RIGHT RANKING, WRONG REASONS: CLOSED-LOOP CAUSAL-CHANNEL FALSIFICATION FOR WORLD-MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models are increasingly evaluated by whether their closed-loop rollouts recover the values or ranking of downstream policies. We show that even exact value and ranking agreement can be causally misleading. An error in the observation channel changes the actions induced by an observation-based policy, while an error in the physical transition model changes the consequences of those actions; the two errors can cancel. We give a finite constructive example in which two policies receive the exact correct values and strict ranking even though both policies take the opposite initial action and the learned physical model swaps the consequences of the two actions. We then introduce Closed-Loop Causal-Channel Falsification for World Models (C3F-WM), a four-corner audit that separates observation-mediated action error, physical-model error, and their cancellation. The empirical protocol combines detector-independent component endpoints, practical error floors, tie-safe calibration ranks, and a continuous train-only support distance. In an independent MetaDrive replication with disjoint train, calibration, and test seeds, the aligned action/support detector is evaluated on 64 units (31 positive, 33 negative). Its area under the receiver operating characteristic curve (AUROC) is 0.999, compared with 0.849 for action-only and 0.911 for physical-only. Seed-cluster bootstrap 95% lower bounds for the two AUROC margins are 0.079 and 0.017. At the calibration-selected threshold, recall is 0.903 at zero false-positive rate. These results demonstrate complementary detection of observation-mediated and physical errors. On 64 held-out scenario clusters, a separately trained three-initialization red–green–blue (RGB) ensemble obtains 2.832/2.828 m current/next gap mean absolute error (MAE) and 0.981/0.981 lead-presence accuracy, demonstrating learned recovery of audit-relevant driving semantics. Together, these findings connect policy-ranking evaluation with channel-specific diagnosis of world-model errors.

1 INTRODUCTION

An action-conditioned predictor is useful as a world model only to the extent that it supports the decision made with it. This observation has motivated a shift from pixel fidelity and one-step prediction error toward closed-loop rollouts, policy value, and policy-ranking agreement. Yet ranking agreement compresses the entire evaluation loop into one ordering. It does not reveal whether the evaluated policy saw the right observation, chose the right action, or reached the right state for the right physical reason.

This ambiguity matters whenever a policy consumes observations rather than latent physical states. Let s, o, and a denote a physical state, observation, and action, respectively; s' denotes the next state. Let $P(s' \mid s, a)$ be the physical transition law, $O(o \mid s)$ the observation channel, and $\pi(a \mid o)$ a reactive policy. Changing O changes the policy's induced action distribution at state s,

$$\bar{\pi}_{\pi,O}(a \mid s) = \sum_{o} O(o \mid s)\pi(a \mid o),$$

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.

even when P is unchanged. A transition-only evaluator that holds the observation channel fixed can therefore roll out a different action law. Conversely, a channel-complete evaluator can induce the right actions in regions where its learned physical model is poorly supported. These two failures may reinforce, oppose, or exactly cancel.

The cancellation case defeats an apparently strong check: exact policy values and their strict ranking. We construct a two-state example where the approximate observation channel exchanges the policies' actions, while the approximate physical model exchanges the consequences of those actions. Both component errors are nonzero, but every policy value is exact. This establishes that ranking correctness alone cannot identify a causally faithful closed-loop evaluator. The minimal witness isolates the cancellation mechanism and establishes the need for a channel-level audit even when downstream values agree exactly.

We propose C3F-WM, a channel-resolved falsification protocol for world-model evaluation. It evaluates four combinations of true or reference versus learned physical and observation components, and it reports three non-substitutable signals: policy ranking, induced actions, and physical occupancy support. A candidate evaluation passes only when all available signals pass. This conjunction gives each causal path its own evidence requirement. Each signal is bound to a specified corner contrast, making the source of a failed audit explicit and directing diagnosis to the affected component.

Our contributions are:

1. We formalize policy-wise cancellation between physical-model error and observation-mediated policy error, and give a constructive witness with exact values and ranking but wrong actions and physics.
2. We convert the cancellation witness into an executable four-corner audit whose action and physical signals are bound to single-channel contrasts. In an analytic suite, ranking-only, action-only, and physical-only checks each miss one of three wrong loops, whereas the predefined conjunction rejects all three and accepts the faithful loop.
3. We provide paired evidence that observation-channel interventions mediate closed-loop actions and policy relations in HighwayEnv and MetaDrive without changing the declared physical mechanism.
4. We evaluate the resulting protocol in an independent MetaDrive replication. The joint action and continuous-support detector satisfies predefined class-balance, point-margin, bootstrap-margin, complementary-stratum, and operating-point criteria.
5. We demonstrate recovery of audit-relevant semantics with a temporal RGB estimator on 64 held-out balanced scenarios under fixed control schedules.

