WHAT ROLLOUT ERROR MAGNITUDE MISSES IN LATENT WORLD-MODEL PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Rollout error is often summarized by its magnitude, but a planner's outcome depends on which action that error causes it to select. We ask whether error direction provides decision-relevant information beyond error magnitude in latent world-model planning. For squared latent-goal cost, an exact decomposition separates residual energy from a signed projection onto the goal displacement. We isolate the projection's value by comparing two oracle scores that use the same candidate actions and each candidate's actual error energy: $M + 2E$ and $N = M + 2E - p$. On 256-action PushT pools from 48 new sources, using the projection raises first-plan physical success from 69/96 to 80/96 for the original predictor and from 63/96 to 79/96 for an independently trained predictor on its own pools; true latent-cost regret also decreases. Additional interventions examine how prediction history and error direction affect action rankings. In Cube-Single, energy-aware selection already recovers 29 of the 30 pools containing a successful action, leaving little additional success to recover; a learned calibrator does not reliably reproduce the PushT diagnostic gain. These results show that goal-related error direction matters beyond the tested energy-aware scores, while leaving its use in a deployable planner unresolved.

1 INTRODUCTION

A visual world model allows a controller to compare imagined action sequences before acting (Hafner et al., 2019; Hansen et al., 2022; Maes et al., 2026). The controller executes the lowest-scored candidate, making action selection the point at which prediction error affects control. Rollout error magnitude describes prediction accuracy, but it does not say whether an error moves a predicted terminal toward or away from the goal, changes the ranking of candidates, or alters the selected action's physical outcome.

Figure 1 illustrates how equal-length errors can change squared goal cost in opposite directions and reverse a pairwise ranking. We ask whether this directional component matters in naturally generated planning pools after accounting for each candidate's actual error energy. An exact cost decomposition yields a controlled comparison: we rescore the same audited actions while keeping the initial state, goal, world model, and continuation policy fixed.

The main result is that the realized goal projection improves fixed-pool selection beyond the tested energy-aware scores. On 48 new PushT sources, the projection increases first-plan success for the original predictor and for a separately trained predictor evaluated on its own candidate pools (Table 1). Earlier discovery and R1 cohorts show the same direction of effect, and true latent-cost regret falls in the new-source evaluation.

We then trace how prediction history and signed cost changes affect the selected action, and test whether directional structure suggests a correction. A prospective CEM audit examines error geometry during search; Cube-Single and a learned calibrator probe the separate limits of candidate coverage and deployment-visible prediction. The principal score comparison uses observed candidate futures and therefore measures information value rather than an online planning method.

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.

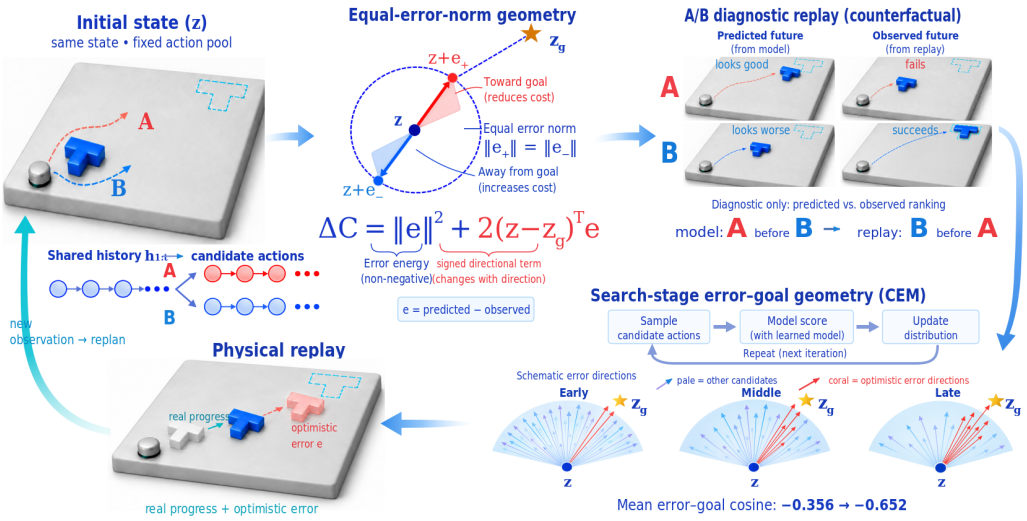

Figure 1: **Directional error in the planning loop.** The equal-norm sketch shows how residual direction changes squared goal cost. The A/B replay is a schematic ranking reversal within a fixed action pool. The lower panels depict physical replay and repeated CEM selection. The fans illustrate different stage populations; their early-to-late mean error–goal cosine (-0.356 to -0.652) comes from endpoint-surviving random-audit actions in P3 and does not establish increased cost underestimation.

2 Fixed-pool setup and diagnostic scores

For candidate action sequence i, let z_i denote the observed terminal latent, $\hat{z}_i$ its world-model prediction, g the goal latent, and $e_i = \hat{z}_i - z_i$ the residual. The planner uses squared latent-goal cost, for which

$$M_i = \|\hat{z}_i - g\|^2, \qquad D_i = \|z_i - g\|^2, \qquad E_i = \|e_i\|^2, \\ M_i - D_i = E_i + q_i, \qquad q_i = 2(z_i - g)^\top e_i. \tag{1}$$

Here E_i is residual energy, whereas the signed term q_i depends on the residual's direction relative to the observed goal displacement. The observed-terminal selector L minimizes D_i. The neutral score $N_i = D_i + E_i$ retains actual energy while removing q_i. Equivalently, with $h_i = \hat{z}_i - g$ and $p_i = 2h_i^\top e_i$,

$$D_i = M_i + E_i - p_i, \qquad N_i = (M_i + 2E_i) - p_i. \tag{2}$$

Since $p_i = q_i + 2E_i$, the same neutral score can be written as $N_i = M_i - q_i = (M_i + 2E_i) - p_i$. Comparing N with $M + 2E$ therefore isolates the realized projection p_i while giving both scores each candidate's actual E_i. We also compare $M + \beta E$ for a fixed coefficient grid. These are oracle diagnostics because E_i and p_i depend on observed futures; the experiment tests this score component, not every possible function of (M, E).

Planning and physical outcomes. The main predictor is the frozen 192-dimensional `quentinll/lewm-pusht` LeWorldModel; a separately trained 384-dimensional public reproduction provides a second predictor. PushT is a contact-rich visual manipulation task (Florence et al., 2022; Chi et al., 2023). Each source starts at expert-cache step 0 and uses that episode's step-25 goal image. CEM runs 30 iterations with 300 proposals and 30 elites, predicting five blocks of five primitive actions. Within a decision, selectors rank the same 256 physically audited unique actions selected from 9,001 generated records. Each predictor generates and audits its own pools.

First-plan success means any hit within 25 steps: the Euclidean difference over agent and block xy coordinates is below 20, and symmetry-aware block orientation error is below $\pi/9$. For the primary terminal latent, physics continues to step 25 while first success is latched; a first-hit absorbing endpoint provides a sensitivity check. In closed-loop evaluation, the controller executes the selected first plan and then uses the same adaptive planner up to step 100. We also report true latent-cost

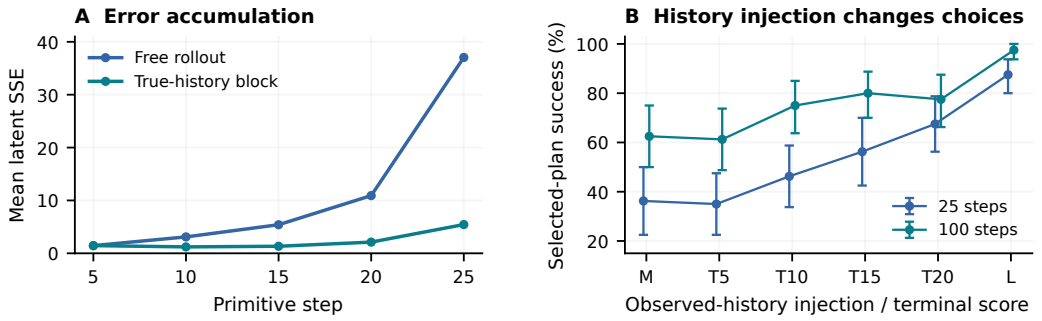

Figure 2: **History affects both error and selection.** (a) Mean SSE on 84 outcome-selected discovery actions; these are descriptive trajectories, not 84 independent sources. (b) Full historical qualified-pool rescoring after observed-history injection at $T5$–$T20$. M uses free prediction and L observed terminals. Bars bootstrap the 40 sources; both horizons use the same 80 decisions.

regret, $D_{\text{selected}} - \min_i D_i$, on the fixed pool. Physical success and latent-cost regret are distinct outcomes; even L need not select a physically successful action. The native CEM final-mean action is a separate reference from the fixed-pool M argmin.

