EVALUATING SEMANTIC STATE RESPONSE AND GENERATED VISUAL FEEDBACK UNDER DISTINCT INPUT CONTRACTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive video systems must do more than predict action-conditioned states: their generated observations may become inputs to later decisions. We introduce State–Feedback Separation (SFS), a contract-based reporting schema for factual semantic response (S), no-anchor generated feedback (G), privileged-reference dependence (A), and transfer stress (T), illustrated by a bounded case study and mapping diagnostic. In inspected, scene-disjoint AI2-THOR FloorPlan5 evaluations, a post-hoc ResNet-50 probe reaches source/target accuracy $0.750/1.000$, but an action-only rule reproduces all 160 target labels. Two distinct no-anchor RGB routes fail their gates: an object-aware predictor scores 0.375–0.500 with nonpositive gain, and a ResNet-50-guided full-resolution editor scores 0.375–0.500 across five seeds, also with nonpositive gain. Both are weak with factual histories, so these results do not isolate damage from feedback. In a separate anchor diagnostic, a distinct editor scores $200/200$ with its factual anchor versus $100/200$ with a blank anchor; one anchor-dropout variant preserves the pattern ($0/5$ seeds meet criterion). Transfer gates fail on Toaster ($0/20$), StoveKnob ($1/20$), and Kettle ($1/20$); oracle-goal inputs raise all tested cells to 1.000, localizing failure to goal inference in this controller pipeline. Both EPIC-KITCHENS routes are negative ($0/5$). On a separate controlled RGB task, post-hoc calibration yields 0.719 success (gain 0.250 over factual copy) across five frozen decoders, but passes only $1/5$ fresh-training seeds. S and G use different systems. The results do not establish a within-checkpoint dissociation, internal causal mechanism, broad cross-architecture or natural-video generality, or a complete interactive world model.

1 INTRODUCTION

Interactive video systems are judged both by what they predict after an action and by whether their generated observations remain useful when fed into a later decision. These are related but distinct capabilities. A semantic label can be correct because it is predictable from the action alone; a generated frame can look plausible yet fail to preserve state under roll-in. A persistent factual reference can make a restoration task solvable while providing information unavailable to a no-reference system. An evaluation score is therefore interpretable only relative to the observations and references its system receives.

Recent benchmarks have broadened world-model evaluation beyond visual quality to include closed-loop control, planning, interaction response, and physical adherence (Qin et al., 2024; Zhang et al., 2026; Shang et al., 2026a;b; Wu et al., 2026; Li et al., 2025). We do not claim to introduce interaction-aware evaluation or to replace these benchmarks. Our narrower concern is how to read evidence across different input contracts: whether a result uses factual or generated observations, whether it receives a privileged reference, and whether its task and visual family are held out.

We ask a contract-specific question: *In persistent-state interactive-video tasks, what do separate evaluations of factual-RGB semantic prediction, no-anchor generated-RGB roll-in, privileged-reference use, and unseen-family transfer reveal?* These evaluations do not compare one predictor under matched factual and generated inputs: S and G use different systems, while A tests reference dependence within one editor and T evaluates that editor's route on unseen object families. We organize

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.

the evaluations as State–Feedback Separation (SFS), a four-axis reporting schema: factual semantic response (S), no-anchor generated feedback (G), privileged-reference dependence (A), and transfer stress (T). The schema makes each input and reference contract explicit and keeps results separate rather than combining them into a score; it is not a new scalar metric or a validated general-purpose benchmark. Figure 1 summarizes the contracts and the bounded evidence profile.

We instantiate SFS in a bounded case study of binary state changes in AI2-THOR, with separate negative routes on EPIC-KITCHENS. FloorPlan5 was inspected during exploratory readout development; accordingly, its S result is post-hoc, and E3/E4 plus anchored-task measurements there are descriptive scene-disjoint diagnostics. A separate controlled reactive-RGB task supplies a bounded mapping example: a post-hoc correction passes on five frozen decoder seeds, while the same correction passes on only one of five fresh-training seeds. This second task broadens the set of tested contracts, but does not validate cross-architecture or natural-video generality.

The case study exposes three limits of the reported scores. Target-state labels follow the action alone; the no-anchor routes are weak even with factual histories; and the positive restoration route receives a privileged state-1 image. The controls quantify these limits rather than identify an internal mechanism. Three unseen-family gates and both natural-video readouts remain negative. Their results motivate reporting the contracts separately, with numerical details in Section 5.

Contributions.

1. **A contract-based reporting schema.** We specify four separately reported evaluation contracts that make explicit the factual observations, generated histories, privileged references, and transfer conditions behind each result. The schema organizes measurements rather than introducing a scalar metric or a validated general-purpose benchmark; it is illustrated by one bounded case study and a separate controlled mapping diagnostic.
2. **An audited diagnostic profile across distinct systems and two controlled tasks.** The AI2-THOR case study shows that an action-conditioned target score can be reproduced from action alone, that a separate no-anchor RGB predictor fails its generated-history gate, and that a distinct editor's positive restoration result depends on a factual anchor under the tested conditions. A second controlled RGB task passes after post-hoc calibration of frozen decoders, while the same correction passes only $1/5$ fresh-training seeds; this is a bounded mapping diagnostic, not evidence of broad portability.
3. **A bounded failure analysis.** Three tested unseen object-family gates and the natural-video route remain negative. Oracle-goal diagnostics rescue the tested transfer cells and localize the measured failure to goal inference in this controller pipeline; they do not identify an internal model mechanism or establish general transfer.

2 Related Work

Recent work already evaluates world models beyond open-loop visual quality. WorldSimBench maps generated videos to control signals in embodied tasks (Qin et al., 2024); World-in-World evaluates online planning and task success in four closed-loop environments (Zhang et al., 2026); WorldArena combines perceptual measures with functional roles such as data generation, policy evaluation, and action planning (Shang et al., 2026a). WorldArena 2.0 extends this line to visuotactile inputs, interactive-RL environments for policy optimization, and evaluation across simulated and real-world robot platforms (Shang et al., 2026b). Omni-WorldBench measures action effects on intermediate state trajectories and final outcomes (Wu et al., 2026), while WorldModelBench evaluates instruction following and physical adherence in generated videos (Li et al., 2025). These works establish that closed-loop utility and interaction response are active evaluation targets; SFS does not claim to introduce either one. Instead, it offers a narrow diagnostic case study whose documented measurements separate factual semantic readout, generated-RGB roll-in, reference dependence, and object-family transfer. The axes are evaluated under distinct contracts, and S and G use different systems.

Appendix Table 6 compares the scope of these evaluations with the diagnostics reported here. The comparison describes each work's stated focus; it does not claim that a diagnostic absent from a paper's headline contribution is absent from every experiment in that paper.

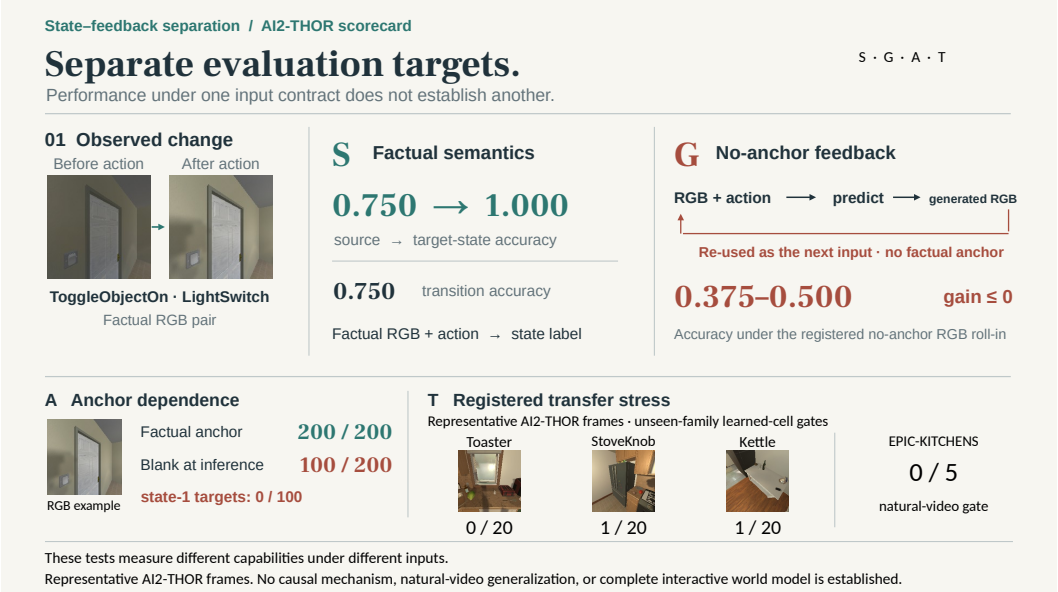


Figure 1: **State–Feedback Separation overview.** The four axes keep factual semantic response (S), no-anchor generated feedback (G), privileged-reference dependence (A), and transfer stress (T) separate. The reported examples summarize audited AI2-THOR results: S source/target/transition accuracy $0.750/1.000/0.750$ (the target label follows action, so 1.000 does not establish RGB contribution); G generated-RGB roll-in accuracy 0.375–0.500 with gain ≤ 0; A factual-anchor performance $200/200$ versus blank-input performance $100/200$, with $0/100$ state-1 targets; T transfer gates of Toaster $0/20$, StoveKnob $1/20$, and Kettle $1/20$, alongside EPIC-KITCHENS routes of $0/5$. The axes use distinct evaluation contracts and are not combined into a score. Values are from audited artifacts. The results do not establish environmental causal mechanisms, natural-video generalization, or complete interactive world models.

