1.000	0/8053	0/16	0.0000
ls20	$44.5 \rightarrow 0.27$	1.01	1.000	0/502	0/24	0.0000
lf52	$62.0 \rightarrow 0.29$	1.01	1.000	0/7399	0/28	0.0000

Only the successful policy leaves the uniform distribution. In Table 1, `cd82`'s world model fits *tighter* than `vc33`'s (image loss 0.065 versus 0.160), yet `vc33` is the only game on which the policy ever clears a level during the 500k-step run. On all five zero-RHAE games, policy entropy remains at the uniform maximum $\ln(4102) = 8.32$ nats at 500k, so the action distribution is still effectively



(a) Open-loop imagination on cd82 from the same state and actions



(b) The trained policy acting

Figure 2: **Imagination and behavior.** **(a)** Open-loop rollout on cd82. The top row is the real board over eight steps of a human replay. The bottom row is the RSSM decode under the same action sequence. The left frame is the copy-last-frame reference. The imagined board settles on one configuration at h=1 and holds it while the real board keeps moving. The window is the median of ground-truth change, not the busiest one. **(b)** Four evenly spaced frames from one trained-model episode per game, seed-0 warm, read left to right and then top to bottom, one color per game as in Figure 1. The board changes in every game, but no game clears a level, so each episode scores RHAE 0. ls20 and lf52 are not shown; Figure 4 dissects their failure mechanisms.

uniform. Only vc33 develops a non-uniform policy, and even there the change is modest: at 500k its entropy is still three quarters of the uniform maximum.

Under the sparse reward $r = \Delta$levels, a run that never clears a level receives no positive reward from which to learn useful action preferences. On the five zero-RHAE games, the immediate bottleneck is therefore reward discovery and exploration rather than exploitation of an already learned policy. vc33 is different: its policy does become non-uniform, but it remains far from committing to a small set of actions.

This dissociation persists across the full sweep (6 games $\times$ 2 arms $\times$ 2 seeds). Image-reconstruction loss falls by $\geq 98\%$ on five of six games; tn36's warm world model is the exception, plateauing at $\sim$6.8 (a $\sim$89% reduction; Appendix B). Yet every game except vc33 clears zero evaluation episodes in every cell. Figure 2 shows the learned policies acting on sb26, cd82, and tn36 without reaching a goal.

Game-specific failures expose distinct policy-side bottlenecks. The agent fails for two distinct reasons on the two games whose boards barely change. **(1)** On games like ls20, the engine offers only the four directional actions, so the unmasked 4102-way interface dilutes every state-changing action $\approx$1025:1 and a near-uniform policy changes the board on only $\approx 0.1\%$ of steps. This is an action-space-dilution confound rather than a world-model failure. **(2)** On games like lf52, every agent episode terminates in a deterministic GAME_OVER at exactly 64 actions, which is the engine's hard level-1 action budget under a policy that never clears (later tiers are 64×5 and 64×10). Because 64 is far below the truncation limit of 1000, these are terminations rather than truncations. The budget is weighted by action type (directional and click moves cost 1, undo costs 20, no-ops cost 0), which desynchronizes the env-step count from the budget count and yields a smooth train-episode

tail from 66 to ≈236 steps with no clustering at multiples of 64 and no level completions. Human episodes on `lf52` take a median of 77 and up to 1604 actions (n=45 across 11 session replays), reach `WIN` four times, abandon the remaining 41 voluntarily, and never trigger `GAME_OVER`. The 64-action ending is therefore a structural property of the game under a non-clearing policy, a level-1 exploration window ≈15× shorter than the 1000-step truncation of `sb26` and `ls20`, and not a horizon of human play. Appendix C gives the board images, the action-budget diagram and the episode-length distributions supporting both diagnoses (Figure 4).

Table 2: Frozen-model probes on the seed-0 warm checkpoints. Level-identity decoding is balanced accuracy from the RSSM latent, from a raw-pixel probe and from an elapsed-time clock; chance is $1/(\#\text{levels})$; † marks decoding above a label-permutation control by CI separation. Rollout columns give per-cell accuracy at horizon H=8 for the open-loop model and for copy-last-frame on human replays.

			Level-identity decode (bal. acc)			Rollout @ H=8	
Game	RHAE	chance	latent	pixel	clock	model	copy
Genuine level structure (latent above control; pixel and clock are not)							
`cd82`	0.000	0.17	**0.43**†	0.37	0.33	0.88	0.92
`tn36`	0.000	0.14	**0.43**†	0.25	0.33	0.42	0.95
Partly visible (latent and the raw-pixel baseline both decode)							
`sb26`	0.000	0.12	0.83†	0.77†	0.32	0.82	0.94
No decodable level structure (latent at control)							
`vc33`	**0.055**	0.14	0.37	0.24	0.25	0.47	0.87
`lf52`	0.000	0.10	0.35	0.27	0.31	0.53	0.86
`ls20`	0.000	0.14	0.23	0.28	0.42	0.58	0.93

Task-relevant information is decodable from the latent state. Table 2 tests whether the low reconstruction loss reflects task-relevant structure rather than visual similarity alone. On `cd82` and `tn36`, level identity can be decoded from the frozen latent state, while the corresponding raw-pixel and elapsed-time probes do not pass the permutation control. In contrast, `sb26` shows strong decoding from both the latent state (0.83) and the raw pixels (0.77), suggesting that much of its level information is already visible in the observation. Level-transition events are also decodable above the permutation control on `cd82`, `sb26`, `vc33`, and `lf52`. These results show that, for some games, the latent representation contains information about level identity beyond what can be decoded from the current image, and that this information is present in games on which the policy never scores.

Multi-step rollouts remain inaccurate. Decodable task information does not imply accurate simulation. At horizon $H = 8$, the world-model rollouts never outperform the copy-last-frame baseline, with the gap ranging from 4.7 percentage points on `cd82` to 53 points on `tn36`. Figure 2 shows the `cd82` case frame by frame. On random-action trajectories, where the board changes little, both the model and the copy baseline reach ≈0.99 accuracy, so rollout errors become pronounced only when the board changes substantially. The contrast is clearest on `tn36`: level identity is decodable from the latent state, yet its rollout fidelity is the worst among the six games. Representing task-relevant information and simulating the environment accurately are therefore separate capabilities, and the competence we find is of the first kind.