Cohorts and inference. Discovery and R1 contain 40 disjoint sources each, evaluated with two optimizer seeds. R1's locked primary comparison concerns top-five removal versus equal-norm shrinkage; its signed-term analyses are secondary. P3 uses 40 further sources for the prospective search audit. The main new-source energy–projection evaluation (P2) fixes another 48 sources before candidate generation; the second predictor uses these same sources but its own pools in a secondary replication. Cube-Single has a separate frozen sample. We average optimizer seeds within source and bootstrap sources 10,000 times (Agarwal et al., 2021). New denotes sources unused in earlier analyses, not a claim about checkpoint training data.

History, subspace, and signed-path experiments use the *historical qualified pools*, which exclude actions that terminate successfully before step 25 (15,578 discovery and 14,652 R1 candidates). The main energy–projection comparison instead uses all 256 audited actions per decision and the dataset-image input. We distinguish these populations when reporting results.

3 From Prediction History to Physical Selection

We first establish that prediction error changes the action selected by the planner and its physical outcome. On historical discovery pools, ranking candidates by observed rather than predicted terminal latents raises first-plan success from 29/80 to 70/80. With the same continuation planner, 100-step success rises from 50/80 to 78/80, a source-paired gain of 35 percentage points [22.5,47.5] with 29 rescues and one harm. Candidate actions and the goal metric remain fixed throughout this comparison.

To examine how earlier prediction errors contribute to this change, compare the free-running history $\hat{H}_t$ with the observed history H_t under identical actions. Their next-step errors admit the exact finite decomposition

$$e_{t+1} = \underbrace{F(\hat{H}_t, a_t) - F(H_t, a_t)}_{\text{changed-history effect } d_{t+1}} + \underbrace{F(H_t, a_t) - z_{t+1}}_{\text{true-history residual } r_{t+1}}. \tag{3}$$

In 42 selected failure/success comparisons (84 candidate actions), repeated prediction from true histories lowers terminal SSE from 37.042 to 5.437 and reverses pair order in 38/42 comparisons (Figure 2a). These outcome-selected pairs motivate a full-pool intervention, rather than providing a population estimate of propagation's contribution.

We then inject observed intermediate history for every candidate and rerun the remaining predictor. Conditioning at $T15$ and $T20$ yields 64/80 and 62/80 successes at 100 steps, compared with 50/80 under free prediction (Figure 2b). Their first-plan successes are 45/80 and 54/80, respectively:

Table 1: First-plan successes on full 256-action PushT audits with dataset-image input and continued-physics endpoints. The first two rows use the same 48 new sources with predictor-specific pools (original: primary; 384-d: secondary). The remaining rows use development sources. $\Delta = N - (M + 2E)$ is the source-paired percentage-point difference with a 95% interval.

Cohort	n	M	$M + 2E$	N	L	Δ [95% CI]
New sources, original	96	41	69	80	85	+11.46 [4.17,18.75]
New sources, 384-d	96	49	63	79	82	+16.67 [8.33,25.00]
Discovery	80	37	63	75	78	+15.00 [7.50,23.75]
R1	80	42	58	71	72	+16.25 [7.50,27.50]
384-d predictor on R1	80	33	46	54	63	+10.00 [3.75,16.25]

the later injection helps immediate selection more, but the earlier injection leads to more eventual successes. This difference separates first-plan quality from subsequent recovery. The intervention complements work on train–rollout distribution mismatch (Ross et al., 2011; Bengio et al., 2015; Talvitie, 2017).

4 Goal projection adds value beyond error energy

Observed-terminal selection removes the entire terminal prediction error. The more specific question is whether goal projection adds value after the scorer is already given the actual error energy. Table 1 answers this question on complete audited pools using dataset-image predictions and continued-physics endpoints. The energy-aware score $M + 2E$ improves on M, and N improves further in all five PushT comparisons.

4.1 Development controls and 100-step consequences

On discovery, R1, and the second predictor's R1 pools, N improves first-plan success over $M + 2E$ by 15.00, 16.25, and 10.00 percentage points, respectively (Table 1). Selecting $\beta \in \{0, 0.5, 1, 2, 4\}$ by discovery first-plan success chooses $\beta = 1$: the five scores select 37, 63, 64, 63, and 61 successes out of 80. With this coefficient fixed, $M + E$ selects 60/80 successes on R1 and 45/80 for the second predictor on R1, versus 71/80 and 54/80 for N; discovery N selects 75/80. Thus a discovery-selected energy coefficient does not close the projection gap on these development pools.

The same comparison lowers true-D regret by 5.02 on discovery, 3.13 on R1, and 1.85 for the second predictor on R1. At 100 steps, successes under M, $M + 2E$, and N are 54, 69, and 79 on discovery; 59, 73, and 76 on R1; and 56, 64, and 70 for the second predictor (each out of 80). Relative to $M + 2E$, N gains 12.5 points on discovery (95% CI [3.75,22.5]), 3.75 on R1 ([0,8.75]), and 7.5 for the second predictor ([1.25,15]). The R1 interval includes zero, so the first-plan effect is more consistent than the 100-step increment.

4.2 Frozen new-source evaluation and predictor replication

The 48 P2 sources were fixed before their candidate pools were generated. For the original predictor, N selects 80/96 successful first plans versus 69/96 for $M + 2E$: 12 decisions are rescued and one is harmed. This is a source-paired gain of 11.46 percentage points (95% CI [4.17,18.75]); true-D regret falls by 2.22 (95% CI [0.96,4.28]). On its own pools for the same sources, the separately trained 384-dimensional predictor yields 79/96 versus 63/96 successes, with 16 rescues and no harms. Its gain is 16.67 points (95% CI [8.33,25.00]), and regret falls by 2.03 (95% CI [1.18,3.06]). These are fixed-pool first-plan results; new-source 100-step continuations were not measured.

Support and input checks. The main comparison includes early successes omitted by the historical endpoint rule. On development sources with simulator-rendered inputs, neutral's 100-step gain over M remains positive under continued and absorbing endpoints: +27.5/+23.75 points on discovery, +23.75/+25 on R1, and +16.25/+16.25 for the second predictor. The native dataset-image input changes most M winners; we therefore use it explicitly in Table 1. A native-batch replay for the

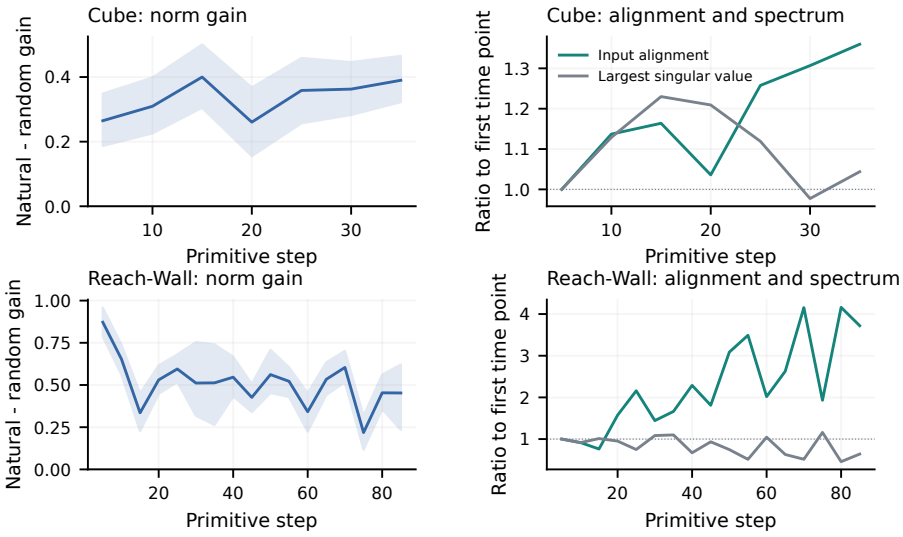

Figure 3: **Propagation depends on direction, not just the largest singular value.** Twenty-source failure cohorts in each setting. Left: natural-minus-equal-norm-random gain with source-bootstrap intervals. Right: input alignment and largest singular value relative to their first time point; ratios are descriptive. Cube uses 256-dimensional Jacobians and a top-5% input subspace; Reach-Wall uses 102,400 dimensions and a top-100 estimate. These panels measure local propagation, not control improvement.

heterogeneous P2 predictor reproduces every CEM cost and both oracle-selector winners in all 96 decisions. Appendices C and K give the input and endpoint checks.

