AUDITING MISALIGNMENT IN MULTI-MODAL WORLD MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

An omnimodal world model can predict one physical outcome and generate another. We distinguish internal alignment with the model's own forecast from external alignment with a physical reference, using matched predictive contracts over events, magnitudes and available timing. Two intervention ladders progressively supply the model's own contract (A) or the physical contract (B), from scene-only prompts to event, magnitude and keyframe instructions. In a matched Super push cohort, $39/40$ jointly valid pairs have correct text but an incorrect measured video endpoint; other source-bound cases exhibit agreement on an incorrect observable. Richer conditioning changes motion occurrence without establishing reliable quantitative repair. We extend the audit to fixed-budget supervision: own-text screening raises push planning success from 0.338 to 0.635. At matched retained action support, the gated rows have lower mean label and fitted-grid errors, but additional task success over the 0.615 control remains unresolved. The resulting framework connects physical forecasting, generation and learning while keeping agreement, successful repair, and downstream benefit as distinct claims requiring distinct evidence.

1 INTRODUCTION

An omnimodal world model can tell two physical stories about the same scene. Its language pathway may predict that a released object will move and stop, while its video pathway generates an object that barely moves. A correct verbal answer is therefore not sufficient evidence that the generated future obeys the same dynamics. Conversely, two outputs can agree and both be wrong. These distinctions matter when generated futures are used for planning or converted into training supervision.

We separate two requirements. **Internal alignment** asks whether the generated video realizes the same checkpoint's explicit text or image–text prediction, T_{text}. **External alignment** asks whether it matches an analytic or recorded physical reference, T_{phys}. Both comparisons use the same measured video observable. Checking the forecast against physics first tells us whether disagreement is a candidate quality signal or whether agreement could preserve a mistaken reference. The two requirements have different meanings even when their targets coincide.

Physical-video benchmarks and understanding–generation benchmarks address related requirements (Bansal et al., 2025; Wang et al., 2026a). We ask how those requirements interact under controlled physical contracts: whether adding a self-derived or physical target closes the corresponding gap, and what the remaining disagreement implies for supervision.

Our central hypothesis is that **expressing a physical prediction and realizing it in video are separable capabilities**. We test not only whether their outputs disagree, but whether supplying the missing prediction closes the gap. This motivates two paired intervention ladders: **A** supplies the model's own predicted contract, whereas **B** supplies the physical contract. Information increases from **T0**, scene and parameter, through **T1**, event, **T2**, magnitude, and **T3**, keyframes/timing. If missing explicit information is sufficient to explain a failure, richer conditioning should improve the corresponding measured alignment. Failure to do so challenges that explanation, not the model's capabilities in all settings.

The resulting questions preserve a single progression: **Q1**, does video match physics? **Q2**, does it match the model's own prediction? **Q3**, does the A ladder reduce internal error? **Q4**, does the B

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

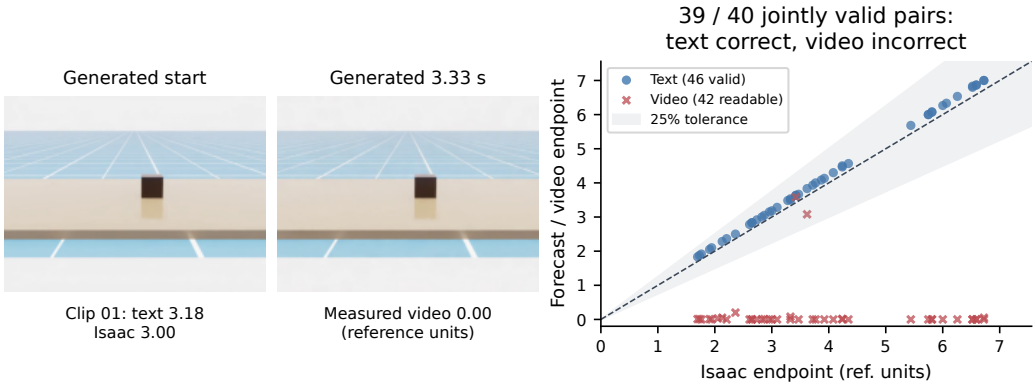


Figure 1: **One checkpoint, incompatible physical outcomes.** Left: unaltered frames from the lowest-ID near-static clip with a passing text forecast in the matched Super push cohort. Right: valid text predictions and readable video endpoints compared with executed Isaac outcomes. In $39/40$ jointly valid pairs, text passes but video fails the endpoint criterion. This diagnoses the baseline gap before testing either conditioning ladder. Unreadable clips are not assigned physical errors; distances use the fixed-scene .06 m reference scale.

ladder reduce external error? We extend that foundation with **Q5**: can the measured discrepancy help screen generated supervision before it damages a downstream learner? Q5 tests an intervention on the training data, not a claim that the generator has been repaired or that improved labels automatically improve control.

The evidence distinguishes these questions rather than giving them a single positive answer. The same Super checkpoint predicts correct push endpoints while $39/40$ jointly valid videos miss them. Nano examples show that agreement with a wrong forecast can also be misleading. Added instructions alter event occurrence, but the A/B results do not establish reliable quantitative repair. In fixed-budget learning, own-text screening raises push success from 0.338 to 0.635. An action-matched analysis finds lower label and fitted-grid errors, but does not resolve an extra task-success gain. Thus the downstream results extend the alignment audit; they do not substitute for Q1–Q4 or identify a latent reasoning mechanism.

Contributions. (1) A physical-contract framework that makes internal and external alignment comparable without conflating their references. (2) A tiered A/B intervention design, with correct and incorrect directives, that tests whether explicit self-derived or physical information closes the corresponding gap. (3) A controlled extension from generated observables to known-action supervision, separating label quality, fitted prediction error, and task success. Source-bound counterexamples, negative repair results, and selection controls specify when each conclusion is warranted. The contribution is this linked experimental audit, not a new filtering algorithm or a claim of universal repair.

2 Related Work

Physical fidelity and cross-modal consistency. VideoPhy and VideoPhy-2 evaluate semantic adherence and physical commonsense in generated video (Bansal et al., 2025; 2026); WorldModelBench evaluates instruction following and physics adherence (Li et al., 2025). Physics-IQ compares generated videos with real-video references and analyzes metric limitations (Motamed et al., 2026). GIR-Bench, RealUnify, and GapEval study consistency or synergy between understanding and generation in unified multimodal models (Li et al., 2026; Shi et al., 2026; Wang et al., 2026a). Our focus is narrower than a general cross-modal benchmark: a matched physical observable, an explicit environment reference, and the consequences of admitting the measured output as supervision. We do not claim that cross-modal consistency or fixed-size synthetic-data mixtures are new evaluation concepts.

Controlling generation and learning from it. Force Prompting and PhyCo learn physically meaningful conditioning, while WMReward uses a latent world model for inference-time physics alignment (Gillman et al., 2025; Narayanan et al., 2026; Yuan et al., 2026). WorldEcho diagnoses action following and WorldSync aligns generation for policy improvement (Chen et al., 2026). Interactive World Simulator studies policy training and evaluation with generated demonstrations, including fixed-size real/generated mixtures (Wang et al., 2026b). We instead leave generation unchanged when testing admission. Cosmos 3 supplies the omnimodal checkpoints (NVIDIA, 2026b;a); OpenVLA and LoRA supply supporting learners (Kim et al., 2025; Hu et al., 2022). Our evidence concerns their declared experimental uses, not a causal localization of physical knowledge inside a shared network.

3 Physics Fidelity Framework

3.1 One measured contract, two alignment references

Let x describe the initial scene and θ the physical intervention, such as restitution, friction, launch speed, or pendulum length. The generated video is $V(x, \theta, p)$ under prompt p. The same checkpoint's explicit prediction is $T_{\text{text}}(x, \theta)$, and the specified physical reference is $T_{\text{phys}}(x, \theta)$. We represent a predictive contract as $C = (z, m, u, K, \mathcal{I})$: event z, measurable magnitude m, units u, available keyframes/timing K, and required scene invariants $\mathcal{I}$. A reader extracts supported video coordinates $C_{\text{video}} = E(V)$. An answer about period cannot validate amplitude; an endpoint match cannot certify the entire trajectory or scene. Object-relative units do not remove the need to qualify scale, perspective, and object identity.

Reasoning fidelity compares the explicit forecast with the physical reference. The two alignment tests then hold the measured video fixed and change only the reference. On a valid common coordinate representation with metric d and fixed positive scales,

$$D_{\text{int}} = d(C_{\text{video}}, T_{\text{text}}), \quad D_{\text{ext}} = d(C_{\text{video}}, T_{\text{phys}}), \quad D_{\text{ref}} = d(T_{\text{text}}, T_{\text{phys}}). \tag{1}$$

The triangle inequality then gives

$$|D_{\text{ext}} - D_{\text{int}}| \leq D_{\text{ref}}. \tag{2}$$

The last inequality is an elementary reference-quality condition, not a learning guarantee. If text and physics coincide, the distances coincide numerically while answering different questions. Empirical separation requires references that differ. Low internal error then need not imply physical accuracy.

Unavailable text and unreadable/censored video receive no continuous error; attempted, text-valid, video-readable, and jointly valid counts remain separate. An invalid attempt can be non-passing without being declared physically wrong. A valid scalar in an incoherent response supports only a labeled component comparison. Reader support and inspection of scene invariants are separate: a numerical extraction does not certify that objects have not duplicated or changed geometry.

3.2 A/B ladders: intervene on the information given to generation

The two ladders share the physical scene and intervention but differ in the source of the supplied outcome. At each tier ℓ, A adds fields from T_{text} and B adds fields from T_{phys} (Table 1). The A question is whether increasing specificity reduces D_{int}; the B question is whether it reduces D_{ext}. The measurable coordinate and its scoring rule stay fixed across the comparison. Adding keyframes to a prompt does not imply that every temporal field has been independently measured.

We also use **W1**, a deliberately incorrect event directive, and **W2**, an incorrect magnitude directive, while holding the physical condition fixed. They test the direction of language responsiveness, not whether following a false instruction constitutes correct physical inference. A forecast is retained as returned; missing fields are not filled from physics, and inconsistent fields restrict the scope of the comparison. In particular, when A and B targets coincide, their effects cannot identify a difference between self-reference and physical-reference information.

3.3 Testbed, measurement, and the downstream extension

The core mechanism audit uses Cosmos3-Super, controlled physical interventions, and declared analytic or recorded Isaac/PhysX references. It varies restitution, friction, speed, and pendulum

Table 1: **Two references, the same information ladder.** A uses the model's own answer; B uses the physical reference. T0 is shared, and the tested A/B event-only targets coincide, so their T1 outputs are not independent replications. These are prompt-information tiers, not policy classes.

Tier	A: self-derived contract	B: physical contract
T0	Initial scene, physical parameter; no stated outcome	
T1	Predicted event z_{text}	Physical event z_{phys}
T2	Event + own magnitude m_{text}	Event + physical magnitude m_{phys}
T3	Event + magnitude + own keyframes K_{text}	Event + magnitude + physical keyframes K_{phys}

length across neutral sweeps and paired prompt tiers. Observables are first-rebound apex, slide/stay and visible incline travel, horizontal stopping distance or projectile range, and pendulum amplitude rather than period. Canonical magnitude acceptance uses $|m - r| \leq \max(.5, .25|r|)$ when the target is above measurement resolution; incline uses regime agreement. The matched-push and admission tolerances are specified separately below, not pooled into a single accuracy measure. The source-bound record also contains a Nano A/B study that includes distinct self and physical targets at T2/T3. Bounce, incline, horizontal push/slide, projectile, pendulum, vial lift, and driving have distinct roles rather than forming seven interchangeable benchmarks (Table 13).

The expanded paired Super push cohort has 48 attempts and executed command references at friction .5. The separate known-action learning bank has 60 readable rows at friction .3, including 23 moving clips. This permits a stronger admission comparison than a nearly all-static bank, but is not a prospective mediation study linking each error in the first cohort to a particular policy failure. Sampling units and reused forecasts are not pooled.

Magnitude measurement is scene-qualified. The canonical reader passes six of six event checks but only three of six 20%-magnitude checks; a separate push validation includes a 21.2% endpoint error. Push uses fixed-scene metric calibration, while its generation prompts omit explicit metric cube width. Analytic targets do not exactly match every recorded simulator trajectory. These qualifications constrain absolute-error attribution; complete reference, visibility and calibration checks are in Appendices D and K.

4 Alignment and Its Response to Conditioning

4.1 Q1–Q2: accurate articulation does not ensure realization

For each matched push command, the model receives the generation prompt plus a stopping-distance question; it is not supplied an Isaac endpoint or a computed answer. Text, image–text, and video are compared with an execution of that command. Passing means at most 25% endpoint error under the declared scale. Among 40 jointly valid text/video pairs, 39 have passing text and a failing measured video endpoint; one passes in both modalities (Figure 1). This supports both gaps in the accurate-forecast regime without treating unreadable video as physically wrong. Because the two references nearly coincide here, this result does not itself separate their judgments.

Across all 48 attempts, text passes on 46, image–text on 47, and video on 2. Text-only answers are valid on 46 attempts and videos readable on 42. Forty of the 42 readable videos travel less than .2 reference units. Thus the primary result is predominantly failure to realize motion, not comprehensive coverage of dynamic physical reasoning. Supplying the starting image to the text query does not remove the discrepancy.

Projectile supplies a separate observable and mechanism. Matched Super range forecasts differ from analytic targets by less than .6% across five speeds, while the 40-video sweep has 29 readable non-launches and 11 degenerate outputs. At 4 m/s, the forecast is 20.40 ball-diameters against an analytic 20.387; three fully tracked example videos have zero measured range. This is further realization-gap evidence, not separation of conflicting references or validation of every low-speed simulator trajectory (Appendix G).

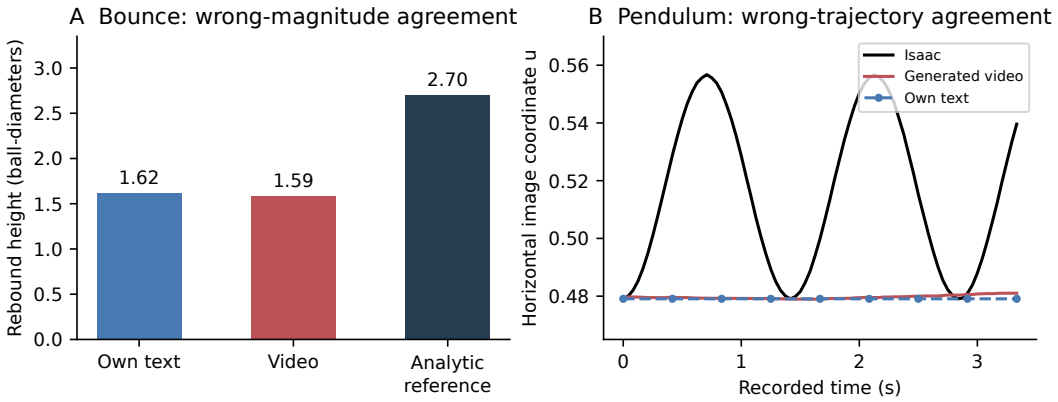


Figure 2: **Two observable-level counterexamples to agreement as a guarantee.** A: a selected moving Nano rebound agrees with its erroneous own *magnitude*, not the analytic target. This is an A/T2 prompt, not a neutral-generation rate or proof that self-conditioning caused agreement. B: in a separate synchronized Nano pendulum case, text and video predict nearly stationary centers while Isaac oscillates. Complete cells, original tolerances, reader sensitivity, and identified artifacts are retained in the appendix.

4.2 When internal and external judgments separate

A wrong forecast changes the interpretation of agreement. In an archived Nano bounce query, restitution .9, drop height .20 m, and ball diameter .06 m imply a first rebound of $.9^2(.20/.06) = 2.70$ ball-diameters. The model predicts 1.62. An A/T2 video conditioned on that own magnitude measures 1.59 (Figure 2A): it passes the original own-reference window and fails the physical-reference window. All 81 frames are tracked; the checked diameter sensitivity range, 1.522–1.903, preserves those decisions.

This comparison concerns magnitude only: the raw textual keyframes are inconsistent with its nonzero rebound. A separate image–text query at restitution .5 forecasts 5.00 while a neutral video measures .80 against a physical .833, giving the reverse magnitude decision. The source modes are explicitly bound; they are not pooled into one text-only accuracy estimate. Both cases were found retrospectively and received unblinded, not independent generated-domain review.

A synchronized Nano panel makes the boundary visible without choosing a new gate. On the same eight non-initial keyframes, two seed-234 pendulum cases have text–video errors .00262/.00234 but video–Isaac errors .02001/.04061 in normalized image coordinates. The forecasts repeat the initial center; the simulator oscillates. A fixed RGB reader reproduces the discrepancy found by SAM2. Both lengths and both seeds remain reported, including a scene-deformation case. The conclusion is narrow but consequential: a measured observable can agree with an unreliable forecast. No population false-admission rate or whole-scene validity certificate follows.

4.3 Q3–Q4: do richer A/B contracts close the intended gap?

The ladders make the information-only explanation testable: compare the same supported observable as event, magnitude and keyframes are supplied, keeping parameter, conditioning image and generation seed fixed. The Super correct/ incorrect-directive subset contains 144 clips across three mechanisms, six conditions (T0–T3, W1–W2), and eight seeds per cell. Its event-level response is clear. At incline friction .2, T0 and T1 each give 0/8 moving clips; T2 and T3 give 3/8 and 6/8. Pendulum W1 suppresses motion to 3/8, versus 6–7/8 under correct tiers. Language can therefore change occurrence in both directions without changing the physical facts. This Super subset tests correct physical and wrong directives: it informs Q4 and language responsiveness, not a differential A-versus-B effect.

That is not equivalent to closing the magnitude gap. A T2 bounce instruction specifying 2.7 diameters gives three moving outcomes between .62 and 2.36, one static outcome, and four invalid

ones. Pendulum T2 gives median amplitude 7.03 despite a four-unit directive. Incline approaches a visible-ramp ceiling, not verification of the entire commanded displacement. Parameter-response curves and complete T0–T3/W1–W2 results remain in Figure 8 and Tables 6–7.

Where self and physical targets differ, the 192-record Nano study includes A and B prompts with different supplied targets in some T2/T3 comparisons (Table 2). It contains four mechanisms, two sources, three instructed tiers, and eight seeds. More specificity does not reliably improve the corresponding acceptance rate. The small cells and inconsistent model keyframes preclude attributing every failure to a lack of generator knowledge; they do rule out treating more explicit conditioning as an automatic solution. Event-only A/B videos are byte-identical in all 32 axis/seed pairs and are not additional independent trials.

Table 2: **Do the A and B ladders improve their intended alignment?** Nano records, eight attempts per cell. A is scored against its own reference; B against physics. Entries give passing counts at T1/T2/T3 under the original observable rule (incline: event agreement), not whole-scene validity. Magnitudes use each mechanism's declared object-relative units. Full cross-reference counts and invalid/readability accounting appear in Table 11.

Mechanism	Own magnitude	Physical	A: internal passes	B: external passes
Bounce	1.62	2.70	0/1/1	0/0/1
Incline	4.70	4.70	2/5/4	2/5/4
Projectile	1.41	20.39	0/0/0	0/0/0
Pendulum	10.00	4.23	2/2/2	4/4/4

A separate 96-generation feedback study reaches the same practical boundary: discrepancy feedback improves over neutral prompting but fails its declared negative-control comparison, with occurrence 0.250 versus 0.3125 for within-mechanism permuted feedback. The control retains physical information, so this does not prove discrepancies are never used. Together, the tier and feedback results fail to establish reliable quantitative repair, rather than invalidating the diagnosis of a gap (Appendix F).

This motivates a different intervention. Instead of asking generation to repair itself, we use its measured agreement to decide which outcomes enter training. Q5 evaluates that extension without assuming the A/B repair questions have a positive answer.

5 Q5: From Alignment Measurements to Training Supervision

5.1 A controlled assay that keeps the commanded action known

Clean rows are $(a, h(a))$, using recorded Isaac outcomes. Generated rows are (a, o), where $o = E(V_a)$ and a remains the command supplied to the rollout. For this scalar specialization of the contract, write $t(a)$ and $h(a)$ for its model and physical outcome references. No action is inferred backward from the video. The primary learner is five-neighbor regression on the scalar command; images enter only through goal perception. A forward learner $\hat{h}$ predicts outcomes; a grid planner selects the command whose prediction is closest to the goal $\hat{g}(I)$ read from a physical image marker:

$$\hat{a}(I) = \arg\min_{a \in \mathcal{A}} |\hat{h}(a) - \hat{g}(I)|, \quad \text{success} \iff |h(\hat{a}) - g| \leq .25g. \tag{3}$$

Push uses 27 previously executed commands and 24 starting/goal configurations (three starts, eight goals). Scoring looks up recorded outcomes under a translation-invariance assumption; it is not fresh execution of each choice. Goal perception is calibrated on these scenes, with split-reader checks in Appendix I. This deliberately small assay isolates the consequence of changing outcome labels; it is not a claim that learned control is necessary for the scalar task.

Fixed initial budget. Every paired arm starts with $N = 44$ rows, of which $n_w = \text{round}(fN)$ are generated and $N - n_w$ are clean. Generated rows *replace* clean rows rather than being added to them. All arms share the clean subset and initial generated draw. Rejection is not replenished, so the final retained dataset may be smaller. We sample with replacement from the fixed banks; repeated occurrences are not new generated clips. The primary comparison is screening versus ingesting all generated outcomes, not a requirement that generation outperform additional clean data.

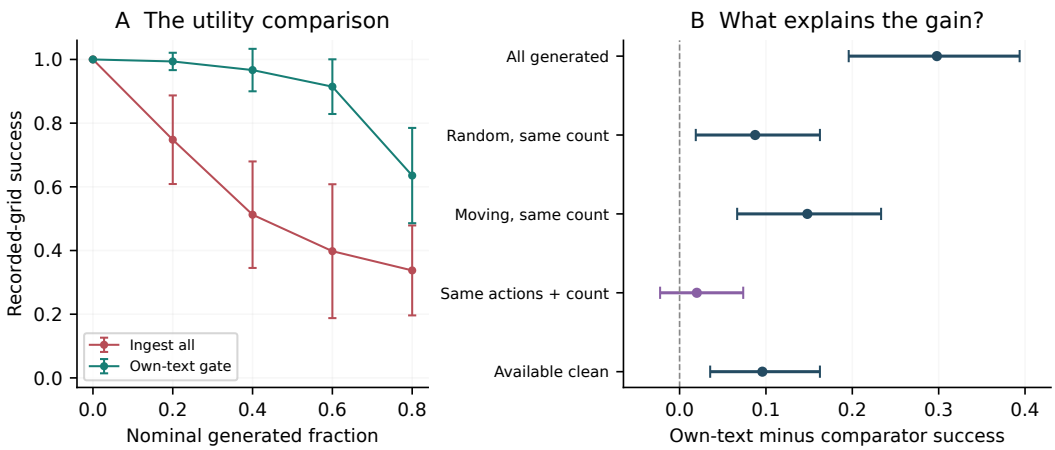


Figure 3: **Filtering utility and its mechanism require different controls.** A: push kNN dose response, with means and draw standard deviations over the same fixed banks. B: gate-minus-comparator success at $f = .8$, with pointwise 95% paired bootstrap intervals over 20 draws. Count/motion baselines are outperformed; after matching retained action multiplicities as well, the remaining difference is unresolved. The action-matched comparison is post hoc and averages every eligible subset within each draw, not ensemble predictions or additional independent training runs.