Distributional and rubric metrics such as FVD (Unterthiner et al., 2019) and VBENCH (Huang et al., 2024) assess clip-level properties. Synthetic diagnostics such as CLEVRER (Yi et al., 2020), Physion (Bear et al., 2021), and CausalWorld (Ahmed et al., 2021) target physical prediction, counterfactual reasoning, or transfer in controlled settings. Probing classifiers (Alain & Bengio, 2017; Belinkov, 2022) motivate the factual-RGB semantic readout used for S; sequence-model work on exposure bias (Bengio et al., 2015; Ranzato et al., 2016) motivates testing generated inputs, while learning using privileged information formalizes the role played here by a factual anchor (Vapnik & Vashist, 2009). We use these strands to define controlled measurements, not to infer a causal or neural mechanism.

We instantiate the diagnostics on AI2-THOR (Kolve et al., 2017) for binary state changes and include a bounded negative route on EPIC-KITCHENS (Damen et al., 2018). The present evidence supports a contract-specific diagnostic report and a failure analysis; it does not establish a general-purpose benchmark, broad benchmark coverage, natural-video generalization, or transfer beyond the tested routes.

3 THE STATE–FEEDBACK SEPARATION PROTOCOL

SFS is a descriptive scorecard over four axes. It is deliberately *not* aggregated into a scalar: the axes use different input contracts and answer different scientific questions.

S — Semantic response. Factual RGB plus action/entity semantics, predicting audited binary states. The two-step rollback is teacher-forced: factual RGB is supplied at every step. This lane measures whether factual observations support action-conditioned semantic prediction (Alain & Bengio, 2017).

G — Generated feedback. At inference, predicted frames replace the rolling RGB history; inputs are RGB history and action/entity conditioning, without a persistent factual anchor or simulator metadata.

Training can use state labels, target RGB, and localization supervision. Offline RGB routes report relative-state accuracy, gain over factual copy, action swap, and source-head checks; live controller cells additionally require task success, initial-state balance, generated-feedback usage, and native alignment.

A — Anchor dependence. A persistent factual state-1 anchor captured at the query pose is added to the input contract, making source-state observability testable. Because the anchor can leak the answer, lane A always reports a paired inference-time ablation: the same frozen checkpoints are re-evaluated with the anchor replaced by a fixed blank tensor, while the scoring reference is unchanged.

T — Transfer stress. Object names are remapped to generic task-family tokens (`openable`/`toggleable`) during training; evaluation uses object families whose images, names, poses, and labels never appear in training—Toaster and StoveKnob for `toggleable`, and Kettle for `openable`, i.e. three unseen families over two action families. Runtime action compatibility is verified first by an independent controller before any transfer claim is attempted.

4 Experimental setup

All live lanes run in AI2-THOR 5.0.0 CloudRendering (Kolve et al., 2017) at 192×192, field of view 90, visibility distance 1.5. Training data cover ten scenes with a scene-disjoint split (eight train scenes, FloorPlan4 validation, FloorPlan5 evaluation). FloorPlan5 was inspected during exploratory readout development. We therefore do not describe any result on this scene as sealed or untouched: S is explicitly post-hoc, and FloorPlan5 measurements in E3/E4 and the anchored task diagnostics are scene-disjoint descriptive checks. E3's 24 unseen starts and the separately evaluated held-out scenes provide the relevant evidence outside FloorPlan5. Protocol files and hashes are preserved as internal project records; no independent public timestamp is available, so "internally specified" below means documented in an internal protocol before the corresponding measurement, not externally timestamped public preregistration. The editor is an anchor-conditioned categorical persistent-state editor with a target-state-gated anchor skip and a full-frame residual RGB decoder. The transfer matrices use five fresh seeds, horizon $H{=}3$, and two initial states $\times$ two first actions $\times$ near/far preconditions per cell. Toaster and Kettle use the development-selected source-state threshold 0.455; StoveKnob uses 0.5 in its baseline, appearance-randomization and oracle-goal records. Thresholds match within each family's paired comparison. The StoveKnob oracle protocol lists 0.455, while execution records report 0.5; we preserve this discrepancy and treat the diagnostic as an archived controller-level check, not exact protocol-conforming confirmation. Reported experimental artifacts have dedicated record-audit scripts, and a cross-artifact closure audit (check count and version reported in Appendix H) ties documents to measurements, including this manuscript and independent recomputation of the offline analyses in Section 5.5. Full protocols, raw rows, and deterministic generators are listed in Appendix H.

5 Results

Table 1 summarizes the full decomposition; Figure 2 visualizes the central semantic-feedback split, anchor ablation, and bounded anchored horizon results. The axes remain separate and are not averaged. Figure-level details and raw counts are in Appendix B.

5.1 S: Factual semantic response is a post-hoc diagnostic

On the scene-disjoint FloorPlan5 evaluation, the frozen ResNet-50 probe reaches source-state accuracy 0.750, action-conditioned target accuracy 1.000, transition accuracy 0.750, and action-swap target accuracy 1.000 across five seeds (audit 11/11). Because the scene had been inspected during earlier exploratory readout development, these values are descriptive evidence rather than an untouched confirmatory result. An action-only rule (OpenObject$\rightarrow 1$, CloseObject$\rightarrow 0$) exactly predicts the 160 collected target labels (audit 160/160; Appendix H), so target accuracy does not establish that RGB contributes; source-state accuracy is the nontrivial visual readout. Transition accuracy also requires a correct source read. Action swap is an input-level perturbation, not an environmental causal test.

Table 1: Capability decomposition on AI2-THOR (all numbers from audited artifacts). Wilson 95% intervals are conditional descriptive intervals computed as if episode rows were independent; they do not model clustering inside seed×scene cells and are not used as confirmatory tests (Section 6). "Anchored" denotes the registered factual state-1 reference contract. The E2/E3 rows are offline reanalyses: they report no live simulator success.

Capability	Input contract	Result	Audit	Status
S: Semantic response	factual RGB + action	src .750 / tgt 1.000 / trans .750	11/11	supported narrowly
G: No-anchor feedback	RGB/action inputs	.375–.500; gain −.125–0	10/10	**not supported**
Anchored roll-in	factual anchor	state 1.000; gain .969	37/37	supported (anchored)
Scene transfer	per-scene anchors, thr. .455	40/40 cells; 744/800=.930	15/15	supported (anchored)
Anchor-drop	blank vs. factual anchor	100/200 vs. 200/200	9/9	diagnostic
Anchor dropout	retrained at p=.5	100/200; 0/5 seeds $\geq$ .65	12/12	persists
T: Toaster transfer	`toggleable` token	0/20 cells; 80/160=.500	7/7	**negative**
T: StoveKnob transfer	`toggleable` token	1/20 cells pass; 82/160=.512	7/7	**negative**
T: Kettle transfer	`openable` token	1/20 cells pass; 84/160=.525	7/7	**negative**
E2: Decision certificate	10 checkpoints; 24 starts	40/40 strict; 0/960 decisions changed	offline	necessity only
E3: Same-input readers	factual RGB + anchor	ridge 19/24, ResNet-18 15/24, pixel 12/24 unseen	28/28	baselines
Natural video	EPIC-KITCHENS	gates 0/5 and 0/5	12/12	**negative**

5.2 G: No-anchor generated feedback fails

For these offline routes, relative-state accuracy thresholds full-image MSE between the prediction and the initial factual frame. FloorPlan4 calibration gives $\tau = 4.17249 \times 10^{-4}$; the factual-image reader scores 30/32 on FloorPlan5. This image-change proxy is distinct from semantic source-head accuracy and live task success. The object-aware predictor and frozen-ResNet-50-guided full-resolution editor both score 0.375–0.500 on generated-history state accuracy, with gains −0.125–0 over factual copy. However, their factual-history predictions are already weak. A post-hoc reanalysis of the archived summaries pairs factual-history prediction and generated roll-in within each seed: mean accuracy changes from 0.463 to 0.450 for the object-aware predictor, and from 0.256 to 0.431 for the ResNet-50 editor (Table 3). Thus, these records show failure to pass the no-anchor gate; they do not isolate a feedback-induced loss of an otherwise strong RGB predictor. The routes also differ in training data and budget, so their comparison does not identify an architecture effect. The ResNet-50 editor's higher action-swap scores (0.781–0.906) do not rescue its failed generated-state, gain, and source-head criteria. The primary audit passes 10/10; the cross-route record audit passes 53/53. Summary arithmetic and the cross-route audit do not rerun inference. Subsequent trained temporal-retrieval and paired action-response repairs remain negative under the full G criteria (Appendix F).