5 Directional interventions and their limits

The diagnostic comparison establishes the value of realized direction, but a correction requires a rule that identifies useful directions before execution. We examine local propagation, matched-norm interventions, and changes in the selected action along signed-score paths.

5.1 Directional amplification in local dynamics

For a local predictor Jacobian $J_t = U_t \Sigma_t V_t^\top$, the gain along the natural error is $G_t = \|J_t e_t\| \,/\, \|e_t\|$. Equal-norm random input directions provide a reference for whether large gain reflects the residual's alignment rather than only the spectrum. On fixed 20-source failure cohorts in Cube and Reach-Wall, natural-direction gain exceeds that reference (Figure 3, left). Top-subspace input alignment rises in 18/20 Cube sources and 20/20 Reach-Wall sources, while the largest singular value is not uniformly increasing (right).

Input alignment increases in both settings. Per five steps, the estimated mean increase is 5.38×10^{-3} in Cube (95% CI $[2.97, 7.59] \times 10^{-3}$) and 2.10×10^{-4} in Reach-Wall (95% CI $[1.67, 2.54] \times 10^{-4}$). Their magnitudes are not directly comparable because the representations and spectral budgets differ. This pattern motivates a candidate-local intervention; a fresh-residual diagnostic also finds accumulation on successful Reach-Wall trajectories (Appendix G).

5.2 Matched-norm controls and the locked R1 test

Let P_i project onto the five leading output singular vectors of the true-history final-block Jacobian, and set $s_i = \|(I - P_i)e_i\|$. Three diagnostic terminal residuals have exactly the same remaining norm:

$$e_i^{\text{remove}} = (I - P_i)e_i, \qquad e_i^{\text{uniform}} = s_i e_i / \|e_i\|, \qquad e_i^{\text{random}} = s_i u_i, \tag{4}$$

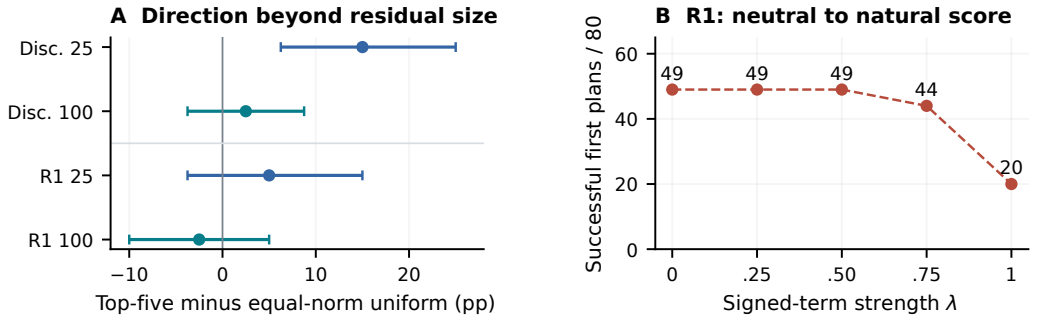

Figure 4: **Direction matters without yielding a confirmed subspace repair.** Both panels use historical qualified pools. (a) Top-five removal minus equal-norm uniform shrinkage, with source-paired intervals; R1's 25-step contrast was locked. (b) R1 post hoc $N + \lambda q$ rescoring uses archived physical labels. Points show evaluated strengths; the dashed segments only guide the eye, not an estimated continuous success curve.

where u_i is sampled uniformly from the unit sphere ($\|u_i\| = 1$), and zero residuals remain zero. On historical discovery pools, removal yields 62/80 first-plan successes versus 50/80 for uniform shrinkage, a +15-point difference [6.25,25]. Five fixed random directions achieve 61–65/80; the advantage is therefore not unique to the leading subspace. Their spherical mean score, $D_i + s_i^2$, gives 63/80.

In the locked R1 test, top-five removal selects 57/80 successful first plans versus 53/80 for equal-norm shrinkage: a gain of 5 percentage points (95% CI [-3.75,15]). At 100 steps, removal instead has 68/80 successes versus 70/80, a difference of -2.5 points (95% CI [-10,5]; Figure 4a). The discovery advantage therefore does not replicate as a confirmed R1 success gain, although the leading subspace contains 63.0% of natural residual energy on discovery. Appendix E reports earlier $T20$ input and $T25$ output arms with their different history templates and residual budgets.

5.3 Signed-score paths and directional references

On the historical discovery/R1 pools, returning from neutral to natural scoring changes 76/80 and 78/80 winners. It causes 32 successful-to-failed first-plan switches in each cohort, versus four and three rescues. The separately measured neutral 100-step outcomes reach 67/80 and 68/80 versus M's 50/80 and 51/80. These effects concern the original qualified support, whereas Table 1 uses the complete audited pools.

The post hoc path $C_i(\lambda) = N_i + \lambda q_i$ follows all candidate score lines as the signed term is restored. R1 success is 49/80 at $\lambda = 0$ and 0.5, 44/80 at 0.75, and 20/80 at one (Figure 4b). Exact line crossings identify a first failed winner in 33 of the 49 neutral-success decisions, at a median strength of 0.878 [quartiles 0.601,0.986]. One decision recovers before the natural endpoint. These crossings show how a continuous score change produces discrete selection changes.

Full-norm shared rotations preserve residual energy and Gram structure, but reduce within-pool signed-cost spread about tenfold. The feasible reassignment reference instead preserves the complete natural q multiset while respecting each candidate's residual-norm bound (Table 2). Natural minus feasible-reference success is +9.33 points [3.18,16.41] on discovery and +2.81 [-6.53,12.29] on R1. Natural assignment is therefore not shown to be especially harmful compared with another realizable assignment of the same signed-cost values. The within-candidate natural–neutral effect remains.

Reducing size and predicting finite response. The separate oracle contraction path $D_i + \alpha q_i + \alpha^2 E_i$ uses complete rendered-input development pools. Halving α raises first-plan success from 36/80 to 62/80 on discovery, from 31/80 to 68/80 on R1, and from 29/80 to 48/80 for the second predictor. Exact path sweeps find physical harm somewhere in only 0–3/80 decisions per cohort and endpoint, despite some nonmonotone true-cost regret. Thus radial accuracy improvement is usually

Table 2: Historical qualified-pool 25-step success (%). Rotation and feasible reassignment average 128 fixed draws using archived labels. Rotation preserves the residual Gram matrix but changes signed-cost spread; feasible reassignment preserves the q multiset and each residual norm, but not the Gram matrix.

Cohort	Natural	Neutral	Shared rotation	Feasible q reassignment
Discovery	36.25	71.25	72.85	26.92
R1	25.00	61.25	57.59	22.19

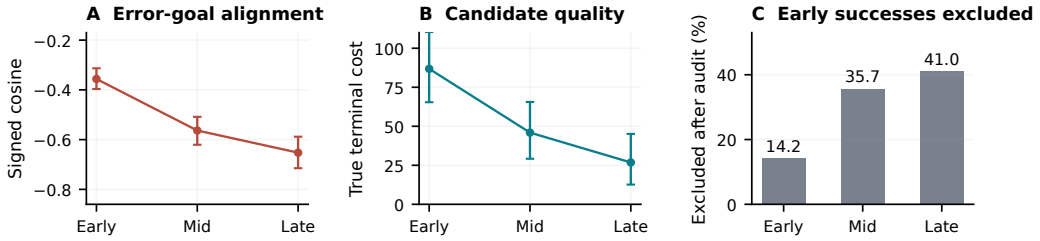

Figure 5: **Search-stage geometry and its population boundary.** P3 has 40 sources and two seeds each. (a,b) Source-averaged statistics of endpoint-surviving random-audit actions, with descriptive source-bootstrap intervals. (c) Raw fractions excluded because success occurred before step 25. Early, middle, and late bins are CEM iterations 0–9, 10–19, and 20–29.

beneficial here. On four inspected R1 pools, a local linear-plus-energy pilot predicts 8/8 winners at quarter amplitude but only 4/8 at natural amplitude.

6 Error geometry during CEM search

The prospective P3 audit fixes 40 new sources, three stage bins (iterations 0–9, 10–19, 20–29), and a source-paired late-minus-early signed-cosine contrast before candidate generation. Across decisions, it samples 10,240 actions at equal probability within each remaining unique-action frame; 7,087 have legal step-25 endpoints, while 3,153 terminate after early success. The analysis concerns these endpoint-surviving random-audit actions, excluding required and deterministically selected low-cost actions.

Mean cosine $c_i = \cos(z_i - g, e_i)$ becomes more negative across early, middle, and late search stages: -0.356, -0.563, and -0.652. The registered late-minus-early difference is -0.296 (95% CI [-0.363,-0.229]). True terminal cost falls from 86.83 to 26.82 over the same stages (Figure 5). Total predicted-minus-true cost changes by -6.04 (95% CI [-23.97,8.10]); this does not resolve an increase in underestimation. Early-success exclusions also rise from 14.19% to 40.99%.