Own-forecast admission. The 60-readable-row push bank has 13 unique command conditions. We query the same Super checkpoint using each exact source prompt plus a stopping-distance question and reuse the answer across identical conditions. Twelve text-only forecasts are valid; the remaining truncated answer rejects its rows rather than receiving an analytic substitute. All valid forecasts pass the local reference check. This extension uses an already inspected bank; its query plan was locally fixed before the extra forecasts, not before the original outcomes.

For admission, $o, t(a), h(a)$ are numerical values in .06 m reference units, and the denominator floor uses that same scale. Parsed forecasts are nonnegative; all valid forecasts in this bank are positive. The scalar gate accepts a readable outcome when

$$q(a, o) = \mathbf{1}\{t(a) \text{ valid}\}\mathbf{1}\{|o - t(a)| \le .30 \max(t(a), 10^{-6})\}. \tag{4}$$

It needs no per-rollout environment answer at admission time, but it still needs forecast inference and a calibrated observable. This reference-relative window is a decision rule, not the common-scale metric in Eq. (1). Even an accepted label only satisfies $|o - h(a)| \le .30 \max(t(a), 10^{-6}) + |t(a) - h(a)|$; reference reliability is a prerequisite, not something acceptance establishes.

5.2 Outcome corruption matters; screening mitigates its cost

At nominal $f = .8$, each draw contains nine clean and 35 generated rows. In Figure 3, ingest-all kNN success is 0.338, compared with 0.635 after own-text admission. The paired difference is $+0.298[+0.196, +0.394]$ over 20 dataset draws. The same rule induces the stated-law gate's acceptance set on this bank: only two of 60 rows pass, both commanded at 1.8 m/s. Thus model forecasting supplies the reference here, but it has no demonstrated advantage over a correct analytic reference.

To test whether the failure comes from outcome labels rather than the sampled actions alone, we replace generated outcomes by recorded Isaac outcomes *at the identical sampled commands*, without changing row counts or planner. Success rises from .338 to .992. This privileged intervention directly diagnoses label damage in the assay; it is neither an affordable data-acquisition method nor proof of why the filter helps. Its role is to separate an outcome-corruption claim from an action-draw claim.

5.3 CONTROLS FOR REMOVAL, MOTION, AND ACTION SUPPORT

The first alternatives are straightforward. A uniform subset matching the retained count reaches 0.548; a moving-only subset with the same count reaches 0.487. Keeping *all* moving clips gives 0.188. Screening therefore beats these particular removal and motion controls, but they leave an important confound: the gate also changes which commands remain.

We address that confound with a retrospective action-and-count-matched sensitivity. For every original draw, preserve the gate's number of retained occurrences at each action and average success over every eligible subset of that draw. Clean rows, action multiplicities, reader, and planner are unchanged; only which outcomes occupy the retained command slots varies. The control reaches 0.615 versus 0.635 for the gate, with difference $+0.020[-0.023, +0.074]$. This does not resolve an additional within-command *task-success* benefit. It also does not establish equivalence: the interval permits a modest benefit or harm. Because it uses the gate's selected action profile, this is a privileged diagnostic, not a proposed deployable selector.

To distinguish numerical supervision from decision performance, a second post hoc diagnostic measures retained-label mean absolute relative error and forward-model mean absolute error over the recorded action grid. Both are lower for the gate at matched action support (Table 3), while the task-success contrast remains unresolved. The grid includes training commands; this is fitted-grid error, not held-out scene generalization. These metrics were requested after the task-level control and are descriptive, not a preregistered mediation test.

Table 3: **Lower observed label and fitted-grid errors, unresolved extra task success.** Same action/count-matched comparison. Error reductions are control minus gate; success gain is gate minus control. Pointwise 95% paired bootstrap intervals condition on the fixed bank. Label error excludes six zero-retention draws without imputation; prediction and task metrics use all 20. Grid error uses .06 m reference units; label relative error is dimensionless.

Metric (paired draws)	Gate	Matched	Reduction / gain [95% interval]
WM label relative error (14)	0.136	0.424	$+0.288[+0.174, +0.400]$
Grid MAE (20; ref. units)	0.960	1.048	$+0.088[+0.018, +0.159]$
Task success (20)	0.635	0.615	$+0.020[-0.023, +0.074]$

The distinction answers Q5 as *local damage mitigation relative to ingest-all*. Gated rows and their fitted predictors have lower errors than the post hoc matched control, but an extra task-success benefit beyond that support is not established. No latent reasoning mechanism is identified by these comparisons. Moreover, screening cannot invent coverage: in an all-generated kNN diagnostic it reduces success from .317 to .075. Available-clean-only success at $f = .8$ is 0.540, while an analytic controller and a post hoc direct-forecast planner both attain 1.000 on the push goals. These are boundaries of a supervision-quality assay, not reasons to present generated video as a necessary controller or substitute for clean expertise.

Supporting learners and actual execution. Small neural and foundation-model studies assess how label damage manifests under other learners, not the alignment distinction itself. In a separate inverse-label OpenVLA experiment, every policy action is executed in Isaac: unfiltered success is 0.075, physics-filtered 0.720, and available-clean-only 0.7075 across eight training blocks and 2,000 executions. The representations use inverse-derived action labels, digitally marked goals, and scene-specific retraining, so these results are not pooled with known-action kNN. Own-text and physics gates retain different counts in that bank, preventing an own-text-specific selectivity claim from its random control. The full execution study and method-sensitive, optimizer-matched driving control are reported in Appendices H and M.

6 DISCUSSION AND LIMITATIONS

Our results connect forecasting, generation, and downstream use while showing that three notions of world-model quality should be evaluated separately. First, *forecasting is not realization*: a correct textual prediction can coexist with an incorrect measured video outcome. Second, *agreement*

is not necessarily correctness: agreement with a reference does not by itself establish physical validity when that reference is imperfect. Third, *utility is not mechanism*: selective admission can improve downstream task success without establishing that numerical consistency alone caused the improvement. The intervention ladders, negative repair results, and action-matched controls provide complementary evidence for these distinctions rather than relying on a single aggregate success metric.

These observations suggest a practical evaluation protocol for generative physical world models: define the relevant observable and forecast regime, evaluate internal and external alignment on common qualified support, and test downstream admission against controls that account for changes in the retained action distribution. The inequalities in Appendix J formalize how reference and prediction errors can propagate into downstream decisions. They provide structural motivation for the evaluation but are not intended as finite-sample guarantees of improved prediction or task success.

The experiments are deliberately controlled to make these distinctions measurable. The primary evidence covers two checkpoints from one model family and a set of controlled physical scenes, complemented by a wider seven-setting diagnostic record. Consequently, the results establish the reported phenomena within the evaluated regime rather than claiming universal behavior across world models or robotic environments. Physical attribution is additionally affected by visibility constraints, scene-scale calibration, retained action support, and differences between analytic and simulator references. The action-matched analysis helps control for support selection, although broader prospective evaluation would further strengthen the causal interpretation of the downstream results.

The known-action downstream evaluation uses recorded action grids and calibrated scene collections, enabling controlled comparisons under matched action support. Fresh open-loop Isaac execution is included in the separate inverse-label study, with a different training representation. Broader fresh-execution generalization, closed-loop control, hardware transfer, and additional model families remain important directions for future evaluation. Qualitative assessments are unblinded and supplement the quantitative measurements; independent human annotation would provide an additional validation layer.

Finally, selective admission introduces additional inference and verification cost. In the known-action study, 13 conditions require 26 mode-specific requests and 31 total attempts, with cached repeated rows excluded from the reasoning-query count. Because we do not benchmark end-to-end latency or GPU-hour consumption across alternatives, our results should be interpreted as evidence about physical consistency and downstream utility rather than computational efficiency.

7 Conclusion

Internal and external alignment ask whether a generated future realizes the model's own contract and whether it matches physics. The A/B ladders test whether progressively supplying either reference closes its intended gap. These questions are distinct from whether screening the resulting data helps a learner. Matched contracts reveal accurate forecasts that are not realized and wrong forecasts that are agreed with. Fixed-budget label interventions show downstream damage and mitigation relative to ingest-all, while stronger selection controls limit what that gain explains. The defensible use of agreement is therefore conditional evidence about supervision—not a certificate of physical correctness, proof of generator repair, or an independently identified learning mechanism.

Reproducibility statement

Detailed experimental protocols, evaluation procedures, implementation settings, and additional results are provided in the appendix to support reproducibility of the reported findings.

Ethics statement

The physical reference is a specified simulator, not verified real-world safety. Filtering is not a deployment certificate: two internally consistent outputs can be wrong, and rejection can remove coverage of difficult conditions.

AI USE STATEMENT

We used large language models for research and analysis design, code, plots, qualitative checks, interpretation of results, language editing, LaTeX support, bibliographic checks, and mathematical consistency. The evaluated world models generated the videos and forecasts studied here. We reviewed all AI-assisted work and take full responsibility for the paper.

REFERENCES

Nahid Alam. Causal physics steering in video world models via concept activation vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5890–5896, 2026.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations*, 2025.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *International Conference on Learning Representations*, pp. 118456–118470, 2026.

Sixiang Chen, Jiaming Liu, Jixian Wu, Yichen Guo, Tinghao Wang, Siyuan Qian, Hao Chen, Jiajun Cao, Jian Tang, and Shanghang Zhang. Do robotic world models really follow actions? Diagnosing and aligning action-conditioned generation for policy learning. *arXiv preprint arXiv:2608.24885*, 2026.

Nate Gillman, Charles Herrmann, Michael Freeman, Daksh Aggarwal, Evan Luo, Deqing Sun, and Chen Sun. Force prompting: Video generation models can learn and generalize physics-based control signals. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-3448.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. In *International Conference on Machine Learning*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P. Foster, Pannag R. Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025.

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.

Hongxiang Li, Yaowei Li, Bin Lin, Yuwei Niu, Yuhang Yang, Xiaoshuang Huang, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Long Chen. GIR-Bench: Versatile benchmark for generating images with reasoning. In *International Conference on Learning Representations*, pp. 60368–60386, 2026.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 948–958, 2026.

Sriram Narayanan, Ziyu Jiang, Srinivasa Narasimhan, and Manmohan Chandraker. PhyCo: Learning controllable physical priors for generative motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41892–41902, 2026.

NVIDIA. Cosmos3-Super model card. Hugging Face model repository, `nvidia/Cosmos3-Super`, 2026a. Accessed September 25, 2026.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI. *arXiv preprint arXiv:2606.02800*, 2026b.

Yang Shi, Yuhao Dong, Yue Ding, Yuran Wang, Xuanyu Zhu, Sheng Zhou, Wenting Liu, Haochen Tian, Rundong Wang, Huanqian Wang, Zuyan Liu, Bohan Zeng, Ruizhe Chen, Qixun Wang, Zhuoran Zhang, Xinlong Chen, Chengzhuo Tong, Bozhou Li, Qiang Liu, Haotian Wang, Wenjing Yang, Yuanxing Zhang, Pengfei Wan, Yi-Fan Zhang, and Ziwei Liu. RealUnify: Do unified models truly benefit from unification? A comprehensive benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22488–22497, 2026.

Chenlong Wang, Yuhang Chen, Zhihan Hu, Dongping Chen, Wenhu Chen, Sarah Wiegreffe, and Tianyi Zhou. Quantifying the gap between understanding and generation within unified multimodal models. *arXiv preprint arXiv:2602.02140*, 2026a.

Yixuan Wang, Rhythm Syed, Fangyu Wu, Mengchao Zhang, Aykut Onol, Jose Barreiros, Hooshang Nayyeri, Tony Dear, Huan Zhang, and Yunzhu Li. Interactive world simulator for robot policy training and evaluation. In *Proceedings of Robotics: Science and Systems*, 2026b. doi: 10.15607/RSS.2026.XXII.018.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, Michal Drozdzal, and Adriana Romero-Soriano. Inference-time physics alignment of video generative models with latent world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16118–16129, 2026.

Geigh Zollicoffer, Kenneth Eaton, Jonathan C. Balloch, Julia Kim, Wei Zhou, Robert Wright, and Mark Riedl. Novelty detection in reinforcement learning with world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80740–80758. PMLR, 2025.

A GUIDE TO THE SUPPLEMENTARY EVIDENCE

The main paper defines the two alignment references and A/B information ladders before evaluating Q1–Q4, then extends the audit to supervision in Q5. The appendices preserve that distinction: shared scene/reader definitions and alignment evidence precede the complete downstream protocols. No statistics are pooled across these studies merely because they use the same model family. Table 4 maps the main questions to their supporting evidence.

Table 4: Where each part of the main argument is expanded.

Main question	Supporting material
Definitions and T0–T3 ladders	Contract schematics below; mechanisms and prompts in Appendix C; reference/reader checks in Appendix D
Q1–Q2: the two gaps	Counterexamples in Appendix E; task-specific sources in Appendix G; matched push records in Appendix H
Q3–Q4: A/B conditioning	Full A/B cells and feedback controls in Appendix F; correct/wrong directive counts in Appendix C
Q5: supervision consequences	Known-action sampling/learners in Appendix I; actual forecasts in Appendix L; action-matched and outcome-level sensitivity in Appendix N
Supporting VLA and proofs	Separate inverse-label executions in Appendix H; optimizer controls in Appendix M; assumptions/proofs in Appendix J
Qualitative inspection	Source-bound galleries in Appendix O; complete executed outcomes in Appendix P

B PREDICTIVE-CONTRACT ILLUSTRATIONS

These schematics distinguish the references and conditioning interventions. Source-bound physical examples, measurements, and complete outcomes are reported in the subsequent appendices. Figure 4 shows the two comparisons for a single generated outcome; Figure 5 shows how the A/B ladders supply each reference at increasing detail.

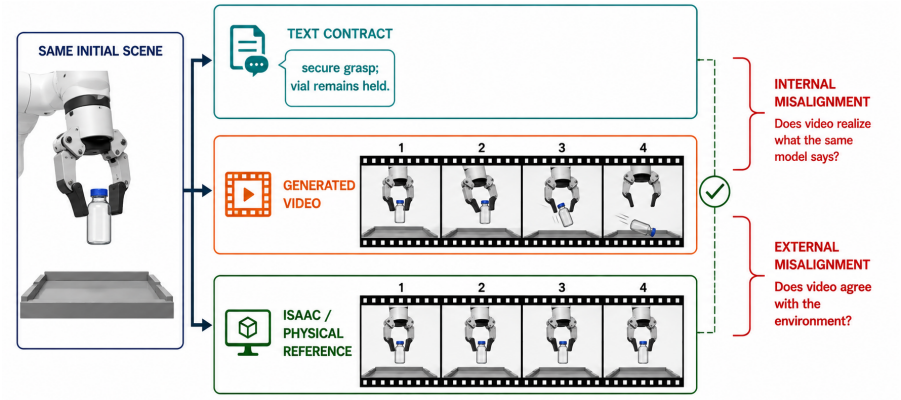


Figure 4: **Internal consistency and external correctness (schematic).** A generated grasp can be compared with the same checkpoint's explicit prediction and a physical reference. This illustration is not an additional evaluated task.

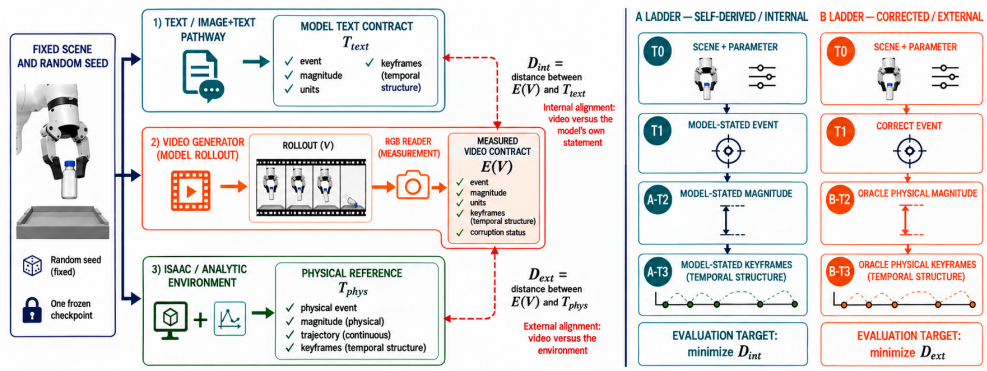


Figure 5: **Two reference ladders (schematic).** The A ladder supplies model-derived event, magnitude, and keyframes; the B ladder supplies physical-reference information. The same measured video is scored against the corresponding target.

C FRAMEWORK, SCENES, AND CONDITIONING PROTOCOLS (Q1–Q4)

C.1 ARCHITECTURE AND INFERENCE AUDIT

Cosmos3-Super couples understanding and generation pathways in the same checkpoint, with distinct stream computations. Attention establishes a route for conditioning, not where numerical physical content is represented or lost. The original image-to-video audit uses temperature-zero text queries and UniPC video sampling at 30 steps, guidance 6.0, flow shift 10, and 24 fps. Native-action driving generation is a different protocol and is not pooled with those settings.

C.2 SCENE AND SOLVER SPECIFICATIONS

The scenes are a 6 cm, 0.2 kg ball dropped 20 cm onto a $50\times50\times10$ cm pedestal (restitution combine mode *max*); a 6 cm, 0.2 kg cube on a 90 cm ramp at 25° with static and dynamic friction equal (combine mode *min*); a 0.2 kg cube given an initial horizontal speed of 1.2 m/s on a flat table with kinetic friction $\mu_k{=}0.3$ (combine mode *min*); and a 10 cm pendulum bob released from rest at 25°. Isaac Sim 6.0.1 uses a 60 Hz physics step, the default TGS solver, and no per-scene solver overrides. The originating Franka vial-grasp case motivates the study but is not counted among these four controlled mechanisms.

Table 5 lists each controlled intervention and its pixel-level observable.

Table 5: Controlled mechanisms, interventions, and pixel observables. Lengths use the stated task normalization; push uses a 0.06 m reference length.

Axis	Scene	θ values	Observable
Bounce	6 cm ball, 20 cm drop	$e \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$	rebound apex (e^2·drop)
Incline	6 cm cube, 25° ramp	$\mu \in \{0.2, 0.35, 0.45, 0.5, 0.6, 0.8\}$	stay/slide; displacement
Push	Cube, flat table ($\mu_k{=}0.3$)	$v_0 \in \{0.6, 0.9, 1.2, 1.5, 1.8\}$ m/s	stopping distance ($v_0^2/2\mu_k g$), 0.06 m units
Pendulum	10 cm bob, 25° release	$L \in \{0.25, 0.5, 0.75, 1.0\}$ m	amplitude (half peak-to-peak)

C.3 TEST SETTING VISUALIZATIONS

We evaluate four parameterized mechanics settings with analytic targets and recorded Isaac references where available, plus a single conditioning frame (Figure 6). The four axes isolate distinct physical mechanisms—restitution, ramp friction, stopping under friction, and oscillation—and each observable uses a declared normalization so it can be read back from generated pixels. Reference agreement and reader accuracy are qualified separately in Appendix D; neither is assumed exact at every parameter setting.

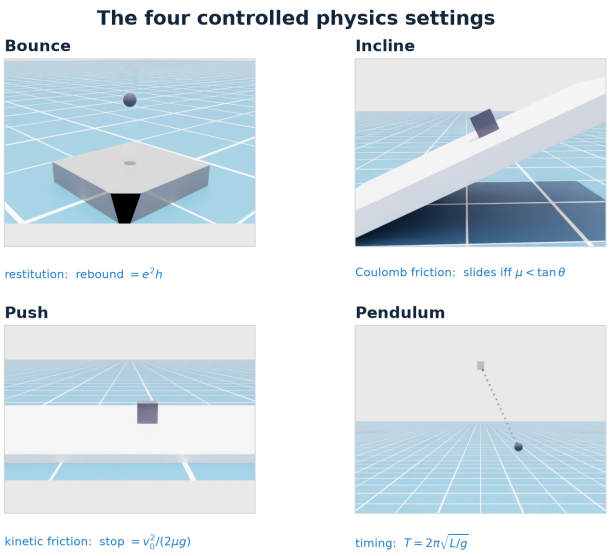


Figure 6: The four controlled physics settings, shown as their Isaac-twin conditioning frames. **Bounce** (restitution, rebound $= e^2 h$): a ball released above a pedestal. **Incline** (Coulomb friction, slides iff $\mu < \tan\theta$): a cube on a $25°$ ramp. **Push** (kinetic friction, stop $= v_0^2/(2\mu_k g)$): a cube given an initial velocity along a flat table. **Pendulum** (oscillation, small-angle $T \approx 2\pi\sqrt{L/g}$): a bob on a rigid rod released from $25°$. Each axis has one pixel-measurable, task-normalized observable (rebound apex, down-slope displacement, stopping distance, swing amplitude).

C.4 Explicit Validation Example

Example verification of T_{phys}: inclined plane. *Scenario:* Consider a cube of width $w = 0.06\,\text{m}$ placed at rest at the top of a ramp of length $0.90\,\text{m}$ inclined at $\alpha = 25°$, with coefficient of friction $\mu = 0.2$. *Solution:* The cube slides when the downslope component of gravity exceeds the maximum frictional resistance, equivalently when

$$\mu < \tan\alpha.$$

Here,

$$0.2 < \tan 25° \approx 0.466,$$

so the cube is expected to slide. Under the standard Coulomb-friction model, its downslope acceleration is

$$a = g(\sin\alpha - \mu\cos\alpha) \approx 9.81(\sin 25° - 0.2\cos 25°) \approx 2.37\,\text{m/s}^2.$$

Starting from rest, the expected displacement is therefore

$$d(t) = \frac{1}{2}at^2.$$

The cube reaches the end of the $0.90\,\text{m}$ ramp after

$$t_{\text{edge}} = \sqrt{\frac{2(0.90)}{2.37}} \approx 0.87\,\text{s}.$$

Expressed in the object-relative units used by our video extractor, the total expected displacement is

$$\frac{0.90}{0.06} = 15$$

cube-widths. Thus, for this scenario, T_{phys} predicts a monotonic downslope trajectory reaching approximately 15 cube-widths of displacement at 0.87 s. The scored camera exposes only part of this displacement, so the video measurement is censored at its visible limit rather than identified with all 15 cube-widths of travel.

C.5 Measurement and validity gating

Figure 7 summarizes the measurement path rather than adding an independent validation result. Tracking failures are separated from static but readable outcomes; broader visual corruption still requires inspection.