5.3 A: The anchor closes a bounded, leaky sub-gate

Under the factual-anchor contract the editor reaches generated state 1.000, gain 0.969, action-swap 0.906, and source invariance 1.000 across five seeds (audit 37/37); with a calibrated threshold, the same contract replicates across four held-out scenes on five fresh seeds (40/40 cells, $744/800 = 0.930$, $[0.910, 0.946]$, audit 15/15). A second internally pre-specified replication on four further held-out scenes, fixed by a read-only scene-inventory rule before any run, passes under the same gate (40/40 cells, $694/800 = 0.8675$, $[0.842, 0.889]$, audit 15/15), extending the positive lane to eight held-out scenes.

The paired ablation quantifies the shortcut. Replacing only the model-visible anchor with a blank tensor—same checkpoints, same held-out rows, same scoring reference—drops performance from 200/200 to $100/200 = 0.500$ ($[0.431, 0.569]$); every state-1 target fails (0/100) while state-0 targets

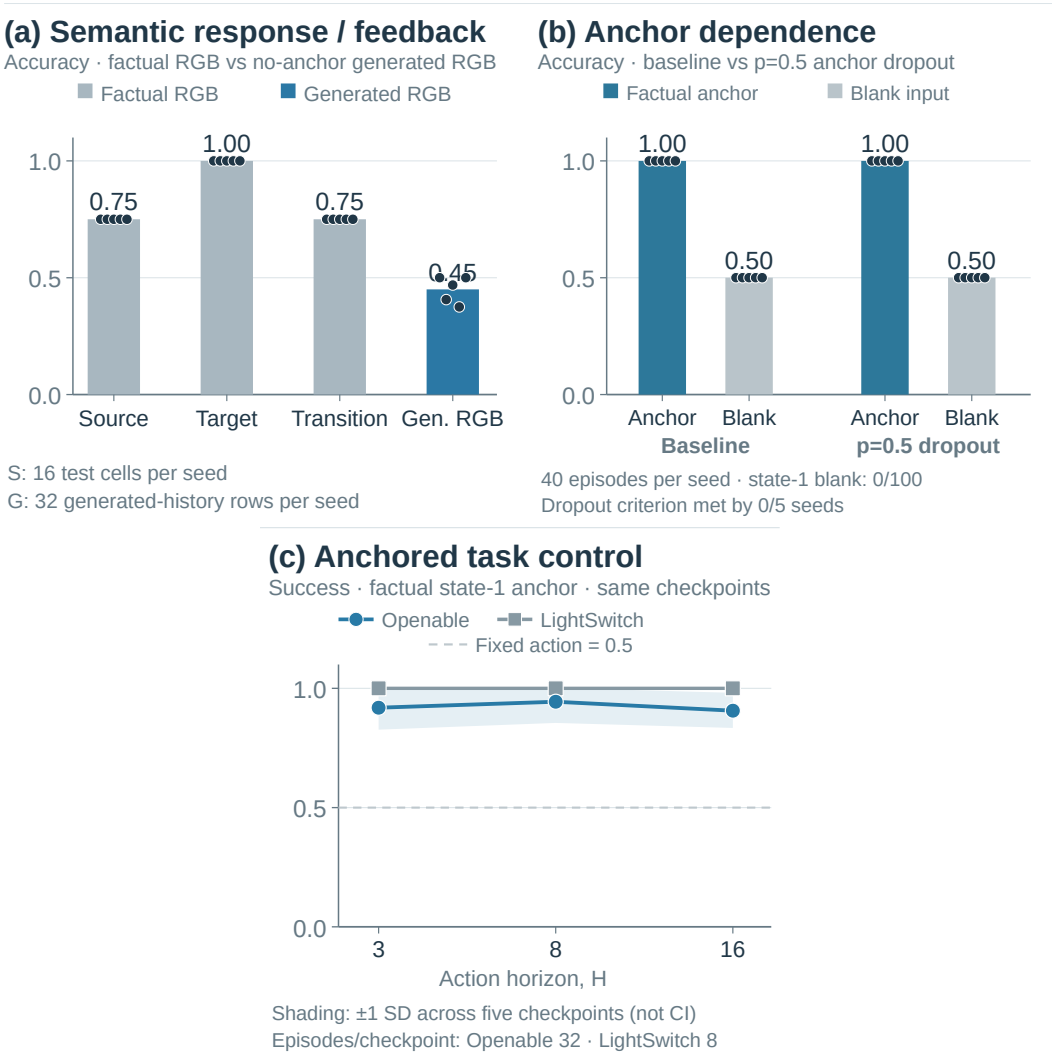


Figure 2: **Audited result profile.** (a) Factual semantic source/target/transition and no-anchor RGB relative-state accuracy under distinct contracts. (b) Factual-anchor and blank-input accuracy before and after $p = 0.5$ anchor-dropout training; bars are five-seed means and dots individual seeds. Each arm contains 40 episodes per seed, with blank state-1 accuracy $0/100$ in aggregate. (c) Factual-anchor task success at $H = 3, 8, 16$ using the same five checkpoints; shading is ± 1 sample SD across seeds, not a confidence interval. The dashed fixed-action control is 0.5. Openable/LightSwitch contain 32 and 8 episodes per checkpoint. All runs use the previously inspected FloorPlan5 scene. Data: `sfs_results_source_data.csv`.

remain perfect $(100/100)$. The scoring threshold is the session-calibrated validation midpoint of the frozen split, and re-scoring the same checkpoints under an explicitly recomputed calibration reproduces these counts seed by seed. The anchored result is therefore real but materially anchor-dependent, which is why we call it *anchored state restoration* and never unrestricted future-frame generation.

A paired training variant blanks the anchor with per-batch probability p=0.5, retaining factual loss targets and fixing data, budget, split and seeds. Each seed records 1,392 training forwards; realized drop fractions are 0.482–0.523 (audit 12/12). Teacher and generated RGB remain 1.000, gain 0.969–1.000, action swap and same-state control 1.000. The blank arm remains $100/200$ with $0/100$ state-1 successes; $0/5$ seeds reach the specified 0.65 criterion. LightSwitch generated-history source

accuracy falls to 0.500–0.875 (openable 0.750–0.9375). This single intervention fails to remove dependence; other rates, objectives, budgets and capacities remain untested.

5.4 T: ACTION COMPATIBILITY DOES NOT IMPLY OBJECT-FAMILY TRANSFER

Toaster passes the runtime gate: in all four transfer scenes an independent controller executes `ToggleObjectOn` / `ToggleObjectOff` from a visible pose and restores the original state; collection audit 8/8. The family-only editor (no Toaster data, generic `toggleable` token, five fresh seeds) then fails everywhere: all 20 learned cells reach 0.500 success with state-0 success 0, gain 0, and native alignment 0.500; $80/160 = 0.500$ episodes ($[0.423, 0.577]$); audit 7/7. The oracle control is 1.000, so the environment and action pair are fine—the learned visual transfer is not.

StoveKnob was fixed by a read-only scene inventory before evaluation; fallback families were allowed only on runtime infeasibility. The same five checkpoints are evaluated without retraining. Runtime restoration passes in all four scenes (collection audit 8/8), and the oracle control is 1.000. Yet $19/20$ learned cells fail; the single passing cell (0.750 success, gain 0.250) does not satisfy the all-cells criterion. Aggregate success is $82/160 = 0.5125$ ($[0.436, 0.589]$), with state-0 $2/80$ and state-1 $80/80$ (transfer audit 7/7).

An archived *oracle-goal* check supplies the true initial state to the controller's restoration goal: same checkpoints, executed threshold 0.5, same 160 episodes, and no intervention inside the model. It lifts every cell to 1.000 ($160/160$; state-0 $80/80$, state-1 $80/80$; diagnostic audit 8/8), including all 19 cells that failed. Its protocol lists 0.455 instead; Section 4 discloses this execution discrepancy. A corresponding Toaster check at its original 0.455 threshold makes all 20 Toaster cells reach 1.000, with $160/160$ (state-0 $80/80$, state-1 $80/80$; diagnostic audit 8/8). These goal-only checks localize the observed failure to controller-level goal inference rather than intervene on the model's internal source-state pathway.