A post hoc geometric analysis helps interpret the angle trend. Let $a_i = \|\hat{z}_i - g\| / \|e_i\|$ and $\rho_i = \cos(\hat{z}_i - g, e_i)$. Then

$$c_i = \frac{a_i\rho_i - 1}{\sqrt{a_i^2 + 1 - 2a_i\rho_i}}. \tag{5}$$

At fixed ρ, reducing predicted-goal distance relative to residual size sends c toward -1. From early to late search, the pooled median a falls from 2.13 to 0.81. The source-paired change in mean $\log a$ is -1.076 (95% CI [-1.269,-0.914]); mean ρ rises by 0.197 (95% CI [0.113,0.280]). The cosine trend is therefore consistent with geometric proximity to the predicted goal. Because candidate populations change across stages, this analysis remains descriptive.

7 LIMITS FROM CANDIDATE COVERAGE AND LEARNED CALIBRATION

7.1 CUBE-SINGLE: RANKING UNDER LIMITED COVERAGE

A separately frozen Cube-Single experiment uses 32 new source episodes, two CEM seeds, and 256 hash-selected actions from each final proposal pool. One decision fails the native-fork validity gate; the complete-pair analysis has 31 sources and 62 decisions. Physical success is any hit within 25 steps at cube–target distance at most 0.04 meters. Only 30 of these fixed pools contain a successful action.

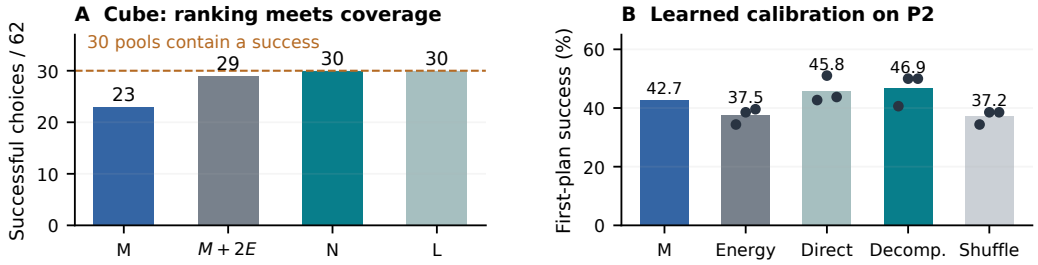

Figure 6: **Two boundaries between information and control gain.** (a) Cube-Single selectors use the same 62 pools; the dashed line is the 30 pools with any successful action, not an online-planning upper bound. (b) Frozen calibration on 96 P2 decisions: bars are means and dots are three training initializations, not confidence limits. Test inputs contain no observed future. Validation did not clear the online gate; paired test intervals are reported in the text.

M, $M + 2E$, N, and L select 23, 29, 30, and 30 successes (Figure 6a). The projection lowers true-D regret by 11.54 [6.45,17.70], while its +1.61-point physical-success gain [0,4.84] falls short of the frozen confirmation criterion; the missing-source worst-case bound is [-1.56,4.69] points (not a confidence interval). Since energy-aware selection already reaches 29 of the 30 pools containing a successful action, most remaining failures require candidate generation rather than reranking.

The Cube result shows that a lower latent-cost regret can coexist with little physical-success headroom in a fixed candidate pool.

7.2 LEARNING A DEPLOYMENT-VISIBLE SCORE

We train energy-only, direct-cost, and decomposed MLPs on 60 old sources, select configurations on 20 separate R1 sources, and evaluate once on the 48 P2 sources. P2 is unseen during calibrator fitting and selection, although its oracle outcomes were already examined. All models use the same deployment-visible actions and predicted latents, a roughly 342k-parameter backbone, three initializations, and training budget. The decomposed score is $M + \beta\hat{E} - \hat{p}$; validation selects $\beta = 1$.

The decomposed model reaches 35% validation success versus 55% for M, failing the prespecified +5-point gate for online continuation. Its fixed-pool test mean is 46.88% versus 42.71% for M, a +4.17-point difference [-8.33,17.01]. Direct cost regression reaches 45.83%, leaving +1.04 points [-3.47,6.25] for decomposition (Figure 6b). Decomposition does improve on energy-only learning by +9.38 points [0.69,19.10], but the failed gate precludes the planned online test.

At $\beta = 1$, both direct and decomposed calibration estimate D; splitting the target adds no deployment-visible input. The decomposed correction error is $(\hat{E} - E) - (\hat{p} - p)$, so accuracy on each component need not yield accurate candidate rankings. This result separates the information carried by the realized projection from what the tested learner can estimate before execution.

8 RELATED WORK

Decision-aware model learning distinguishes predictive fit from control value (Farahmand et al., 2017; Lambert et al., 2020; Grimm et al., 2020), and recent theory examines one-sided cost under-

estimation on planner-reachable states (You et al., 2026). Closest to our empirical setting, DMA audits latent–real ranking and search stages, TRM learns a terminal metric, and D-JEPA diagnoses local inversions before learning a correction (Wang et al., 2026; Li et al., 2026b; Liu et al., 2026a). ASAR separates candidate coverage from selection, while ARC-Bench studies how replanning affects first-plan rankability (Li et al., 2026a; Zhang and Li, 2026). We hold the predictor, squared-goal metric, and candidate actions fixed to isolate the selection value of the realized error projection after accounting for error energy. The accompanying intervention, search, and calibration studies delimit what this diagnostic implies for correction. Appendix M discusses the broader predictor and planning literature.

9 Conclusion

In fixed PushT candidate pools, the realized goal projection of rollout error improves physical action selection beyond the tested scores given each candidate's actual error energy. The effect persists on new sources and with a separately trained predictor. History interventions and signed-score paths show how prediction errors alter selected actions; matched-norm controls do not establish a specific directional repair. Search-stage geometry, Cube-Single coverage, and learned calibration further separate diagnostic information from gains available to a deployed planner. A practical correction must estimate useful directional information before execution and act on candidate pools with recoverable success.

AI Use Statement

AI tools were used to assist with experimental design and manuscript editing. The authors have reviewed all content and take full responsibility for the final manuscript.

Ethics Statement

This work uses publicly available models and datasets, involving no human subjects or personally identifiable information. As our research focuses on diagnostic evaluation and interpretability, we foresee no direct negative societal or ethical impacts.

Reproducibility Statement

All mathematical formulations, experimental settings, and hyperparameters are detailed in the main text and appendix. The models and data used are publicly accessible. Complete code and execution scripts will be open-sourced upon publication.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron C. Courville, and Marc Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *NeurIPS*, 2021.

Kavosh Asadi, Dipendra Misra, and Michael L. Littman. Lipschitz continuity in model-based reinforcement learning. In *ICML*, PMLR 80:264–273, 2018.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv:2506.09985*, 2025.

Jiaxin Bai and Jiaxuan Xiong. Temporal-Distance JEPA: Plan-aware representation learning for latent world model predictive control. *arXiv:2607.25337v2*, 2026.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv:2511.08544*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*, 2015.

Amirhosein Chahe, Siwei Cai, and Lifeng Zhou. Latent world models with monotone planning costs for image-goal navigation. *arXiv:2608.09073*, 2026.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems XIX*, 2023. doi:10.15607/RSS.2023.XIX.026.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *NeurIPS*, 2018.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 134:19–67, 2005.

Marc Peter Deisenroth and Carl E. Rasmussen. PILCO: A model-based and data-efficient approach to policy search. In *ICML*, pp. 465–472, 2011.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize." *Management Science*, 68(1):9–26, 2022.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep RL: A case study on PPO and TRPO. In *ICLR*, 2020.

Amir-Massoud Farahmand, André Barreto, and Daniel Nikovski. Value-aware loss function for model-based reinforcement learning. In *AISTATS*, PMLR 54:1486–1494, 2017.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov decision processes. In *UAI*, pp. 162–169, 2004.

Pete Florence, Corey Lynch, Andy Zeng, Oscar A. Ramirez, Ayzaan Wahid, Laura Downs, Adrian Wong, Johnny Lee, Igor Mordatch, and Jonathan Tompson. Implicit behavioral cloning. In *CoRL*, PMLR 164:158–168, 2022.

Yuntian Gao and Xiangyu Xu. Fast LeWorldModel. *arXiv:2606.26217*, 2026.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *ICML*, PMLR 97:2170–2179, 2019.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *NeurIPS*, 2020.

David Ha and Jürgen Schmidhuber. World models. *arXiv:1803.10122*, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *ICML*, PMLR 97:2555–2565, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020.