2 Related work

2.1 Decision-centric world-model evaluation

Recent work argues that world models for embodied agents should be evaluated by counterfactual action fidelity, closed-loop validity, value prediction, policy ranking, exploitability, and uncertainty rather than visual plausibility alone (Yu et al., 2026). Policy-in-the-loop systems likewise alternate generated observations and policy actions (Ma et al., 2026). C3F-WM adopts the decision-centric premise but asks whether a correct downstream ranking was produced through the correct observation-to-action and action-to-state paths.

2.2 Policy-aware physical-model error

Value-equivalent models deliberately preserve Bellman updates for selected policy/function sets rather than all transition details (Grimm et al., 2020); partial variants trade fidelity for scalable, robust planning (Alver & Precup, 2023). Our result is complementary: agreement on the values or ranking of a finite policy set cannot identify which closed-loop component produced that agreement. Logged-data off-policy evaluation instead targets policy value estimation without treating a learned simulator's internal causal channels as the audit object (Jiang & Li, 2016).

World-model policy evaluation has been observed to underestimate values for in-distribution actions and overestimate values for out-of-distribution actions while retaining policy rankings in the paper's first version (Quevedo et al., 2025). Its accepted WorldGym version evaluates real-robot vision–language–action (VLA) policies through video-world-model rollouts and reports both success-rate correlation and relative-ranking preservation (Quevedo et al., 2026). C3F-WM asks the next diagnostic question: whether that ranking is supported by the intended causal channels. Policy-aware simulator learning directly optimizes worst-policy value gaps and relates adversarial policy search to an error Markov decision process (Error-MDP) (Dann et al., 2026). Offline and robust policy evaluation also study support mismatch, conservative values, and calibrated predictive uncertainty (Malik et al., 2019). Those methods target physical-model reliability or policy learning. C3F-WM instead studies an evaluation identifiability problem created by the interaction of physical error with an observation-induced action law. Its defining design is the combination of single-channel corner contrasts and a predefined conjunction: together, they expose compensation across the closed loop and assign discrepancies to the corresponding causal paths.

2.3 Observation interventions and driving world models

Known-axis benchmarks isolate visual shifts while holding latent dynamics and reward fixed (Cherepanov et al., 2026); closed-loop world-model benchmarks prioritize embodied task success (Zhang et al., 2026); and CG-World records explicit observation, action, and mechanism intervention branches (Cai et al., 2026). WorldModelBench instead evaluates whether generated videos obey instructions and physical constraints (Li et al., 2025). More abstractly, generative models can pass existing diagnostics while encoding incoherent implicit world models that fail on related tasks (Vafa et al., 2024). Building on this motivation for stronger diagnostics, we study the causal support of policy rankings through the paired four-corner relation needed to detect cancellation between observation-mediated and physical errors.

3 Closed-loop evaluation through two channels

We consider a finite-horizon partially observed controlled process with state space $\mathcal{S}$, observation space $\mathcal{O}$, action space $\mathcal{A}$, physical transition P, observation channel O, reward $r(s, a)$, initial-state distribution ρ, and reactive policy π. At time t, the process samples $s_0 \sim \rho$, $o_t \sim O(\cdot \mid s_t)$, $a_t \sim \pi(\cdot \mid o_t)$, and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. For a positive integer horizon H, the undiscounted value is

$$J(P, O, \pi) = \mathbb{E}_{\rho,P,O,\pi}\left[\sum_{t=0}^{H-1} r(s_t, a_t)\right].$$

A learned evaluator contains an approximate physical model $\widehat{P}$ and an explicit or implicit observation model $\widehat{O}$. We compare the true and approximate loops through the hybrid $(\widehat{P}, O)$:

$$\begin{aligned} J(P, O, \pi) - J(\widehat{P}, \widehat{O}, \pi) = {} & \underbrace{J(P, O, \pi) - J(\widehat{P}, O, \pi)}_{\Delta_P(\pi)} \\ & + \underbrace{J(\widehat{P}, O, \pi) - J(\widehat{P}, \widehat{O}, \pi)}_{\Delta_O(\pi)}. \end{aligned}$$