3.3 Ablation Study

The main results are observational. We add two ablations for which the two failures predict different outcomes, and point to a third that is reported in the appendix.

Cross-game world-model pretraining. If the world model were what limits the agent, starting from a better world model should improve performance. The warm arm of Section 3 tests this: its world model is pretrained on all 340 public human replays, and its image-reconstruction loss at the first logged training window is 20–30× lower than the cold arm's, which confirms that the

pretrained model is loaded and fits better. Table 6 reports the paired RHAE. Only `vc33` is ever non-zero, and there the per-seed sign of Δ disagrees (s0 favors warm by 0.014, s1 favors cold by 0.002). With only two seeds we make no variance estimate; what we can say is that the sign of the effect is not consistent across them, so the premise that a better initial world model improves play is *not supported*. Both arms sit two orders of magnitude below human parity and at the random floor (RHAE exactly 0.0000) on every game but `vc33`, whose departure from random is real and measured against the random-policy evaluation above.

Supervised imitation with Tiny Recursive Model. If exploration and policy learning are what limit the agent, a learner that needs neither should score on the same games, through the same interface and under the same metric. We test this with a supervised arm, Tiny Recursive Model (TRM) (Jolicoeur-Martineau, 2025) trained per game on the human replays alone to predict the next 32 human actions as a plan, which we call plan cloning. TRM was the basis of the winning system in ARC-AGI-2, NVARC, so this arm also shows how a strong architecture performs on ARC-AGI-3 when it learns from human play instead of from reward. The prediction holds (Table 7): from scratch, the supervised arm scores on `tn36` and `lf52`, games on which no RL run in this paper ever scores, and its combined RHAE more than triples the DreamerV3 cold arm (0.0184 versus 0.0049). With exploration removed, a learner scores exactly where the diagnosis places the RL failure. Cross-game pretraining also transfers for this supervised arm in a way it never did for the RL arm: a plan-cloning pretrain pooled over all 25 public games moves `ls20` and `cd82` off the cold arm's zero floor, at some cost to `lf52`, for a modest aggregate gain (0.0204 versus 0.0184); only `sb26` remains zero in every supervised cell. The supervised arm is still far from human play, with zero RHAE on several games, and because it changes the model and the training method together, it locates the failure without isolating a single cause. A TRM trained as a world model, the one-step-lookahead controllers built on it, and the interaction of transfer with the `ls20` action mask are reported in Appendix E.

Factored action representation. Within DreamerV3 itself, representing an action as a type plus click coordinates, with the environment and reward unchanged, lets the agent find reward and leave the uniform policy on `cd82` and `sp80`, although it does not broadly improve final RHAE. This third ablation, run across all 25 public games, is reported in Appendix F.

4 Related Work

Agents on ARC-AGI-3. ARC-AGI-3 was released in March 2026 with 25 public interactive games, with humans solving every environment and leading AI systems scoring far below them (Foundation, 2026), continuing the program of measuring skill acquisition rather than skill (Chollet, 2019). The living survey (Vahdati et al., 2026) covers the approaches reported across the three benchmark generations, program synthesis, neuro-symbolic and neural alike, documents the sharp drop in scores on ARC-AGI-3, and finds no world-model approach on it. Lee et al. (2024) give the one direct precedent for model-based RL on this task family, DreamerV3 beating PPO on multi-step ARC-AGI-1, and Ying et al. (2025) argue that ARC-AGI-3-style games are the testbed for adaptive world models built during play from abstract prior knowledge. We ask whether a latent world model learned from pixels and human replays alone can act on the benchmark, and where it fails if it cannot.

World models and their evaluation. DreamerV3 (Hafner et al., 2025) is the world-model agent that this work uses unmodified, and Plan2Explore (Sekar et al., 2020) extends the Dreamer family with exploration that does not depend on environment reward, which makes it the natural next experiment for the failure we find. The supervised ablation follows Jolicoeur-Martineau (2025). The distinction between competence and performance comes from developmental psychology, where a missing behavior was long read as a missing ability until later work traced it to factors that affect only performance (Firestone, 2020; Mitchell, 2026; Rane et al., 2025); probing a trained representation for task variables, against controls, is the standard test for competence, which we follow.

5 DISCUSSION

We interpret the results in terms of competence and performance and state what they do and do not establish.

Competence and performance. *The central claim of this paper is a dissociation between* competence *and* performance *(Firestone, 2020; Mitchell, 2026; Rane et al., 2025) on ARC-AGI-3*: a system can possess the internal capacity a task requires and still fail the task, because the failure lies in the process that turns that capacity into action. ARC-AGI-3 was built to test whether a system can enter an unfamiliar interactive world, infer the rules through play, and act on them (Vahdati et al., 2026; Ying et al., 2025), and it scores agents only on the levels they complete. Such a score cannot, on its own, tell apart an agent that never built a model of the world from one that built a useful model but cannot use it, and the two call for opposite remedies, better representation learning in one case and better exploration or policy learning in the other.

Table 3: Competence by test, policy behavior, and performance, seed-0 warm. Reconstruction: image loss falls by $\geq 98\%$ (Table 1). Level identity: latent decodes above pixel, clock and permutation control (Table 2); *pixel also decodes. Level transition: decodes above permutation control. Simulation: rollout beats copy-last-frame at $H{=}8$. Non-uniform: policy entropy below its uniform maximum at 500k (normalized entropy below 1 in Table 1).

	Competence				Policy	Performance
Game	Reconstruction	Level identity	Level transition	Simulation	Non-uniform	RHAE
vc33	✓	×	✓	×	✓	0.055
sb26	✓	✓*	✓	×	×	0
cd82	✓	✓	✓	×	×	0
tn36	×	✓	×	×	×	0
ls20	✓	×	×	×	×	0
lf52	✓	×	✓	×	×	0

Where the DreamerV3 agent fails. Table 3 lists the outcome of each competence test per game. The agent's reconstruction and dynamics losses converge on these environments, yet on five of six games its policy never leaves the uniform distribution it was initialized with. Three comparisons place the failure on the policy side. First, image-reconstruction loss falls by $\approx$99.7–99.8% on both vc33 and cd82, and the world model fits cd82 *more tightly* than vc33, the one game on which the agent ever clears a level, yet the cd82 policy stays uniform; a failure that is worse where single-step prediction is more accurate cannot be a failure of single-step prediction. Second, the warm arm's 20–30$\times$ lower initial reconstruction loss ends at the same RHAE 0 as the cold arm on five games (Section 3.3). Third, tn36's warm model never brings its image loss below 1, yet it too scores RHAE 0, so a poorly fitting and a nearly perfect world model yield the same zero performance. *The model represents the single-step structure of these environments well and, on two games, their level structure (Table 2), but the agent cannot act in them.* The supervised ablation shows the same dissociation from the other side: a model with no world model at all, imitating human plans, achieves the best score in this paper, while the agent whose model fits these environments most tightly scores zero. Measuring competence and performance separately is what lets us place the failure after representation, in exploration and policy learning. This placement, more than the RHAE values, is our main result.