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

Nicklas A. Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *ICML*, PMLR 162:8387–8406, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *ICLR*, 2024.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI*, 2018.

Jiaming Hu, Yan Zheng, and Tian Wang. SCALE: State-calibrated latent embeddings for JEPA planning in the right geometry. *arXiv:2608.16287v2*, 2026.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *NeurIPS*, 2019.

Łukasz Kaiser, Mohammad Babaeizadeh, Piotr Miłoś, Błażej Osiński, et al. Model-based reinforcement learning for Atari. In *ICLR*, 2020.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-based offline reinforcement learning. In *NeurIPS*, 2020.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *L4DC*, PMLR 120:761–770, 2020.

Liangyu Li, Shengzhi Wang, Libin Qiu, Mingliang Xiong, and Qingwen Liu. World model control by trajectory reachability metrics. *arXiv:2605.22164v2*, 2026b.

Liangyu Li, Qingwen Liu, Mingqing Liu, and Wen Fang. Action from adjacent set in physical space outperforms the best prediction in world models. *arXiv:2607.23602v2*, 2026a.

Andy Zeyi Liu, Haoran Sun, Lucas Baker, Randall Balestriero, and John Sous. Semigroup-JEPA: Latent dynamics consistency for zero-shot physics generalization. *arXiv:2609.10464*, 2026b.

Shuaijun Liu, Chengyu Wu, Qifu Wen, Feiyang You, Chenglong Zhang, Shuyang Hao, Xi Lin, and Ningxin Su. D-JEPA: A decision-aligned latent world model. *arXiv:2609.24749v1*, 2026a.

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

Thomas M. Moerland, Joost Broekens, Aske Plaat, and Catholijn M. Jonker. Model-based reinforcement learning: A survey. *Foundations and Trends in Machine Learning*, 16(1):1–118, 2023.

Shaowu Pan and Karthik Duraisamy. Long-time predictive modeling of nonlinear dynamical systems using neural networks. *Complexity*, 2018:4801012, 2018. doi:10.1155/2018/4801012.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, PMLR 15:627–635, 2011.

Reuven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1:127–190, 1999.

James E. Smith and Robert L. Winkler. The optimizer's curse: Skepticism and postdecision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

Xinyuan Song and Zekun Cai. Understanding rollout error in graph world models. *arXiv:2606.27780*, 2026.

Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4):160–163, 1991.

Erik Talvitie. Self-correcting models for model-based reinforcement learning. In *AAAI*, 31(1):2597–2603, 2017.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *Transactions on Machine Learning Research*, 2026.

Lloyd N. Trefethen and Mark Embree. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, 2005.

Claas A. Voelcker, Anastasiia Pedan, Arash Ahmadian, Romina Abachi, Igor Gilitschenski, and Amir-Massoud Farahmand. Calibrated value-aware model learning with probabilistic environment models. In *ICML*, PMLR 267:61745–61768, 2025.

Jiawei Wang, Yushen Zuo, Ke Rui, Yichun Feng, and Minglei Li. Decision-metric alignment in latent world models: Diagnostics and action-conditioned objectives for MPC planning. *arXiv:2608.18746v1*, 2026.

Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *AAAI*, pp. 1658–1665, 2019.

Hanzhe You, Yonggang Zhang, Maohao Ran, Zhiqin Yang, Zhenyuan Zhang, Wei Xue, Jun Song, Xinmei Tian, and Yike Guo. A control theory of predictability in latent world models. *arXiv:2607.10362*, 2026.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In *NeurIPS*, 2020.

Zhengshu Zhang and Zhiyuan Li. ARC-Bench: Closed-loop replanning masks broken action ranking in frozen JEPA world models. *arXiv:2609.05461*, 2026.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *ICLR*, 2021.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, PMLR 267:79115–79135, 2025.

A SELECTOR OUTCOMES AND ENDPOINT CONTROLS

Tables 3 and 4 give the historical qualified-pool outcomes and the later complete-audit endpoint sensitivity. The dataset-image energy–projection comparison appears in Table 1.

Table 3: Endpoint-qualified fixed-pool PushT selection and physical success (80 decisions per cohort). M is the original predicted-future score. L and neutral use observed futures; Haar additionally averages 128 post hoc full-norm shared rotations. The Haar-selected winners' 100-step continuation was not measured.

Cohort	Steps	M	L	Neutral	Haar mean
Discovery	25	36.25%	87.50%	71.25%	72.85%
	100	62.50%	97.50%	83.75%	—
R1	25	25.00%	67.50%	61.25%	57.59%
	100	63.75%	85.00%	85.00%	—

Table 4: PushT selection on the complete *audited* 256-action support per decision (40 sources, two optimizer seeds; 80 decisions per cohort). M_{qual} is the original qualified-pool baseline. M_{all} keeps its score definition but admits earlier successes. N_{cont} and N_{abs} use observed futures under the two new terminal semantics; all 100-step counts are closed physical outcomes or already determined by an observed first-plan success.

Cohort	Steps	M_{qual}	M_{all}	N_{cont}	N_{abs}
Discovery	25	29/80	36/80	70/80	68/80
	100	50/80	53/80	75/80	72/80
R1	25	20/80	31/80	73/80	69/80
	100	51/80	56/80	75/80	76/80
Second predictor	25	24/80	29/80	53/80	52/80
	100	52/80	55/80	68/80	68/80

B SEARCH-STAGE GEOMETRY AND AUDIT POPULATION

The fixed-pool analysis motivates a search-stage comparison of signed error geometry. P3 uses 40 episodes disjoint from discovery and R1, with two optimizer seeds per episode. The sources, CEM stage bins (iterations 0–9, 10–19, 20–29), primary statistic, expected direction, and sample size were fixed before candidate generation.

Population and estimand. Candidates are grouped by effective-action hash within each decision before physical audit. A repeated action would inherit the stage and predicted latent of its first stable binding; all 720,080 generated P3 records have distinct hashes within decision. Required actions and then low-predicted-cost deterministic actions occupy up to 128 of the 256 audit slots. Remaining slots are sampled without replacement at equal probability from the remaining unique actions.

The primary analysis uses random-audit actions with a legal observed step-25 endpoint. Of 10,240 sampled actions, 7,087 qualify. Early/middle/late bins contain 3,200/3,520/3,520 actions before this filter and 2,746/2,264/2,077 afterward. Every excluded action (454/1,256/1,443 by stage) succeeded early and terminated, so the exclusion rate rises from 14.19% to 35.68% to 40.99%. All 80 decisions retain early and late observations. The estimand describes endpoint-surviving actions in the sampled remainder, rather than appearance-weighted CEM proposals or the native planner's selected action.

For candidate i, write $d_i = z_i - g$ and $c_i = d_i^\top e_i / (\|d_i\| \, \|e_i\|)$. For source s, seed k, and bin b, we average c_i over qualified random-audit actions to obtain $\bar{c}_{skb}$. The registered estimate is

$$\hat{\Delta} = \frac{1}{40} \sum_{s=1}^{40} \frac{1}{2} \sum_{k \in \{9127, 42017\}} (\bar{c}_{sk,late} - \bar{c}_{sk,early}). \tag{6}$$

A 10,000-draw bootstrap resamples the 40 sources. The near-zero-denominator branch was not used by the 7,087 primary candidates; its numerical rule is recorded in Appendix B.1. The estimate applies to endpoint-qualified random-audit actions.

Table 5: P3 audited effective actions, counted within decision, before and after the legal 25-step endpoint filter. The 7,087 qualified random-audit actions form the primary analysis population; Table 6 gives their stage-specific attrition.

Audit class	Audited	Endpoint-qualified
Required	4,800	2,520
Deterministic	5,440	2,806
Equal-probability random	10,240	7,087
Total	20,480	12,413

Result. Among endpoint-qualified random-audit actions, mean signed cosine is -0.356, -0.563, and -0.652 in early, middle, and late stages (Figure 5). The registered source-paired late-minus-early contrast is -0.296 [-0.363,-0.229], meeting the frozen directional gate. Over the same stages, true terminal cost falls from 86.83 to 26.82 and the 25-step success fraction rises from 0.029 to 0.358. The growing exclusion of early successes limits this result to the endpoint-qualified population.

The stage contrast concerns relative alignment. The late-minus-early raw cross term is -13.05 [-47.64,13.89], error norm changes by +0.175 [-0.492,1.025], and total predicted-minus-true cost changes by -6.04 [-23.97,8.10]; none shows a resolved increase in underestimation. Geometric proximity provides one explanation for the cosine trend. For $a, b > 0$ and orthogonal unit vectors u, w, set $\hat{z} - g = au$ and $e = bw$. Then $c = -b/\sqrt{a^2 + b^2}$ becomes more negative as a decreases, although the residual remains bw and total cost error remains $-b^2$.