Here $\Delta_P(\pi)$ is the physical-channel value discrepancy, $\Delta_O(\pi)$ is the observation-channel value discrepancy, and $\Delta_{\text{total}}(\pi) = \Delta_P(\pi) + \Delta_O(\pi)$ is the total value error. Large opposing component errors can therefore produce a small total error. We say that policy π has a cancellation witness at a nonnegative tolerance ϵ if

$$|\Delta_P(\pi) + \Delta_O(\pi)| \le \epsilon, \quad |\Delta_P(\pi)| > \epsilon, \quad |\Delta_O(\pi)| > \epsilon, \quad \Delta_P(\pi)\Delta_O(\pi) < 0.$$

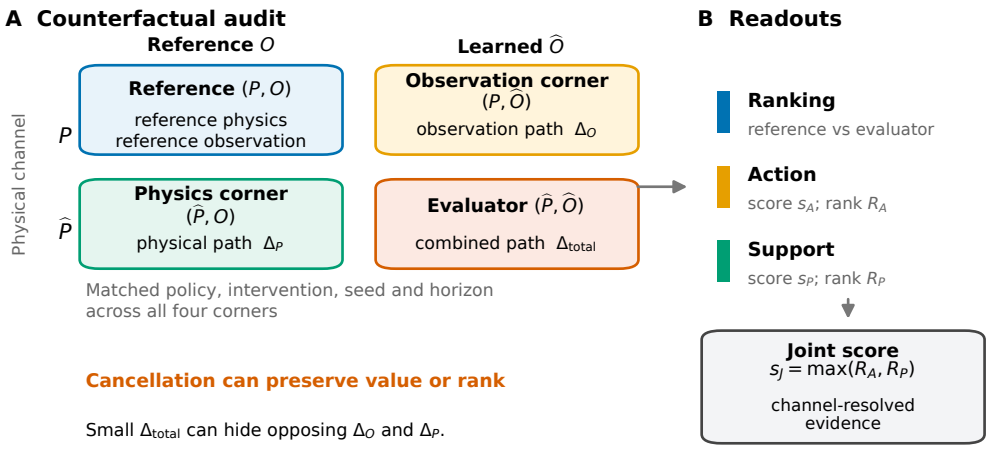

Figure 1: C3F-WM four-corner audit. (A) Crossing reference or learned physical (P) and observation (O) channels isolates observation-path (Δ_O), physical-path (Δ_P), and total (Δ_{total}) discrepancies under matched policies, interventions, seeds, and horizons. (B) Ranking disagreement is reported separately from action score s_A and physical-support score s_P; their calibration ranks R_A and R_P form the joint score $s_J = \max\{R_A(s_A), R_P(s_P)\}$. Opposing component errors can preserve total value or policy rank.

3.1 Proposition 1: Correct ranking does not identify loop fidelity

There exist $P \neq \widehat{P}$, $O \neq \widehat{O}$, and two policies π_A, π_B such that

$$J(P, O, \pi_i) = J(\widehat{P}, \widehat{O}, \pi_i) \quad \text{for } i \in \{A, B\},$$

and hence both loops have the same strict ranking, while $\bar{\pi}_{\pi_A, O}(\cdot \mid s_0) \neq \bar{\pi}_{\pi_A, \widehat{O}}(\cdot \mid s_0)$ and likewise for π_B, where s_0 is the initial state and A, B index the two policies.

Proof by construction. Let the initial state be bad with probability one, use horizon two, and let reward be zero in bad and one in good, independent of action. Under true physics, action 1 moves bad to good and action 0 remains bad; the approximate physics exchanges these consequences. The true observation at bad is symbol 1 and the approximate observation is symbol 0. Policy A maps symbols 0 and 1 to actions 0 and 1, respectively, while policy B uses the opposite mapping. The true loop therefore gives $J(P, O, \pi_A) = 1$ and $J(P, O, \pi_B) = 0$. The approximate observation exchanges the policies' initial actions and the approximate physics exchanges their consequences, so the approximate loop also gives the same values. For π_A, $(\Delta_P(\pi_A), \Delta_O(\pi_A)) = (+1, -1)$; for π_B, $(\Delta_P(\pi_B), \Delta_O(\pi_B)) = (-1, +1)$. Thus both component discrepancies are nonzero, the induced initial actions differ for both policies, and their sums vanish, which proves the claim.