Two policy-side failures. The failure on the policy side has two distinct causes. On ls20, the unmasked interface leaves the four state-changing actions outnumbered $\approx$1025:1, which is a property of the action *interface*. On sb26, cd82, tn36 and lf52, clicks do change the board, and the failure is instead *exploration under sparse reward*: a uniform policy that never clears a level receives no reward and therefore never changes. Both causes come after representation, but they call for different remedies, masking the interface in one case and encouraging exploration in the other. The supervised arm's mask test (Appendix E) suggests that the dilution on ls20, although real as arithmetic, is not what limits that arm, since masking alone leaves it at zero while cross-game pretraining clears the game unmasked. For the RL agent, games in which no level is ever cleared

give no positive reward under $r = \Delta\text{levels}$, so the matching test is exploration that does not depend on environment reward, for which Plan2Explore (Sekar et al., 2020) is the Dreamer-family baseline.

Consequences. This is the first model-based RL result on ARC-AGI-3 with a learned latent world model, and it adds detail to the $\sim 3\times$ drop that the survey reports (Vahdati et al., 2026): a model that fits these environments still cannot act in them with an unmodified policy and a realistic number of environment steps, which is the "static representation vs. on-the-fly task model" gap of Ying et al. (2025). It also qualifies the Lee et al. (2024) precedent: DreamerV3's ARC-AGI-1 success does *not* carry over to this interactive setting, where the rules must be inferred. More generally, a score based only on completed levels is necessary but not sufficient to characterize an interactive agent, since the same zero can mean that the agent has no world model or that it has one and does not use it. Interactive benchmarks would benefit from reporting a measure of competence alongside performance.

Limitations. We study unmodified DreamerV3 at one model size and one training budget, with and without one form of pretraining, on six of the 25 public games at two seeds. Our claims concern that configuration and do not imply that model-based RL cannot solve ARC-AGI-3. With two seeds we report no variance estimate, the number of evaluation episodes per run is small, and RHAE human baselines cover 70.5% of levels. The wrapper provides a per-game action mask, but we do not apply it, which confounds the `ls20` result. Open-loop rollouts never beat copy-last-frame, so the competence claim covers single-step fit and linear decodability, not simulation. The supervised ablation changes model and training method together, so it locates the failure without isolating one cause.

6 Conclusion

We built the first Gymnasium- and DreamerV3-compatible environment wrapper for ARC-AGI-3, ran unmodified DreamerV3 on six public games, and measured what its world model learned separately from what its policy did. The agent learns more about the games than it uses: reconstruction and dynamics losses converge and level identity is decodable from the latent state in games on which the agent never scores, while the policy stays uniform on five of six games. Ablations that pretrain the world model across games and that replace the reinforcement learner with a supervised imitation model place the failure in exploration and policy learning rather than in representation. On ARC-AGI-3, competence and performance come apart, and the benchmark score alone cannot tell which one is missing. We release the wrapper, replay loader and RHAE tool for work on both problems.

AI use statement

In this work, we used generative AI tools for code development, analysis scripts and language editing of the manuscript. We did not use generative AI tools to design the experiments, generate results, or write the claims. All numbers were produced by the authors' scripts from logged data, and the authors reviewed and take responsibility for all content.

Ethics statement

This work trains and evaluates agents on a public benchmark of abstract games and uses the benchmark's publicly released human replay corpus. It involves no human-subject data collection and no personally identifiable information, and we foresee no direct ethical concerns beyond those general to reinforcement-learning research.

Reproducibility statement

The wrapper, replay loader, RHAE tool, probe code and launch scripts are released as an anonymized repository with the submission. Every run is seeded and named by configuration, game and seed, with the exact configuration saved next to its log directory, and the human-baseline file and the per-run evaluation logs behind every table are included. Section 2 and Section 3 state the training budget, the arms, the metric and the evaluation protocol.

REFERENCES

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020.

ARC Foundation. Arc-agi-3: A new challenge for frontier agentic intelligence. *arXiv preprint arXiv:2603.24621*, 2026.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Alexia Jolicoeur-Martineau. Less is more: Recursive reasoning with tiny networks. *arXiv preprint arXiv:2510.04871*, 2025.

Jihwan Lee, Woochang Sim, Sejin Kim, and Sundong Kim. Enhancing analogical reasoning in the abstraction and reasoning corpus via model-based rl. *arXiv preprint arXiv:2408.14855*, 2024.

Melanie Mitchell. On evaluating cognitive capabilities in machines (and other "alien" intelligences). *AI: A Guide for Thinking Humans*, January 2026. URL `https://aiguide.substack.com/p/on-evaluating-cognitive-capabilities`. Published January 14, 2026.

Sunayana Rane, Cyrus F Kirkman, Graham Todd, Amanda Royka, Ryan MC Law, Erica Cartmill, and Jacob Gates Foster. Position: Principles of animal cognition to improve llm evaluations. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

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.

Sahar Vahdati, Andrei Aioanei, Haridhra Suresh, and Jens Lehmann. The arc of progress towards agi: A living survey of abstraction and reasoning. *arXiv preprint arXiv:2603.13372*, 2026.

Lance Ying, Katherine M Collins, Prafull Sharma, Cedric Colas, Kaiya Ivy Zhao, Adrian Weller, Zenna Tavares, Phillip Isola, Samuel J Gershman, Jacob D Andreas, et al. Assessing adaptive world models in machines with novel games. *arXiv preprint arXiv:2507.12821*, 2025.