An internally specified appearance-domain-randomization variant then tests whether that bottleneck can be repaired. The only training change is that each example's history frames are perturbed (brightness/contrast jitter plus one gray occlusion rectangle) with per-example probability 0.5; data, budget, and the five paired seeds are fixed, and evaluation is unperturbed. The variant still fails both transfer gates ($2/20$ Toaster cells pass, $88/160 = 0.550$, $[0.473, 0.625]$; $6/20$ StoveKnob cells pass, $104/160 = 0.650$, $[0.573, 0.720]$; audits 7/7 each), and its development gate fails only the generated-source-state sub-check, with two of five seeds falling to 0.500. State-0 success rises from $2/80$ to $56/80$ on StoveKnob and from $0/80$ to $8/80$ on Toaster. The StoveKnob gain comes with state-1 regression from $80/80$ to $48/80$, yielding a net improvement of 22 episodes out of 160. However, the clustered reanalysis finds no reliable contrast in any family, and the registered all-cells gates remain failed (Appendix A).

Kettle extends the stress test to the openable action family. The inventory fixes the only eligible unseen openable type, episode grid and threshold 0.455. A documented feasibility revision replaces inaccessible FloorPlan15 with the next feasible scene, FloorPlan22; the partial collection is retained. Baseline and appearance-randomization arms fail: $1/20$ cells and $84/160 = 0.525$ ($[0.448, 0.601]$) versus $2/20$ and $88/160 = 0.550$ ($[0.473, 0.625]$). State-0 success is $4/80$ versus $8/80$, with state-1 $80/80$ in both arms; collection audit 8/8 and transfer audits 7/7 each. The oracle control is 1.000. A separately specified oracle-goal diagnostic uses the same episodes, checkpoints and threshold (re-running the same 20 baseline learned cells, no internal-state intervention) and supplies the true initial state only to the controller's goal. It lifts all 20 Kettle cells to 1.000 ($160/160$ episodes, state-0 $80/80$, state-1 $80/80$; diagnostic audit 8/8 remote and local). All three tested families therefore show the same controller-level goal-inference failure pattern across both action families; this is pipeline localization, not an internal-state causal intervention.

5.5 OFFLINE CHECKS: A CERTIFIED INVARIANCE AND A NON-GENERATIVE READER

The analyses below are offline: they replay archived checkpoints and observations without calling a renderer or simulator, and report no live simulator success. E2/E3 were specified in internal protocol records before analysis, and the later ResNet-18 architecture-breadth control was separately documented before measurement; these records are not independently timestamped and are listed in Appendix H. These analyses do not alter a registered live lane, threshold, or historical artifact.

E2: affine dominance certificate and history intervention. For a fixed goal d and the implemented two-action pair, a checkpoint's action score is $q_a(p, d) = (1 - p)\,T[a, 0, d] + p\,T[a, 1, d]$, where T is its softmax transition table and p the source-state probability. This is affine in p, and the runner verifies every table entry exceeds 10^{-6} so the implementation's 10^{-6} probability clamp is inactive and the expression is affine on the closed interval; strict dominance at both endpoints $p \in \{0, 1\}$ then certifies strict dominance for every $p \in [0, 1]$, with no sampling grid and no grid-based inference (Appendix I). Across 10 archived generic-family checkpoints (two training variants $\times$ five seeds), two entity types and two goals, all 40 certificates are strict, with minimum endpoint margin 0.0346 against a 10^{-7} tolerance; endpoint ties and sign crossings are recorded as *no certificate*, never as invariance, so the test can fail.

We then intervened on the history the controller actually consumes. On 24 archived starts (8 per family $\times$ three families) and six three-frame input arms (1,440 rows), the initial learned goal is frozen from the initial factual RGB plus the anchor, one frame is generated with the trajectory's registered first action, and the next decision is recomputed with the newest generated frame replaced by four non-alias alternatives (blank, repeated initial, repeated anchor, spatially flipped generated frame). The controller does consume that frame—generated and identical-alias arms agree exactly, with no mismatch recorded—yet no compared decision changes: 0 of 960 under the learned goal and 0 of 960 under the true goal, while the source-state probability moves in 840 of the 960 comparisons (max absolute change 0.97386). A full-forward versus registered-scoring-path check, one candidate pair per checkpoint, agrees to $\leq 5.96 \times 10^{-8}$, inside the 10^{-6} tolerance required before the exact path is used. The reading is a negative necessity result: on these archived decisions, with the initial goal fixed, the certified action choice does not depend on the newest generated frame. It is not evidence that generated feedback is usable, it makes no live-success claim, and it holds for the generic-family checkpoints only.

E3: same-input non-generative readers. Because the initial goal is computed from the initial factual RGB and the factual state-1 anchor, we ask whether a non-generative reader of the same two inputs can read the initial state. A ridge least-squares reader on the 14 anchor-contrast features (first transition per trajectory only; 80 training rows from eight scenes, 10 FloorPlan4 validation rows, 10 previously inspected FloorPlan5 diagnostic rows; $\lambda = 1$; standardization from training scenes only; threshold 0.689 chosen on validation balanced accuracy and frozen before any unseen scoring; no held-out-family tuning) reaches $9/10$ on held-out FloorPlan5 and $19/24$ on the 24 unseen starts. FloorPlan5 was previously inspected during exploratory readout development, so its E3 score is descriptive rather than untouched confirmatory evidence. A deliberately parameter-free comparator—mean absolute RGB difference from the anchor at most $1/255$ implies state-1, otherwise state-0—reaches $9/10$ and $12/24$. Matched learned readers on the *same* 24 starts and 10 checkpoints pool $131/240 = 0.546$, but that denominator reuses 24 images across 10 checkpoints and is not 240 independent scenarios. A simple non-generative reader of the same observable inputs at least matches learned reading here, so the initial reading gate is not specific to generation; had it failed we would have retained that as a negative. A separately internally specified architecture-breadth control freezes the convolutional backbone of an independently trained ResNet-18 spatial observer, discards its original head, and fits the same ridge rule to a 2,048-dimensional signed contrast of pooled initial and anchor features. With the same eight-scene/FloorPlan4/FloorPlan5 split and the same 24 unseen starts, it reaches $9/10$ on the previously inspected FloorPlan5 diagnostic and $15/24$ unseen (independent audit 28/28). This is weaker than the hand-engineered ridge result but above its internally specified descriptive chance count; it shows that readability is not unique to the 14-feature family. It remains an offline readout built from the same observable pair, not a second video generator or live task success.

5.6 Natural video remains negative

Two EPIC-KITCHENS forecasting readouts use frozen ImageNet ResNet-18 and ResNet-50 encoders and heads trained on 13,644 EPIC rows. The unseen-participant validation subset has 1,065 rows across 22 videos and two derived kitchen clusters. The gate requires both verb and noun top-1 accuracy to exceed training-majority predictions by 0.10, with at least four of five seeds passing; both routes achieve $0/5$ (feature audit 12/12). Appendix G gives accuracies, baselines and limitations. This is a separate forecasting check, not transfer of the AI2-THOR model to natural video.

5.7 CONTROLLED REACTIVE RGB: A BOUNDED CALIBRATION DIAGNOSTIC

A separate controlled reactive RGB benchmark uses a three-decision receding-horizon policy: after each action, the selected candidate's generated last three frames become the next policy history. The policy receives no privileged simulator inputs; native state is used only for the final task audit. Applying the fixed inference-time correction $\lceil \text{step} + 0.5 \rceil$ to five frozen ceiling checkpoints yields mean task success 0.719 and paired gain 0.250 over factual-copy across 32 episodes per seed (16 per confirmation family); family success is $0.750/0.6875$, versus 0.688 for fixed action 3 and 0.875 for the oracle (independent audit 11/11). Relative to the best fixed action, the improvement is only $1/32 = 0.03125$ per seed; seeds reuse the same 32 scenes. This is a post-hoc inference-time calibration of fixed checkpoints: placing the same rule inside fresh training passes only $1/5$. It is a distinct controlled result, not a change to the registered AI2-THOR G result or evidence of general learned feedback.

6 STATISTICAL SCOPE

Across the six registered learned arms (960 episodes), success equals correct restoration-goal reading on every audited row; the three oracle-goal diagnostics raise each tested family to 1.000. This is a finite-sample identity and a controller-pipeline localization, not identification of an internal or environmental mechanism. Appearance-randomization contrasts are descriptive: clustered tests are non-significant and all crossed-bootstrap intervals include zero (Appendix A). Wilson intervals are conditional row-wise descriptions and do not model seed$\times$scene clustering. The original threshold-0.5 transfer result is retained; the 0.455 threshold was selected on development scenes only.