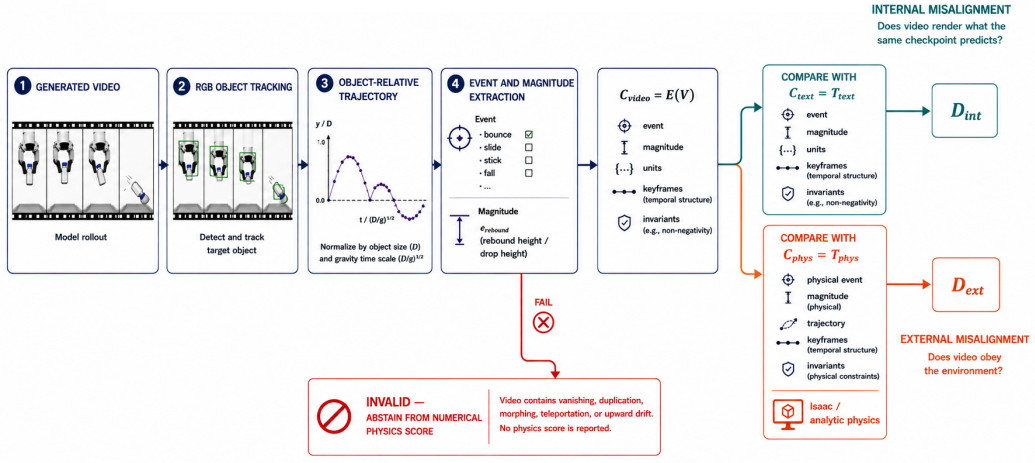


Figure 7: **One measured video contract, two alignment tests (schematic).** The frozen RGB reader maps a rollout to event, magnitude, units, keyframes, and invariants. Comparing that same contract with T_{text} yields D_{int}; comparing it with T_{phys} yields D_{ext}. Reader-detected tracking failures are unscorable; the wider corruption modes illustrate validity concerns rather than a guarantee of automatic detection. Static but measurable clips retain their outcomes.

C.6 Steering context

Steering and aligning generated physics. Related approaches include force prompting (Gillman et al., 2025), controllable physical priors (Narayanan et al., 2026), inference-time physics alignment (Yuan et al., 2026), rather than universal success of prompt-only correction. Separately, concept-vector steering changes physical-plausibility judgments in VideoMAE (Alam, 2026); interpretability work examines physical representations (Joseph et al., 2026). World models also support reinforcement-learning control (Hafner et al., 2025) and novelty detection (Zollicoffer et al., 2025). Our crossed correct/incorrect directive design is complementary: before steering, it measures what the text channel actually controls, event occurrence, in both directions, and what it does not, magnitude on an unmodified checkpoint.

C.7 Contract matching and directive inventory

The neutral sweep reported in the main text comprises 120 bounce, incline, and pendulum clips plus 40 push clips. The common directive comparison uses three mechanisms, six directive types, and eight seeds per cell (144 clips). A separate A/B diagnostic contains 32 records: model-reference and physical-reference arms at four seeds each for bounce, incline, pendulum, and projectile. The projectile records are not included in the push sweep. This diagnostic is distinct from the 144-clip correct/incorrect-directive comparison. Where the text and physical targets coincide, the two

references cannot identify a differential effect of internal versus external information; we therefore make no powered A-versus-B superiority claim.

Reference correctness requires matching conditions and observables. For the 1.2 m/s push probe, the text answer is 4.08 reference units, compared with 4.077 under the continuum law and 3.91 at the Isaac endpoint. The grasp probe predicts the hold event, but its 16 cm rise differs from the 11.8 cm reference. Period-only pendulum answers and bounce probes at a different drop height do not supply matched magnitude forecasts for the parameter sweep. The per-example 48-clip comparison in the main text uses its own joined requests and commanded-action executions rather than importing accuracy counts from these separate probes.

Figure 8 shows the measured parameter responses for the neutral sweeps; readable motion and agreement with the reference remain distinct outcomes.

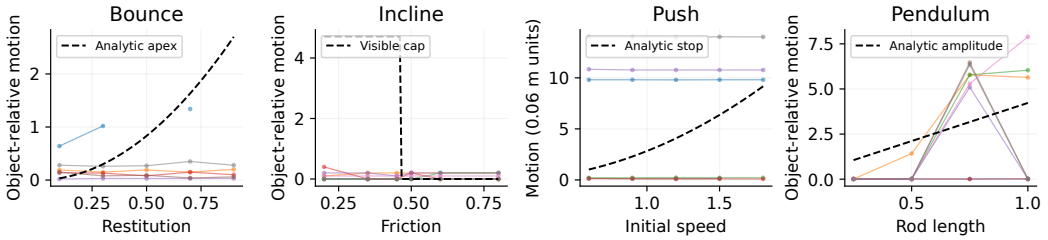


Figure 8: **Physical parameter response beyond the primary paired cohort.** Solid lines are generation seeds; dashed targets are analytic, except the incline visible-ramp cap. Canonical corruption labels exclude upward-drift bounces from valid traces. Push is a separate 40-clip bank, not a renamed projectile axis. Simulator/reference discrepancies and reader support qualify the magnitude comparisons (Appendix D).

C.8 Complete occurrence counts

Tables 6 and 7 report the complete neutral-sweep and directive-cell occurrence counts, respectively. A moving event is not a correct magnitude; these counts must be read alongside the reference and reader qualifications in Appendix D.

Table 6: Neutral-prompt sweep: eight seeds per setting. Dynamic denotes an observed moving event, not a correct magnitude. Invalid includes degenerate or release-exceeding outcomes under the axis-specific scorer.

Mechanism	Settings	Clips	Dynamic	Static	Invalid
Bounce	5	40	3	25	12
Incline	6	48	0	48	0
Push	5	40	15	14	11
Pendulum	4	32	10	22	0

Table 7: Event occurrence under correct and incorrect directives, eight seeds per cell. T0: condition only; T1: event; T2: magnitude; T3: keyframes; W1/W2: incorrect event/magnitude. Occurrence is independent of which reference is used to assess accuracy.

Axis	T0	T1	T2	T3	W1	W2
Bounce	2/8	1/8	3/8	1/8	2/8	3/8
Incline	0/8	0/8	3/8	6/8	0/8	1/8
Pendulum	6/8	6/8	7/8	7/8	3/8	7/8

C.9 ADDITIONAL GUIDANCE DIAGNOSTIC

Guidance affects swing occurrence but not amplification Regenerating the $L{=}1$ neutral condition on Pendulum with a change of guidance (a hyperparameter which controls sensitivity to the prompt during video generation) to 3.0 and 9.0 (four seeds each) leaves seed 234 swinging with an over-large amplitude at all three scales. Seed 345 swings at guidance 6.0 but freezes at 3.0 and 9.0; seeds 123 and 456 remain frozen. Thus the gain persists conditional on a rendered swing. In the over-amplified clips the bob's pixel diameter varies by only 16–26% across the swing (median 19–22 px), while the measured amplitude exceeds the physical maximum by $\sim$90%; a depth excursion large enough to inflate the amplitude that much would roughly double the apparent bob size.

D REFERENCE AND READER QUALIFICATION

The physical reference and the video reader require separate checks. Recorded simulator motion need not equal an ideal analytic trajectory, and correct event detection need not imply accurate magnitude measurement. The canonical audit binds both checks to saved source traces. Two bounce references deviate from the e^2 height relation by 13.8% and 5.7%; the incline traces agree on all four slide/stay cases. Projectile range has 87.1% relative error at 1.5 m/s and 22.9% at 2.5 m/s, versus at most 4.8% at the higher tested speeds. These are reference discrepancies, not generated-model errors.

Pendulum is especially important for interpreting magnitude. At $L = .25\,\text{m}$, the saved Isaac amplitude is 1.675 bob-diameters versus the analytic 1.057, a 58.5% discrepancy. At $L = 1\,\text{m}$, the corresponding values are 4.191 and 4.226 (0.82%). The short-length analytic curve therefore cannot be treated as an independently verified simulator endpoint. Missing motion can still be identified where both references predict motion; quantitative magnitude comparisons are anchored to the explicitly stated reference. The synchronized 12-degree Nano panel below uses its own recorded center trajectories, not these 25-degree traces.

On six lossless Isaac feeds, the canonical reader recovers all six event classes, but only three magnitudes fall within the declared 20% tolerance. The pendulum $L = 1$ feed reads 3.207 versus 4.23; the incline slide reads 4.5 versus 15 because of limited visibility, and the sparse projectile feed at 2.5 m/s reads 5.245 versus 8.1. Passing feeds are bounce, a static incline, and the 4 m/s projectile. These checks support event-level qualification more strongly than universal magnitude calibration. They neither license a generated-domain error bound nor justify treating every extracted object as a valid scene.

Table 8 reports every canonical lossless-feed check, including the magnitude failures.

Table 8: All six canonical lossless-feed checks. Every event class is correct; magnitude qualification is only three of six. These are feed-specific checks, not six independent generated-domain validations.

Mechanism	Feed	Read value	Reference	Within 20%
Bounce	bounce_e09	3.050	2.700	Yes
Incline	incline_slide	4.500	15.000	No
Incline	incline_stay	0.000	0.000	Yes
Projectile	proj_v40	18.781	20.000	Yes
Projectile	proj_v25	5.245	8.100	No
Pendulum	pendgt_L100	3.207	4.230	No

The separate 15-entry validation record in Table 24 and the goal-reader split checks measure different aspects of perception.

For the primary mechanism curves, bounce, incline and pendulum now use the single corruption-aware canonical score definition; the separately generated push bank keeps its declared stopping-distance reader. The canonical bounce sweep contains three dynamic, 25 static, six corrupted and six degenerate clips. Its T2 cell contains three dynamic, one static and four invalid clips. This prevents upward drift from being counted as a successful rebound. The gallery retains source-bound

illustrative traces; a matching re-extraction of an illustration does not independently establish reader accuracy.

E ARCHIVAL COUNTEREXAMPLES: AGREEMENT WITHOUT CORRECTNESS

This is retrospective evidence mining, not a new independent test cohort. The source archives contain a 576-record Nano scalar study (536 distinct video hashes) and a 64-record synchronized panel. Case selection followed inspection of their saved scores. The complete cell/panel records are retained so that selected examples are not mistaken for prevalence estimates. Source hashes, raw forecasts, and the primary example clips accompany the manuscript. No additional model queries or simulator executions are counted here.

E.1 MATCHED SCALAR OBSERVABLES AND UNCHANGED ADMISSION RULES

The scalar contract explicitly requests horizontal pendulum amplitude rather than period, with unit `bob-diameters`. The source text-only Nano response returns magnitude 10 and nine frame-indexed positions. Those positions are also retained; they are not replaced by the physical trajectory. The supplied geometry gives $L\sin(25^\circ)/d = 4.22618$ at $L = 1\,\text{m}$ and $d = .10\,\text{m}$. All eight neutral videos in Table 9 share the exact base physical prompt and conditioning image used to construct the forecast query. One forecast is shared across seeds; there are not eight independent reasoning trials.

Table 9: Complete archived neutral scalar cell, using the original automatic reader and rule $|m-r| \leq \max(.5, .25|r|)$. A scalar pass is not a full-scene validity label. An unblinded assistant review flags an extra bright blob at seed 234 and a duplicate bob at seed 345; those rows remain visible but do not support clean-case prevalence claims.

Seed	Measured amplitude	Against text 10	Against physics 4.226
123	5.11	Fail	Pass
234	5.12	Fail	Pass
345	7.54	Pass	Fail
456	5.05	Fail	Pass
567	4.14	Fail	Pass
678	3.55	Fail	Pass
789	5.74	Fail	Fail
890	8.05	Pass	Fail

Seeds 567 and 890 have full 81-frame tracks and no lost runs. Re-extraction with the frozen reader reproduces 4.14 and 8.05. Uniformly spaced frames and measured extrema were inspected; no duplicate object or camera-motion artifact was identified at those sampled frames. This is an unblinded computational and visual check, not independent human annotation or an uncertainty certificate. The same relative rule gives different absolute half-widths around text and physics (2.50 and 1.0565). These widths are shown in the main figure rather than hidden behind binary labels.

Normalization sensitivity matters. Replacing the median observed bob diameter by its 5th or 95th percentile changes seed 567's amplitude to 3.919–4.394, still inside the physical window and outside the text window. For seed 890 the range is 7.365–8.654, whose lower endpoint misses the text window. Thus the reported original-rule decision is reproducible, but the false-admission label is not measurement-uncertainty certified. A parameter-matched neutral wording variant measures 9.50, with diameter sensitivity 7.985–10.334; it is supplied as a secondary illustration, not pooled into the exact-prompt cell.

Figure 9 retains the complete scalar windows and the synchronized pendulum example for context; it is not a second independent experiment.

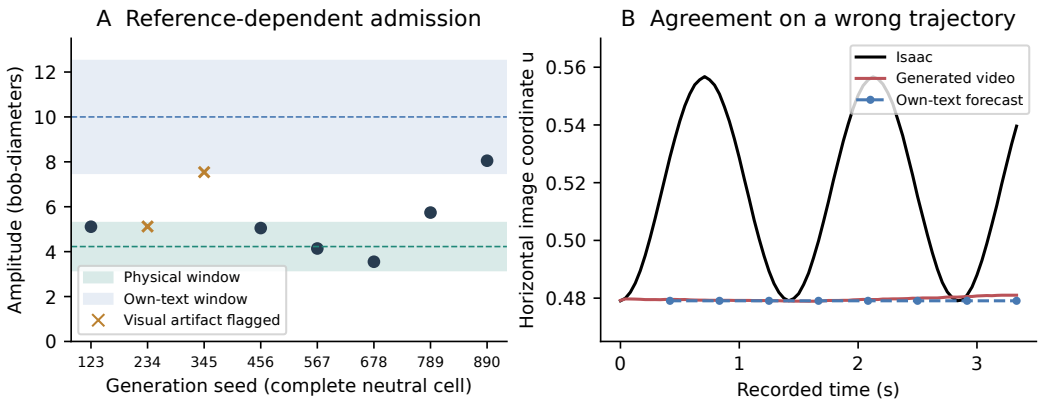


Figure 9: **Complete scalar-pendulum comparison retained for context.** The same eight-seed archive and synchronized example underlying Appendix E. Crosses mark inspected artifacts. Relative windows, selection and diameter sensitivity remain explicit; this is not a second independent experiment.

E.2 A CONTINUOUS COMPARISON ON IDENTICAL FUTURE KEYFRAMES

The synchronized panel uses Nano's own raw JSON forecasts of the visible blue bob center, in normalized image coordinates. It provides the same initial image and physical conditions to text forecasting and video generation, and records 81 Isaac states at 24 Hz. All three errors below use exactly frames 10,20,30,40,50,60,70,80; frame zero is excluded. The means are $\frac{1}{|K|}\sum_{k\in K}\|T_k - V_k\|_2$, $\frac{1}{|K|}\sum_{k\in K}\|V_k - G_k\|_2$, and $\frac{1}{|K|}\sum_{k\in K}\|T_k - G_k\|_2$ on the same support K. They do not compare eight-frame text error with an 80-frame video average.

Table 10 reports the four joint-keyframe comparisons; Figure 10 shows all corresponding length–seed cases.

Table 10: All four synchronized Nano pendulum cases. Distances are normalized image-coordinate errors, not meters. Each case has eight jointly valid future keyframes. No new correctness threshold is selected for these comparisons.

Length (m)	Seed	Text–video	Video–Isaac	Text–Isaac
0.25	123	0.06063	0.04718	0.02061
0.25	234	0.00262	0.02001	0.02061
0.5	123	0.02360	0.03973	0.04113
0.5	234	0.00234	0.04061	0.04113

For seed 234 at both lengths, the model's raw coordinates remain at the initial center, as does the video approximately, while the simulator bob oscillates. A fixed RGB component reader (blue $>$ red$+20$, blue $>$ green$+10$, blue > 60; largest connected component with at least six pixels) detects the bob in all 81 frames of each of the four generated clips and both references. For the two seed-234 clips, its external errors are .019947 and .040483; maximum differences from SAM2 centers are .001203 and .001019. This is a second implementation check, not a calibrated uncertainty bound. Source video, initial image, forecast, mask, clock, and model/configuration hashes were checked.

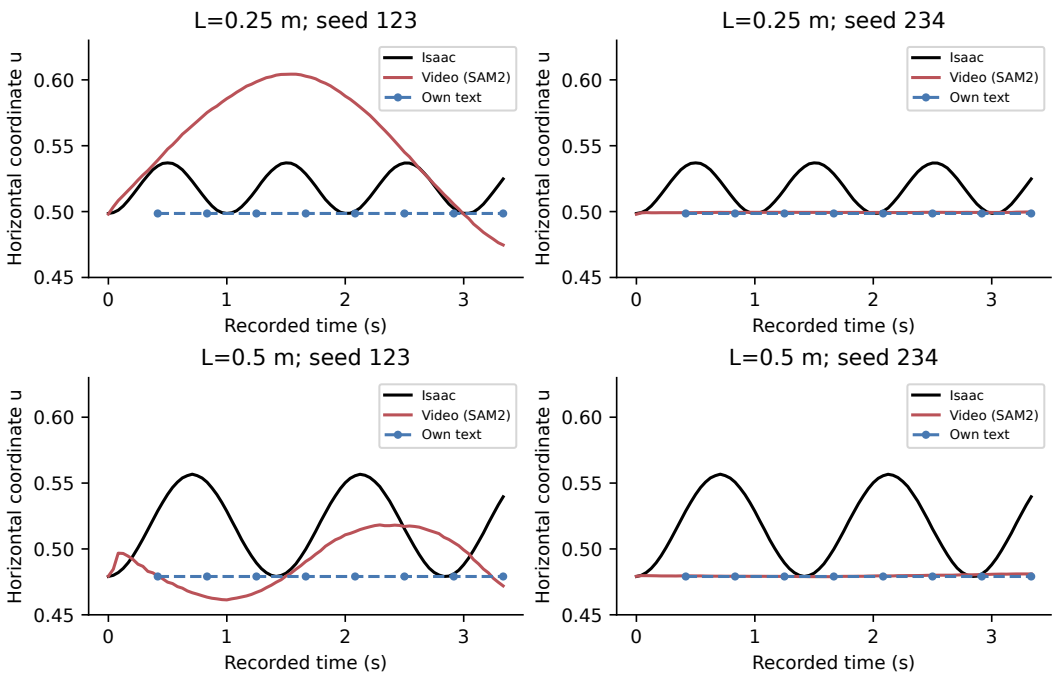


Figure 10: **Complete synchronized pendulum comparison.** Both lengths and both seeds are shown, including seeds that do not exhibit low internal error. The $L = .5\,\mathrm{m}$, seed-123 video deforms/rotates the support surface and is not a clean full-scene example. The seed-234 claims concern the visible bob-center trajectory, not all scene invariants.

E.3 Quarantined candidates and model coverage

The current Super push bank's incorrect image–text answer is not an agreeing-wrong example under its original gate. In another archive, an answer `5/6` was parsed by a coarse numeric regex as 5; raw inspection identifies a parsing error rather than faulty physical reasoning. A Super restitution answer genuinely uses e rather than e^2, but its apparently passing paired video has an event-unclear/letterbox visual audit and is not a primary example. Some automatically static incline clips add a second moving cube. Their original-object trajectory is measurable, but they cannot be described as physically correct rollouts. All such exclusions remain in the audit records.

The full synchronized archive contains 16 videos each for Nano, Super, Edge and LTX. Only Nano supplies eight valid own-trajectory contracts under the original parser; Super and Edge supply zero of eight, and LTX uses a Nano teacher rather than its own text stream. External tracking coverage differs across models, and generated-domain review remains incomplete. This inventory is not evidence of an architecture ranking or four-model internal alignment.

F Self-reference, physical-reference and discrepancy-feedback interventions

The Nano scalar archive supplies 192 A/B records: four mechanisms, two target sources, three prompt levels, and eight seeds. A uses the actual model contract; B uses the declared physical contract. Unlike coincident-target diagnostics, references genuinely differ for bounce, projectile and pendulum. The original rule and complete automatic outcomes appear below; they are descriptive and subject to the scalar reader's qualifications.

For pendulum, own-text targets amplitude 10 versus physical 4.22618. Increasing A specificity gives internal pass counts 2,2,2 and external counts 4,3,5; B gives internal counts 2,1,3 and external counts 4,4,4. Thus target separation makes the question identifiable, but does not establish successful internal or external repair. These are not powered equivalence tests.

Table 11: Complete archived A/B ladder. Counts are out of eight generation seeds. Measurable does not certify scene coherence. Event-only A/B videos are identical in all 32 axis/seed pairs and must not be counted as independent replications. No superiority claim is made from these small cells.

Mechanism	Source	Level	Measurable	Internal pass	External pass
Bounce	A	1	3/8	0/8	0/8
Bounce	A	2	3/8	1/8	0/8
Bounce	A	3	3/8	1/8	0/8
Bounce	B	1	3/8	0/8	0/8
Bounce	B	2	4/8	1/8	0/8
Bounce	B	3	3/8	1/8	1/8
Incline	A	1	8/8	2/8	2/8
Incline	A	2	8/8	5/8	5/8
Incline	A	3	8/8	4/8	4/8
Incline	B	1	8/8	2/8	2/8
Incline	B	2	8/8	5/8	5/8
Incline	B	3	8/8	4/8	4/8
Projectile	A	1	4/8	0/8	0/8
Projectile	A	2	7/8	0/8	0/8
Projectile	A	3	5/8	0/8	0/8
Projectile	B	1	4/8	0/8	0/8
Projectile	B	2	6/8	0/8	0/8
Projectile	B	3	4/8	0/8	0/8
Pendulum	A	1	8/8	2/8	4/8
Pendulum	A	2	8/8	2/8	3/8
Pendulum	A	3	8/8	2/8	5/8
Pendulum	B	1	8/8	2/8	4/8
Pendulum	B	2	8/8	1/8	4/8
Pendulum	B	3	8/8	3/8	4/8

Discrepancy-feedback negative control. A separate protocol generates 96 videos: four mechanisms, eight seeds, and three feedback arms (signed feedback, error-permuted, sign-flipped). The neutral and one-shot baselines reuse existing clips. Its declared pass rule requires improvement over neutral and both negative controls, not merely neutral. All selected physical targets require motion, so the saved event metric is the dynamic-occurrence rate. The reported error is hybrid: capped relative magnitude error for bounce/projectile and binary non-occurrence error for incline/pendulum, with invalid outputs assigned unit error. It is not a four-mechanism magnitude metric or an estimate of unseen error on invalid frames.

Table 12 reports the event, hybrid-error, and invalid-rate summaries for every feedback arm.

Table 12: Recorded discrepancy-feedback analysis. Lower normalized error is better. Columns follow the frozen analysis contract rather than a new shared metric invented during mining. The error-permuted control prevents interpreting a neutral-only improvement as a discrepancy-specific repair effect.