A WORLD MODEL PRELIMINARIES

A *world model* is a learned simulator of an environment's dynamics: given a compact summary of the current situation and a candidate action, it predicts the next situation together with the observation, reward, and episode-continuation signal that would follow (Hafner et al., 2025). An agent so equipped can *imagine* rollouts entirely in this learned latent space and improve its policy against them, decoupling the expensive environment-interaction budget from the cheap policy-improvement budget. This is exactly the capability ARC-AGI-3 stresses: the rules are unstated and must be inferred through active play, and a world model is the natural prior from which to build a task-specific model on the fly (Ying et al., 2025). DreamerV3 realizes this with a Recurrent State-Space Model (RSSM): at each step t the agent holds a *model state* $s_t = (h_t, z_t)$, trains an encoder, decoder, and reward/continue heads against real frames (the *reconstruction loss*), and forces a dynamics prediction $\hat{z}_t$ to match the encoder's latent z_t (the *dynamics KL loss*); the actor $\pi_\theta(a_t \mid s_t)$ and critic are trained *only* on rollouts the dynamics head imagines forward from s_t, so the environment is never touched during policy improvement. The full component-by-component definition is given in Table 4, and a schematic of the two training phases is given in Figure 3. This separation is what makes our negative result diagnostic: the reconstruction and dynamics losses measure whether the world model *fits*, whether $\hat{x}_t, \hat{r}_t, \hat{c}_t$ track reality, while RHAE measures whether the *policy*, trained on imagined rollouts, can act. When the former converge to their floors while the latter stays ≈ 0, the bottleneck sits on the policy side: on most games the actor never explores enough to generate the imagined rollouts that would improve it.

At each step t the agent holds a *model state* $s_t = (h_t, z_t)$ and learns the components in Table 4, all sharing parameters ϕ. Figure 3 draws only the stochastic z_t, with the deterministic h_t residing inside the RSSM box.

Table 4: The DreamerV3 world-model components, all sharing parameters ϕ. The raw frame is $x_t \in \{0, \ldots, 255\}^{64\times64\times3}$; the latent z_t is a vector of 32×32 categorical variables.

Symbol	Component	Role in plain terms
h_t	Recurrent state	Deterministic memory; carries history forward
z_t	Stochastic latent	Compressed summary of the current frame x_t
$s_t = (h_t, z_t)$	Model state	Everything the agent plans on at step t
$h_t = f_\phi(h_{t-1}, z_{t-1}, a_{t-1})$	Recurrence	Update memory from the last state and action
$z_t \sim q_\phi(z_t \mid h_t, x_t)$	Encoder	Compress the observed frame into a latent
$\hat{z}_t \sim p_\phi(\hat{z}_t \mid h_t)$	Dynamics	Predict that latent *without* seeing x_t
$\hat{x}_t \sim p_\phi(\hat{x}_t \mid s_t)$	Decoder	Reconstruct the frame from s_t
$\hat{r}_t, \hat{c}_t \sim p_\phi(\cdot \mid s_t)$	Reward / continue	Predict reward and episode-continue flag
$a_t \sim \pi_\theta(a_t \mid s_t)$	Actor	Choose an action; trained only in imagination

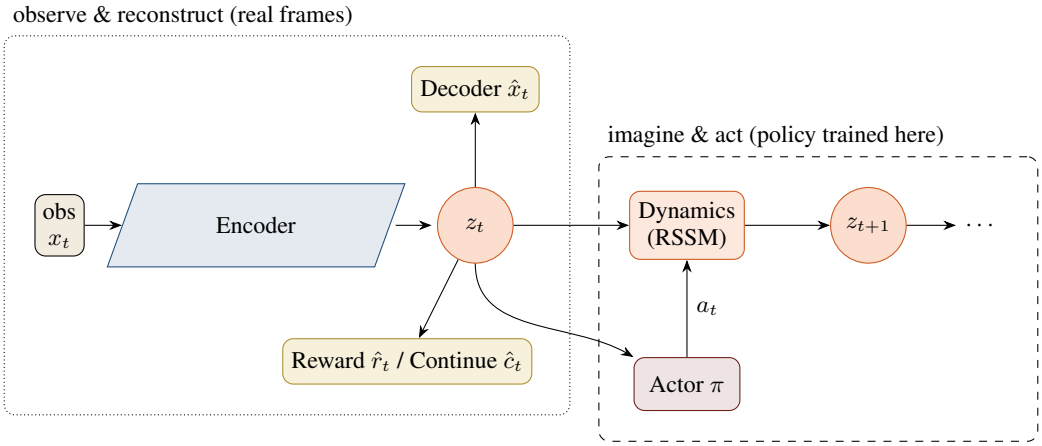


Figure 3: The DreamerV3 world model, in two phases. *Left (dotted):* a real frame x_t is encoded to a stochastic latent z_t, from which decoder, reward, and continue heads reconstruct the observation, the supervision that makes the model *fit the world*. *Right (dashed):* starting from z_t, the dynamics head rolls the latent forward under actions from the actor, and the policy is improved *only* on these imagined rollouts, the locus of our controller bottleneck.

B FULL SWEEP: ALL 24 RUNS

Table 5 lists every cell of the paired sweep behind Table 6 and Table 1.

Table 5: Full sweep, all 24 runs (6 games × {cold, warm} × 2 seeds). Final-step values: reconstruction loss (`train/loss/image`), dynamics KL (`train/loss/dyn`), policy entropy and normalized entropy (`train/ent`, `rand/action`), training- and eval-phase level-clears, and post-hoc RHAE. Training episode counts range 502–9,839 and eval counts 16–28 per run, so raw clear-counts are not comparable across rows; RHAE is the normalized metric. **Bold** marks the four cells that deviate from the otherwise-uniform off-vc33 pattern (see notes).