7 LIMITATIONS

Semantic shortcut and cross-model comparison. In S, the binary target is determined by Open/Close action, so its 1.000 score does not establish visual contribution. S and G use separate systems; we do not claim a within-checkpoint dissociation. **Privileged anchor.** The positive route receives a factual state-1 image; one registered training-time dropout variant (p=0.5, five paired seeds, audit 12/12) leaves the blank-anchor arm at $100/200$ and $0/5$ seeds at criterion. This single intervention does not exclude under-training or establish structural dependence. **Simulator and sample limits.** Live lanes use AI2-THOR, fixed binary objectives, and simulator-specific families; each transfer family uses four scenes and five seeds. FloorPlan5 was inspected during exploration, so E3/E4 and anchored-task measurements from that scene are descriptive diagnostics rather than untouched confirmatory evidence. The pooled $131/240$ reader result reuses 24 starts across 10 checkpoints. Clustered contrasts are sensitivity analyses, not coverage guarantees, and non-significance is not equivalence. **Architecture and transfer.** Positive anchored results use one editor. The two negative G routes use different training data and budgets; neither is a controlled architecture ablation. E3/E4 readers are not additional generators. Toaster, StoveKnob, and Kettle are bounded stress tests of this route, not an impossibility theorem. Oracle-goal interventions localize failure to goal inference in this controller, not to environmental causal mechanisms or generated-pixel content. **Offline and natural video.** E2/E3 are offline; the certificate is conditional on a fixed goal, two actions, and ten checkpoints. The controlled reactive RGB positive result depends on post-hoc inference-time calibration of frozen checkpoints and did not replicate at the same rate when embedded in fresh training. EPIC-KITCHENS is a bounded negative route with unavailable public test labels.

8 DISCUSSION

Interactive video benchmarks should report semantic probes, generated-observation roll-in and reference dependence separately. Our distinct systems motivate this separation; they do not demonstrate within-model dissociation. A factual reference supports bounded restoration, while transfer remains negative. The controlled RGB calibration passes only $1/5$ fresh-training seeds, leaving learned reliability open. The SFS scorecard and released audits make these contracts inspectable.

AI USE STATEMENT

Generative AI assisted research discussion and study design, record organization, implementation of training, evaluation and audit scripts, experiment execution under human direction, statistical interpretation, mathematical claims and proof drafting, figure design and production, manuscript drafting and editing, and internal review. The evidence-traceability workflow was informed by Scientific Agent Skills (Kassis et al., 2026). The authors remain responsible for checking code, results, citations, derivations and claims against archived artifacts and for final verification of all AI-assisted work. No independent verification of the AI-assisted derivations is claimed. Experiments produced records through deterministic scripts; no LLM generated synthetic datasets.

ETHICS STATEMENT

Evaluations use standard simulators and existing public datasets, with no new human-participant recruitment or data collection and no autonomous deployment on physical robots. We address the risk of overclaiming world-model competence from semantic scores through explicit input contracts and controls.

REPRODUCIBILITY STATEMENT

Appendices H and F index protocols, outcome records, training budgets and audits. The supplement includes manuscript source, figures and plotting inputs, training/evaluation code, prediction traces and audit scripts. It excludes model checkpoints, raw experimental images and reader feature vectors: records support count and certificate recomputation, but the package alone does not reproduce training, model inference or reader fitting. Internal protocol timestamps are not independently verified; the StoveKnob oracle threshold discrepancy is disclosed in Section 4 and Appendix G. Original input hashes and anonymized-copy hashes are both retained. Audits establish record consistency rather than independent experimental validation.

REFERENCES

Ossama Ahmed, Frederik Träuble, Julius von Kügelgen, Francesco Locatello, Niels Görnitz, Bernhard Schölkopf, Yoshua Bengio, and Stefan Bauer. CausalWorld: A robotic manipulation benchmark for causal structure and transfer learning. In *International Conference on Learning Representations*, 2021.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2017.

Daniel M Bear, Elias Kubilius, Hsiao-Han Fan, Sicheng Wang, Tomer Ullman, Joshua B Tenenbaum, and Daniel LK Yamins. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Ziqi Huang, Xiaojie Zhang, Xiaowei Liu, Ruizheng Xia, Wenyu Shi, Chuang Gan, Fangcheng Zhang, Chi-Man Chan, et al. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Timothy Kassis, Vinayak Agarwal, Yuhuan He, Darshil Patel, and Aubrey M. Brueckner. Scientific agent skills: A library of procedural knowledge for research agents. *arXiv preprint arXiv:2609.00065*, 2026. doi: 10.48550/arXiv.2609.00065.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Aniruddha Gordon, Yuke Zhu, et al. AI2-THOR: An interactive 3d simulation platform for visual reasoning. *arXiv preprint arXiv:1712.05474*, 2017.

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, Datasets and Benchmarks Track*, 2025.

Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, Lei Bai, Wanli Ouyang, and Ruimao Zhang. WorldSimBench: Towards video generation models as world simulators. *arXiv preprint arXiv:2410.18072*, 2024.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *International Conference on Learning Representations*, 2016.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Xin Zhang, Yinzhou Tang, Chen Gao, Wei Wu, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, and Yong Li. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026a.

Yu Shang, Yinzhou Tang, Yiding Ma, Zhuohang Li, Lei Jin, Weikang Su, Xin Jin, Zhaolu Wang, Ziyou Wang, Xin Zhang, Haisheng Su, Weizhen He, Wei Wu, Haoyi Duan, Gordon Wetzstein, Xihui Liu, Dhruv Shah, Zhaoxiang Zhang, Zhibo Chen, Jun Zhu, Yonghong Tian, Tat-Seng Chua, Wenwu Zhu, Chen Gao, and Yong Li. WorldArena 2.0: Extending embodied world model benchmarking on modality, functionality and platform. *arXiv preprint arXiv:2605.17912*, 2026b.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenge. In *NeurIPS Workshop on Deep Generative Models*, 2019.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 2009.

Meiqi Wu, Zhixin Cai, Fufangchen Zhao, Xiaokun Feng, Rujing Dang, Bingze Song, Ruitian Tian, Jiashu Zhu, Jiachen Lei, Hao Dou, Jing Tang, Lei Sun, Jiahong Wu, Xiangxiang Chu, Zeming Liu, and Kaiqi Huang. Omni-WorldBench: Towards a comprehensive interaction-centric evaluation for world models. *arXiv preprint arXiv:2603.22212*, 2026.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

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.

A Detailed failure analysis and statistical reading

The paired lanes localize the bottleneck to the controller's restoration goal rather than to action execution: semantic targets are predictable from the action alone; a privileged witness makes a narrow restoration contract solvable; removing the witness collapses exactly the state-1 targets; and three unseen object families spanning two action families fail despite executing their actions. Internally specified oracle-goal interventions—which supply the true initial state to the controller's goal only—lift all three families to 1.000, so the failure sits in goal inference from the initial factual

RGB plus the anchor, not in execution; that is a localization inside our own pipeline and not a claim about what the generated pixels contain. A registered appearance-randomization variant partially repairs state-0 reading (StoveKnob $2/80 \rightarrow 56/80$) without passing the transfer gates, and the third family shows the same state-1-perfect/state-0-collapsed signature with no material DR repair. A $4/8$ binary cell has Wilson interval $[0.215,\ 0.785]$; we report exact denominators everywhere and treat small cells as bounded diagnostics, and every Wilson interval in this manuscript is a conditional descriptive interval that treats rows as independent and does not model seed×scene clustering, so it is neither independence-aware nor confirmatory. The original threshold-0.5 transfer result ($38/40$) is preserved as a negative measurement; the calibrated threshold 0.455 was selected on development scenes only.

An exact decomposition, and a generic information bound. Let X_0 be the observable input from which the controller infers the initial state—the initial factual RGB together with the factual state-1 anchor—let g be that inference, $R = \{g(X_0) = S_0\}$ the event that the restoration goal is read correctly, and Y the event of task success. Total probability gives the exact identity

$$\Pr[Y] \;=\; \Pr[R]\,\Pr[Y \mid R] \;+\; \Pr[\neg R]\,\Pr[Y \mid \neg R], \tag{1}$$

which assumes nothing about independence. Across the six registered learned arms (960 episodes) the audit finds no row with success but no correct read and no row with a correct read but no success, so $\Pr[Y \mid R] = 1$ and $\Pr[Y \mid \neg R] = 0$ and (1) collapses to $\Pr[Y] = \Pr[R]$: a finite-sample fact about these audited rows, not a modelling assumption. Consistently, the three oracle-goal interventions—which set the controller's goal to the true initial state and touch nothing else—lift every cell of all three families to 1.000.

We deliberately do not convert this into an information ledger. The standard bound for a balanced binary S_0 read from any observable X_0 states that for any decision rule,

$$\Pr[\widehat{S}_0(X_0) = S_0] \;\le\; \tfrac{1}{2} + \sqrt{\tfrac{\ln 2}{2}\, I(S_0; X_0)}, \tag{2}$$

with I in bits (proof in Appendix I). We state it as a generic proposition only: it constrains every rule and every observable, and we estimate no mutual information, report no empirical bits, and claim neither that information is absent from generated history nor that a higher measured accuracy proves more information was transferred. What remains are measured counts: the three baseline families (Toaster, StoveKnob, Kettle) read state 1 on $80/80$ episodes against state-0 reads of 0, 2, and 4 of 80—a constant or near-constant state-1 reading—while the appearance-randomization variants move StoveKnob state-0 reads to $56/80$ and its state-1 reads to $48/80$, a genuine but gate-insufficient change of read behaviour, and Toaster and Kettle state-0 reads to $8/80$. The clustered reanalysis below supplies the conditional reference calculations; the earlier independence-based calculations remain in historical artifacts but are not used for inference.