Arm	Event occurrence	Hybrid error	Invalid rate
error permuted	0.3125	0.6791	0.2188
feedback	0.2500	0.7419	0.2188
neutral existing	0.0938	0.8985	0.1562
sign flipped	0.0312	0.9688	0.2812

Feedback-minus-neutral event occurrence is +.15625 with seed-block bootstrap interval [.03125,.28125], but feedback-minus-permuted is -.0625 [-.125,0]. Hybrid error is also higher with feedback than with permuted feedback (.74190 versus .67905). The frozen verdict is therefore FAIL. The permuted arm retains the correct contract preamble and permutes feedback within mechanism; it is not a semantically irrelevant prompt. The result fails to establish instance-specific use of the supplied discrepancy rather than proving its absence. The unit of inference is eight seed blocks crossed with four mechanisms, not 96 independent training runs. The one-shot and recorded-Isaac-reference sensitivity analyses are retained as secondary, not used to rescue the primary result. The visual

screen also identifies duplicate/vanish artifacts in five of six feedback incline slides. This study concerns regeneration, not the closed-loop execution of a robotic policy; it supplies no closed-loop robotics validation.

Relation to downstream learning. The interventions act on different objects. A/B and feedback change the *generation prompt*; admission changes which generated outcome labels a learner receives. Failure to establish generator repair does not preclude selective admission, and successful admission does not demonstrate generator repair. The known-action outcome-repair diagnostic isolates label corruption; count/motion controls isolate selection; clean-only and all-generated tests expose sparse support. The VLA experiments test consequences of supervision, not the empirical independence of internal and external alignment.

G TASK BREADTH: DISTINCT MECHANISMS AND DISTINCT EVIDENTIAL ROLES

The evidence is not a pendulum-only study. Seven physical settings contribute, but their experimental depth and supported claims differ (Table 13). Horizontal push/slide and inclined sliding are separate interventions; projectile is not relabeled as push. Simulator-only rolling, lever and center-of-mass fixtures do not increase the alignment-task count because matched model forecasts and generated videos are not established.

Table 13: **Seven settings, not seven interchangeable benchmarks.** Counts refer to named studies and must not be pooled across shared clips, forecasts, conditions or seed blocks.

Setting	Evidence	Supported role
Bounce	Canonical 40-clip sweep; Nano A/B cells at eight seeds	Missing/incorrect rebounds; selected wrong-magnitude agreement
Incline	48 neutral clips; correct/wrong directive cells	Friction-insensitive events and prompt responsiveness; duplicated objects exclude apparent clean reverse cases
Projectile	Five matched Super range forecasts; 40 videos	Correct analytic forecasts with non-launching video, not separated references
Push/slide	48 matched attempts; 60-row known-action bank	Per-example discrepancy and selective downstream admission
Pendulum	32-clip sweep; scalar and synchronized Nano cells	Magnitude/trajectory errors; conditional internal/external separation
Vial lift	Two own-forecast prompts $\times$ eight Super seeds	Contact-maintenance realization illustration; no calibrated reference separation
Driving	15 command-response clips; 48 additional clips; forward-learning controls	Native command/flow sensitivity and conditional learner consequences; no validated metric text/video comparison

G.1 BOUNCE: A MOVING WRONG-MAGNITUDE AGREEMENT CASE

The raw Nano contract for $e = .9$ predicts magnitude 1.62 ball-diameters, rather than the analytic 2.70. The A/T2 prompt inserts that actual model forecast. Seed 456 measures 1.59, with landing at frame 9 and first post-impact apex at frame 15. Re-extraction with the frozen, source-bound reader reproduces the score and 81/81 tracking. Under the unchanged tolerance, the own window is [1.12,2.12] and the physical window [2.025,3.375]. Substituting the tracked diameter's 5th/95th percentile for its median yields 1.522–1.903, still in the own window and outside the physical one. This sensitivity interval is not a certified generated-domain measurement bound. Sampled impact/apex/settling frames show one coherent ball, without identified duplicate or disappearance; the review is unblinded, not independent human annotation.

The full A/T2 cell has eight attempts, three measurable clips, one own-reference pass and no physical passes. The B/T2 cell has eight attempts, four measurable, one own-reference pass and no physical

passes. B/T2/seed 456 is a distinct video and measures 1.54 despite its physical 2.70 directive. These cells do not show that self-conditioning uniquely caused agreement or that physical conditioning repaired the generator. The complete six-cell bounce ladder is in Table 11.

Importantly, the raw text contract's nine keyframe positions are all zero, contradicting its nonzero magnitude. We therefore use only its magnitude component and the magnitude-only A/T2 intervention. We do not call the entire text contract coherent or the entire video internally consistent. The A/T3 clip is not promoted as full-trajectory agreement.

Reverse direction with explicit interface provenance. At $e = .5$, a matched Nano image–text contract predicts 5.00 ball-diameters, whereas the neutral seed 456 video measures .80 and the physical reference is .833. Both base prompt and conditioning-image hash match. The same frozen reader reproduces a visible rebound and 81/81 tracking; diameter sensitivity .777–.983 stays inside the physical window [.333,1.333] and far outside the own-image–text window [3.75,6.25]. This is a magnitude-only, retrospectively identified reverse case; its text keyframes are not physically coherent.

The source scorer used a filename pattern that matches both text-only and image–text contract files and keyed them by scene ID. Its recorded target 5.0 therefore agrees with the image–text response, not the text-only response 1.5. We retain the exact interface identity rather than calling 5.0 stale or silently pooling modes. Against text-only 1.5, the same video also nominally fails the own window, but its sensitivity maximum is only .017 from that window, making that alternative comparator less robust. The complete eight-seed neutral cell has three measurable clips, one dynamic, two static, one corrupted and four degenerate; the second automatic physical pass is a static scalar-floor case and is not promoted as a correct bounce. The Super $e = .7$ candidate remains excluded for its low-confidence, event-unclear, letterboxed video.

G.2 Projectile: correct range forecasts, missing launch events

The Super matched scalar queries and generated videos share the same base physical prompts. Forecast error is below .6% relative to the analytic range at all five speeds (Table 14). The entire 40-video sweep contains 29 readable/static and 11 degenerate outputs, with no numerical passes against either reference. Degenerate outputs are non-passing attempts, not observed zero-range physical trajectories.

Table 14: Projectile cohort, eight generation seeds per speed. Range is in ball-diameters. Text accuracy here is against the analytic reference; recorded simulator discrepancies at low speed remain disclosed in Appendix D. No video passes either range gate.

Speed (m/s)	Text range	Analytic range	Readable	Degenerate
1.5	2.86	2.867	7/8	1/8
2.5	8.01	7.964	5/8	3/8
3.5	15.60	15.609	6/8	2/8
4	20.40	20.387	6/8	2/8
5	31.90	31.855	5/8	3/8

At 4 m/s, seeds 567, 678 and 890 have 81/81 tracked frames and measured range 0.0; the selected raw frames show a stationary ball rather than a launch. A separate archived sparse-keyframe screen marks 16 of the 40 clips high-confidence, anomaly-free non-launches, including those three. That screen is model-based with incomplete evaluator identity, not human ground truth. Event absence and full tracking support the example; they do not validate fine high-speed trajectory measurements, which are field-of-view limited in other feeds. The text and physical references nearly coincide, so projectile strengthens cross-modal dissociation, not the distinction between two conflicting references.

G.3 VIAL/GRASP: AN EXPLICITLY INSTRUCTED HOLD NEED NOT BE REALIZED

A staged run names Cosmos3-Super for the text query and subsequent generator loads. The retained response tail predicts that a 1.2 N-per-finger grasp holds and the vial rises 16 cm. Two prompts insert this outcome as magnitude or trajectory instructions, with eight seeds each. Model identity is supported by the served ID and local path in the staged logs, not a saved checkpoint digest. The task-level forecast is reused; there are not 16 independent reasoning queries.

The endpoint reader measures signed upward image displacement divided by frame height, not meters. Table 15 reports all 16 endpoints. Fifteen are below the existing .1 threshold; the maximum is .1021 image heights. The custom-hook Isaac positive-control endpoint is .4479. These are outcome examples, not synchronized or metrically interchangeable trajectories: the clips have different durations, and the reference uses an explicit contact-capacity hook. The selected A/T2/seed 123 video shows an empty gripper lifting while the vial remains near the pedestal.

Table 15: All 16 own-forecast-conditioned vial endpoints, re-read from the stored first/last decoded frames. Units are fractions of image height. These measurements do not establish correct metric lift or a population failure rate.

Own-forecast tier	Seed	Endpoint rise
2	123	0.0292
2	234	0.0146
2	345	0.0104
2	456	0.0021
2	567	0.0354
2	678	0.0187
2	789	0.0125
2	890	0.0312
3	123	0.1021
3	234	0.0104
3	345	0.0333
3	456	0.0021
3	567	0.0167
3	678	0.0208
3	789	0.0146
3	890	0.0083

The archive mixes per-finger, total-interface and measured-average force conventions. At this particular 1.2 N condition, either 1.4×1.2 or $2\times1.4\times1.2$ exceeds the 1.18 N weight, so the qualitative hold prediction is robust to that ambiguity. Exact rise is not: the requested lift is 16 cm, recorded end-effector rise 14.97 cm, and vial rise 11.8 cm. An older scorer maps both text and physical references to the same .448 normalized target; it cannot establish internal/external separation. We do not reuse that mapping as evidence of two independently calibrated reference magnitudes.

A pooled 101-outcome inventory has maximum recorded rise .0521 image heights, not .0521 m. It lacks complete per-row seed/media binding and is not treated as 101 matched triplets. Nor do we claim grasp generation universally fails: a separate campaign has three successful hold generations and three release abstentions under another scene/reader protocol. That study has no explicit own-model forecast and is not pooled with the present alignment evidence.

Figure 11 juxtaposes the source-bound bounce, projectile, and vial/grasp examples described above.

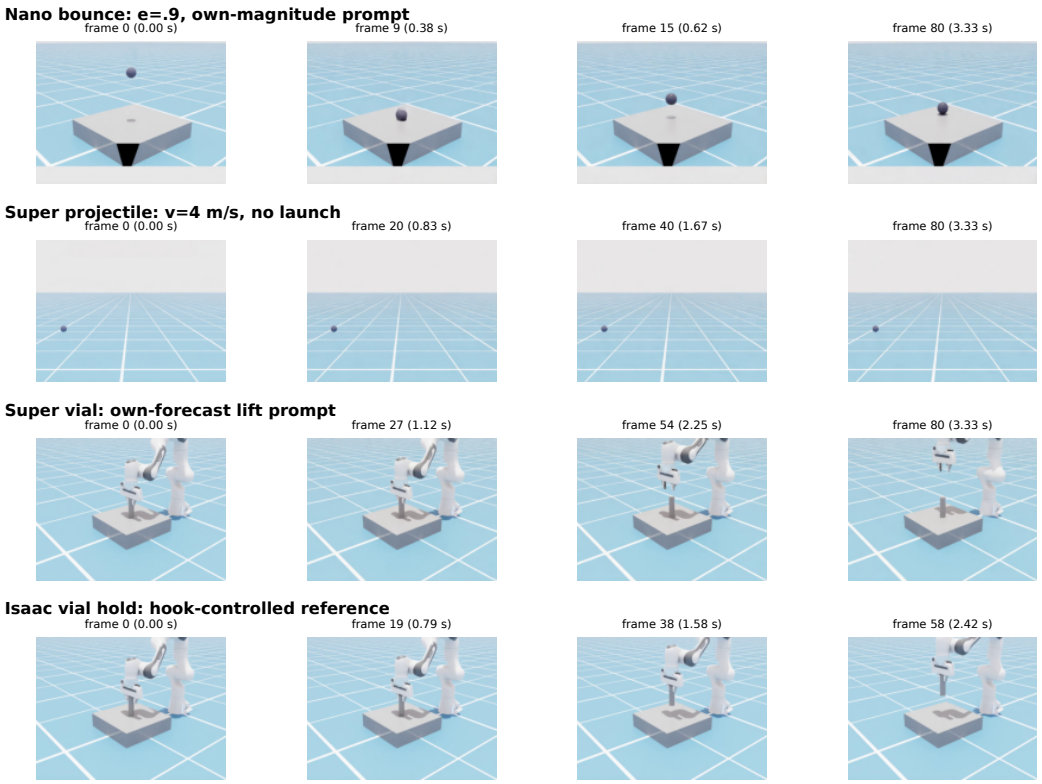


Figure 11: **Source-bound examples beyond pendulum.** A moving wrong-height bounce; a projectile non-launch despite a correct range forecast; a generated vial-lift instruction that leaves the vial near its start; and the custom-hook Isaac hold positive control. Each row retains its own encoded times. Rows are not synchronized physical trajectories, a random sample, or additional model runs.

G.4 DRIVING: COMMAND RESPONSE WITHOUT AN UNSUPPORTED METRIC CONVERSION

Fifteen AV clips cross five native command levels with three seeds. The actual mean-absolute forward-coordinate ratio is 2.400, while its signed-mean ratio is 1.386; neither is a speed ratio. Flow endpoint ratios within seeds are 1.061, 1.033 and 1.130. A separate 48-clip bank yields Pearson correlation .03981 between mean-absolute native coordinate and optical flow; Spearman is .17119. A retrospective clip-bootstrap interval for Pearson correlation is [-.1815,.3006], so zero causal action effect is not established.

Figure 12 shows the two driving banks on their declared native-coordinate and flow scales.

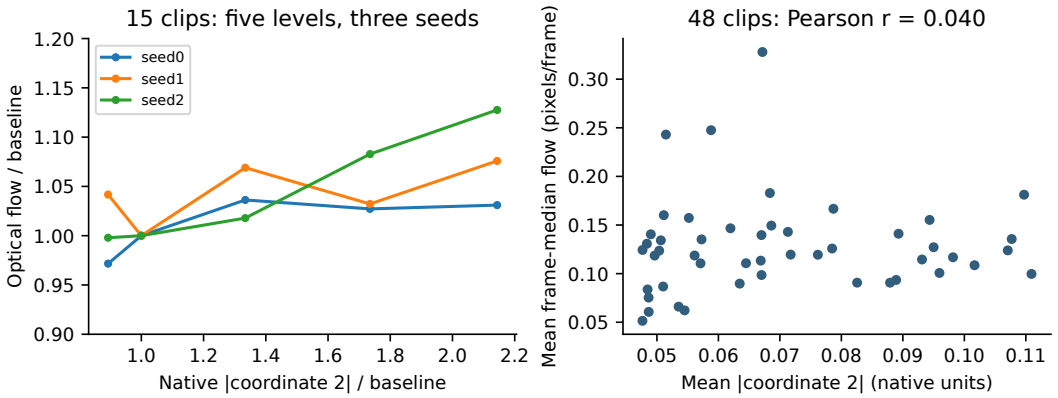


Figure 12: **Driving contributes a command-response diagnostic.** Actual native coordinates and flow are reported without identifying them as m/s or metric distance. The two panels are separate 15- and 48-clip banks. Weak observed association is not proof of zero causal action sensitivity.

Five text outputs solve constant-speed word problems $d = VT$, but the archive retains only parsed answers and the native action/flow conversion is not validated. The 48 generated clips share 832×480 geometry and 33 frames at 10 Hz (3.2 s first-to-last span), rather than the 1.5 s textual task. Driving therefore supports native-channel sensitivity and the explicitly normalized learning assays, not a new calibrated text/video/physics counterexample. The corrected simulator law $d = 1.5V - .1962$ does not repair that conversion.

H MATCHED-CONTRACT PUSH STUDY: PROTOCOL AND DETAILED RESULTS

This study measures the internal gap on the clips used to construct OpenVLA training labels. It uses inverse-derived action supervision and scene-specific retraining, complementing the known-action forward-model experiment. Quantities below are regenerated from frozen requests, videos, policy actions, and Isaac state logs; statistical comparisons remain within their respective protocols.

H.1 A CHANGED SCENE, MATCHED REQUESTS, AND AN EXPLICIT DENOMINATOR

The matched-contract push setting changes friction from .3 to .5, table and cube appearance, and lighting while keeping camera geometry and initial position fixed. Forty-eight new commands are drawn uniformly from $[1.00, 2.05]$ m/s, rounded to two decimals, with fixed generator seeds 8000–8047. A single new conditioning image is used throughout. The 0.06 m reference length is a normalization convention, not a claim about the rendered cube's visible width.

For each generated clip V_i, the same Cosmos3-Super model is given its exact generation prompt plus a stopping-distance question, either without or with the conditioning image. The requests give the commanded speed and friction, but no numerical stopping-distance answer or Isaac result. Temperature is zero; the parser takes the last `ANSWER = <number>` marker and allows at most two identical-request retries. All responses, including unsuccessful attempts, are retained. The command is separately executed in Isaac to obtain the reference $h(a_i)$, not scored merely against the continuum formula.

Table 16 separates forecast validity from endpoint accuracy for all three modalities.

Table 16: Matched forecasts, scored against the commanded-action Isaac endpoint. Passing means at most 25% relative error. Unreadable/invalid outputs do not pass; this does not identify their unobserved physical error.

Modality	Valid / attempted	Passing / attempted
Text only	$46/48$	$46/48$
Image + text	$48/48$	$47/48$
Generated video	$42/48$	$2/48$

The all-attempt count $45/48$ combines correct text with *non-passing* video, including six unreadable videos. The stricter jointly valid comparison is $39/40$: correct text and a readable video with an incorrect stopping distance. The remaining jointly valid pair passes in both modalities. Of the other two readable videos, one passes and one fails, but both lack a valid text-only answer. Thus the dissociation does not depend on treating unreadable video as physically wrong. Forty of the 42 readable clips have travel below 0.2 reference units; the result is primarily a near-static failure mode, not evidence covering all kinds of dynamic error. The image-conditioned text result also shows that providing the starting image does not resolve the cross-modal gap here.

Figure 13 places the matched-modal cohort beside its separate inverse-label policy study.

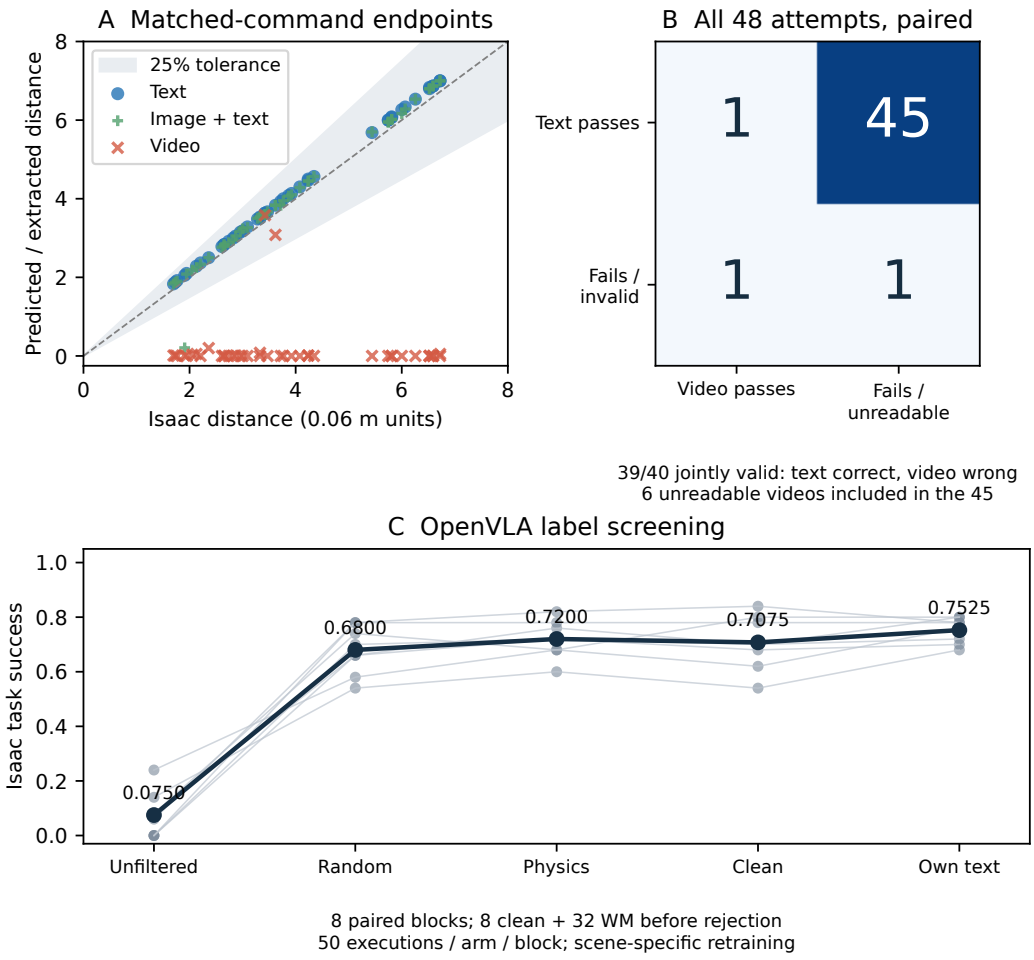


Figure 13: **Matched-modal cohort and supplementary inverse-label study.** A–B show forecasts and generated endpoints for the 48-clip friction-0.5 cohort; unreadable outputs are separated from jointly valid pairs. C shows the associated goal-to-action inverse-label OpenVLA experiment, with all actions executed in Isaac. It is not the known-action forward-model control in the main text. Eight training blocks determine uncertainty.

H.2 TRAINING REPRESENTATION AND ADMISSION RULES

Unlike the forward-model rows in Section 5, this protocol reconstructs a scalar action from a generated endpoint. Writing $w = .06\,\text{m}$, $g_0 = 9.81\,\text{m/s}^2$ and $\mu = .5$, its generated training row is

$$(I(g_i), \widetilde{a}_i)\,, \qquad g_i = \frac{a_i^2}{2\mu g_0 w}, \quad \widetilde{a}_i = \text{clip}_{[.4,2.6]}\sqrt{2\mu g_0 w\, E(V_i)}. \tag{5}$$

Here $I(g_i)$ is the Isaac starting image with a *digitally drawn red goal stripe*; it is not a generated frame or the physical marker used in the known-action study. The intended goal remains the continuum-law stopping distance of the command, while the pseudo-action comes from the video's observed distance. A near-static video therefore supplies a small action for a distant intended goal. This is a test of a specified pseudo-labeling pipeline, not proof that every use of generated video corrupts control.

The clean action for goal g uses the step-corrected inverse

$$a_c(g) = \frac{\mu g_0 \Delta t + \sqrt{(\mu g_0 \Delta t)^2 + 8\mu g_0 w g}}{2}, \quad \Delta t = 1/60\ \text{s}. \tag{6}$$

This inverts the discrete-law approximation $d = a^2/(2\mu g_0) - a\Delta t/2$; the actual simulator endpoints, not an assumption of exact analytical integration, determine success. Clean and generated labels therefore use different inverses. Forty clean candidate goals span $[.5, 11]$ reference units. Each of eight blocks samples eight clean rows and 32 distinct readable generated rows without replacement. The five arms share this initial draw. They train OpenVLA-7B LoRA (rank 16, scale 32, no LoRA dropout) for 240 AdamW steps at learning rate $2 \cdot 10^{-4}$. Initialization and shuffle seeds are paired. Only the first action-token coordinate controls the scalar push speed; this is not a seven-degree-of-freedom manipulation policy.