The construction is stronger than a ranking reversal: even exact per-policy values do not identify either component. It motivates component-wise falsification rather than interpreting total value or rank error alone.

4 C3F-WM: A four-corner falsification audit

4.1 Four loops

For each policy and intervention, C3F-WM evaluates, when the benchmark permits, the following loops:

The reference channels are supplied by simulator access or paired held-out trajectories. C3F-WM uses these references for offline world-model benchmarking and model selection, with explicit control over the evaluated channels.

Physical channel	Observation channel	Role
(P)	(O)	reference true loop
(P)	$\widehat{O}$	observation-error isolation
$\widehat{P}$	(O)	physical-error isolation
$\widehat{P}$	$\widehat{O}$	evaluated world-model loop

Table 1: Four closed-loop combinations used by C3F-WM.

4.2 Three falsification axes

For a policy pair, the ranking axis records reference-versus-evaluator rank disagreement. The action axis measures paired mean absolute command difference, which operationalizes disagreement between induced action laws. The physical axis measures the fraction of model-rollout state-action rows outside a support set fitted only on training occupancies. We use standardized nearest-neighbor distance and freeze its threshold as the 99th percentile of leave-one-training- seed-out distances.

Let s_A denote paired mean absolute command difference and s_P the continuous physical-support score. Calibration converts them to empirical cumulative distribution function (CDF) ranks $R_A(s_A)$ and $R_P(s_P)$. The primary empirical joint score is

$$s_J = \max\{R_A(s_A), R_P(s_P)\}.$$

It has no learned test-time weights and treats either causal-path failure as sufficient to identify a component discrepancy. Ranking disagreement is reported as a separate third axis and does not enter the primary label or detector, preventing endpoint-and-score overlap. The analytic certificate is the conjunction of all three axes. A binary operating point for the empirical joint score is selected on calibration under a maximum 10% false-positive rate.

The component endpoints and detector variables measure distinct quantities along the same causal paths. For a fixed policy, the observation endpoint is $e_O = |J(P, O, \pi) - J(P, \widehat{O}, \pi)|$, and the physical endpoint is $e_P = |J(P, O, \pi) - J(\widehat{P}, O, \pi)|$. Observation and physics labels threshold e_O and e_P, whereas the corresponding detectors use paired command discrepancy and train-only occupancy distance. Thus the evaluation tests whether path diagnostics predict downstream component error. A seed-block negative control in Section 6.4 measures the dependence of this prediction on the alignment between diagnostics and their corresponding endpoints.

4.3 Interpreting audit decisions

A rejected audit identifies a discrepancy in the observation-to-action or action-to-state path underlying the tested ranking, providing a component-level target for investigation. A passing audit records consistency across the tested causal obligations for the chosen interventions, policies, support representation, and reference simulator.

5 Experimental design

5.1 Exploratory paired channel evidence

We first measure whether observation-only interventions can change closed-loop behavior while preserving the declared physical mechanism. In HighwayEnv, 180 nominal versus `gap +15 m` pairs share seed, policy, environment configuration, initial physical state, and 40-step horizon. In MetaDrive 0.4.3, 27 central processing unit (CPU), headless episodes pair nominal, gap-bias, and ego-speed-bias channels across three seeds and three semantic controllers (Leurent, 2018; Li et al., 2023).

5.2 ENDPOINT AND SCORE DESIGN

Labels are defined only by independently measured observation and physical value errors. For component $c \in \{O, P\}$, its endpoint threshold is the maximum of the calibration 75th percentile and $10^{-4}+10^{-3}v_{\mathrm{ref}}$, where v_{ref} is the maximum of 1 and the absolute calibration reference values for the evaluated policy pair. Ranking is secondary stratification and never enters a detector score. Given n calibration scores $x_1, \ldots, x_n$, a score x receives the strict empirical rank $R(x) = n^{-1} \sum_{i=1}^{n} \mathbf{1}[x_i < x]$, where $\mathbf{1}[\cdot]$ is the indicator function; tied minima therefore map to zero.

The physical score summarizes train-standardized nearest-neighbor distances continuously: it is the trajectory's 95th distance percentile divided by the train-only 99th-percentile cross-seed threshold. This formulation preserves resolution when binary unsupported occupancy is constant.