Clustered paired reanalysis. Episodes inside a seed×scene cell form a cluster rather than independent draws, so the three paired appearance-randomization contrasts (Toaster, StoveKnob, Kettle; 160 matched episodes and 20 seed×scene cells each, all rows and the original thresholds retained) were recomputed with conditional reference calculations. Mean per-cell gains in task success are 0.05, 0.1375 and 0.025. Exhaustive two-sided sign-flip reference p-values—conditional on sign symmetry and exchangeability of the paired units, hence not experimental randomization tests—are 1.0, 0.25, 1.0 at the seed level (32 assignments over five seeds) and 0.5, 0.125, 1.0 at the scene level (16 assignments over four scenes); Holm-adjusted across the three families within each grouping, none is rejected at 0.05. Paired crossed seed/scene bootstrap percentile intervals (10,000 replicates, fixed seed, both arms sharing draws) are $[0.0,\ 0.2]$, $[0.0,\ 0.325]$ and $[0.0,\ 0.1]$, all containing zero. With four scene clusters that are not a random sample of environments these are sensitivity analyses rather than coverage guarantees, and non-significance is not equivalence: all three registered all-cells gates remain failed.

Table 3: Archived within-route history controls. TF predicts RGB from factual history; GR feeds predictions back. Entries are means over five seed summaries, each scoring 32 linked transitions on inspected FloorPlan5. Δ is the paired GR-minus-TF difference. These are descriptive summary-level calculations, not image rescoring. Epochs and training examples differ between routes; this is not an equal-budget architecture ablation.

Route	Epochs	Train examples	TF	GR	Δ
Object-aware	100	256	0.463	0.450	-0.013
ResNet-50 editor	20	384	0.256	0.431	$+0.175$

B FULL CAPABILITY MATRIX

Table 2: Extended capability matrix with raw binary counts and Wilson 95% intervals (rows regenerated from audited artifacts by `scripts/build_paper_result_matrix.py`).

Capability	Contract	Primary result	Raw count (95% WI)	Status
Factual semantic response	factual RGB + action/entity, teacher-forced	source 0.750; target 1.000; transition 0.750; swap 1.000	—	supported narrowly
No-anchor generated RGB	RGB history only	0.375–0.500; gain -0.125–0	—	not supported
Anchored generated roll-in	+ persistent factual state-1 anchor	state 1.000; gain 0.96875; swap 0.90625	—	supported (anchored)
Calibrated fresh scene transfer	per-scene anchors; threshold 0.455; $H{=}3$	40/40 cells; openable 0.750–1.000; LightSwitch 20/20 at 1.000	744/800 = 0.930 (0.910, 0.946)	supported (anchored)
Additional held-out scenes	same anchored contract; threshold 0.455	40/40 cells; eight held-out scenes across both replications	694/800 = 0.8675 (0.842, 0.889)	supported (anchored)
Anchor-drop dependence	blank vs. factual anchor at inference	factual 200/200; blank 100/200; blank state-1 0/100	100/200 = 0.500 (0.431, 0.569)	diagnostic
Toaster unseen transfer	generic toggleable token; Toaster test-only	0/20 learned cells; state-0 success 0; gain 0; alignment 0.500	80/160 = 0.500 (0.423, 0.577)	negative
StoveKnob unseen transfer	generic toggleable token; threshold 0.5	1/20 cells; state-0 2/80	82/160 = 0.5125 (0.436, 0.589)	negative
Kettle unseen transfer	generic openable token; threshold 0.455	1/20 cells; state-0 4/80	84/160 = 0.525 (0.448, 0.601)	negative
Natural video (N1)	participant-disjoint RGB readout	N1 gate 0/5; ResNet-50 route 0/5	—	negative

C SFS SCORECARD

Table 4: State–Feedback Separation scorecard (generated by `scripts/build_sfs_scorecard.py` from audited artifacts). Descriptive only; axes are never averaged into a scalar.

Axis	Definition	Observed result	Audit	Status
S	factual RGB + action-conditioned semantic state	src 0.750; tgt 1.000; trans 0.750; swap 1.000	11/11	supported narrowly
G	no-anchor generated-RGB roll-in	relative-state 0.375–0.500; gain -0.125–0	10/10	not supported
A	paired factual-anchor vs. fixed blank input	factual 200/200; blank 100/200; blank state-1 0/100	9/9	anchor-dependent diagnostic
T	generic task-family tokens on three unseen families	Toaster 0/20; StoveKnob 1/20; Kettle 1/20 cells; thresholds 0.455/0.5/0.455	8/8, 7/7	negative

D INTERNALLY SPECIFIED CONTRACTS AND FALSIFICATION OUTCOMES

Table 5: Hypotheses, falsifying outcomes, and observed diagnostics.

Lane	Hypothesis	Falsifying outcome	Observed
Factual semantic	factual RGB + action predicts audited target	below semantic gate	target 1.000, transition 0.750 (narrow positive)
No-anchor RGB	generated RGB preserves state without reference	at/below copy baseline	0.375–0.500, gains ≤ 0 (negative)
Anchored	factual witness makes restoration testable	state/gain controls fail	state 1.000, gain 0.969 (narrow positive)
Anchor-drop	performance should not hinge on the reference	large drop without anchor	200/200 $\rightarrow$ 100/200; state-1 0/100 (dependence)
Anchor-dropout training	anchor dropout at $p = 0.5$ removes the dependence	blank arm fails to reach recovery criterion	100/200 again; 0/5 seeds ≥ 0.65 (dependence persists)
Toaster	generic family token transfers	balance/gain/alignment fail	0/20 cells (negative)
Natural video	frozen RGB forecasting meets the unseen-participant gate	gates below threshold	0/5 and 0/5 (negative)

E BENCHMARK SCOPE COMPARISON

Table 6: Scope of selected world-model evaluations. The final column states the additional diagnostic reported in this case study, not a claim of exclusive novelty.

Work	Stated evaluation focus	Additional diagnostic in this study
WorldSimBench (Qin et al., 2024)	Generated-video perception and video-to-control evaluation in embodied tasks	Action-only target-label audit and factual-anchor removal on a binary restoration task
World-in-World (Zhang et al., 2026)	Online planning and task success in closed-loop environments	Separately scored source-state readout and generated-RGB roll-in, with their distinct-system boundary
WorldArena / 2.0 (Shang et al., 2026a;b)	Perception and task utility; 2.0 adds visuotactile, interactive-RL, and cross-platform robot evaluation	Paired anchor-dependence measurement and registered object-family stress tests
Omni-WorldBench (Wu et al., 2026)	Interaction effects on intermediate state evolution and final outcomes	Row-level artifact audits and a post-hoc action-only label check for the controlled probe
WorldModelBench (Li et al., 2025)	Instruction following and physical adherence in generated-video judgments	A bounded task-level case study with explicit no-anchor and privileged-reference contracts
SFS (this study)	Registered binary-state diagnostics in AI2-THOR; separate S/G systems; limited EPIC-KITCHENS negative route	Not a general-purpose benchmark; no within-checkpoint S/G dissociation or broad external validation

F EXPLORATORY NO-ANCHOR RGB DEVELOPMENT

These experiments extend the no-anchor diagnostic with actual training and saved-checkpoint replay. They use the same previously inspected FloorPlan5 scene and its 32 transitions; they are exploratory development, not new untouched confirmation. TF predicts from factual history, GR feeds generated frames back, and swap changes the requested action while retaining the history. All predictions use three RGB history frames plus the existing action and entity inputs. Native labels, masks and future targets are restricted to training or evaluation; no renderer, future target or privileged anchor enters inference. The historical full G criteria and thresholds remain unchanged.

F.1 TEMPORAL RETRIEVAL IMPROVES FACTUAL-HISTORY READOUT BUT FAILS ACTION CONTROL

A learned convex mixture of the three existing chronological history images and synthesized RGB was compared with an RGB-only control at seed 3601. Both arms train 384 native rows for 100 epochs (batch 8, AdamW learning rate 10^{-4}), with identical edit-content losses and allocated heads. Selection uses FloorPlan4 TF accuracy, then TF RGB MSE, then earliest tie; GR and FloorPlan5 never select an epoch. The first history-mixture candidate passes its fixed factual-history prerequisite, triggering the four previously listed candidate-only seeds. This is one paired comparison plus four candidate replications, not a five-seed architecture ablation.

Table 7: Complete temporal-history development record. Entries are correct readouts out of 32 per condition on the same diagnostic transitions. Val. is FloorPlan4; TF, GR and swap are FloorPlan5. Initial factual copy scores 16 in every run. Epochs are selected by validation alone.