The physics filter keeps a row if $|E(V_i) - g_i| \le .30 g_i$. The own-text gate instead keeps it iff

$$q_{\text{text}}(i) = \mathbf{1}\{t_i \text{ valid and positive}\}\mathbf{1}\{|E(V_i) - t_i| \le .30 t_i\}, \tag{7}$$

where t_i is the model's parsed text-only answer in the same units. All valid answers here are positive. *This admission decision* needs no Isaac endpoint or analytic stopping law. However, the enclosing pipeline still uses the physics law for goal construction and inverse pseudo-labels, and calibrated pixels for E. It is not a physics-free learning method. The physics and own-text gates keep respectively 2 and 1 of the 42 readable pool clips; both keep one and reject 40, with one physics-only acceptance ($\kappa = 0.656$). That disagreement is clip 42: its video passes the external accuracy test, but its text answer is invalid, so the text gate rejects it. High agreement therefore mostly reflects shared rejection, not broad agreement on diverse useful rollouts. The random arm matches the *physics* gate's count, not the text gate's smaller count. A count-controlled selectivity claim for the text gate is not established.

What internal admission can and cannot certify. Let the true endpoint be $h > 0$, the text answer $t > 0$, and the read video endpoint o. If $|o - t| \le \tau t$ and $|t - h| \le \epsilon_t h$, then the triangle inequality and $t \le (1 + \epsilon_t)h$ give

$$\frac{|o - h|}{h} \le \tau + (1 + \tau)\epsilon_t. \tag{8}$$

For an underlying endpoint differing from o by at most $\epsilon_E h$, add ϵ_E to the bound. Internal agreement is useful only insofar as the reasoner and reader are accurate; two agreeing modalities can both be wrong. In particular, a 30% gate and a text answer within 25% imply only a 62.5% worst-case endpoint bound, not a 25% certificate or policy-success guarantee. The downstream result is therefore empirical, not a consequence of choosing these tolerances.

H.3 EXECUTION, INDEPENDENT UNITS, AND UNCERTAINTY

Each trained policy predicts one scalar velocity for each of 50 goal values in $[1.05, 8.45]$. All 2,000 actions are executed in a headless Isaac scene at $\mu = .5$ and the same initial position. Forty-eight additional executions provide command truth for the modal comparison. The executor uses simulator state with no camera or feedback policy loop; appearance changes enter training observations, not the numerical success test. Success is $|d/w - g| \leq .25g$. The policies are *retrained* using the changed scene and its friction. The scene is new relative to the friction-0.3 pool, not withheld from training these policies. Goals are distinct from the clean scalar grid, although rounding goal-stripe pixel coordinates can create image aliases.

Table 17 reports arm-level success; Table 18 reports paired block differences and intervals.

Table 17: Separate inverse-label OpenVLA confirmation: means over eight blocks; 50 simulator-scored goals per block and arm. The fixed budget is 40 rows *before rejection*; rejected rows are not replaced.

Arm	Success
Unfiltered	0.0750
Random, physics-filter count	0.6800
Physics filter	0.7200
Available clean only	0.7075
Own-text filter	0.7525

For paired block differences $\delta_b = J_{b,A} - J_{b,B}$, intervals resample the eight complete blocks, not individual goal executions. We reproduce the reported 10,000-resample bootstrap and additionally enumerate its exact empirical distribution using multinomial weights $8!/(\prod_b n_b!)\, 8^{-8}$ for every $\sum_b n_b = 8$. The table uses exact empirical percentile intervals; they are pointwise, descriptive, unadjusted for multiple contrasts, and conditional on one generated bank and setting. They do not represent eight independent scene replications.

Table 18: Paired mean differences and exact empirical 95% bootstrap intervals. Neither a zero-containing interval nor a near-zero mean establishes equivalence.

Contrast	Difference [interval]
Physics − unfiltered	$+0.6450[+0.5525, +0.7300]$
Physics − random	$+0.0400[+0.0025, +0.0725]$
Physics − clean	$+0.0125[-0.0325, +0.0475]$
Own text − unfiltered	$+0.6775[+0.6075, +0.7425]$
Own text − physics	$+0.0325[-0.0025, +0.0700]$
Own text − clean	$+0.0450[+0.0000, +0.0925]$

The large recovery relative to unfiltered ingestion is supported. The smaller physics-minus-random gain is uncertainty-method sensitive: a paired t interval over the same eight blocks is $[-0.0047, +0.0847]$, crossing zero. The own-text-minus-clean t interval is $[-0.0161, +0.1061]$; the bootstrap lower endpoint is exactly zero. We claim neither improvement over available clean data nor equivalence between filters, nor general non-inferiority. The local protocol's P3 rule (interval includes zero and lower endpoint at least $-.10$) is reproduced as a documented decision rule, not an equivalence test or a community standard.

Paired seeds are not identical training outcomes. The text gate retains no generated row in 4 blocks. Those text-filter and clean-only arms have identical training rows and seed-derived schedules, yet their saved actions differ in all four blocks; success differs in 2 of them ($-.06$ and $+.02$). This exposes residual run-level variability despite seed pairing; the records do not identify its cause. The analysis therefore does not assume bitwise deterministic retraining or attribute every small paired contrast solely to retained examples. Resolving that variability and adding a text-count-matched random arm would be required for a sharper text-filter selectivity claim.

H.4 Protocol amendments and instrument limits that remain disclosed

This is an *amended* local protocol, not an untouched preregistration. The supplied dated record and earlier hash file document the following changes; local hashes do not independently establish a public registration timestamp.

1. The initial color choice washed out under tonemapping. Table/cube colors were changed before extractor or generated-output inspection, according to the run record. Camera geometry was unchanged.
2. The first extractor check failed: the 0.8 m/s inverse-label error was -11.1% and the 2.4 m/s clip was unreadable. A second check increased source capture rate from 15 to 30 fps but still failed at 2.4 m/s. The recorded stop distance there is about 9.5 reference units; near the image edge the dark visible top face exceeds the reader's 3,000-pixel area cap.
3. The run proceeded after narrowing commands to $[1.00, 2.05]$ m/s, rather than following the earlier stop rule. All four in-band validation points passed the 10% inverse-speed-label tolerance (worst 7.3%). This is a label-space check, not a 7.3% bound on distance error; the square-root conversion changes relative errors. Generated overshoots beyond about eight reference units may be unreadable and excluded from policy training.
4. Text requests initially truncated before the answer marker. The token limit increased from 1,024 to 4,096 with otherwise unchanged requests, according to the logs, before comparison against Isaac truth. Two text-only answers remain invalid. The earlier truncated attempt is not treated as an additional independent prediction.

The prompt says "to the right" while the physical motion is image-left; the endpoint reader uses absolute travel and therefore does not audit direction. The inherited clipping counter uses an obsolete 2.0 m/s ceiling and is not used in our analysis; the training normalization clips at 2.6 m/s. All these restrictions apply to both the modal evidence and its interpretation as a filter signal. They cannot be removed by increasing the number of executions at the same scene.

H.5 Complete clip-level outcomes

Figure 14 shows the declared illustrative strata, while Table 19 reports every clip in generation order. Missing measurements remain unavailable rather than being imputed as zero. The examples illustrate the classes; the complete table supplies the denominator.

Table 19: All 48 clip identities in generation order. Distances use the 0.06 m reference unit; — means unavailable, not a zero displacement. Truth is the separately executed commanded action.

ID	Isaac	Text	Image + text	Video
00	2.05	—	2.18	0.03
01	3.00	3.18	3.23	0.00
02	1.76	1.92	1.90	0.00
03	3.77	4.00	3.95	0.00
04	6.59	6.87	6.87	0.00
05	3.72	3.93	3.83	0.00
06	1.94	2.10	2.08	0.00
07	2.96	3.15	3.17	0.00
08	3.42	3.63	3.60	—
09	2.61	2.78	2.77	0.00
10	6.72	7.00	7.00	0.05
11	5.44	5.68	5.70	0.00
12	5.81	6.08	6.02	0.00
13	3.47	3.67	3.67	0.00
14	4.24	4.47	4.45	0.01
15	4.24	4.50	4.45	0.01
16	2.21	2.37	2.33	0.00
17	6.06	6.33	6.28	—
18	3.62	3.83	3.83	—
19	3.09	3.28	3.27	0.00

ID	Isaac	Text	Image + text	Video
20	2.83	3.00	2.95	0.00
21	4.35	4.57	4.57	0.00
22	2.65	2.83	2.83	0.00
23	2.65	2.83	2.80	0.00
24	1.73	1.88	1.88	0.01
25	2.87	3.05	3.05	0.01
26	6.59	6.87	6.87	0.00
27	3.28	3.48	3.50	—
28	1.91	2.05	0.21	0.01
29	3.92	4.13	4.10	0.00
30	3.42	3.63	3.60	3.58
31	6.72	7.00	7.00	0.00
32	6.52	6.80	6.80	0.00
33	5.81	6.07	6.02	0.00
34	2.74	2.92	2.88	0.00
35	3.33	3.53	3.53	0.00
36	5.75	6.00	5.95	—
37	5.75	6.00	5.95	0.00
38	6.00	6.27	6.17	0.00
39	3.33	3.50	3.53	0.08
40	1.70	1.83	1.83	0.01
41	4.08	4.30	4.28	0.00
42	3.62	—	3.83	3.08
43	2.36	2.50	2.50	0.20
44	6.26	6.53	6.53	0.00
45	2.13	2.28	2.25	0.05
46	6.52	6.83	6.80	0.00
47	3.87	4.08	4.07	—

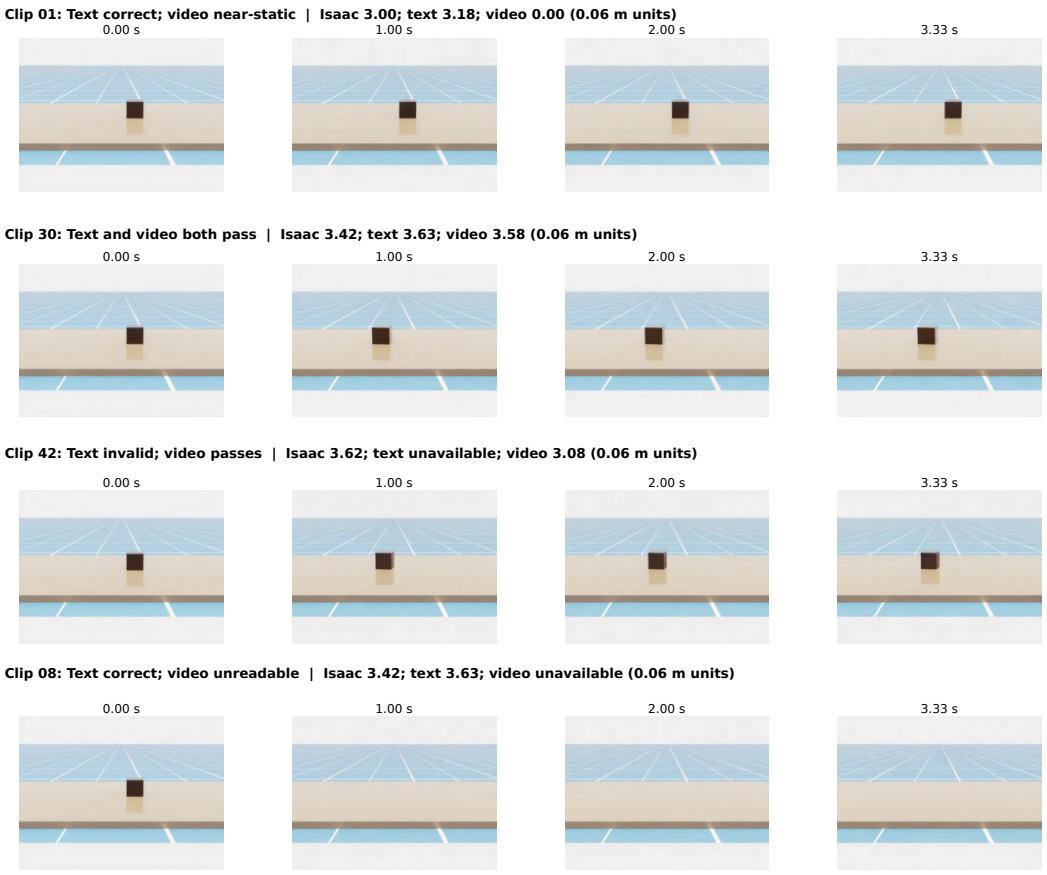


Figure 14: **Actual generated frames, including counterexamples and reader failure.** The lowest-ID clip is chosen in each of four declared strata: static video with correct text, jointly passing text/video, passing video with invalid text, and unreadable video. Thus both readable dynamic clips appear, not just failures. Full frames at indices 0, 24, 48, and 80 are shown without motion retouching. The unreadable row illustrates an instrument failure and is not assigned a physical error. This deliberately stratified figure is illustrative, not a prevalence estimate; all 48 videos, prompts, and values are available in the offline gallery.

I KNOWN-ACTION PROTOCOL, COMPLETE RESULTS, AND CONTROLS

I.1 OBSERVED INPUTS, MEASURED OUTCOMES, AND EVALUATION SCOPE

The forward-model protocol preserves the action supplied to each rollout. Training targets are outcomes, not actions reconstructed by inverting video motion. Push uses 27 measured speeds on [.55,1.85] m/s in .05 increments; driving uses 21 recorded speeds on [.4,2] m/s in .08 increments. Each task has 24 starting images: push start positions .10/.16/.22 m and goals 1.5–8.5 in unit increments; driving start positions .10/.25/.40 m and goals 2/2.7/3.4/4.1/4.8/5.5/6.2/7. The reference lengths are .06 m and .36 m, respectively. Their use as numerical units is not a measurement of the visible object's width or length.

Goal markers are physical rendered geometry with no collider. A fixed color segmentation extracts target and object centroids; a quadratic in horizontal pixel separation maps these to goal distance. An odd/even scene split gives held-out $R^2 = .999814$ for push and .999555 for driving, with mean absolute errors .02566 and .02935 reference units. The reported experiment then refits the reader on all 24 scenes before evaluation. It is therefore not a held-out perception test. The learned dynamics module receives only action, not pixels; perception is kept separate to isolate outcome-label contamination.

All candidate actions were executed at the oracle's reference start position; their final state displacements form $h(a)$. The policy chooses only among these actions, and scoring performs an exact table lookup. The oracle-optimal residual is at most .142 units for push and .122 for driving, with an attainable within-tolerance grid action at every evaluated goal. Applying the same outcome table to other starting positions assumes translation invariance over the uniform surface. No fresh 24-scene rollout is executed for every trained planner. Clean training outcomes and evaluation share this oracle, so the study does not test simulator-to-hardware transfer or unseen physical dynamics.

Figure 15 shows starting-scene variation and the physical goal markers used by the calibrated reader. These are configurations within the stated task scenes, not additional environments.

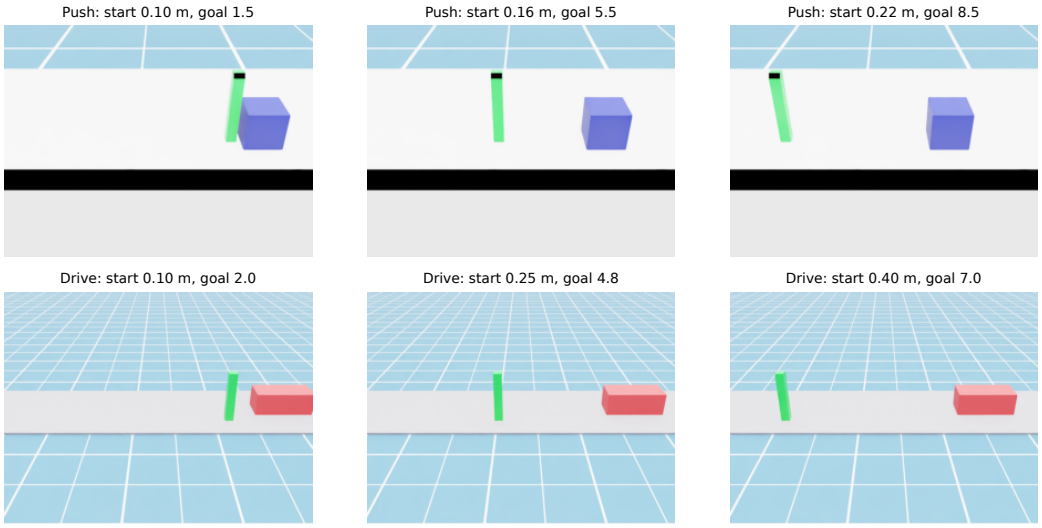


Figure 15: **Goal markers are part of the rendered scene.** Six illustrative starting observations from the 48-image collection: three positions per task, with different physical target locations. The markers are not image overlays. These examples show configuration variation, not variation in friction, camera model, or task family. The goal reader is a calibrated color-centroid mapping, separate from the learned scalar forward model.

I.2 CANDIDATE BANKS AND SAMPLING

The push pool contains 60 readable rows at 13 distinct commands: 29 from the 40-clip sweep and 31 from a 32-clip dense bank. Thus 12 attempted generations are excluded as unreadable before the training budget is sampled. The dense bank was independently re-extracted without overwriting the source data; all 31 readable (a, o) pairs match exactly, and the existing-sweep pair multiset matches its per-clip records. The aggregate pool lacks original row-level identifiers, so equal (a, o) rows can have indistinguishable source assignments; clip-level records and hashes are retained without inventing unique identities. The measured means are 6.0376 for readable sweep rows and 2.8929 for readable dense rows. Both contain small-motion and large-motion clusters; these subgroup means must not be substituted for the locations of individual modes.

At each nominal f, the counts are $n_w = \mathrm{round}(44f)$ and $n_c = 44 - n_w$. Sampling uses replacement from the fixed banks, with seed $7000 + b$ for repetition b. Unfiltered and filtered arms share that draw. The actual source fractions are therefore 9/44, 18/44, 26/44, and 35/44 at the nonzero nominal settings; examples need not be unique. The same repetition seeds are reused across settings, so five settings do not supply five times as many independent seed blocks. The retained fraction is reported as retention, not "leakage" or a measured false-acceptance rate.

I.3 LEARNERS AND COMPLETE TABLES

The kNN estimator averages outcomes of the five nearest commanded actions (or all rows if fewer than five remain). The planner evaluates it at every recorded action and chooses the first minimizer. If the training set is empty, the implementation falls back to the median grid action. This fallback is specified rather than silently treating an empty model as accurate.

The MLP has architecture $1 \to 64 \to 64 \to 1$ with tanh hidden activations. Each arm standardizes actions/outcomes using that arm's training set, resets Torch seed $9000 + b$, and takes 600 full-batch Adam updates at learning rate .003. Its reported ten-run means/standard deviations are archived aggregates. An independent CPU recreation of the high-contamination setting using the same data/initialization seeds yields .4583 unfiltered and 1.000 filtered, compared with reported GPU means .446 and 1.000. It preserves the qualitative effect but is not an exact numerical reproduction of the producing GPU run. The recreation is not counted as ten extra independent experiments; original per-seed GPU predictions are unavailable in the supplied aggregate file.

Tables 20 and 21 report the complete kNN and archived push-MLP results.

Table 20: Completed kNN results independently recomputed from frozen images, outcome grids, and pools. Means ± population standard deviations over 20 dataset draws. Paired intervals condition on these fixed inputs. Driving is subject to the normalization qualifications below.

Study	Nominal f	Clean/WM	Unfiltered	Filtered	Difference [95%]
Push kNN	0.0	44/0	1.000 ± 0.000	1.000 ± 0.000	$+0.000$ $[+0.000, +0.000]$
Push kNN	0.2	35/9	0.748 ± 0.139	0.994 ± 0.027	$+0.246$ $[+0.181, +0.308]$
Push kNN	0.4	26/18	0.512 ± 0.167	0.967 ± 0.067	$+0.454$ $[+0.375, +0.533]$
Push kNN	0.6	18/26	0.398 ± 0.210	0.915 ± 0.086	$+0.517$ $[+0.421, +0.602]$
Push kNN	0.8	9/35	0.338 ± 0.141	0.635 ± 0.150	$+0.298$ $[+0.196, +0.394]$
Drive kNN	0.0	44/0	0.998 ± 0.009	0.998 ± 0.009	$+0.000$ $[+0.000, +0.000]$
Drive kNN	0.2	35/9	0.919 ± 0.064	0.973 ± 0.050	$+0.054$ $[+0.021, +0.087]$
Drive kNN	0.4	26/18	0.810 ± 0.111	0.921 ± 0.068	$+0.110$ $[+0.060, +0.160]$
Drive kNN	0.6	18/26	0.742 ± 0.097	0.863 ± 0.087	$+0.121$ $[+0.065, +0.177]$
Drive kNN	0.8	9/35	0.660 ± 0.207	0.758 ± 0.087	$+0.098$ $[+0.017, +0.181]$

Figure 17 connects generated outcome labels, learned forward predictions, and recorded-grid decisions. It visualizes the same known-action assay rather than adding fresh policy executions.

Table 21: Reported push MLP results, ten runs; means ± population standard deviations. These are a scalar forward model, not a vision CNN or OpenVLA.

Nominal f	Clean/WM	Unfiltered	Filtered
0.0	44/0	1.000 ± 0.000	1.000 ± 0.000
0.2	35/9	0.900 ± 0.086	1.000 ± 0.000
0.4	26/18	0.617 ± 0.175	1.000 ± 0.000
0.6	18/26	0.575 ± 0.118	1.000 ± 0.000
0.8	9/35	0.446 ± 0.185	1.000 ± 0.000

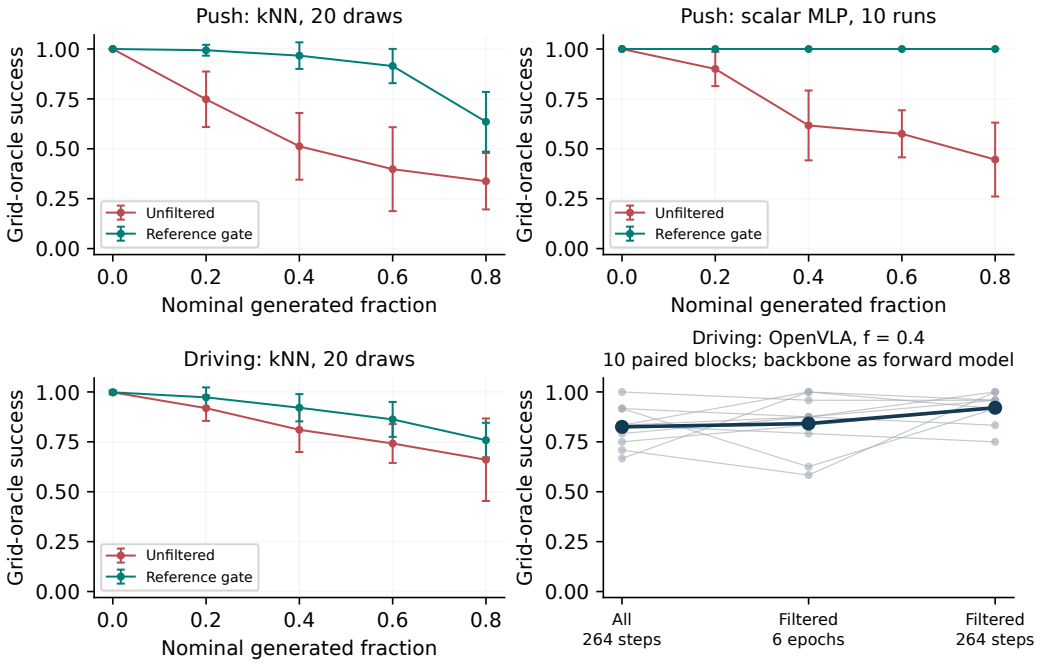


Figure 16: **Supporting forward-learner comparisons.** Task/learner pairs are shown separately, not averaged into a fictitious full task–policy grid. The first three panels use the analytic gate; actual own-text selection induces the same push kNN acceptance set. The OpenVLA driving panel is a distinct one-fraction exact-update control, with method-sensitive uncertainty and normalization-dependent grid scoring, not fresh policy executions.