5.3 INDEPENDENT REPLICATION

The MetaDrive 0.4.3 CPU/headless replication uses 12 train, 8 calibration, and 16 test seeds unused by earlier C3F-WM runs. Four fixed controllers form two predeclared policy pairs: observation specialists and mixed-sensitivity controllers. Train channels are nominal, `gap +5 m`, and `ego speed +2 m/s`; calibration and test both use nominal, `gap +10 m`, and `ego speed +3 m/s`. Each rollout has 40 decisions at 10 hertz (Hz). A fixed 15-parameter ridge linear delta model with coefficient 10^{-6} represents $\widehat{P}$; declared additive biases implement O exactly.

Each evaluation unit is one seed–target-intervention–policy-pair tuple. Thus the test has $16 \times 2 \times 2 = 64$ units; units sharing a seed are kept in the same bootstrap cluster. For each unit, the reference is (P, O), the observation counterfactual is $(P, \widehat{O})$, the physics counterfactual is $(\widehat{P}, O)$, and the combined evaluator is $(\widehat{P}, \widehat{O})$. The primary label is the union of thresholded reference-to-observation-corner and reference-to-physics-corner value errors. The action detector compares commands in the first two corners; the physical detector uses continuous support of the third. Their joint score is the maximum of calibration-derived strict ranks. The combined corner and ranking are excluded from the label and primary detector.

5.4 STATISTICAL ANALYSIS

The analysis criteria were set before test evaluation. Calibration must contain both classes, at least eight distinct scores per axis, and a joint vector different from either axis. Test criteria require at least 8 positives and negatives, at least 4 observation-only and physics-only positives, joint AUROC margins of at least 0.05 over both axes, positive 95% seed-bootstrap lower bounds, and recall at least 0.75 with a false-positive rate (FPR) at most 0.10 at the calibration-selected operating point. Uncertainty uses 2,000 seed-cluster resamples. The 0.05 margin is the predefined operational minimum for this replication. Average precision is reported descriptively as a secondary metric. The semantic-channel replication uses simulator-generated data and is reproducible on a CPU.

5.5 LEARNED RGB OBSERVATION STRESS TEST

We test compatibility with a learned visual component. MetaDrive 0.4.3 renders lossless 84-by-84 RGB for four fixed control schedules. A prediction-independent census assigns 64/32/64 train/calibration/test scenario seeds, balanced across low and high lead presence; model selection uses only the calibration split. The estimator concatenates two frames, uses 32/64/128-channel convolutions and a 256-dimensional latent, and predicts current plus action-conditioned next semantics. Three calibration-selected initializations (1981/1993/1997) form a arithmetic ensemble. Evaluation criteria require current/next gap MAE at most 4 m (upper 95% bound at most 6 m), speed MAE at most 2 m/s, and presence accuracy at least 0.90 (lower bound at least 0.85); intervals use 5,000 scenario-cluster resamples. The experiment evaluates learned recovery of current and next-step driving semantics from RGB observations under fixed scripted control schedules.

Detector	AUROC	Average precision
Action only	0.8490	0.8853
Physical support only	0.9110	0.9272
Joint C3F-WM	**0.9990**	**0.9980**

Table 2: Detector performance in the independent MetaDrive replication.

6 RESULTS

6.1 OBSERVATION INTERVENTIONS MEDIATE ACTIONS AND RELATIONS

In the HighwayEnv pairs, every one of 180 source pairs changes at least one command; 92.17% of 7,200 paired steps differ. Across 990 within-seed policy pairs, 324 relations change, each from a nominal tie to an intervention-directional ordering. These changes demonstrate observation-mediated action differences and tie resolution; the strict-reversal count is zero.

In the exploratory MetaDrive breadth assay, `gap +15 m` changes 320/360 commands (88.89%) and 2/9 policy relations. `ego speed +5 m/s` changes 240/360 commands (66.67%) with all relations retained. True time-to-collision (TTC) risk values remain zero across this assay.

6.2 ADDING A CHANNEL MODEL ALONE IS INSUFFICIENT

An earlier held-out MetaDrive comparison fitted an exact affine observation channel but used a shared linear physical model. Transition-only ranking accuracy was 4/4 per policy class, while channel-complete accuracy was 3/4. A diagnostic analysis found 12.5% unsupported channel-complete occupancy and correlation 0.817 between unsupported fraction and absolute value error. This result shows why explicit observation modeling and physical-support assessment are complementary.