The cached latents allow a direct geometric check. With $h_i = \hat{z}_i - g = d_i + e_i$, $a_i = \|h_i\| \,/\, \|e_i\|$, and $\rho_i = h_i^\top e_i / (\|h_i\| \, \|e_i\|)$, Eq. 5 relates predicted-goal proximity to the measured cosine. In a post hoc analysis of the same 7,087 actions, the pooled median a_i falls from 2.13 to 0.81 between early and late stages. Source-paired mean $\log a_i$ changes by -1.076 [-1.269,-0.914], while ρ_i changes by +0.197 [0.113,0.280]. Thus the cosine shift co-occurs with approach toward the predicted goal relative to residual size, even as error alignment with the predicted displacement moves in the opposite signed direction. The nonlinear identity and changing candidate populations prevent attributing a fraction of the shift to either component. Early and late bins contain different actions, so this audit measures a change in the candidate distribution encountered by search rather than a trajectory of one action's error.

B.1 Sampling protocol and estimator

The third cohort contains 40 disjoint episode identities fixed before any new candidate generation, with optimizer seeds 9127 and 42017. The same I30/P300/E30/H5/K5 CEM and cap-256 audit produced 720,080 candidate records and 20,480 audited action memberships. Effective actions are grouped by action hash within decision; this run had no repeated effective-action hash, a check made against the archived candidate records. The stable first ID would otherwise determine both stage and the predicted latent used for its representative. Required actions (native selection, stage rank-zero or final-stage top-k) enter first. The lowest-predicted-cost remaining actions then fill deterministic slots to 128; both classes have inclusion probability one and are excluded from the primary comparison. The remaining slots are randomly drawn without replacement from the rest of that decision's unique-action frame, with equal nonzero inclusion probability within that frame. It is *not* equal probability over all 9,001 generated records, nor an appearance-weighted sample of CEM proposals. Of the random audited actions, 3,153 are excluded because they *succeeded* and simulation terminated before a legal 25-step endpoint; the stage-specific breakdown is in Table 6. No truncation or failed replay accounts for these exclusions.

For source s, optimizer seed k, and fixed bin b, let A_{skb} be its qualified random-audit actions and $\bar{c}_{skb} = |A_{skb}|^{-1} \sum_{i \in A_{skb}} c_i$. All 80 decisions have nonempty early and late bins. The registered source contrast and estimate are

$$D_s = \frac{1}{2} \sum_{k \in \{9127, 42017\}} (\bar{c}_{sk,late} - \bar{c}_{sk,early}), \qquad \hat{\Delta} = \frac{1}{40} \sum_{s=1}^{40} D_s. \tag{7}$$

Table 6: P3 random-audit stage coverage and post hoc geometry from existing cached actions. $a = \|\hat{z} - g\| \,/\, \|e\|$ and $\rho = \cos(\hat{z}-g, e)$. The last two rows pool qualified candidates for description; inferential contrasts in the main text are source-paired.

	Early	Middle	Late
Random audited	3,200	3,520	3,520
Legal endpoint	2,746	2,264	2,077
Terminated before 25	454	1,256	1,443
Median a (qualified)	2.13	1.04	0.81
Mean ρ (qualified)	0.085	0.188	0.333

The 95% interval resamples the 40 source values 10,000 times with seed 20260923. The implementation sets c_i to zero when $2\,\|d_i\|\,\|e_i\| \leq 10^{-12}$ and clips finite results to $[-1, 1]$; none of the 7,087 qualified random records required the zero-denominator branch. The registered gate was $\hat{\Delta} \leq -0.10$ with an interval upper endpoint below zero. The paired result is -0.296 [-0.363,-0.229]. Raw cross term, $\|e_i\|$, total cost error, true terminal cost, and candidate success are secondary. The (a_i, ρ_i) identity and summaries in Section 6 were added post hoc. Their paired intervals use the same two-seed-within-source aggregation and 10,000 source bootstrap draws, with seed 20260924; stage medians and means in Table 6 are pooled descriptions. The estimand retains the audit and endpoint qualifications defined above.

C Reproduction and numerical validation

Discovery, R1, and P3 use the same original checkpoint, encoded goal cost, candidate generator, and endpoint-qualification rule; Appendix J describes the second predictor. The checkpoint-weight SHA-256 prefix is `48938400ae3464c9`, and the corrected full-file expert-cache prefix is `b6ebd9ac94bbe9e3`; complete hashes appear in the R1 protocol and its immutable erratum. Discovery contains 15,578 qualified actions and R1 contains 14,652; R1's six original selectors yield 285 unique continuation branches with complete 100-step outcomes. Discovery uses its original 40 episode identities. R1 takes the first 40 distinct bank-ordered identities after excluding discovery, and P3 takes the next 40 after excluding both, without reading source-bank outcomes. A full-file dataset hash exposed a typo in the frozen R1 metadata; the erratum corrects that hash while preserving sources, arms, estimands, and stopping. Legal endpoints require 25 physical steps, five rendered block endpoints, and finite encoded terminal costs. Early success stops the historical audit and thus excludes those actions from the original qualified pools.

The retrospective matched-norm and cross-term analyses reproduce 560 frozen winners across top-five removal, uniform matching, and five random realizations. The maximum selected-cost difference is 4.0×10^{-13} and the largest matched squared-norm discrepancy is 2.3×10^{-13}. The largest discovery free-cost reconstruction difference, 0.001307, passes the prespecified 0.05 tolerance. Winner parity checks near-tied rankings more directly: all 560 intervened selections match their archives, and the R1 score-path endpoints match both archived natural and neutral winners in all 80 decisions (Table 12). The R1 protocol and dataset-hash erratum, P3 protocol, and P3 execution manifest are separate indexed artifacts in the result bundle. P3 protocol and execution digests are `2267147da9203ea4` and `81b279856f824512`; they identify versions and the prescribed generation sequence but do not independently timestamp the freeze. World-model weights remain frozen in the diagnostic comparisons; Appendix L describes the separately trained calibration heads.

The later full-audited-pool dual-input experiment has protocol digest `8c2380c65b8b9359` and a separate 100-step closure digest `ba140294b4c57bd9`. Independent readback verifies all 240 decisions, 61,440 candidate scores, 226 newly selected continuation branches, and 496 trace references. The frozen new-source protocol has digest `b09169599e37a4f5`; its v3 execution digest is `cc242003a6683183`, with 96 decisions per predictor and no source replacement. Earlier v1/v2 attempts stopped on deterministic resume-metadata errors after one decision and were excluded; v3 reran every source. A post hoc native-batch check for the heterogeneous predictor (digest `68815b6effcb0d55`) reproduces all 9,001 original CEM scalar costs per decision before recomputing E, N, and $M + 2E$ from native-width latents. Its oracle-selector winners match the

frozen fixed-pool choices in all 96 decisions. The companion evidence bundles contain full digests, row/action hashes, source identities, and input metadata.

D HISTORY SUBSTITUTIONS AND ERROR PROPAGATION

Equation 3 compares two histories under the same fixed actions. It is an exact finite difference; the latest-latent Jacobian below is a local derivative with other history frames fixed. Table 7 reports the outcome-selected 84-candidate cohort. True-history single-block prediction improves terminal accuracy for 83/84 candidates and reverses pair order in 38/42 failure/success comparisons. Because the 84 actions come from 42 selected comparisons, these aggregates do not yield a source-level interval or a population error share.

Table 7: Mean latent SSE in the outcome-selected 84-candidate teacher-forcing cohort.

Endpoint	$T5$	$T10$	$T15$	$T20$	$T25$
Free rollout	1.453	3.102	5.406	10.914	37.042
True-history single block	1.453	1.213	1.332	2.117	5.437

E DIRECTIONAL CONTROLS AND LOCKED R1 OUTCOMES

For the latest-latent Jacobian with earlier history and actions fixed, $J_t = \partial F(H_t, a_t)/\partial z_t = U_t \Sigma_t V_t^\top$. We call $G_t = \|J_t e_t\| / \|e_t\|$ the *norm gain* plotted in Figure 3; the squared gain used in the spectral identity is G_t^2. For the top-k *input* subspace, alignment is $A_{t,k} = \left\|V_{t,1:k}^\top e_t\right\|^2 / \|e_t\|^2$. Thus

$$G_t^2 = \frac{\|J_t e_t\|^2}{\|e_t\|^2} = \frac{\sum_j \sigma_{t,j}^2 (V_t^\top e_t)_j^2}{\|e_t\|^2}. \tag{8}$$

This is a local latest-latent sensitivity, not the complete finite response of every history frame; local spectral response and multistep amplification need not coincide (Trefethen and Embree, 2005). A selected earlier PushT failure cohort had mean top-five input alignment 0.125 at $T10$ and 0.186 at $T20$; it is hypothesis-forming, not a balanced failure-versus-success estimate.