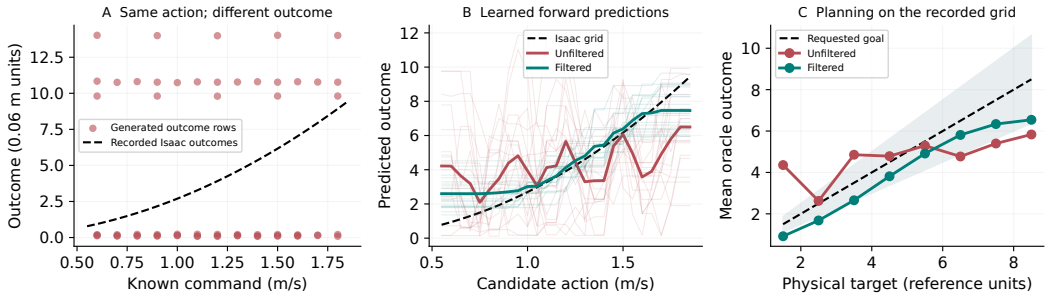


Figure 17: **Current known-action pipeline.** The commanded action remains the input. Generated outcomes corrupt the forward-model target; no action is recovered from video. At nominal $f = .8$, filtering changes the learned forward predictions and the resulting action choices. The final panel reports outcomes from the recorded Isaac grid, averaged over repetitions and the configurations sharing a goal. It is not a video or fresh execution of each policy decision.

I.4 Additional CPU-only analyses, clearly separated from the reported runs

Reanalysis preserves the published draws, fit rules, and evaluation for the two original arms. It adds controls without collecting data or changing any commanded action: random-same-count selects a uniform subset of the drawn WM rows using seed $100000 + b$; available-clean-only fits the identical clean subset; and simulator-reference filtering substitutes recorded $h(a)$ for the stated law while retaining the same 30% window. The last control has privileged simulator information and is not a deployment baseline or an upper bound on finite-data policy success.

A fixed-input outcome-repair control keeps every action occurrence and all 44 rows of the unfiltered draw, but replaces each generated outcome with its recorded Isaac value. This uses additional privileged outcome supervision, so it is a causal diagnostic of the pipeline, not a competing method under the clean-data budget. It separates outcome-label corruption from changes in action marginal or sampling density. At nominal $f = .8$ it scores .9917 for push and .9938 for the normalized driving study, versus .3375/.6604 unfiltered. The paired improvements are .6542 [.5896,.7146] and .3333 [.2438,.4250]. The action roster is unchanged; the intervention changes only the outcome column. For kNN this also preserves the neighbor identities and averaging weights at every grid query.

As a sanity baseline, directly planning with the stated equation and the same pixel-read goal, without learning from any rows, scores 1.000 for push and .875 for driving. These elementary tasks can often be controlled from known laws; we do not claim that learning or generated demonstrations are necessary to solve them. Their role is to expose propagation of measured outcome errors through a controlled learning pipeline, not to outperform an analytic controller.

Table 22 reports these additional controls for both tasks at nominal $f = .8$.

Table 22: Additional kNN controls at nominal $f = .8$, preserving the original 20 draws. Only U/F were in the original two-arm study; R/C/S/O are added analyses. S uses a simulator-reference gate; O repairs outcomes at unchanged inputs.

Task	U	F	R	C	S	O
Push	0.338	0.635	0.548	0.540	0.635	0.992
Drive	0.660	0.758	0.715	0.594	0.742	0.994

For push, F−R is .0875 [.0188,.1625] and F−C is .0958 [.0354,.1625]. This conditional kNN analysis supports beneficial retention in this particular sampled-support regime, unlike the older inverse-label experiment. It is not a universal synthetic-data advantage or the criterion for the main mitigation claim. For driving, F−R is .0438 [−.0375,.1396], so selector superiority is not established by that interval. Its F−C contrast is .1646 [.1167,.2125] under the assumed normalization. Reference-filter changes are discussed below.

The intervals use 100,000 percentile-bootstrap resamples of the complete paired 20-repetition vectors with analysis RNG seed 20260925. Targets are not resampled as independent trained policies. These are pointwise, uncorrected, conditional intervals; the fixed banks, selected settings, and reused seeds limit broader inference. The new controls were added after inspecting the original results and are identified as diagnostic reanalysis, not preregistered confirmation.

A further reader-split sensitivity check fits the goal reader only on the 12 even-indexed scenes and evaluates only the other 12, without changing the physics-model training draws (Table 23). It preserves the mitigation ordering. The images were already part of the original collection; this is not an independently collected held-out benchmark or a claim of general visual understanding.

Table 23: Split-reader sensitivity, 12 evaluated configurations, 20 kNN draws.

Task	Nominal f	Unfiltered	Filtered
Push	0.4	0.517	0.988
Push	0.8	0.354	0.650
Drive	0.4	0.800	0.979
Drive	0.8	0.642	0.812

I.5 ALL-GENERATED BOUNDARY: SUPPORT AND FALLBACK MATTER

The main mixed-budget sweep stops at nominal $f = .8$. A separate kNN boundary calculation at $f = 1$ reproduces push .3167 unfiltered versus .0750 filtered. The only passing pool rows share action 1.8, so filtering can leave a constant forward estimate over the whole grid. The first-argmin tie-break then chooses the smallest grid action. Empty sets instead trigger the median-action fallback. The result therefore depends on action-support collapse and planner conventions, not only the number of rows. Driving's normalized boundary analysis likewise gives .5083 versus .2563; it remains subject to its mapping assumptions.

We do not use the archived MLP $f = 1$ value as evidence of statistical starvation: five of its ten draws retain one row, and default unbiased standardization of one sample can produce NaN, which `clamp_min` does not repair. The reported $f \leq .8$ settings contain at least nine clean rows. This edge case must be resolved before a neural all-generated boundary claim is made.

I.6 DRIVING: REFERENCE MISMATCH AND UNVALIDATED NORMALIZATION

The completed grid fits $h(V) \simeq 4.16659V - .54491$ in .36 m reference units, whereas the stated constant-speed law is $r(V) = 1.5V/.36$. The absolute difference is .5447–.5453 units at the stored precision. As a fraction of the text-law prediction it is approximately 6.5–32.7%; as a fraction of the measured outcome it is approximately 7.0–48.5%. Denominators must be stated. This is a mismatch between the constant-speed word problem and the velocity-reset/frictional simulator protocol, not necessarily a failure to solve the word problem.

The declared generated pool assigns $V = .8\,s$ from coordinate scale s and maps Farnebäck flow through the older `ego_matched_final` fit. This does not establish metric velocity from a normalized pose trajectory, or metric distance from flow across different scenes/depths. The older calibration also predates the capture-timing correction. The AV clips span 3.2 s, while the conversion projects their flow statistic to a nominal 1.5 s outcome rather than measuring a duration-matched metric trajectory. Sharing a mapping across arms does not validate it: changing it can change which rows pass the reference gate and how they influence a model trained on the physical action grid. We therefore report the resulting scores as normalization-dependent, without a five-fold physical speed claim or a calibrated native-action replication claim.

Under the declared mapping, the stated-law gate keeps five of 15 rows, while the same-tolerance simulator-reference gate keeps six. Their high-contamination successes are .7583 and .7417; the paired difference .0167 has interval [−.0063,.0417]. More accurate references do not mechanically guarantee better finite-data learning. For push, both gates retain the same two source rows and produce identical results. These comparisons illustrate why the internal and external references are distinct even when they coincide in some tasks.

J MATHEMATICS FOR KNOWN-ACTION OUTCOME LEARNING

J.1 CONTRACTS, REFERENCES, AND READER ERROR

On a fixed common observable space with positive scales $s_j(x)$, define

$$d_x(C, C') = \max\{\mathbf{1}[z \neq z'], \max_j |m_j - m'_j|/s_j(x)\}. \tag{9}$$

This is a metric on the represented coordinates (a pseudometric if other fields are retained). Write $D_{\text{text}} = d_x(T_{\text{text}}, T_{\text{phys}})$ and use the same observable scales for both comparisons. The triangle inequalities imply

$$|D_{\text{ext}} - D_{\text{int}}| \leq D_{\text{text}}, \qquad D_{\text{ext}} \leq D_{\text{int}} + D_{\text{text}}.$$

The same coordinates and scales must be used for all references; missing or censored measurements are not imputed as zeros. If an accepted scalar training outcome o satisfies $|o - r(a)| \leq \tau s(a)$ and the reference error obeys $|r(a) - h(a)| \leq \eta_r(a)$, then

$$|o - h(a)| \leq \tau s(a) + \eta_r(a). \tag{10}$$

If the reader differs from the video's underlying measurable outcome o_v by at most ϵ_E, the corresponding physical-outcome bound adds ϵ_E. These conditional bounds do not establish that every reference or reader satisfies a small error bound in generated imagery.

J.2 Forward prediction error propagates into planning

Proposition 1 (Finite-grid planning bound)**.** *Let $\mathcal{A}$ be a finite, nonempty action grid and $h : \mathcal{A} \to \mathbb{R}$ the recorded outcome oracle. Suppose $\max_{a\in\mathcal{A}} |\hat{h}(a) - h(a)| \le \epsilon$ and $|\hat{g} - g| \le \eta$. If $\hat{a}$ minimizes $|\hat{h}(a) - \hat{g}|$ on the grid, then*

$$|h(\hat{a}) - g| \le \delta_{\mathcal{A}}(g) + 2\epsilon + 2\eta, \quad \delta_{\mathcal{A}}(g) = \min_{a\in\mathcal{A}} |h(a) - g|.$$

If a scene's true execution function h_x differs from the stored oracle by at most ν over the grid, the bound for $|h_x(\hat{a}) - g|$ adds ν.

Proof. Choose $a^* \in \arg\min_a |h(a) - g|$, which exists on a finite grid. Then

$$\begin{aligned} |h(\hat{a}) - g| &\le |h(\hat{a}) - \hat{h}(\hat{a})| + |\hat{h}(\hat{a}) - \hat{g}| + |\hat{g} - g| \\ &\le \epsilon + |\hat{h}(a^*) - \hat{g}| + \eta \\ &\le \epsilon + |\hat{h}(a^*) - h(a^*)| + |h(a^*) - g| + |g - \hat{g}| + \eta \\ &\le 2\epsilon + \delta_{\mathcal{A}}(g) + 2\eta. \end{aligned}$$

The scene-transfer statement follows by one more triangle inequality. No assumption that filtering reduces ϵ is made in the proposition. □

A sufficient condition for the task's relative-tolerance success is $\delta_{\mathcal{A}}(g) + 2\epsilon + 2\eta \le .25g$ for $g > 0$; this can be conservative. When a model is undefined because all rows were rejected, the proposition does not certify the fallback policy. It also does not prove that the underlying dynamics are translation-invariant: that assumption is isolated in ν rather than silently removed.

J.3 Selection changes bias and support, not just dataset size

Let the clean action marginal be $p_c(a)$, the generated marginal $p_w(a)$, and source fraction f. Assume clean conditional mean $h(a)$ and generated conditional mean $h(a) + b(a)$, with finite second moments. At zero-probability source/action pairs the corresponding unused bias can be set to zero. The source posterior is

$$\rho_f(a) = \frac{f p_w(a)}{(1-f)p_c(a) + f p_w(a)}.$$

Thus the unrestricted squared-loss population minimizer is $m_f(a) = h(a) + \rho_f(a)b(a)$. If the gate accepts generated rows with probability $r(a) = \mathbb{P}(q = 1 \mid a, \mathrm{WM})$ and their accepted conditional bias is $b_q(a)$ (set to zero if $r(a) = 0$), its minimizer is

$$m_q(a) = h(a) + \rho_q(a)b_q(a), \qquad \rho_q(a) = \frac{f p_w(a) r(a)}{(1-f)p_c(a) + f p_w(a) r(a)}.$$

These formulas apply where the relevant denominators are positive. They follow by conditioning on source and acceptance; squared-loss minimizers are conditional means. For a clean evaluation distribution $\mu(a)$ supported where both predictors are defined, let $R_c(m) = \mathbb{E}_{a\sim\mu, Y_c|a}(Y_c - m(a))^2$. The risks differ by

$$R_c(m_f) - R_c(m_q) = \mathbb{E}_{a\sim\mu}\left[\rho_f(a)^2 b(a)^2 - \rho_q(a)^2 b_q(a)^2\right],$$

because the cross term with clean zero-mean residual vanishes. Selection need not reduce the expression: it can retain worse bias or eliminate action support. This accommodates unequal source action distributions, as in the actual banks. It is an explanatory population identity, not a model of the finite-sample kNN or MLP optimizer and not a task-success guarantee.

K MEASUREMENT SCOPE AND DATASET COVERAGE

The original alignment figures retain the source-bound 160-clip sweep: 120 bounce/incline/pendulum records plus the separate 40-clip push sweep. Additional projectile and Nano records are outside this sweep. The reader's per-axis qualification, visibility censoring, force convention, and period-versus-amplitude distinction remain necessary; strong fit on selected simulator renders does not establish universal generated-domain validity. The known-action study's goal reader is a different perception component from the generated-video outcome reader, and their accuracy statistics are not interchangeable.

The supplementary inverse-label OpenVLA experiments reconstruct action targets, whereas the main study preserves known commands and learns outcomes. The two representations and evaluation procedures are reported separately. The small neural forward model is a scalar MLP, not a vision CNN; the OpenVLA backbone is an additional driving forward learner. Known-action results cover push and driving. The physical-example video galleries remain useful illustrations of the original audit; they are not videos of the forward-model planners. Frozen numerical inputs, starting images, source hashes, and reanalysis scripts support reproduction of this revision without changing the live experiment directory. Models, simulator images, and public-release approval are separate requirements for a full regeneration release.

K.1 PER-RENDER OUTCOME-READER CHECKS

Table 24 reports all 15 entries in the frozen validation record. Trajectory fits and stopping-endpoint errors are different metrics; they are not pooled into a single reader score. Two projectile renders lose tracking as motion leaves the field of view or becomes blurred. They are reported as such rather than assigned a favorable fit on the remaining frame. The static incline render is correctly measured as zero motion. The short-travel push case has 21.2% endpoint error, illustrating why a strong fit on larger motions does not certify all small displacements. No bounce validation row is present in this particular record; calibration fits reported elsewhere should not be substituted for a held-out bounce check.

Table 24: Complete saved validation entries. Trajectory R^2 uses tracked frames; tracked/present is a visibility/detection count. Push rows report stopping endpoints in .06 m reference units and relative endpoint error. Dashes denote an inapplicable or unavailable statistic, not zero error.

Mechanism	Condition	Tracked / present	Trajectory R^2
Pendulum	$L = .25$ m	80/80	0.9633
Pendulum	$L = .50$ m	80/80	0.9780
Pendulum	$L = .75$ m	80/80	0.9852
Pendulum	Base scene	60/60	0.9902
Projectile	$v = 1.5$ m/s	7/7	0.8541
Projectile	$v = 3.5$ m/s	2/8	–
Projectile	$v = 5.0$ m/s	1/11	–
Incline	$\mu = .35$	14/15	0.9997
Incline	$\mu = .45$	40/40	0.9997
Incline	$\mu = .60$	80/80	–
Push condition	**Reference endpoint**	**Read endpoint**	**Relative error**
$v = .6$ m/s	0.938	0.739	21.2%
$v = .9$ m/s	2.171	2.218	2.2%
$v = 1.5$ m/s	6.165	6.492	5.3%
$v = 1.8$ m/s	8.926	9.449	5.9%
Base scene	3.913	3.912	0.0%

These checks qualify the measurement under the tested rendering conditions. They do not transfer a uniform error bound to generated appearances, eliminate visibility censoring, or validate native metric control in driving. Re-extracting archived videos with the same reader establishes computational reproducibility, not an independent accuracy assessment.

L KNOWN-ACTION MODEL-FORECAST ADMISSION: PROTOCOL AND CONTROLS

This extension compares actual model-derived admission with motion screening and count-matched removal on the known-action push bank. These controls do not match the retained action profile; the additional sensitivity in Appendix N leaves the within-command benefit unresolved. It reuses a fixed, inspected bank: 60 readable rows from 72 attempted clips at 13 commands. The analysis plan was frozen before the additional model queries, but the bank and baseline outcomes were already known. It is a retrospective extension, not an independently generated confirmation cohort. Every source video hash and the command/outcome multiset were checked, and all 72 videos are included in the evidence archive.

L.1 REQUESTS, PARSING, AND REFERENCE PROVENANCE

Each request contains the exact generation prompt followed by a question asking for stopping distance in meters. It contains no measured or computed outcome. We use the same Cosmos3-Super checkpoint identifier and original conditioning frame, with text-only and image–text modes, temperature zero, a fixed seed, and a 4,096-token output budget. Identical conditions share a cached answer: there are 26 distinct mode-specific requests, not 120 independent forecasts for 60 rows. Up to two identical-request retries are permitted when the last `ANSWER = <number>` marker is unparsable. All raw responses and attempts are retained; no answer is filled from the physical law.

Text-only yields $12/13$ valid conditions and image–text $13/13$, across 31 request attempts. The text-only response at 1.6 m/s remains truncated after the fixed retry budget. Its four associated rows, including one moving clip, are rejected. The 12 valid text conditions and 13 image–text conditions are all within the 25% endpoint tolerance against the recorded Isaac outcomes. Validity and accuracy remain separate: a truncated response is an unavailable forecast, not a demonstrated numerical reasoning error.

The forecasts are served with tensor parallel size one; the matched-modal friction-0.5 cohort used size two. These are additional queries, not a bitwise inference replay. Two infrastructure startup failures occurred before any forecast was collected (CLI quoting and the path to an installed compiler utility); neither changed the requests, model, precision, or analysis thresholds. Model inference is an additional cost, even though no per-rollout simulator answer is needed by the gate.

The generation prompts specify speed and friction but do not give an explicit numeric cube width. Absolute magnitude comparisons therefore use the declared fixed-scene calibration. They do not establish calibrated native metric control in arbitrary generated scenes. Parameter-response diagnostics and downstream label effects are interpreted under this setup.

L.2 PRESERVED LEARNING PROCEDURE AND MATCHED CONTROLS

All arms preserve the original 20 draws, initial $N = 44$ budget, replacement sampling, clean subset, five-neighbor predictor, calibrated goal reader, and recorded action grid. The five nominal fractions are $0, .2, .4, .6, .8$; no fraction is chosen after inspecting the additional forecasts. Dataset seeds are $7000 + b$. One WM permutation, seeded $100000 + b$, supplies random subsets at each gate's retained count. Moving-count controls use the same order restricted to moving rows first. A static fallback is defined if too few moving rows exist, but it is never used in these comparisons.

The motion criterion is inherited from the reader: travel at least .6 reference units. It is not tuned on downstream success. The own-text gate uses the same 30% relative window as the stated-law comparison, with a 10^{-6} denominator floor for a zero forecast. Invalid forecasts reject rows. A zero numeric forecast is retained as a valid prediction rather than silently discarded for disagreeing with physics. No such zero forecast occurs among the parsed answers here.

Table 25 reports success and retained-row counts for the complete nominal-$f = .8$ comparison.

Uniform random and motion-count controls have exactly the same retained-row count as their corresponding gate in every draw. The main contrasts use text-only forecasts; image–text is secondary. Pointwise 95% intervals use 100,000 resamples of the 20 paired blocks, seed 20260925, and condition on the fixed bank, reader, and grid. Neither duplicated training rows nor reused forecasts

Table 25: Push kNN at nominal $f = .8$, with nine clean and 35 generated rows before admission. Means over 20 paired draws; retained counts refer to generated rows after screening. Counts need not be integers after averaging over draws.

Arm	Success	Mean retained WM rows
Unfiltered	0.338	35.00
Motion-only	0.188	13.05
Random, text-count matched	0.548	1.20
Moving, text-count matched	0.487	1.20
Available clean only	0.540	0.00
Own-text admission	0.635	1.20
Image–text admission	0.635	1.85

are counted as independent generated examples. Every admitted index, forward prediction, selected action, and achieved outcome is archived.

L.3 PROXY QUALITY AND ITS BOUNDARY

The bank contains 37 static and 23 moving rows. Under the 25% external endpoint criterion, two rows are correct. Actual text admission accepts both and rejects the other 58; its accepted indices match the stated-law gate. This describes which source rows are admitted, not an independently identified planning mechanism or a precise population precision estimate from only two accepted examples. Numerical internal and external errors, evaluated on the same recorded-truth scale, have Spearman correlation .937 across 56 rows with valid text forecasts and 1.000 on the 22 valid moving rows. These are fixed-bank descriptive associations, not independent evidence that internal agreement certifies physical accuracy.

The image–text gate accepts one additional row at 1.7 m/s. Its forecast is .500 m, the recorded outcome .47694 m, and the measured video endpoint .6468 m. The forecast-relative 30% window admits that video, whereas the recorded-reference window rejects it; external endpoint error is 35.6%. This is an operational boundary-sensitivity example involving reference error and relative tolerances, not a powered study of severe reasoning failures. The image–text gate's mean kNN success equals text-only success, which also shows that one more externally inaccurate label does not mechanically imply a lower aggregate task score.

For comparison, in the friction-0.5 bank of Appendix H, the physics gate accepts exactly the two clips classified as moving, and the text gate accepts exactly the moving clip with a valid text answer. That bank therefore cannot identify an admission advantage over those simpler rules. The 60-row bank differs: 21 of its 23 moving outcomes fail the primary reference. That fact distinguishes the gate from a motion predicate, but the planning advantage still depends on retained action support (Appendix N).

Direct forecasts as an attainability diagnostic. A post hoc check plans directly over the 12 commands with parsed text forecasts, using the same goal reader and recorded oracle, without fitting a dynamics model or consuming new model requests. It attains 1.000 on the 24 goals; the analogous 13-command image–text diagnostic also attains 1.000. These are different controllers, not additional learned-policy repetitions. They make the scope of the learning assay explicit: the scalar task does not require video-derived supervision, and the experiment measures the consequences when such supervision is nevertheless admitted. This finding does not establish that filtering is preferable to direct forecasting when the latter suffices for the task.