6.3 ABLATIONS MOTIVATING INDEPENDENT ENDPOINTS

An ablation with rank disagreement in both endpoint and score attained AUROC 0.976 but contained direct endpoint–score overlap. Removing the shared indicator made the label independent of the detector. HighwayEnv value differences below 10^{-7} further motivated practical error floors. Binary unsupported occupancy became constant under matched intervention strengths, motivating the continuous support score in Section 5.2. A separate 12-seed analysis reached joint AUROC 0.978 but contained only three negatives, and its joint-minus-action bootstrap interval included zero. Together, these ablations identify endpoint independence, numerical floors, continuous support, and class balance as the conditions required for the replication below.

6.4 INDEPENDENT CONTINUOUS-SUPPORT REPLICATION

Calibration contains 12 positives and 20 negatives, including four in each positive causal stratum. The independent test contains 64 units: 31 positives and 33 negatives; positives comprise 10 observation-only, 7 physics-only, and 14 both-component errors.

The joint AUROC margins are +0.1500 over action-only and +0.0880 over physical-only, both above the predefined +0.05 criterion. All 2,000 seed-cluster resamples contain both classes. The 95% margin intervals are `[0.0788, 0.2273]` and `[0.0170, 0.2366]`; both lower bounds exceed zero.

As a cluster-sensitivity analysis, we recomputed all metrics after deleting each test seed in turn. The joint-minus-action AUROC margin ranges from 0.1341 to 0.1707, and the joint-minus-physical margin ranges from 0.0722 to 0.1016; all 16 deletions retain both classes and positive margins. This complementary sensitivity analysis agrees with the seed-cluster bootstrap intervals and the predefined primary criteria.

We also evaluate a seed-block alignment negative control. It keeps endpoint labels fixed and cyclically shifts the paired action/physical score blocks across seeds, preserving class balance, score

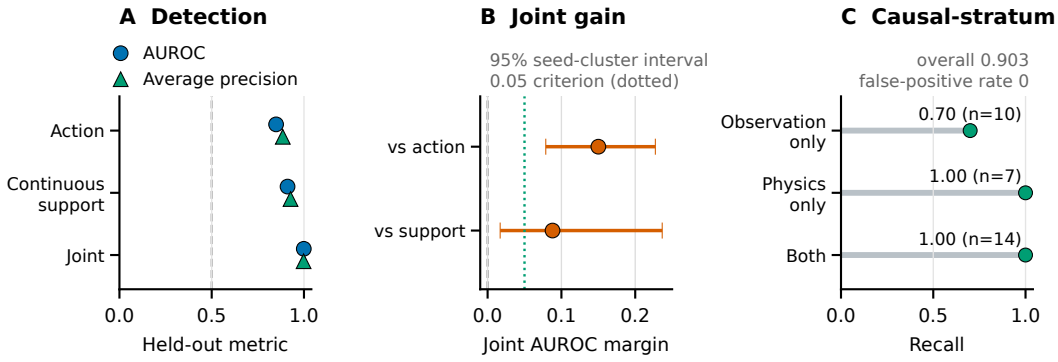

Figure 2: Independent continuous-support MetaDrive replication (64 test units: 31 positive, 33 negative). (A) Area under the receiver operating characteristic curve (AUROC) and average precision for the aligned action, support, and joint detectors. (B) Joint-minus-single-axis AUROC margins with 95% seed-cluster bootstrap intervals; the dotted line marks the predefined 0.05 criterion. (C) Recall at the calibration-selected threshold by causal stratum (overall 0.903; false-positive rate, 0).

marginals, score dependence, and each four-unit block while breaking only seed-level alignment. Across all 15 nonidentity shifts, joint AUROC has median 0.532 and range 0.349–0.652, versus 0.999 when aligned; 0/15 shifts reach the aligned value. This descriptive alignment analysis isolates the contribution of correct seed-level pairing. Inferential uncertainty is quantified by the seed-cluster bootstrap.

6.5 Operating point and causal strata

At the calibration-selected threshold 0.78125, test recall is 0.9032 and the false-positive rate is 0.0. Recall is 0.70 for observation-only failures and 1.0 for physics-only and both-component failures. The result satisfies all five evaluation criteria—class balance, complementary strata, point margins, bootstrap lower bounds, and operating point—and demonstrates complementarity between aligned action and physical-support signals in this MetaDrive setting.

6.6 Learned RGB stress test

The held-out evaluation contains 64 scenario clusters (16,384 transitions). Ensemble current/next gap MAE is 2.832/2.828 m (95% intervals [2.015, 3.763]/[2.019, 3.739]); ego-speed MAE is 0.074/0.065 m/s, lead-speed MAE 0.228/0.197 m/s, and presence accuracy 0.981/0.981 (lower bounds 0.967/0.968). Figure 3 shows every initialization and four representative slots selected without reference to model predictions.