The $T20$ input intervention and $T25$ endpoint intervention use different controls. Let $e_{20} = \hat{z}_{20} - z_{20}$ and P project onto the latest-input top-five directions. At $T20$, free history is $(\hat{z}_{10}, \hat{z}_{15}, \hat{z}_{20})$; removal uses $(\hat{z}_{10}, \hat{z}_{15}, \hat{z}_{20} - Pe_{20})$; retention uses $(z_{10}, z_{15}, z_{20} + Pe_{20})$. Uniform and random controls share the true earlier prefix of retention. Thus removal and retention are not symmetric ablations of one history. At $T25$ the output residual is changed directly and no predictor call follows; its top-five projector uses the left singular vectors of the true-history final-block Jacobian. The outcomes of every arm appear in Table 10.

The later candidate-wise removal-matched controls in Eq. 4 use a different remaining-error budget from the original starred controls. Each isotropic direction is generated from a candidate-ID and repetition-specific SHA-256 seed before outcome aggregation. The five draws are averaged within a decision and then a source. Table 11 gives the complete PDF-resident R1 selector summary.

Post hoc R1 score path. For every qualified candidate, the intercept C_i^{neutral} and slope q_i define one affine cost line. We follow the lower envelope of all lines, with lexicographic candidate-ID ties, to find the first $\lambda \in [0, 1]$ at which a neutral-success pool selects a failed action. Table 12 reports selected 25-step outcomes at fixed illustrative strengths. These are deterministic re-rankings of the same archived pool, not new physical executions or 100-step continuations. Reconstructed natural and neutral winner IDs both match the saved cross-term analysis in 80/80 decisions; the natural IDs also match the archived free selector in 80/80. The maximum R1 free-cost reconstruction difference is 2.68×10^{-5}. The first-harm median in the text uses exact line crossings, not the grid in Table 12; no crossing was unresolved as a numerical near-tie.

Table 8: Candidate-wise matched-remaining-norm $T25$ intervention. Random denotes five fixed draws (reported as a range) on discovery but one fixed draw on R1. Differences are source-paired percentage points with 95% source bootstrap intervals; only the R1 25-step top-five–uniform contrast was locked.

Cohort	Steps	Top-five	Uniform	Random	Top-five−uniform
Discovery	25	62/80	50/80	61–65/80	+15 [6.25,25]
	100	68/80	66/80	73–75/80	+2.5 [-3.75,8.75]
R1	25	57/80	53/80	54/80	+5 [-3.75,15]
	100	68/80	70/80	67/80	−2.5 [-10,5]

Table 9: Natural cross-term consequences in fixed qualified pools. The discovery decomposition was defined before its secondary application to R1; R1's locked primary endpoint was a different comparison. Intervals for failure-minus-success gaps resample sources.

Cohort	Mixed pools	Failure−success q [95% CI]	Winner changes	Harm / rescue
Discovery	75/80	-64.25 [-88.58,-43.40]	76/80	32 / 4
R1 (secondary)	69/80	-49.81 [-64.69,-36.81]	78/80	32 / 3

F CUBE-SINGLE SELECTION AND FORK VALIDATION

We froze 32 Cube-Single expert-data episodes absent from 1,049 historical Cube episode identities. Two planner seeds (11, 29) run one $T0$ solve each, with a goal image 50 dataset rows after the start, 30 CEM iterations, 300 proposals, 30 elites, and 25 primitive actions. A candidate-ID hash selects 256 of the final 300 before costs or outcomes are read. M, $L = D$, $N = D + E$, and $M + 2E$ share this pool, initial state, and goal. Sources, not seeds or candidates, are the bootstrap units.

With `terminate_at_goal=False`, task success is any `info["success"]` within 25 steps at cube–target distance ≤ 0.04 meters; `terminated` is not its label. We continue every plan to $T25$, encode its physical image for D and E, and separately encode the first-hit image for an absorbing sensitivity. Every decision has a native-mean fork gate on terminal pixels and cube position; invalid forks remain in the audit without replacement.

The inherited fork dropped the reset seed, sampled a random task, took two unseeded `action_space.sample()` warmup physics steps, then restored only HDF5 qpos/qvel and goal. Two runs of the first 256-action pool changed three any-hit labels and up to 0.0191 meters in terminal cube position despite both passing the native-mean gate. A separately frozen canonical reset fixes the official default task ID, a persistent seeded warmup-action Box, and the environment RNG before setting the same HDF5 state and goal. All 256 initial integration states, 25-step trajectories, endpoint pixels, and labels then agree bitwise across two same-process and one independent-process runs, with the native gate still passing. The formal experiment uses only this disclosed physics reconstruction.

All 64 planned decisions complete. One fork fails its native cube-position tolerance (0.03146 meters versus a 0.001-meter limit); its source pair is omitted without replacement, leaving 31 sources and 62 decisions. Only 30/62 fixed pools contain any successful action. L and N select all 30 covered successes, $M + 2E$ selects 29, and M selects 23. The primary $N - (M + 2E)$ success gain is +1.61 points [0,4.84], one rescue and no harms, with a missing-source worst-case bound [-1.56,4.69]; the frozen positive-transfer criterion fails. True D regret nevertheless falls by 11.54 [6.45,17.70]. This separates a transferable latent-ranking benefit from physical success limited by the particular fixed pool and oracle energy comparator, not by an online-planner bound.

G CROSS-SETTING DIRECTIONAL PROPAGATION

On constant-size failure cohorts, natural-direction gain exceeds equal-norm random gain, while input alignment can rise without monotonic growth of the largest singular value (Figure 3). Cube input alignment rises in 18/20 sources and Reach-Wall in 20/20, yet gain is not a sufficient failure criterion. The 256-dimensional Cube and 102,400-dimensional Reach-Wall spectral budgets are not pooled.

Table 10: Original PushT direction interventions (80 discovery decisions per arm). Columns give selected first-plan and 100-step closed-loop success. The starred controls match the norm of *retained* top-five error; the removal-matched experiment below uses a different budget. $T20$ arms differ in earlier history as specified in Appendix E.

	$T20$ input		$T25$ output	
	25-step	100-step	25-step	100-step
Free error	29	50	29	50
True reference	54	62	70	78
Retain top-five	33	53	39	64
Remove top-five	56	66	62	68
Retain random-five	54	65	69	78
Uniform shrink (*)	48	61	36	56
Random direction (*)	52	64	58	69
Retain cost gradient	42	57	26	51

Table 11: All six R1 selectors on the same 80 decisions (40 independent sources, two optimizer seeds). Successes are decision counts; only top-five minus uniform at 25 steps was the locked primary contrast. C_{mean} is the deterministic spherical-direction expectation.

Selector	25-step success	100-step success
Free prediction	20/80	51/80
Observed terminal L	54/80	68/80
Top-five removal	57/80	68/80
Uniform, same remaining norm	53/80	70/80
One frozen random direction	54/80	67/80
C_{mean}	56/80	68/80

Cube gain and spectra use two trajectories per source in the 20-source failure confirmation cohort. Reach-Wall directional gain uses two trajectories per source, but its top-100 spectral estimates use one. The ratios in Figure 3 are descriptive and have no confidence band; natural-minus-random gain retains its saved source-bootstrap intervals. Success, failure, and random trajectory summaries outside the constant-size failure cohort are not treated as one balanced trend. A separate one-percent fresh-residual experiment found substantial accumulation on successful Reach-Wall trajectories too, so error magnitude and local gain cannot be read as a general task-failure classifier.

H Finite-response prediction pilot

The exploratory sign pilot selects the first two already inspected R1 sources in frozen order and both optimizer seeds, giving four complete candidate pools and 969 candidates. With observed $T10/T15$ held fixed, it perturbs observed z_{20} by $\pm\lambda e_{20}$ for $\lambda \in \{0.25, 0.5, 1\}$. At the same anchor, the signed linear-plus-energy predictor uses $\|b\|^2 + 2s\lambda b^\top Je + \lambda^2 \|Je\|^2$, where $b = F(z_{20}) - g$. Its predictions were hashed before the finite responses were evaluated. Whole-vector sign reversal preserves input and linearized output energy, but finite nonlinear response energies differ.