Arm	Seed	Epoch	Val. TF	TF	GR	Swap
RGB-only	3601	74	18	16	16	16
History mix	3601	96	24	24	20	7
History mix	3613	50	26	30	26	2
History mix	3631	59	23	24	25	8
History mix	3643	58	23	26	25	6
History mix	3659	79	26	23	21	4

Across the five mixture models, TF is $127/160$, GR $117/160$ and swap $27/160$; only $2/5$ seeds meet both validation and diagnostic TF prerequisites. These are five models evaluated on the same 32 transitions, not 160 independent observations. An action-free middle-history copy already scores TF $23/32$, GR $16/32$ and swap $9/32$. In $24/32$ transitions, the target state occurs in factual history. All eight unseen-target rows are first transitions, where TF and GR inputs coincide. Historical retrieval can therefore raise factual-history readout without competent action-sensitive synthesis; these records do not establish general feedback damage or novel-state feedback stability.

F.2 FOUR PAIRED ACTION-RESPONSE REPAIRS FAIL THE FIXED PREREQUISITES

The initial native same-history plan contains 64 source-state pairs, but preflight finds only eight with exactly identical network RGB histories; it is rejected before training. Revised studies use those eight unique pairs across two training scenes and four entities, each repeated eight times per epoch. Equality is at the 192×192 float32 network-input resolution, not a claim about hidden simulator states. Each arm trains for 100 epochs, with 384 ordinary native rows and 64 additional pair exposures: 512 supervised images and 64 optimizer updates per epoch, batch 8, AdamW learning rate 10^{-4}. Repeated exposures are not independent samples. Budgets are equal within each pair; they differ from the temporal study.

Response loss. Both arms concatenate the action one-hot directly into all three output heads and receive paired image reconstruction. Only the extra open-minus-close response loss differs (weight 0 versus 0.5), normalized by the native target-response mean squared magnitude with a 10^{-5} floor.
Constructed context. Both arms use response weight 0.5; one preserves native histories and the other constructs four past-frame patterns from the current image and the native opposite-state training target, preserving the latest factual input. Training labels select the opposite-state image. This is

supervision-derived augmentation, not a newly observed simulator trajectory. **Warm context.** Both arms start from the designated first temporal checkpoint (seed 3601); the added action columns are zero-padded, with all other tensors unchanged. Both share its 100-epoch pretraining cost before 100 further epochs. **Output routing.** Both warm arms allocate two identically initialized branches per output head and use constructed contexts; one averages branch logits and the other selects by the action one-hot, retaining upstream action FiLM. This changes conditional capacity and optimization dynamics, not routing alone. Warm/cold comparisons do not isolate an equal-total-budget architecture effect.

All eight checkpoints are selected on FloorPlan4 alone: maximize the minimum of TF and action-swap accuracy, then their sum, then minimize TF RGB MSE, with earliest exact tie. This differs from the preceding TF-only selector. Expansion requires validation TF, validation swap and diagnostic TF each at least 24 correct out of 32; no GR score selects an epoch.

Table 8: All eight completed action-response repair runs. Every entry is a correct count out of 32; copy is 16 throughout. Budgets and initialization match within each adjacent pair. All eight fail the fixed expansion prerequisites. Different seeds and warm/cold costs prevent pooling these as independent replications or attributing cross-pair differences to architecture.

Study / arm	Seed	Epoch	Val. TF	Val. swap	TF	GR	Swap
Response / off	3701	61	18	18	17	16	16
Response / on	3701	87	20	17	21	17	14
Context / native	3801	66	19	17	16	15	17
Context / constructed	3801	61	19	22	14	11	16
Warm context / native	3901	61	19	17	21	16	8
Warm context / constructed	3901	37	20	19	15	14	13
Output heads / shared	4001	83	19	19	16	13	19
Output heads / routed	4001	62	19	19	17	14	14

None $(0/8)$ passes; no additional seeds are launched. Post-hoc replay on all 256 linked training rows gives action-swap counts ranging from 80 to 142, so the gap cannot be attributed solely to held-out scene transfer. This does not identify a general optimization or capacity cause. The competent action-sensitive RGB predictor required for a stronger feedback comparison remains missing.

F.3 Verification, data follow-up and release boundary

All six temporal weights were replayed on GPU, matching 82 recorded fields each (492 total); all eight repair weights match 83 each (664 total). Frozen input checks cover 3619 files per cold replay and 3621 per warm/routed replay, including initialization. Original backbone/source-head tensors are unchanged. Gradient preflights check finite nonzero decoder/action response and routing semantics; warm initialization equivalence holds within 2×10^{-6} on checked training contexts. Local weight and fixed-grid fingerprints match their records. These audits verify the saved runs, not additional independent observations or scientific validity.

A separate read-only audit of all 384 native training rows identifies 12 unique identical-history/opposite-action groups across three training scenes. A fingerprinted pair bank is prepared but not used to train or select any reported model. No exact RGB/entity input group has conflicting source-state labels in this inspected training set. This narrow absence of exact-pixel label conflict does not establish general observability, feature separability or the cause of failure.

The supplement includes raw metrics, prediction traces, replay records, training-fit summaries, protocols and this deterministic table generator. It excludes weights and raw images, so the released records support recounting and consistency checks rather than end-to-end model reruns. Exact study and protocol paths are indexed in the released reports: `artifacts/temporal_rgb_mixture_study_v1/REPORT.md`, `artifacts/action_response_rgb_study_v1/REPORT.md`, and `artifacts/ai2thor/native_training_rgb_pair_coverage_v1/results.json`. The appendix generator is `scripts/build_rgb_development_appendix.py`. No full G success, broad causal mechanism, natural-video generalization or complete interactive world model follows from this development batch.

G EXECUTION CONTRACTS AND NATURAL-VIDEO REPORTING

These tables are derived from archived records. Source-state thresholds are matched within each family, but differ across families. The StoveKnob oracle protocol specifies 0.455 while execution records use 0.5. This discrepancy was found during manuscript review; the original protocol and execution records are preserved. That diagnostic is an archived controller-level check, not exact protocol-conforming confirmation.

Table 9: Executed transfer thresholds and initial-state success counts. Each arm contains 80 episodes per initial state. Baseline/DR/oracle use the same executed threshold within a family.

Family	Threshold	Baseline (s0,s1)	DR (s0,s1)	Oracle (s0,s1)	DR total
Toaster	0.455	0/80; 80/80	8/80; 80/80	80/80; 80/80	88/160
StoveKnob	0.5	2/80; 80/80	56/80; 48/80	80/80; 80/80	104/160
Kettle	0.455	4/80; 80/80	8/80; 80/80	80/80; 80/80	88/160

The EPIC-KITCHENS measurements train forecasting heads on $13{,}644$ rows from 495 training videos (67,217 official training annotations available); the unseen-participant subset contains 1,065 validation rows from 22 videos, with only two derived kitchen clusters. Three frames precede action start by 1.0, 0.5 and 0.1 seconds. The original frozen ResNet-18 head trains for 20 epochs; the second route uses a frozen ResNet-50 encoder and a temporal MLP. These are separately trained EPIC readouts, not simulator-to-video transfer of the AI2-THOR model.

Table 10: Unseen-participant forecasting accuracy ranges over five seeds. Both verb and noun must exceed their training-majority baselines by 0.10; the declared route gate requires at least four of five seeds to pass.

Encoder	Verb top-1	Noun top-1	Seeds passing
ResNet-18	0.1549–0.2188	0.0451–0.0742	0/5
ResNet-50	0.2225–0.2516	0.0714–0.0967	0/5

The verb/noun majority baselines on this subset are 0.1587/0.0366. The recorded kitchen bootstrap resamples only two clusters; its intervals are sensitivity summaries, not reliable population coverage guarantees. Public hidden-test labels are unavailable, so file coverage is not test-set performance. Both negatives are confined to these readouts, training data and gates.

The reproducible table source is `scripts/build_submission_reporting_addendum.py`; its record-level output is `artifacts/submission_reporting_addendum_v1/results.json`. The original StoveKnob protocol is retained, with a dated disclosure in `docs/SUBMISSION_REPORTING_DEVIATIONS_V1.md`.

H ARTIFACT AND AUDIT INDEX

The separate controlled reactive RGB calibration diagnostic is traceable to `docs/REACTIVE_TRAJECTORY_CEILING_MARGIN_CALIBRATION_CONFIRMATION_PROTOCOL.md`, its result at `artifacts/reactive_trajectory_generated_history_policy_rgb_object_warp_ceiling_margin_confirmation_v1/REPORT.md`, and the independent 11/11 audit at `artifacts/reactive_trajectory_generated_history_policy_rgb_object_warp_ceiling_margin_confirmation_audit_v2/REPORT.md`. The fresh in-training comparison is at `artifacts/reactive_trajectory_generated_history_policy_rgb_object_warp_ceiling_margin_v1/REPORT.md`; these artifacts remain outside the registered AI2-THOR result matrix. The architecture-distinct no-anchor editor is documented at `artifacts/ai2thor/factorial_resnet50_state_editor_v1/REPORT.md`; the cross-route evidence consistency audit is `artifacts/research_gap_evidence_audit_v1/REPORT.md` (53/53 checks; record consistency rather than independent model-score recomputation). The paired history summaries and their reproducible

arithmetic are at `artifacts/no_anchor_history_controls_v1/results.json` and `scripts/analyze_no_anchor_history_controls.py`.