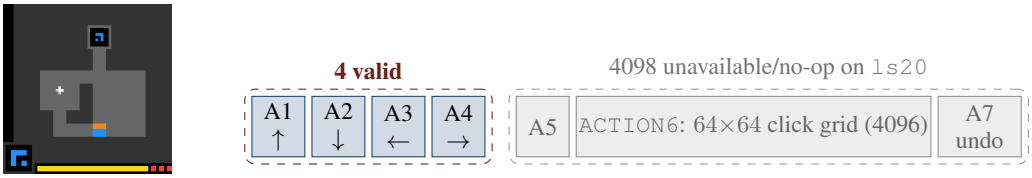
Game	Arm	Seed	Reconstruction loss (final)	Dynamics KL (final)	Policy entropy (nats)	Normalized entropy	Training clears	Eval clears	RHAE
vc33	cold	0	0.661	1.290	6.481	0.779	33	2	0.0411
vc33	cold	1	0.178	1.119	6.372	0.766	17	3	0.0182
vc33	warm	0	0.160	1.244	6.276	0.754	17	4	**0.0548**
vc33	warm	1	0.135	1.143	6.517	0.783	17	2	0.0166
sb26	cold	0	0.992	1.139	8.319	1.000	0	0	0.0000
sb26	cold	1	3.058	1.635	**8.282**	**0.995**	**1**	0	0.0000
sb26	warm	0	0.742	1.549	8.319	1.000	0	0	0.0000
sb26	warm	1	1.904	2.375	8.319	1.000	**1**	0	0.0000
cd82	cold	0	0.183	1.001	8.319	1.000	0	0	0.0000
cd82	cold	1	0.185	1.002	8.319	1.000	0	0	0.0000
cd82	warm	0	**0.065**	1.033	8.319	1.000	0	0	0.0000
cd82	warm	1	0.062	1.162	8.319	1.000	0	0	0.0000
tn36	cold	0	1.052	1.001	8.319	1.000	0	0	0.0000
tn36	cold	1	1.003	1.001	8.319	1.000	0	0	0.0000
tn36	warm	0	**6.797**	1.014	8.319	1.000	0	0	0.0000
tn36	warm	1	**7.280**	1.014	8.319	1.000	0	0	0.0000
ls20	cold	0	0.388	1.139	8.319	1.000	0	0	0.0000
ls20	cold	1	0.338	1.007	8.319	1.000	0	0	0.0000
ls20	warm	0	0.266	1.006	8.319	1.000	0	0	0.0000
ls20	warm	1	0.295	1.049	8.319	1.000	0	0	0.0000
lf52	cold	0	0.309	1.001	8.319	1.000	0	0	0.0000
lf52	cold	1	0.360	1.001	8.319	1.000	0	0	0.0000
lf52	warm	0	0.290	1.006	8.319	1.000	0	0	0.0000
lf52	warm	1	0.274	1.007	8.319	1.000	0	0	0.0000

Notes. Lower reconstruction loss and dynamics KL mean a tighter world-model fit; normalized entropy 1.000 means the policy is exactly uniform (entropy at the ceiling $\ln 4102 = 8.319$). **Bold** flags the four off-vc33 deviations: (a) sb26 cold-s1/warm-s1 each log one training clear under a still-uniform policy, yielding no eval clear (chance, not commitment); (b) tn36's warm world model alone fails to reach the <1 reconstruction floor (6.8–7.3), so its dissociation is weaker. Eval-clears and RHAE have no exceptions: only vc33 is non-zero.

C GAME-SPECIFIC FAILURE ANALYSIS

Figure 4 gives the board images, the action-interface diagram for `ls20`, and the agent and human episode-length distributions for `lf52` behind the two game-specific failures of Section 3.2.

(a) `ls20`: only 4 of 4102 action indices change the board



`ACTION1–4` are the four directional moves (the only state-changing actions on `ls20`); `ACTION5` is a parameter-less no-op and `ACTION7` is undo; `ACTION6`'s 4096 grid cells are not offered by the engine on this game at all (the click action appears in none of the 7,605 logged `ls20` frames), masked-unavailable, not merely inert. Hence $4/4102 = 0.098\%$ of indices are valid $\Rightarrow$ a uniform policy changes the board $\approx$1 step in 1025 (4102/4).

(b) `lf52`: every agent episode ends in deterministic `GAME_OVER` at exactly 64 actions; humans of comparable length are never forced out

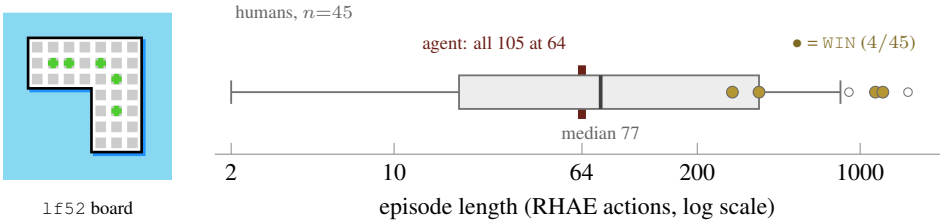


Figure 4: Error analysis of the two games whose boards barely change (cf. Figure 2). **(a)** `ls20`'s engine offers only the four directional actions; the 4096-cell click grid is never made available and the remaining types are no-ops, so a near-uniform policy changes the board on only $\approx 0.1\%$ of steps, an action-space-dilution confound, not a world-model failure. **(b)** `lf52` episode lengths on a common timeline, per episode and in RHAE action units. Every agent eval episode terminates at exactly 64 actions with zero variance (105/105; train episodes floor at 65 with a thin tail to $\approx$236, none approaching the 1000-step truncation budget, so these are terminations, not truncations: the engine enforces a hard 64-action level-1 budget and sets `GAME_OVER` when a non-clearing policy exhausts it, confirmed from the engine source). Human episodes (n=45 across 11 session replays; three 0-action reset artifacts dropped) take a comparable median of 77 actions but a far wider spread (max 1604), reach `WIN` four times, and *never* trigger `GAME_OVER`. The contrast is thus not episode length (those are comparable) but determinism and outcome: the agent exhausts the same fixed action budget at the same step every episode under a uniform policy (entropy $\ln 4102$, Table 1), gaining nothing from repetition, whereas humans vary, sometimes win, and are never forced out.

D DreamerV3 and TRM RHAE Score Table

E TRM world-model arms: full grids

These runs are post-hoc but protocol-matched: training recipe, episode-grouped split, unmasked interface, and 30-episode RHAE evaluation are identical to the Table 7 cold arm, and all values were recomputed from the raw per-episode logs. Table 8 reports the open-loop rollout fidelity of the discrete TRM world model against copy-last-frame. The *wm* controller scores each candidate action by one forward step of the frozen world model (change-mask, level-clear, and next-frame heads); *hybrid* rescales the cloned policy's action probabilities by the same one-step scores (Table 9). Table 10 gives the paired cold versus warm rollout fidelity.

Table 6: Paired RHAE at 500,000 env-steps for stock DreamerV3 `size12m` on 6 public games $\times$ 2 seeds (s0, s1), cold (from scratch) versus warm (cross-game WM-pretrained); Δ = warm$-$cold. $\bar{x}$ is the per-game mean over the two seeds within an arm, and Combined is the mean of the six per-game $\bar{x}$ values; per-level scores are capped at 1.15.