The balanced RGB ensemble meets every predefined semantic-recovery criterion, showing that a compact temporal estimator can recover the driving quantities used by the audit from held-out visual observations.

7 Limitations and broader impacts

The primary C3F-WM evidence uses one simulator, 16 test seeds, semantic car-following states, four hand-specified controllers, additive sensor biases, and a small linear physical model. The RGB extension adds one compact convolutional architecture, three initializations, four scripted control schedules, and 64 held-out scenario clusters in the same simulator. Its measured endpoint is semantic recovery under scripted controls. Uncertainty resampling accounts for shared seeds. Paired reference trajectories define the audit's offline setting. The four-dimensional nearest-neighbor support representation offers a tractable physical-error diagnostic; sensitivity to feature scaling and richer representations are important directions for further evaluation.

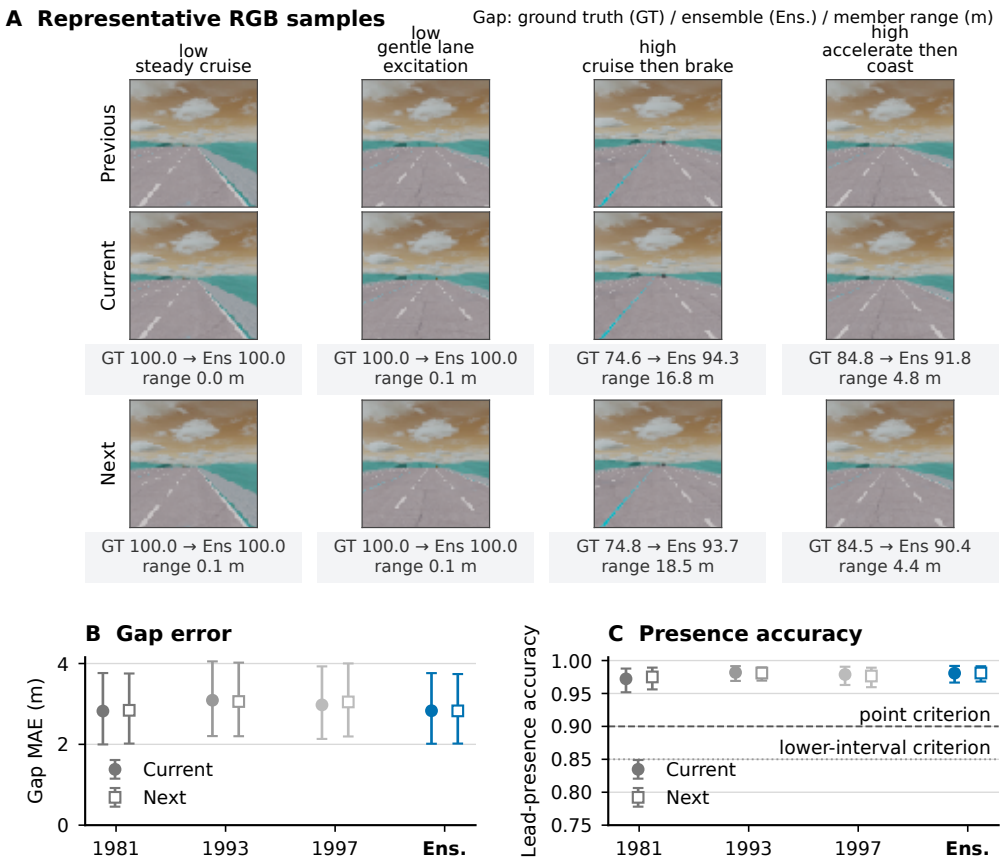

Figure 3: Learned red–green–blue (RGB) stress test on 64 held-out MetaDrive scenario clusters. (A) Four prediction-independent previous/current/next slots with reference-to-ensemble gap and three-member range. (B) Gap mean absolute error (MAE) and (C) lead-presence accuracy with 95% scenario-cluster intervals; panel C also shows predefined criteria. Gray denotes individual initializations and blue their arithmetic ensemble. The experiment measures recovery of driving semantics under fixed scripted controls.