Key artifacts (all paths relative to the released repository): factual semantic transition and audit (`resnet50_semantic_transition_v1`, `_audit_v1`); the post-hoc action-only label check (`resnet50_semantic_action_only_audit_v1`); no-anchor predictor (`rgb_transition_predictor_v7` + audit); anchored editor (`anchored_rgb_persistent_editor_v3_five_seed` + follow-up audit); calibrated fresh-seed transfer (`anchored_multiscene_fresh_seed_replication_threshold0455_v1` + audit, 56 source-inference mismatch rows retained as diagnostics) and its internally pre-specified new-scene extension (`anchored_multiscene_new_scenes_replication_threshold0455_v2` + audit, protocol `AI2THOR_ANCHORED_MULTISCENE_NEW_SCENE_REPLICATION_PROTOCOL.md`); anchor-drop ablation (`anchored_anchor_drop_ablation_v1` + audit); anchor-dropout training variant with its paired ablation and independent contract audit (`anchored_rgb_anchor_dropout_editor_v1`, `anchored_anchor_drop_ablation_dropout_v1`, `anchored_anchor_dropout_training_audit_v1`, protocol `AI2THOR_ANCHORED_RGB_ANCHOR_DROPOUT_TRAINING_PROTOCOL.md`); Toaster protocol, collection, editor, transfer and audits, plus the second-family scene inventory, protocol, collection, transfer, audits and internally pre-specified oracle-goal diagnostics for both families (artifacts retain their `oracle_source` names; `unseen_toggle_family_*`, with the StoveKnob run at `_v2`, the StoveKnob diagnostic at `transfer_oracle_source_v1`, the Toaster diagnostic at `transfer_oracle_source_toaster_v1`, protocols `AI2THOR_UNSEEN_TOGGLE_FAMILY_ORACLE_SOURCE_DIAGNOSTIC_PROTOCOL.md` and `_V2.md`, and the internally pre-specified appearance-randomization intervention `*_appearance_dr_v1` with its editor, transfer, audits and protocol `AI2THOR_UNSEEN_TOGGLE_FAMILY_APPEARANCE_DR_PROTOCOL.md`); the third-family openable extension (protocol, collection, both registered transfer arms and audits at `unseen_openable_kettle_*`, with the retained feasibility-revision artifact `_partial_floorplan15_infeasible`, its internally pre-specified oracle-goal diagnostic rev 2 at `oracle_source_v1` with audit, protocol `AI2THOR_UNSEEN_OPENABLE_KETTLE_PROTOCOL.md`); the registered source-state information analysis with its independent audit (`source_state_information_analysis_v1`, `_audit_v1`, protocol `SOURCE_STATE_INFORMATION_ANALYSIS.md`); the offline strengthening batch registered in `docs/SFS_STRENGTHENING_PROTOCOL_V1.md`, comprising the crossed-design reanalysis `artifacts/sfs_crossed_design_v1/results.json` (script `scripts/analyze_sfs_crossed_design_v1.py`) and the decision-certificate and reader-baseline run `artifacts/sfs_feedback_certificate_v1/results.json` with its decision rows `rows_e2.json` and `initial_readings.json` (script `scripts/evaluate_sfs_feedback_certificate_v1.py`), together with the immutable pre-strengthening snapshot of this manuscript at `artifacts/sfs_manuscript_pre_strengthening_v1/main.tex`; EPIC-KITCHENS N1 forecasting and ResNet-50 transfer; and the cross-artifact closure audit `research_closure_consistency_v98` (130/130), which also checks that this manuscript cites the audit version being produced and that its abstract matches the text released for the submission field.

I PROOFS: INFORMATION BOUND AND ENDPOINT DOMINANCE CERTIFICATE

Information bound (generic). Let X_0 be the observable input—the initial factual RGB together with the factual state-1 anchor—let S_0 be the balanced binary initial state, and let $p(X_0) = \Pr[S_0 = 1 \mid X_0]$; let $\hat{S}_0(X_0)$ be any (possibly randomized) decision rule. Then $\Pr[\hat{S}_0 = S_0] = \frac{1}{2} + \mathbb{E}\big[\Pr[S_0 = \hat{S}_0(X_0) \mid X_0] - \frac{1}{2}\big] \le \frac{1}{2} + \mathbb{E}\,|p(X_0) - \frac{1}{2}|$, with equality for the MAP rule $\hat{S}_0 = \mathbf{1}\{p(X_0) > \frac{1}{2}\}$. Pointwise, Pinsker's inequality applied to $\mathrm{KL}\big(\Pr[S_0{=}\cdot \mid X_0] \,\big\|\, \Pr[S_0{=}\cdot]\big)$ (in nats) gives $|p(X_0) - \frac{1}{2}| \le \sqrt{\mathrm{KL}(X_0)\,/\,2}$, and Jensen's inequality for the concave square root yields $\mathbb{E}\,|p(X_0) - \frac{1}{2}| \le \sqrt{\mathbb{E}\,\mathrm{KL}(X_0)\,/\,2} = \sqrt{\ln 2 \cdot I(S_0; X_0)\,/\,2}$, since S_0 is uniform over $\{0, 1\}$ and $\mathbb{E}\,\mathrm{KL}(X_0) = \ln 2\, I(S_0; X_0)$ when the mutual information is expressed in bits. Rearranging gives Equation (2). The bound holds for an arbitrary rule and an arbitrary observable, and is used in the text as a generic proposition: no mutual information is estimated from data and no empirical bit value is reported. Its inversion would give $I(S_0; X_0) \ge 2(a - \frac{1}{2})^2 / \ln 2$ bits for a population accuracy $a \ge \frac{1}{2}$;

we record the inversion only for completeness and draw no information-repair or information-increase claim from it, because the inequality is one-directional and a is a finite-sample estimate. Equation (1) is the law of total probability and requires no assumption beyond measurability.

Endpoint dominance certificate. Fix a goal state d and two implemented actions a, b of one checkpoint, and let T be that checkpoint's softmax transition table, so the probability action a assigns to d at source-state probability $p \in [0, 1]$ is $q_a(p, d) = (1 - p)\, T[a, 0, d] + p\, T[a, 1, d]$. The difference $\Delta(p) = q_a(p, d) - q_b(p, d)$ is affine in p and therefore determined by its endpoints: $\Delta(p) = (1 - p)\, \Delta(0) + p\, \Delta(1)$, a convex combination of $\Delta(0)$ and $\Delta(1)$. Hence if $\Delta(0) > 0$ and $\Delta(1) > 0$ then $\Delta(p) \geq \min\{\Delta(0), \Delta(1)\} > 0$ for every $p \in [0, 1]$: a strictly dominates b uniformly in p, and symmetrically when both endpoint margins are negative. No sampling grid is used, and no finite grid can replace this endpoint argument, since an affine function can change sign inside $[0, 1]$ only by differing in sign at the endpoints. Two implementation conditions are verified by the runner. First, every entry of T must exceed 10^{-6}: the implementation applies a 10^{-6} lower clamp to the mixed probability, and when all entries exceed 10^{-6} that clamp is inactive, so Δ is exactly affine rather than piecewise affine and the convex-combination argument applies verbatim. Second, a margin is declared strict only when both endpoint margins exceed the numerical tolerance 10^{-7} with the same sign; an endpoint tie ($|\Delta(0)| \leq 10^{-7}$ or $|\Delta(1)| \leq 10^{-7}$) or opposite endpoint signs is recorded as *no certificate*, never as dominance and never as invariance. The certificate is conditional on the fixed goal, the implemented two-action pair and the checkpoint tested; it says nothing about other action pairs, other goals, or the information content of generated frames.

J SFS BEYOND AI2-THOR: MAPPING NOTES

SFS axes are environment-agnostic in form: S asks whether factual observations support the environment's action semantics; G asks whether self-generated observations survive roll-in; A asks which privileged references carry the result; T asks whether task-level abstractions transfer to unseen visual families. On our controlled reactive RGB benchmark, the analogous profile is consistent: factual readouts pass, post-hoc inference-time calibration passes on $5/5$ frozen decoder checkpoints (success 0.719, gain 0.250 over factual copy) on new scene families, whereas embedding the same correction in fresh training passes only $1/5$ seeds. We report this as a mapping diagnostic only; the registered claims remain those of the AI2-THOR lanes.