	Cold		Warm			
Game	s0	s1	s0	s1	Cold $\bar{x}$	Warm $\bar{x}$
vc33	0.0411	0.0182	0.0548	0.0166	0.0297	0.0357
sb26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
cd82	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
tn36	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ls20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
lf52	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Combined (mean of 6)					**0.0049**	**0.0059**
Random baseline			0.0000			
Human			1.0000			

Table 7: Paired RHAE for the supervised TRM plan-cloning arm, same protocol as Table 6 with 30 evaluation episodes per cell.

	Cold		Warm			
Game	s0	s1	s0	s1	Cold $\bar{x}$	Warm $\bar{x}$
vc33	0.0314	**0.0766**	0.0579	0.0152	**0.0540**	0.0366
sb26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
cd82	0.0000	0.0000	0.0645	0.0000	0.0000	0.0323
tn36	0.0357	0.0411	0.0229	0.0357	0.0384	0.0293
ls20	0.0000	0.0000	0.0035	0.0411	0.0000	0.0223
lf52	0.0133	0.0222	0.0018	0.0021	0.0178	0.0019
Combined (mean of 6)					**0.0184**	**0.0204**
Random baseline			0.0000			
Human			1.0000			

Table 8: Open-loop rollout fidelity of the discrete TRM world model at horizon 8 (per-cell accuracy on held-out episodes, s0 / s1) against copy-last-frame on the same windows. The TRM model exceeds copy in 11 of 12 cells; the RSSM exceeds it in none (Table 2; windows differ, so values are not directly comparable across tables).

	TRM-WM		Copy			
Game	s0	s1	s0	s1	TRM $\bar{x}$	Copy $\bar{x}$
vc33	0.9488	0.9866	0.8198	0.8936	0.9677	0.8567
sb26	0.9882	0.9711	0.9744	0.9633	0.9797	0.9689
cd82	0.9940	0.9733	0.9232	0.9199	0.9837	0.9216
tn36	0.9703	0.9358	0.9539	0.9229	0.9531	0.9384
ls20	0.9555	0.9648	0.9627	0.9528	0.9602	0.9578
lf52	0.9625	0.9501	0.8826	0.8247	0.9563	0.8537
Combined (mean of 6)					**0.9668**	**0.9162**

Table 9: RHAE of the world-model-based controllers, cold protocol of Table 7 (30 episodes per cell). Neither composition approaches the model-free plan-cloning arm (0.0184); hybrid rescoring degrades the cloned prior (`vc33` s1: $0.0766 \rightarrow 0.0005$).

	wm		hybrid			
Game	s0	s1	s0	s1	wm $\bar{x}$	hybrid $\bar{x}$
vc33	0.0166	0.0110	0.0548	0.0005	0.0138	0.0277
sb26	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
cd82	0.0000	0.0548	0.0000	0.0000	0.0274	0.0000
tn36	0.0000	0.0000	0.0006	0.0000	0.0000	0.0003
ls20	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
lf52	0.0000	0.0000	0.0008	0.0012	0.0000	0.0010
Combined (mean of 6)					**0.0069**	**0.0048**

Table 10: Paired open-loop rollout fidelity at horizon 8: the per-game TRM world model trained cold vs. warm-started from a 25-game cross-game pretrain under the identical recipe, probed on identical held-out windows within each seed. Warm-starting does not improve fidelity. † marks warm cells whose finetune early-stopped with frame exact-match pinned at 0, leaving a near-pretrain best checkpoint; their deficits are a checkpoint-selection artifact, not evidence of harmful transfer. Excluding them, the mean warm$-$cold difference is ≈ 0.000.

	Cold		Warm				
Game	s0	s1	s0	s1	Cold $\bar{x}$	Warm $\bar{x}$	Copy $\bar{x}$
vc33	0.9488	0.9866	0.9339	0.8793†	0.9677	0.9066	0.8567
sb26	0.9882	0.9711	0.9865	0.9153†	0.9797	0.9509	0.9689
cd82	0.9940	0.9733	0.9117†	0.9716	0.9837	0.9417	0.9216
tn36	0.9703	0.9358	0.9603	0.9566	0.9531	0.9585	0.9384
ls20	0.9555	0.9648	0.9518	0.9840	0.9602	0.9679	0.9578
lf52	0.9625	0.9501	0.9416	0.9666	0.9563	0.9541	0.8537
Combined (mean of 6)					**0.9668**	**0.9466**	**0.9162**

F CONTROLLER-SIDE EXTENSION: STRUCTURING THE INTERACTION SPACE

The paired cold versus warm study of Section 3.3 establishes that a large representational head start does not, by itself, produce control. We next ask a complementary question: whether the controller is being asked to learn through an unnecessarily hostile interface. This extension broadens the cold baseline to all 25 public games, then changes only the representation through which the actor expresses an action. The purpose is not to introduce task knowledge or a solver, but to test whether a stationary, compositional control grammar makes the sparse reward signal easier to discover and retain.

F.1 TWENTY-FIVE-GAME COLD BASELINE

We first ran stock, unforked DreamerV3 (`size12m`) from a cold start for 500k environment steps on every public game (seed 0, native level-clear reward, no warm-start and no replay pre-population). The 25 games divide by engine affordance into six button-only, six click-only, and thirteen mixed-control games. Only four agents scored non-zero RHAE, and all four were click-only games (Table 11). Every scorer cleared exactly one level; the other 21 games cleared none.

Table 11: Non-zero cells in the 25-game cold baseline. The remaining 21 games all score exactly zero and clear no level. Normalized entropy is the actor's entropy relative to its uniform floor; 1.000 is therefore exactly uniform, not an empirical fraction of random actions.

Game	Control class	RHAE	Levels	Norm. entropy
`ft09`	click	0.115	1	0.901
`vc33`	click	0.055	1	0.771
`lp85`	click	0.041	1	0.920
`r11l`	click	0.018	1	0.998
Other 21 games	button/click/mixed	0	0	1.000

The association with controller commitment is exact in this sweep: normalized policy entropy is below 1.000 on precisely the four non-zero games and is exactly 1.000 on all 21 zero-scoring games. Meanwhile, reconstruction loss falls sharply on every game, including all 21 failures. This larger sweep therefore reproduces the original competence–performance dissociation and sharpens its controller-side signature: performance appears if and only if the actor departs its uniform prior. The observation that all four scorers use a click interface also points to the spatial action parameterization as a useful intervention target, although the single-seed sweep alone cannot establish a causal class effect.