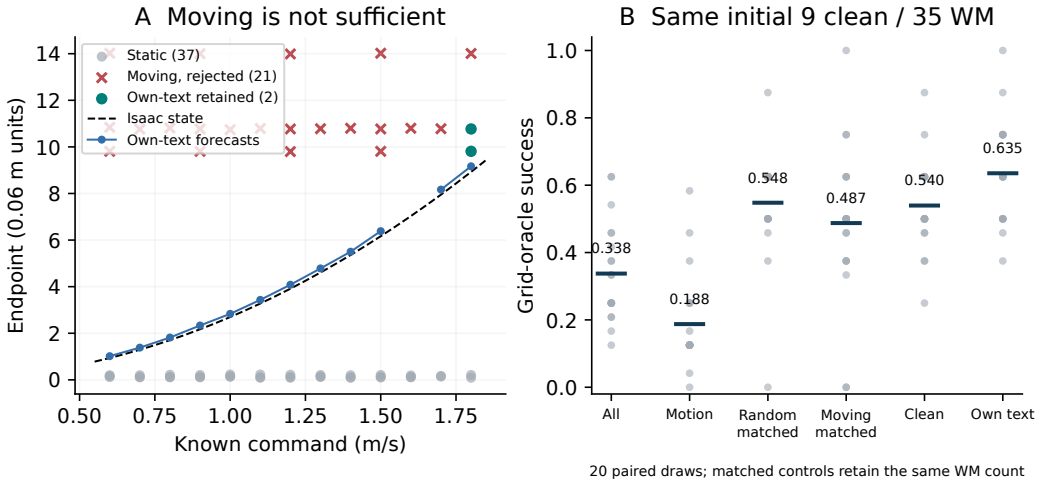


Figure 18: **Original count and motion controls, with their scope explicit.** The primary gate keeps two of 60 readable source rows, both moving and at 1.8 m/s. It outperforms the plotted fixed-bank baselines, but this panel does not match action composition. Table 28 supplies that additional control; it leaves the within-command benefit unresolved.

M OpenVLA forward prediction: optimizer-matched control

This is a targeted control at driving $f = .4$, not a complete step-matched sweep. It retains the commanded action as input and trains an OpenVLA-7B LoRA backbone to predict the outcome from the starting image and candidate speed. The first coordinate of its native action-token vocabulary is repurposed for the outcome range $[0, 10]$ reference units. The planner sweeps the recorded action grid; the experiment is not native seven-degree-of-freedom policy execution.

Ten blocks use dataset seeds $7000 + b$ and initialization/schedule seeds $9000 + b$. Each draw has 26 clean and 18 generated rows. The stated-law gate and the driving command/flow normalization are unchanged. Every training row receives a randomly drawn real starting image; physical outcome is assumed independent of the goal marker and initial position over this uniform scene. All scene images are hash-bound to the original collection.

Both final arms receive 264 optimizer updates, the unfiltered six-epoch budget. The filtered arm is evaluated at its six-epoch checkpoint (192 updates on average), then continued to 264 on the same model, optimizer, and random-number trajectory. Intermediate evaluation preserves the training RNG state. LoRA rank is 16, scale 32, dropout zero; AdamW learning rate is $2 \cdot 10^{-4}$. Every training row/scene choice, step number, loss, grid prediction, and decision is retained. Independent checks reconstruct the full schedules and score every decision against the recorded oracle.

Table 26 separates the filtered trajectory before and after continuation to the exact unfiltered update budget.

Table 26: Driving OpenVLA control at $f = .4$: ten paired blocks and 24 grid-scored goals per checkpoint. Filtered checkpoints share one training trajectory; they are not additional independent blocks. Driving remains normalization-dependent.

Arm/checkpoint	Optimizer updates	Success
Unfiltered	264	0.825
Filtered, epoch-matched	192 mean	0.842
Filtered, step-matched	264	0.921

Filtered-minus-unfiltered at matched steps is $+0.096[+0.013, +0.175]$ under paired bootstrap; the paired-t interval is $[-0.004, +0.195]$. Continuing the filtered trajectory gives $+0.079[-0.008, +0.183]$; its paired-t interval is $[-0.037, +0.195]$. These pointwise intervals condition on the fixed scene/bank and ten blocks. The matched-step point estimate is encouraging but method-sensitive, and the continuation interval includes zero. The control removes a budget difference by design; it does not prove that optimization alone explains every difference in the full sweep.

Complete epoch-matched sweep, retained for transparency. The separately produced sweep uses six epochs at every fraction. Its unfiltered/filtered means at $f = 0, .2, .4, .6, .8$ are respectively .958/.958, .858/.829, .800/.804, .637/.529, and .400/.438, with ten repetitions per setting. These curves are not uniformly positive and their filtered arms have fewer updates. The f=0 filtered score reuses the unfiltered evaluation; its equality is not an independent determinism test. The targeted rerun's unfiltered mean .825 differs from the sweep's .800, so it is reported as a separate controlled run rather than an exact training replay. Per-block uncertainty is computed from the controlled run, not by pooling the two runs as twenty independent datasets.

M.1 Five-fraction sweep with approximately matched updates

Table 27 and Figure 19 report a separate complete sweep with ten paired blocks at each fraction. It preserves the known-action driving representation and analytic gate. The three arms are unfiltered at six epochs, filtered at six epochs, and filtered with the epoch count set to $\mathrm{round}(264/n_b)$, where n_b is the retained training-row count in block b. The third arm therefore approximates, but does not exactly match, 264 updates. We name it the rounded-epoch arm rather than adopting the source file's `stepmatched` identifier as a literal guarantee. The filtered arms are separately fitted with the same seed; they are not checkpoints on one continued trajectory as in the exact-update control above.

Table 27: Full driving OpenVLA sweep. Means over ten paired blocks per fraction. Unfiltered always uses 264 updates; the final column reconstructs the rounded-epoch arm's update range from the saved data and declared sampling seeds. At $f = 0$ the same evaluation is reused across arms.

f	Unfiltered	Filtered, six epochs	Filtered, rounded epochs	Updates
0.0	0.967	0.967	0.967	264–264
0.2	0.879	0.850	0.946	246–280
0.4	0.800	0.817	0.900	270–280
0.6	0.692	0.529	0.800	252–275
0.8	0.408	0.429	0.708	253–270

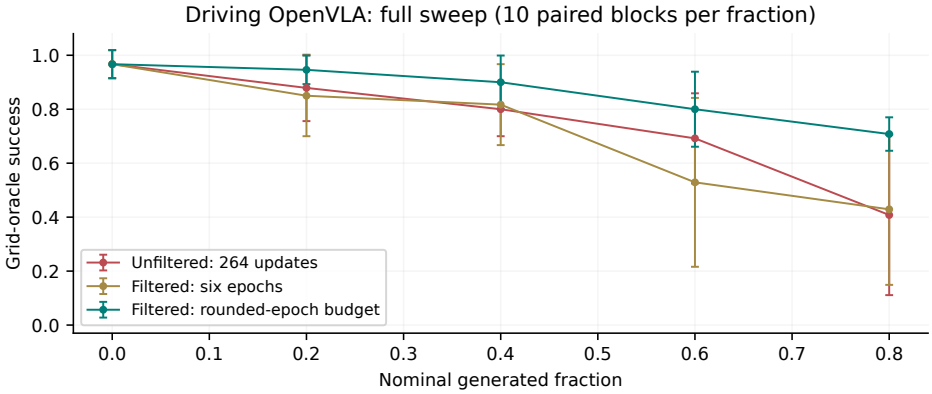


Figure 19: **Data rejection and training budget must be distinguished.** Means and block standard deviations for each fraction. The six-epoch filtered curve is not uniformly better than unfiltered training; the rounded-epoch curve uses approximately the unfiltered update budget. Curves are not pooled with the targeted exact-update run, and fractions reuse dataset seeds.

The archived source contains per-block success rates rounded to three decimals, not complete action-decision traces. Offline checks verify aggregate means, standard deviations, and paired differences within that precision; schedules are reconstructed from the declared code and seeds rather than verified from step logs. The source and result hashes accompany the archive. The targeted exact-update experiment above retains full schedules and decisions and is the basis of the main text's exact-budget comparison. Neither sweep tests own-text admission on driving, and both inherit the command/flow normalization.

N Action-support sensitivity and complete primary controls

N.1 The question the count controls leave open

Uniform- and moving-count controls preserve how many generated rows the gate retains but not the actions attached to those rows. In this bank, the two accepted source rows both have command 1.8 m/s. Better downstream performance could therefore reflect the selected command profile, the selected outcome values at that command, or their interaction with the clean subset and planner. Source-fixed outcome repair establishes that corrupted outcomes can damage the unfiltered assay; it does not identify which of these pathways explains the filtering gain.

N.2 Exact within-draw action and count matching

This sensitivity was designed after inspecting the fixed-bank results. It is an explicitly post hoc reanalysis, not a locally preregistered or independent confirmation. It uses no new model queries, generated clips, simulator executions, or gradient-based training runs.

For draw b, let W_b be its multiset of 35 generated row occurrences and C_b its nine clean rows. Let $k_b(a)$ be the own-text gate's retained occurrence count at command a. Define

$$\mathcal{S}_b = \{S \subseteq W_b : \#\{i \in S : a_i = a\} = k_b(a) \text{ for every } a\}. \tag{11}$$

Here subsets are over *draw positions*: duplicate source rows remain distinct sampled occurrences. For every $S \in \mathcal{S}_b$, retain the same clean rows, preserve the selected occurrences' original order, fit the same five-neighbor predictor, and score the same recorded grid and goal reader. The control is the exact expected success of uniform subset selection,

$$J_b^{\text{act}} = \frac{1}{|\mathcal{S}_b|} \sum_{S \in \mathcal{S}_b} J(C_b \cup S). \tag{12}$$

It averages *scores of separate subset policies*, not their predictions or actions. It is not an ensemble policy with additional inference capability. When the gate retains no generated rows, the single eligible subset is empty and both arms use the same clean rows. On this bank all nonzero $k_b(a)$ occur at 1.8 m/s, but the enumeration handles arbitrary command strata.

The action profile comes from the gate itself, so this is a privileged mechanism diagnostic rather than a cheaper substitute for the gate. Conditioning on that profile also removes any benefit the gate provides by choosing which actions to retain. The remaining contrast asks whether the particular outcomes selected *within* those commands improve the tested planner.

Table 28 reports every original draw, its retained count, and the exact action-matched subset expectation.

Table 28: Every original dataset draw at $f = .8$. The matched control averages all eligible occurrence subsets exactly, not a favorable sampled subset. The 20 draws remain the units of paired uncertainty; subset counts are not extra replications.

Draw	Retained WM	Eligible subsets	Gate success	Matched expectation
0	0	1	0.4583	0.4583
1	1	2	0.6250	0.6250
2	3	35	0.7500	0.7429
3	1	5	1.0000	0.5750
4	3	4	0.3750	0.5833
5	1	3	0.5000	0.5000
6	1	1	0.7500	0.7500
7	1	2	0.6250	0.5000
8	0	1	0.6250	0.6250
9	2	10	0.6250	0.5250
10	0	1	0.6250	0.6250
11	0	1	0.5000	0.5000
12	1	5	0.7500	0.7750
13	0	1	0.5000	0.5000
14	3	1	0.7500	0.7500
15	2	10	0.5000	0.5250
16	1	1	0.7500	0.7500
17	0	1	0.6250	0.6250
18	1	1	0.5000	0.5000
19	3	4	0.8750	0.8750

Mean success is 0.635 for the gate and 0.615 for the matched control. Gate minus control is $+0.020[-0.023, +0.074]$, using 100,000 paired draw resamples and seed 20260926. This pointwise interval is conditional on the bank, reader, recorded oracle, choice of $f = .8$, and inherited draw schedule. It is neither an equivalence test nor a population guarantee. Original gated scores, row indices, goal calibration, and source hashes are reproduced; a second implementation separately enumerates and scores every eligible subset.

N.3 LABEL ERROR, PREDICTION ERROR, AND TASK SUCCESS ARE SEPARATE OUTCOMES

This diagnostic was requested after observing the action-matched task contrast. It reuses exactly the same subset enumeration, not a different or expanded bank. For a retained WM set S with at least one row, define

$$L(S) = \frac{1}{|S|} \sum_{i \in S} \frac{|o_i - h(a_i)|}{|h(a_i)|}, \qquad F(C_b \cup S) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} |\hat{h}_{C_b \cup S}(a) - h(a)|. \tag{13}$$

All recorded denominators $h(a_i)$ are nonzero. Six draws have no retained WM rows; L is undefined there, not zero, leaving 14 paired draws. All 20 draws have clean rows and a defined predictor, so F and task success use 20. F is in .06 m reference units and includes training commands in the fixed grid. For the control, L and F are computed for each subset and then averaged; F uses subset-specific predictors. Averaging predictions before scoring would define a different control.

The gate has lower observed label relative error (0.136 versus 0.424) and fitted-grid MAE (0.960 versus 1.048). Table 3 reports paired intervals, with 100,000 resamples and fixed seeds 20260927/20260928 for these two descriptive endpoints. Task success and its interval are unchanged. The observation supports numerical improvement in this inspected bank; it does not turn a mean-error reduction into a uniform-error bound, an independently confirmed policy gain, or causal mediation. Per-draw rows and the order of analysis are retained in the supplementary scripts.

Dose response, motion controls, and supporting learners. Figures 18 and 16 retain the complete comparisons underlying the main-text interpretation. Their favorable count/motion contrasts do not by themselves isolate an effect independent of the retained action profile.

O ILLUSTRATIVE VIDEOS AND MOTION DIAGNOSTICS

The following source-bound examples make the failure taxonomy inspectable. They are selected illustrations, not a random sample, prevalence estimates, or additional experiments. Full video frames are shown without geometric editing; frame indices and encoded-video times remain explicit. Complete recorded clips, selection notes, and hashes are available in the offline visual supplement. The numerical audit and complete execution tables, rather than these examples, support the aggregate claims.

Observable failure modes. Figure 20 contrasts corrupted, static, magnitude-inaccurate, and unreadable examples. These categories motivate separate checks for event occurrence, measurement validity, and numerical agreement.

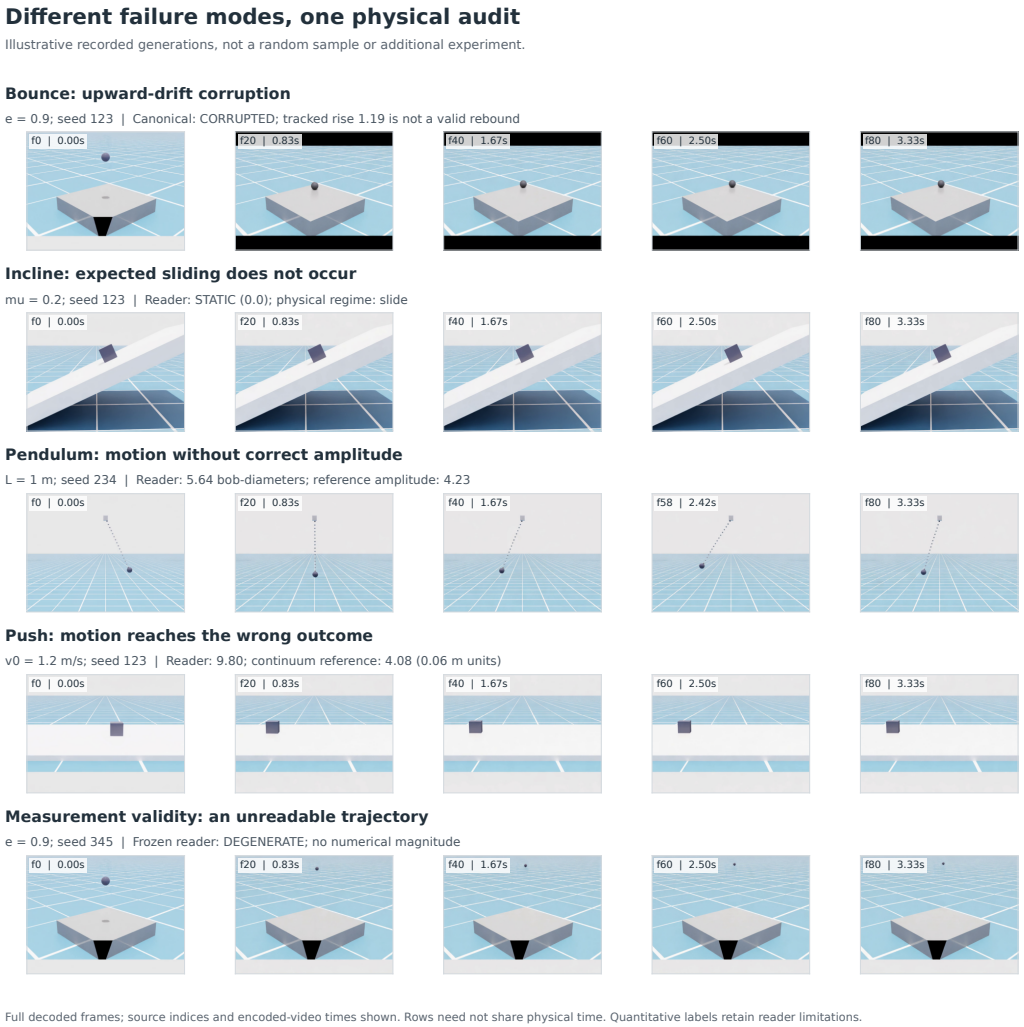


Figure 20: **Distinct observable failure modes.** An upward-drift corruption, failure to slide, excessive measured pendulum amplitude, wrong push displacement, and an unreadable trajectory. Labels come from the frozen reader records; they are not independent human judgments. The changing letterbox/framing in the bounce example remains visible and is a reader qualification, not hidden by cropping. Pendulum keyframes include its measured extrema; the other rows use five evenly spaced stored-frame indices. Encoded-video time is not necessarily validated physical time.

Matched-seed interventions. Figure 21 compares matched seeds under changed speed prompts or richer conditioning. It illustrates why inducing motion and recovering the correct physical response are separate tests.

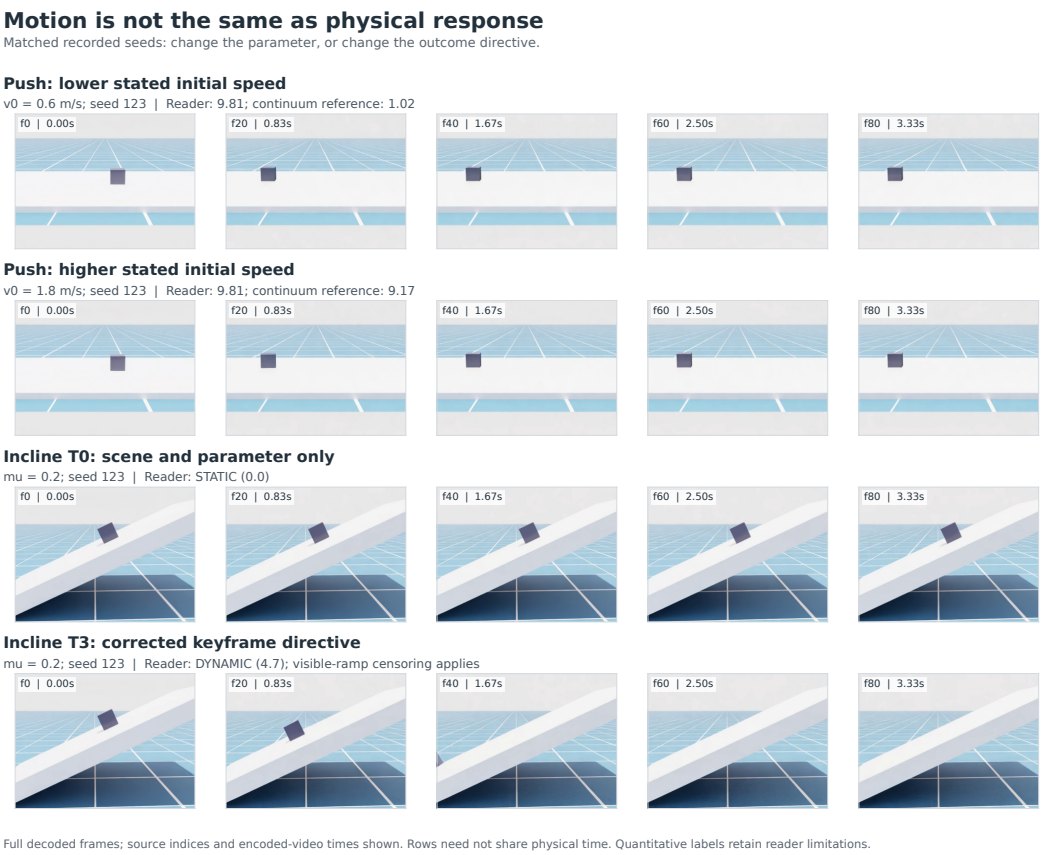


Figure 21: **Visible motion and physical response are different.** Top pair: the same recorded push seed at initial-speed prompts .6 and 1.8 m/s; both reader outcomes are 9.81 reference units although the continuum references differ. Bottom pair: the same recorded incline seed under parameter-only T0 and corrected-keyframe T3. The latter elicits visible motion, but the 4.7-unit reader outcome is subject to visible-ramp censoring and does not establish the full 15-unit travel. These examples illustrate the quantitative intervention tests, not a universal inability to respond to commands.

Grasp and driving context. Figure 22 retains the motivating grasp reference and native-action driving examples. Their differing durations and uncalibrated command coordinates prevent a synchronized metric comparison.

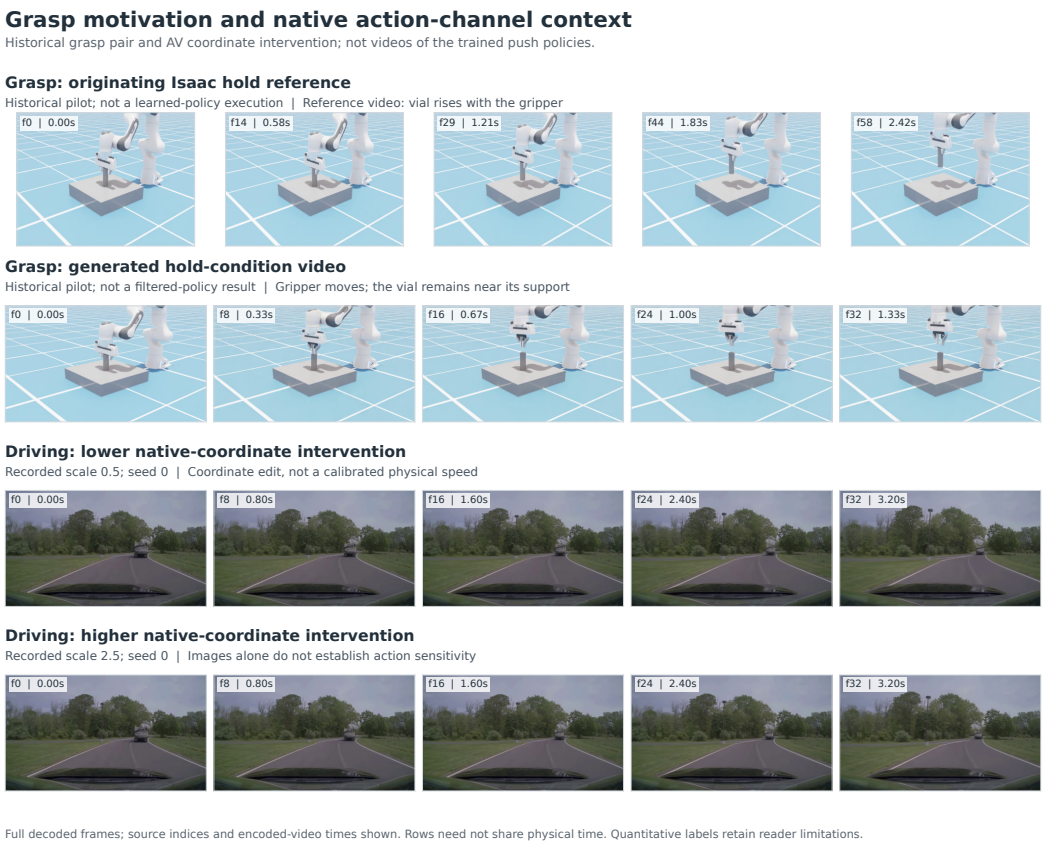


Figure 22: **Motivating grasp and native-channel context.** The grasp rows are the Isaac hold reference and generated hold-condition video. They have different encoded durations and are not time-synchronized policy comparisons; neither is a video of the trained push policy. The AV rows show two native-coordinate interventions, not a calibrated five-fold speed change. Visible ego-motion by itself does not establish action sensitivity or metric distance. These are not videos of the current planners; their scores use recorded action-grid outcomes, not newly filmed policy rollouts.