The independent split separates training, calibration, and test seeds. The Highway relation changes arise from ties, and its replication produces numerical-tolerance-sensitive labels; intervention-strength extrapolation can also reduce class balance. These observations motivate additional engines, nonlinear world models, learned policies, richer corruptions, and support representations.

C3F-WM contributes diagnostic evidence for model selection in controlled driving simulation. By exposing compensating errors and resolving them by channel, it helps researchers identify which model components warrant improvement.

8 Conclusion

Correct policy rankings can arise for the wrong reasons when observation and physical errors cancel. C3F-WM exposes this identifiability gap through a four-corner audit with diagnostics bound to single-channel contrasts. Its analytic witness proves the gap, while independent MetaDrive results show complementary component detection and accurate recovery of audit-relevant semantics from RGB observations. The method provides a channel-resolved basis for diagnosing world-model errors and interpreting policy rankings.

Reproducibility statement

The anonymous supplement contains the analytic construction, implementation, pinned environment, configurations, exact seed splits, calibration procedures, statistical routines, and machine-readable results underlying every reported table and figure. Sections 3–6 specify the assumptions, interventions, models, thresholds, and cluster-bootstrap procedures. The supplement also provides verification tests and portable environment specifications for reproducing the experiments on independent hardware.

Artificial intelligence use statement

Generative artificial intelligence (AI) tools assisted with language editing and code development. The authors verified the resulting content and take responsibility for the paper.

References

Safa Alver and Doina Precup. Minimal value-equivalent partial models for scalable and robust planning in lifelong reinforcement learning. In *Proceedings of the 2nd Conference on Lifelong Learning Agents*, volume 232 of *Proceedings of Machine Learning Research*, pp. 548–567, 2023. URL https://proceedings.mlr.press/v232/alver23a.html.

Yiming Cai, Fangjie Yu, Meiqing Yu, Ziyue Shi, Pengfei Yuan, and Yong Guo. CG-World: A large-scale world-state dataset and protocol for world models, 2026. URL https://arxiv.org/abs/2607.26452. arXiv preprint, version 1.

Egor Cherepanov, Daniil Zelezetsky, Alexey K. Kovalev, and Aleksandr I. Panov. KAGE-Bench: Fast known-axis visual generalization evaluation for reinforcement learning, 2026. URL https://arxiv.org/abs/2601.14232. arXiv preprint, version 2.

Christoph Dann, Yishay Mansour, and Mehryar Mohri. Theoretical foundations and effective algorithms for policy-aware simulator learning, 2026. URL https://arxiv.org/abs/2605.29032. arXiv preprint, version 3.

Christopher Grimm, André Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 652–661, 2016. URL https://proceedings.mlr.press/v48/jiang16.html.

Edouard Leurent. An environment for autonomous driving decision-making. https://github.com/Farama-Foundation/HighwayEnv, 2018.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1834. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4ec03ed08a3fcb59e1c815b5598beff1-Abstract-Datasets_and_Benchmarks_Track.html.

Quanyi Li, Zhenghao Peng, Lan Feng, Qihang Zhang, Zhenghai Xue, and Bolei Zhou. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3461–3475, 2023. doi: 10.1109/TPAMI.2022.3190471. URL https://doi.org/10.1109/TPAMI.2022.3190471.

Chong Ma, Taiyi Su, Jian Zhu, Jianjun Zhang, Zitai Huang, Yi Xu, and Hanli Wang. PiL-World: A chunk-wise world model for VLA policy-in-the-loop evaluation, 2026. URL https://arxiv.org/abs/2606.05773. arXiv preprint, version 1.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4314–4323, 2019. URL https://proceedings.mlr.press/v97/malik19a.html.

Julian Quevedo, Percy Liang, and Sherry Yang. Evaluating robot policies in a world model, 2025. URL https://arxiv.org/abs/2506.00613v1. arXiv preprint, version 1; later versions use the title WorldGym and an expanded author list.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, 2026. URL https://iclr.cc/virtual/2026/poster/10008029.

Keyon Vafa, Justin Y. Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0846. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/2f6a6317bada76b26a4f61bb70a7db59-Abstract-Conference.html.

Yang Yu, Shiyuan Zhang, Yifei Sheng, Haoxiang Ren, and Haoxin Lin. How should world models be evaluated for embodied decision-making? a decision-making-centric position, 2026. URL https://arxiv.org/abs/2606.15032. arXiv preprint, version 2.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal Patel, Paul Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-in-world: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/5b4263be85820683d78675cc18d2efc7-Abstract-Conference.html.