F.2 FROM 4102 UNRELATED LABELS TO A COMPOSITIONAL ACTION

In the flat interface used above, DreamerV3 chooses one categorical index from 4102 possibilities. Five indices encode parameter-less actions, $4096 = 64 \times 64$ indices separately encode every possible screen coordinate for `ACTION6`, and one index encodes undo. Thus "click at $(20, 30)$" and "click at $(20, 31)$" are statistically unrelated output labels, despite sharing an action type and neighboring coordinates. A controller must learn each spatial choice separately from an exceptionally sparse level-clear reward.

We replace this output with the stationary factorization

$$a_t = (\text{type}_t, x_t, y_t),$$

where the type head ranges only over action types available in the current game and the two coordinate heads each range over 64 positions. Every original action remains representable and the mapping is lossless; the native environment, reward, observation, evaluation protocol, DreamerV3 architecture, and 500k-step budget are unchanged. The same grammar applies across all 25 games, while each per-game agent is still trained independently: this is parameter sharing across coordinates within an agent, not transfer of actor weights between games. For a human analogy, the flat controller resembles a panel of 4102 unexplained buttons, whereas the factored controller exposes

an ordinary pointer with an action type and two coordinates. It specifies *how an action is expressed*, never which action solves the game.

We compare three arms:

- **A0 (flat control):** the cold 4102-way baseline.
- **B1 (factored control):** the cold (type, x, y) representation, with no exploration hint.
- **B2 (factored + candidate prior):** B1 plus a training-only, pixel-derived exploration prior. With probability ϵ, the behavior action is redrawn from plausible interactive cells; ϵ decreases linearly from 0.3 to zero over the first 200k steps and is identically zero thereafter. The prior is absent from evaluation, does not shape reward, does not clone human behavior, and never makes an action unreachable.

All factored-control runs are cold because the pretrained world-model checkpoint's action-conditioned dynamics input is structurally tied to the 4102-dimensional flat action. This choice is also empirically conservative: the paired experiment above found no consistent warm-start benefit. The A0 comparison remains matched within the same cold, 500k-step protocol.

F.3 Pre-registered pilot: factorization ignites cd82, the prior does not

The first controller-side test used {vc33, cd82, sb26}, two seeds per arm, and a gate frozen before observing any B1 or B2 GPU result. The primary hypothesis required B2 to exceed A0 in median RHAE on at least two of three games, with a strict majority of B2 seeds clearing at least one level in every qualifying game. Table 12 reports the complete endpoint.

Table 12: Pre-registered controller pilot, two seeds and 500k steps per cell. Parentheses show the median. The primary B2 gate fails: only vc33 qualifies, versus two games required.

Game	A0 flat	B1 factored	B2 factored + prior
vc33	.0411, .0182 (.0297)	.0248, .0548 (.0398)	.0140, .0548 (.0344)
cd82	0, 0 (0)	.0548, .0548 (.0548)	.0548, 0 (.0274)
sb26	0, 0 (0)	0, 0 (0)	0, 0 (0)

The primary gate fails $1/3$. The candidate prior is not supported: B2 has a lower median than B1 on both vc33 and cd82, and changes nothing on sb26. The strongest secondary signal is instead the stationary representation alone. B1 makes cd82 non-zero in 2/2 seeds where A0 is zero in 2/2, while also modestly increasing the vc33 median. On sb26, all arms remain at the reward floor; factorization cannot help when the relevant successful trajectory is never encountered.

F.4 Balanced cd82 follow-up: pooled signal, inconclusive holdout

We followed the B1 cd82 signal with four additional B1 seeds and four matched A0 seeds, yielding a balanced six-versus-six comparison. Flat A0 succeeds in $0/6$ seeds; factored B1 succeeds in $4/6$, with median RHAE 0 versus 0.0548. A directional Fisher exact test on all twelve runs gives $p = 0.0303$.

This pooled number is not an independent confirmation: cd82 was selected after seeds 0–1 produced the promising B1 $2/2$ versus A0 $0/2$ contrast. On the four genuinely new seeds alone, B1 is $2/4$ versus A0 $0/4$ ($p = 0.2143$, one-sided). We therefore treat the balanced result as a strong pooled, single-game signal whose post-selection holdout remains inconclusive, not as a confirmed general effect.

The internal diagnostics nevertheless localize what changed. Across seeds, median final image loss is lower for A0 than B1 (0.1805 vs. 2.8213), so factorization does not buy control by improving visual reconstruction. A0's imagined maximum return remains exactly zero and every actor stays at its uniform floor. In contrast, all six B1 agents develop positive imagined return (median maximum 1.6399) and depart the matched factored-uniform entropy floor. Most importantly, all $6/6$ B1 agents clear level 1 at least once during training, including the two seeds that finish with zero evaluation clears. Factorization therefore reliably breaks the all-zero reward/controller cold start on this

game. Under the original periodic evaluation it did not always appear to consolidate the discovered behavior; the frozen-checkpoint audit below tests that interpretation directly.

F.5 Stratified generalization: ignition transfers, endpoint performance does not

To test whether the cd82 mechanism generalizes, we froze a second experiment before observing its new GPU results: one button-only game (g50t), one click-only game (ft09), and one mixed game (sp80); A0 versus B1; three seeds per cell; 500k steps. Cross-class generalization required B1 to have a higher successful-seed count on at least two of three games. Table 13 separates final evaluation from reward discovery during training.

Table 13: Stratified A0–B1 generalization. "Train" and "eval" count seeds with at least one level clear. The frozen endpoint gate fails because B1 has a higher evaluation success count only on ft09.

		A0 flat		B1 factored	
Game	Class	Train	Eval	Train	Eval
g50t	button	0/3	0/3	0/3	0/3
ft09	click	3/3	1/3	3/3	2/3
sp80	mixed	0/3	0/3	3/3	0/3

The formal generalization gate fails: final success count improves only on ft09, from $1/3$ to $2/3$. Its A0 median remains zero, while B1's median is 0.00304 (the exact non-zero B1 scores are 0.115 and 0.00304). Neither arm succeeds on g50t, an important boundary case showing that even a small exposed button grammar does not solve hard exploration.