Measured motion traces. Figure 23 shows the readers' trajectories for the preceding illustrations. Gaps, extrema, and censoring remain visible rather than being replaced by idealized motion.

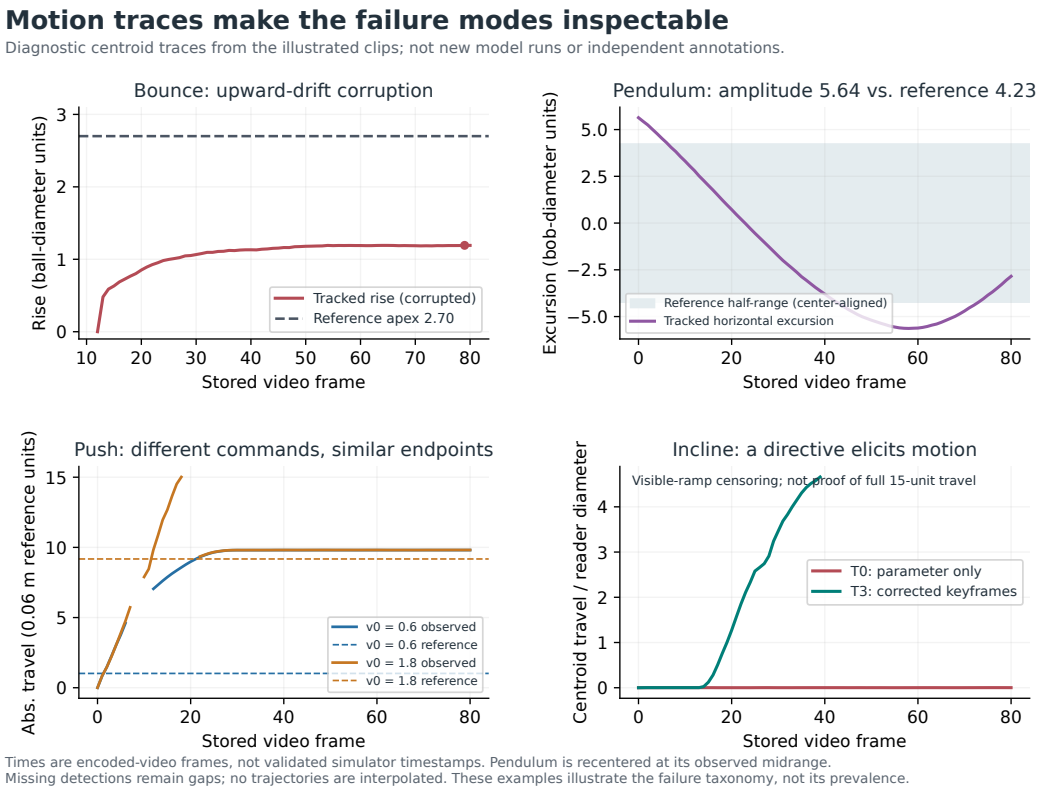


Figure 23: **Diagnostic motion traces of the illustrated clips.** Bounce shows tracked rise in a canonically corrupted clip, not a valid rebound; pendulum shows horizontal excursion recentered at its observed midrange, compared with the physical half-range. This is an amplitude comparison, not an independently synchronized phase prediction. Push uses the reader's horizontal/size-validity mask; rejected detections remain gaps. Similar scored push endpoints do not imply identical intermediate trajectories. Incline T3's trace ends at the visible limit rather than certifying the complete physical trajectory. The scalar bounce, pendulum, and push observables reproduce the frozen values at their reported precision. These diagnostics are not a second independent validation of the reader.

Selection and source binding. Sweep examples use the first recorded qualifying seed at the stated condition: bounce $e = .9$, seed 123 (canonical upward-drift corruption), and seed 345 (unreadable); incline $\mu = .2$, seed 123; pendulum $L = 1$, seed 234 (dynamic); push $v_0 = 1.2$, seed 123. The intervention pairs use seed 123, and the AV pair uses released seed 0. The grasp rows are the originating hold-reference pair. The manifest records all 13 full video hashes and all 65 displayed-frame hashes. The sweep and Nano collections have overlapping filenames, so numeric records are matched by both arm and clip ID, not filename alone. Images are illustrative evidence of the reported cases, not a substitute for the all-seed summaries.

P COMPLETE EXECUTION EVIDENCE ACROSS TWO PUSH SETTINGS

This appendix complements the matched-contract and example galleries with complete outcome maps and continuous endpoint diagnostics. These are *descriptive views of existing executions*, not additional experiments. For comparison, the $\mu = .3$ execution confirmation is shown separately from the $\mu = .5$ matched-contract study. Their episodes are not pooled to inflate the apparent number of independent replications.

P.1 MATCHED-CONTRACT CONFIRMATION: ALL OUTCOMES AND ERROR SIZES

Figure 24 shows every policy episode in the five-arm, $\mu = .5$ confirmation of Appendix H. No block or goal is selected for visual impact, and failed episodes are not omitted. Rows use training seeds 9000–9007 in their original order; columns use the same 50 target values in increasing order. The 48 commanded-action truth episodes used for the text/video comparison are not included in these 2,000 cells. For block b, arm A, and target g_j in reference units, each cell is classified using the same relative endpoint error as the reported score:

$$e_{bAj} = \frac{d_{bAj}/w - g_j}{g_j}, \qquad c_{bAj} = \begin{cases} -1 & e_{bAj} < -.25, \\ 0 & |e_{bAj}| \le .25, \\ +1 & e_{bAj} > .25, \end{cases} \qquad w = .06 \text{ m}. \tag{14}$$

Thus color identifies undershoot, success, or overshoot rather than merely distinguishing good-looking and bad-looking videos. The actions are physically executed, but the policy remains a one-command open-loop controller.

Figure 25 adds endpoint magnitude information. For each arm, thin curves show all eight policies and the thick curve their mean achieved travel. A mean curve inside the tolerance band does not imply every policy succeeds. The lower panel reports each block's mean absolute endpoint error,

$$\mathrm{MAE}_{bA} = \frac{100}{50} \sum_{j=1}^{50} |d_{bAj} - .06 g_j| \quad \text{in cm}, \tag{15}$$

using recorded simulator displacement in meters. Horizontal marks average the eight block-level errors; no episode-independent confidence interval is implied. The continuous diagnostic reveals failure severity that a thresholded success score hides, but it is a secondary analysis of the same observations.

Mean endpoint error falls from 25.09 cm unfiltered to 4.20 cm with the physics gate and 4.06 cm with the own-text gate. All 8 paired physics-filter blocks have smaller MAE than their unfiltered counterparts. Descriptively, the recovery concerns error magnitude as well as crossing the binary success threshold; it is not merely a change in how many endpoints lie just inside that threshold.

Table 29 gives the complete under/success/over counts and block-level endpoint MAE for each arm.

Table 29: Complete matched-contract accounting, regenerated from state logs. Each arm has 400 executions across eight training blocks. MAE is the mean of the eight block-level absolute endpoint errors, not a new independent test.

Arm	Under	Success	Over	MAE (cm)
Unfiltered	370	30	0	25.09
Random, physics-filter count	64	272	64	6.11
Physics filter	48	288	64	4.20
Available clean only	23	283	94	4.66
Own-text filter	30	301	69	4.06

The same-budget random and available-clean arms remain visible in both plots. This guards against attributing recovery to the filter solely because fewer generated labels were used. As already noted, the random arm matches the physics gate's count, not the own-text gate's count. The maps neither remove that limitation nor resolve the identical-data training variability. Their purpose is transparency about *which goals fail and by how much.*

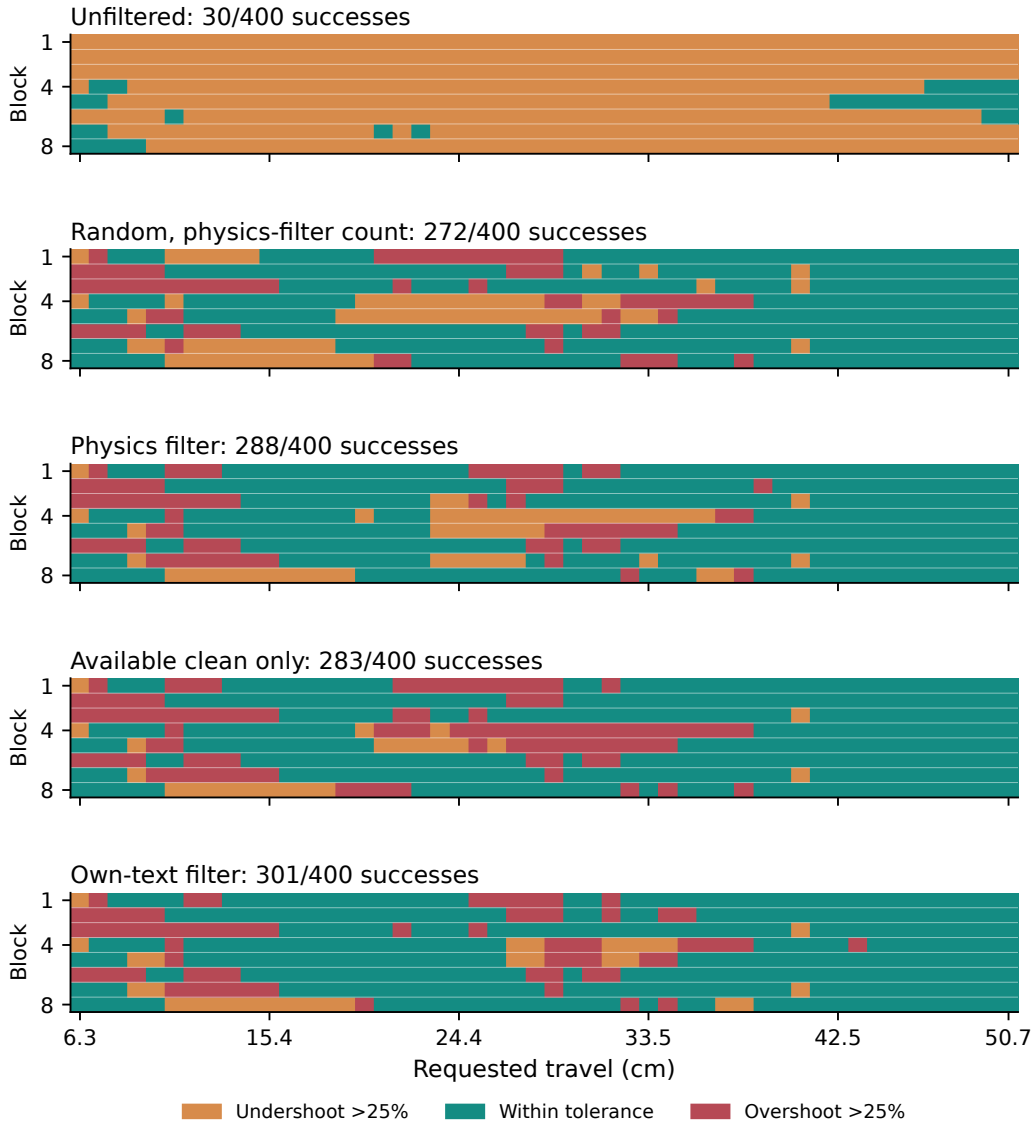


Figure 24: **All 2,000 policy executions in the matched-contract confirmation.** Every cell is one state-scored Isaac execution. Fixed block and goal order allow direct within-block comparison; no best seed is selected. Orange denotes undershoot, teal success, and red overshoot relative to the 25% tolerance. Eight blocks, not 2,000 independent trained policies, are the replication units. The fixed initial budget is eight clean plus 32 generated rows; rejection never replaces rows. This map concerns the inverse-label, new-scene-retrained OpenVLA experiment, not known-action grid planning.

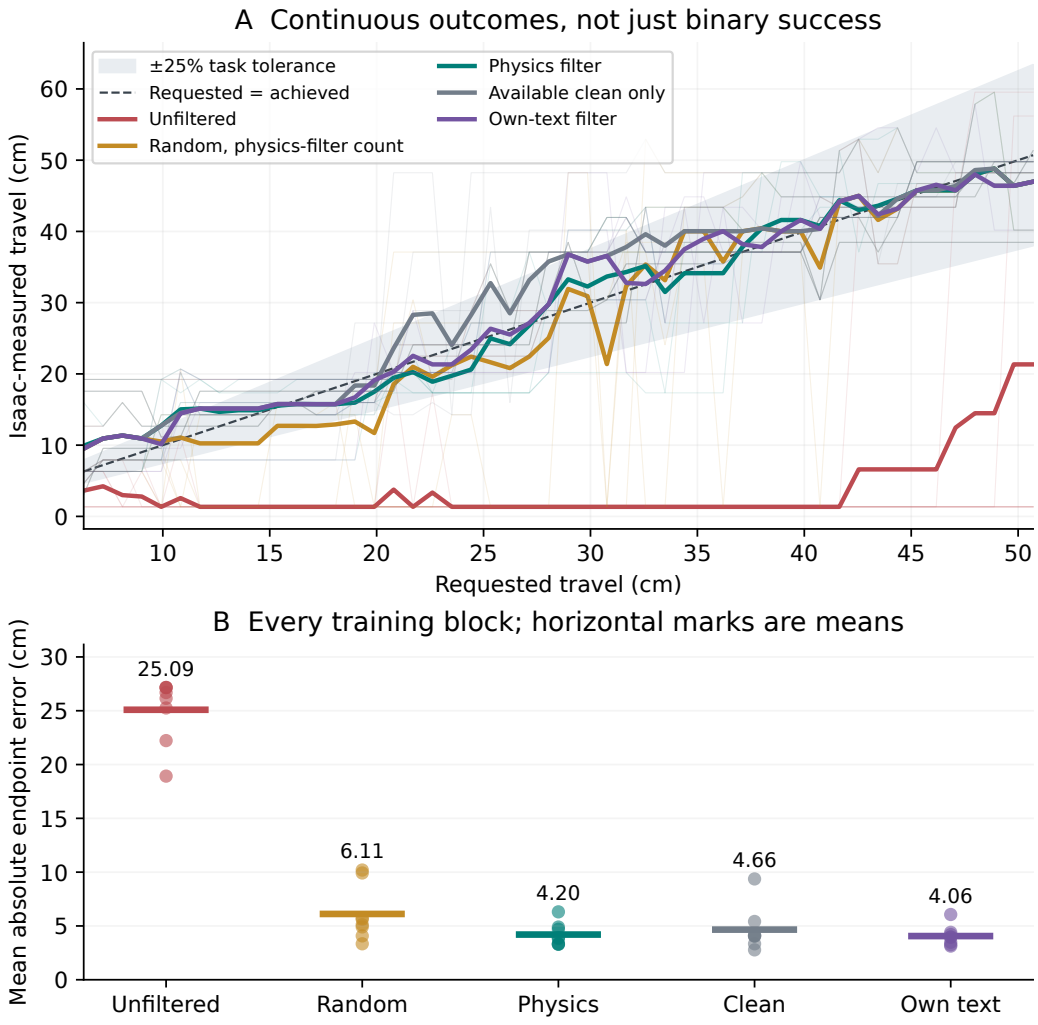


Figure 25: **How far policies move, and how large their errors are.** Top: recorded physical travel versus requested travel for all five arms; thin lines are individual training blocks, thick lines block means, and the shaded region is the task tolerance. Bottom: each point is one policy's 50-goal MAE in centimeters; horizontal marks and numeric labels show eight-block means. These continuous summaries use the same 2,000 executions as Figure 24, not additional evidence from a new run. A correct mean trajectory does not certify every individual output.

P.2 PUSH EXECUTION STUDY AT FRICTION 0.3

Figures 26–27 show the friction-0.3 execution study. They concern the $\mu = .3$ setting, an 80-clip attempted pool (55 readable, 25 unreadable), eight training blocks with seeds 7000–7007, and four policy arms. Five readable pool clips meet the 30% physics gate; individual blocks retain one to four of their 32 generated candidates. All arms begin with the same eight clean examples within a block, train for 240 optimizer steps, and are evaluated on 50 goals by actual Isaac execution. The figure does not show the later own-text arm, and the 1,600 executions are *not* pooled with the 2,000-execution matched-contract study to inflate precision.

How to read the four-panel evidence story. Panel A uses unaltered frames from the reference 1.2 m/s condition and the first listed sweep seed, rather than selecting a clip by downstream success. The saved text probe gives 4.08 reference units while the video reader gives 9.80. Applying the continuum inverse to that endpoint produces a 1.86 m/s pseudo-action instead of preserving the commanded 1.20 m/s action. The intended goal still comes from the command's analytical stopping distance; thus the illustrated inconsistency becomes a mismatched goal/action label. The dashed text-target marker is an annotation projected through reader calibration, not an object generated in the video or a measured Isaac path. Unlike the later matched-contract experiment, this is a matched-condition text probe, not a fresh text forecast for every pool clip.

Panel B shows all 55 readable outcomes and the implemented acceptance region. It visualizes the rule, not an independent validation of its physical accuracy; unreadable clips are not treated as zero motion. Panel C shows measured travel for each trained unfiltered/filtered policy and their means. Panel D keeps random-same-count and available-clean-only controls visible; its error bars are standard errors across eight blocks. No intervention changes the world model itself. The illustrative clip explains label construction; it does not establish that this particular clip caused the aggregate downstream effect.

What the complete map adds. The companion heatmap displays all 1,600 outcomes, keeping goal and block order fixed across arms. This exposes overshoot, undershoot, and difficult goals that can disappear in average curves. The friction-0.3 study's filtering gain is evidence for mitigation in its specified inverse-label pipeline, not a claim of generated-data benefit over clean-only, per-example own-text filtering, or broad robotic transfer. Clean labels use a step-corrected inverse whereas generated labels use the continuum inverse; the admission rule and transformations use known physics. Neither figure should be substituted for the known-action results or treated as another set of independent scenes.

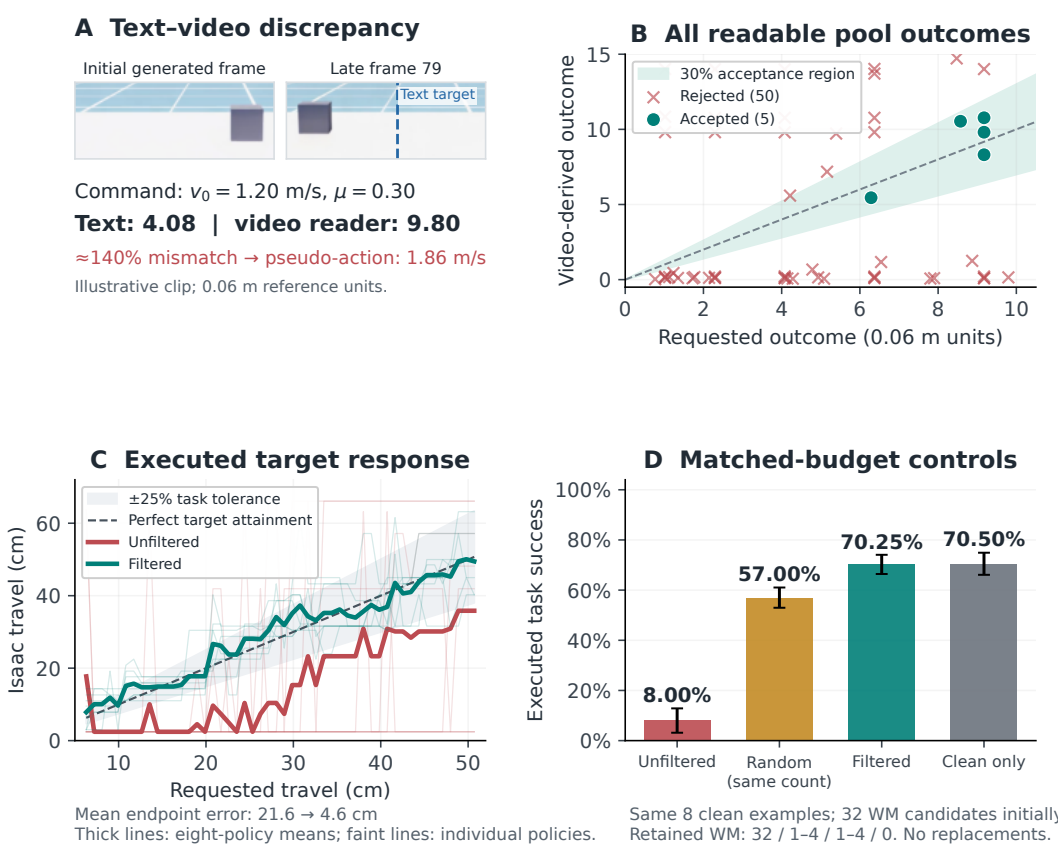


Figure 26: **Push label screening at $\mu = .3$: from a measured discrepancy to supervision and executed behavior.** A: source-bound actual frames and an explicitly annotated text target illustrate the inverse-label conversion. B: all 55 readable outcomes from 80 attempted clips, including five accepted examples. C: individual and mean Isaac-measured trajectories. D: success with matched available clean data; bars show block-level standard errors. This figure reports the four-arm study at $\mu = .3$, not the matched-contract experiment at $\mu = .5$. See Appendix P.2 for the transformations and limits.



Figure 27: **Push execution study at $\mu = .3$: every executed outcome.** Each cell represents one of the 1,600 state-scored Isaac episodes, with the same eight training blocks and 50-goal order across four arms. Colors use the 25% success tolerance, separating under- and overshoot. No successful subset or best seed is selected. This is the complete view of the friction-0.3 experiment, not 1,600 independent policy-training repetitions or additional scenes.