The mixed-control sp80 result is more informative than its zero endpoint suggests. A0 never clears a level in training or evaluation. B1 produces 16, 16, and 32 positive training episodes across seeds 0–2, yet all three final policies score zero in evaluation. Thus the controller-side *mechanism* generalizes beyond cd82: factorization can make the reward/controller loop ignite without yielding reliable terminal performance. What does not generalize is a benchmark improvement.

F.6 Frozen final-checkpoint audit: small evaluation can hide reliable skill

The preceding endpoint counts came from the evaluation episodes emitted during training. Because success is sparse, we froze a reliability audit before rerunning any checkpoint: restore each final policy without further learning and sample 100 complete episodes from DreamerV3's native stochastic evaluation policy. Table 14 reports the 23 of 24 cells completed at the time of writing; the missing cd82 B1 seed 0 is excluded (its one-episode restoration smoke cleared, but is not a valid 100-episode estimate).

Table 14: Frozen final-checkpoint reliability. "Seeds" counts checkpoints with at least one clear; "episodes" is descriptive because episodes within a checkpoint are not independent training replicates. RHAE lists per-seed values.

Game	Arm	Successful seeds	Clear episodes	Per-seed RHAE
cd82	A0	0/6	0/600	0,0,0,0,0,0
cd82	B1	5/5 complete	500/500	.0548,.0548,.0548,.0548,.0548
ft09	A0	2/3	183/300	.1150,0,.00002
ft09	B1	3/3	290/300	.1148,.0134,.00255
sp80	A0	0/3	0/300	0,0,0
sp80	B1	2/3	140/300	0,.1917,.1377

This audit changes the mechanistic interpretation. The sp80 B1 checkpoints previously labeled 0/3 at evaluation are not uniformly unconsolidated: seeds 1 and 2 clear 72/100 and 68/100 episodes (95% Wilson intervals 62.5–79.9% and 58.3–76.3%), while A0 remains 0/300 (upper bound 1.26%). On cd82, all five complete B1 cells clear 100/100, whereas A0 clears 0/600 (upper bound 0.64%).

ft09 similarly exposes a reliability difference hidden by binary labels: A0 is seed-fragile (100%, 0%, 83%), while B1 is consistently high (100%, 97%, 93%). Thus periodic small-sample evaluation can produce false-negative endpoint labels; final-policy competence must be separated from both training reward discovery and the chance of observing it under a limited evaluation budget.

F.7 Interpretation: ignition, expression, and evaluation reliability

Taken together, these experiments refine the competence–performance account. The original negative result is not repaired by a more structured action space, and the pre-registered B2 and cross-class gates both fail. We therefore do not claim that factorization solves ARC-AGI-3 or broadly improves RHAE. Instead, the intervention separates two controller failures that a terminal score conflates.

First, there is *ignition*: encountering reward, learning a non-zero reward model, assigning positive return to some imagined futures, and moving the actor away from its uniform prior. B1 changes this process on cd82 ($6/6$ training-success seeds versus A0 $0/6$) and sp80 ($3/3$ versus $0/3$). Second, there is *expression*: retaining and sampling the discovered behavior under held-out evaluation. The original periodic evaluation observed it in only $4/6$ cd82 B1 agents and $0/3$ sp80 agents, but the frozen-checkpoint audit shows near-deterministic final skill in the completed cd82 cells and reliable skill in two sp80 seeds. The remaining heterogeneity is across training seeds, not simply within-policy intermittency.

This result narrows the actionable bottleneck. Better reconstruction is not sufficient; A0 can fit cd82 more tightly while remaining uniformly inactive. Better action structure is also not sufficient; it can produce positive imagined return and real training clears while leaving final RHAE at zero under the original periodic evaluation. The frozen final-checkpoint diagnostic resolves that ambiguity: substantially more stochastic episodes reveal reliable behaviors that the original protocol missed. We retained 32 intermediate sp80 checkpoints to determine when that reliability emerges or disappears. The present evidence supports the narrower conclusion that factoring the interaction space helps DreamerV3 both discover and, in some seeds, reliably express useful behavior; small evaluation budgets can conceal that competence.

F.8 A supervised deconfound: the statistical channel in isolation

Table 15: Supervised head ablation under the cold protocol of Table 7; two seeds per cell, median in parentheses. The spatial-map column is the Table 7 cold arm itself; sp80 is excluded from the combined score.

Game	Flat (A0 grammar)	Factored (B1 grammar)	Spatial map
vc33	.0269, .0548 (.0408)	.0315, .1247 (.0781)	.0314, .0766 (.0540)
sb26	0, 0 (0)	0, 0 (0)	0, 0 (0)
cd82	0, 0 (0)	0, 0 (0)	0, 0 (0)
tn36	.0016, .0016 (.0016)	0, 0 (0)	.0357, .0411 (.0384)
ls20	0, 0 (0)	0, 0 (0)	0, 0 (0)
lf52	0, 0 (0)	0, 0 (0)	.0133, .0222 (.0178)
Combined	.0071	.0130	.0184
sp80	0, 0 (0)	0, 0 (0)	0, 0 (0)

To separate the statistical benefit of action-space structure from its effect on exploration, we repeat the supervised protocol of Table 7 with two parameter-matched alternative output heads: a coordinate-factored head in the B1 grammar and a slot-shared monolithic 4102-way softmax that treats actions as unrelated labels. Table 15 reports the full grid. With exploration and the world model removed entirely, combined RHAE degrades monotonically as structure is removed ($0.0184 \rightarrow 0.0130 \rightarrow 0.0071$, two seeds per cell), and the gains are coverage-shaped: of the three games the spatial map scores on, the monolithic head retains only vc33 and a trace of tn36, and the factored head only vc33. Part of the factored-action effect thus survives in pure behavior cloning, consistent with the sample-efficiency reading; the exploration-geometry and dynamics-conditioning channels of the RL result are outside the reach of this ablation. We additionally ran all three heads on sp80, one of the two games where the factored actor ignites in RL (Table 13; the other, cd82, is

already in Table 7): under this cold behavior-cloning protocol every head scores zero on both games, so the supervised representation effect and the RL ignition occur on disjoint games, consistent with the ignition itself being driven by exploration geometry rather than the statistical channel isolated here.