Exact full-pool winner counts for the signed predictor are 8/8, 5/8, and 4/8 as λ increases; the first-order cost-gradient baseline gives 6/8, 3/8, and 1/8. At $\lambda = 1$, all four $-e$ winners are exact, while none of the four $+e$ winners appears in the predicted top five. One bounded development round compared output-space secant and polynomial extrapolations from $|\lambda| \le 0.5$. The best rule reached 6/8 exact winners but no single rule passed the frozen natural-$+e$ exact/top-five gate, so no new-source predictor confirmation followed.

I Full-norm rotations and score references

The analysis reuses the exact qualified 25-step discovery and R1 pools (40 sources, two optimizer seeds each), with no new simulator trajectory or model prediction. We first locked 128 source-specific Haar-orthogonal draws before evaluating outcomes. For each draw one matrix R acts on *all*

Table 12: R1 post hoc full-pool interpolation $C_i^{\text{neutral}} + \lambda q_i$, with all success labels from existing 25-step physical audits. Harm and rescue are relative to the neutral winner on the same decision.

λ	Winner changes	Harm	Rescue	Successful first plans
0	0	0	0	49/80
0.25	10	2	2	49/80
0.50	28	5	5	49/80
0.75	53	11	6	44/80
1.00	78	32	3	20/80

candidate residuals in a source's two pools, and candidates are rescored by

$$C_i^{(R)} = \|d_i\|^2 + \|e_i\|^2 + 2d_i^\top Re_i. \tag{9}$$

The complete natural error norm and every residual–residual inner product are preserved. The largest relative numerical deviations are below 1.8×10^{-15} for squared norm and 1.6×10^{-14} for sampled Gram entries. We retain lexicographic candidate-ID tie-breaking, average the two seeds within source, and bootstrap 40 sources 10,000 times for natural-minus-reference intervals. The reference distribution itself has no justified rotation-invariance null.

Table 13: Archived 25-step first-plan success in post hoc score references. Rotated and permuted entries are means over 128 preassigned draws, not integer success counts. The full-pool q permutation was designed after the rotation-scale audit.

Cohort	Natural	Neutral	Shared Haar R	Full-pool q permutation
Discovery	29/80	57/80	58.28/80	20.38/80
R1	20/80	49/80	46.07/80	18.68/80

Natural minus shared-rotation success is -36.60 points [-49.85,-23.23] on discovery and -32.59 [-45.42,-19.35] on R1 (paired source bootstrap). But Haar rotation does *not* hold fixed the realized cost-perturbation amplitude: the median ratio of natural to Haar-expected within-pool centered q standard deviation is 9.65 on discovery and 10.31 on R1, and is above one in every decision. Thus Table 13's first reference separates natural orientation from a generic high-dimensional orientation but also reduces cross-candidate score noise.

After observing this limitation, we added a second, explicitly post hoc *score-only* reference that randomly permutes the complete observed q_i multiset across candidates within each pool. It exactly matches the empirical signed-cost mean, variance, and tails but not residual realizability: a new audit finds that 36.91% (discovery) and 37.43% (R1) of assigned candidate values violate $|q_i| \leq 2\,\|d_i\|\,\|e_i\|$. Every one of the 10,240 permutation-pool draws per cohort has at least one violation. Natural minus this permutation's mean success is +10.78 points [1.61,20.57] on discovery and +1.65 [-8.57,12.76] on R1. A stress test permuting q_i only within joint true-cost/residual-norm quartiles produced very low success (5.16% discovery, 0.67% R1), but its oracle-defined strata and broken residual realization also make it unsuitable as a causal null.

We therefore evaluated two additional post hoc references. Independent candidate-wise $e_i \mapsto \pm e_i$ is always realizable, retaining each $\|e_i\|$ and $|q_i|$, but does not keep the centered pool-wide q variance fixed: the median natural standard deviation divided by the sign-reference expected standard deviation is 0.68 on discovery and 0.76 on R1. Its mean 25-step success is 35.66% and 29.45%, versus natural 36.25% and 25.00%.

The more stringent constrained swap reference begins at the observed assignment and proposes $20n$ random pairwise q exchanges per pool and draw, accepting only when both resulting values satisfy their candidates' geometric bounds. Each draw retains the *exact* natural q multiset and per-candidate residual norm; in 192 dimensions, a residual with that norm and any bounded dot product exists. Across 128 fixed draws, 24.6–25.7% of proposed swaps are accepted and 88.1–89.4% of candidates end with a changed q assignment. Its mean 25-step success is 26.92% and 22.19%; natural minus the reference is +9.33 points [3.18,16.41] on discovery and +2.81 [-6.53,12.29] on R1. Finite mixing of this chain is unverified, and neither reference preserves the residual Gram matrix or represents a

trained world model. These randomized references use archived 25-step labels only; neutral winners have separately measured 100-step outcomes in Table 3.

J SECOND-PREDICTOR PROTOCOL AND OUTCOMES

We use the public `MasonJK99/lewm-pusht vanilla_vits/fs5` reproduction checkpoint (SHA-256 prefix `202497efe285cdcb`; 154,822 training steps). Its ViT-Tiny encoder projects to 384 latent dimensions and its predictor differs from the original 192-dimensional frozen model. We load only the Lightning state dictionary with safe tensor reconstruction and strict key matching, then use the same 224-pixel ImageNet evaluation transform and PushT action scaling. This is a capacity-and-training replication rather than a same-architecture second seed. The 40 R1 source episodes and two optimizer seeds are reused, but candidate pools, predicted/observed latent endpoints and physical audit are recomputed for this checkpoint. Of 20,480 audited unique actions, 15,755 have a legal 25-step endpoint and 4,725 are excluded by the existing endpoint rule; every one of these exclusions is an early success. We execute each distinct M/L/neutral winner under its own saved first-plan RNG-after to at most 100 primitive steps.

Table 14: Independent 384-dimensional PushT predictor, 40 R1 sources and two optimizer seeds. Counts are actual physical successes out of 80; neutral and L use observed futures. Paired percentage-point intervals resample sources 10,000 times.

Steps	M	Neutral	L	Neutral−M [95% CI]
25	24/80	41/80	50/80	+21.25 [7.5,35]
100	52/80	66/80	71/80	+17.5 [6.25,30]

At 100 steps, neutral rescues 15 M failures and harms one M success. On the exactly matched 40 sources/two seeds, its gain over M is 3.75 points smaller than the original model's R1 gain, with a paired interval [-20,12.5] for this cross-predictor difference. The feasible same-distribution q swap reuses the fixed 128-draw, $20n$ proposal rule on the new pools; 33.8% of swaps are accepted and 89.5% of candidate assignments change. Its mean archived 25-step success is 23.67%, compared with 30.00% natural, a natural-minus-reference contrast of +6.33 points [0.16,12.87]. This secondary post hoc reference supports the same limitation on special natural-alignment harm. The source episodes are reused from R1.

K EARLY-SUCCESS ENDPOINT SENSITIVITY

This retrospective analysis keeps the original candidate generation, source episodes, two optimizer seeds, physical start and goal, effective action, frozen checkpoint, and downstream continuation protocol. The original endpoint audit executed 256 distinct effective actions per decision and stopped on `terminated=True`, which the PushT implementation sets exactly when the task's four-coordinate position and symmetry-aware angle thresholds are met. All 4,902/5,828/4,725 actions excluded from discovery/R1/second-predictor qualified pools were early successes, not failed rollouts or truncations. Their first-hit time is already a valid first-plan success label; their natural-stopping trajectory has no observed $T25$ latent. Table 15 separates coverage from scoring.

Table 15: Audit support per cohort (80 decisions, 20,480 physically audited unique-action memberships). "Successful pool" counts decisions with at least one 25-step successful action; native and M_{all} are distinct selectors. Early winners are selected actions already successful before step 25.

Cohort	Early hits	Successful pool qual/all	Native 25	Native early	M_{all} early
Discovery	4,902	75/79	38/80	26/80	18/80
R1	5,828	69/80	42/80	33/80	23/80
Second predictor	4,725	56/66	32/80	25/80	18/80

The new main endpoint runs all 25 original physical actions even after the first success, recording images at steps 5, 10, 15, 20, and 25 and retaining a sticky first-hit success label. The predefined

sensitivity endpoint instead uses the actual image rendered at first success as an absorbing state for remaining block endpoints. Neither endpoint is mislabeled as the original stopped trajectory. For originally qualified actions both reuse archived observed latents; early-success actions have both new observed terminal encodings. Predicted actions, the original goal image, and M_{all} remain fixed across endpoint semantics. The neutral score is recomputed from the corresponding true latent and the same candidate's natural residual, so $M = L + E + q$ is checked per action.

In the historical qualified-pool and rendered-input sensitivity analyses, M uses the simulator-rendered initial image; Table 1 instead uses the dataset image employed by native CEM. Substituting dataset-image input into the historical scoring chain reproduces only 10/80 R1 and 4/80 second-predictor old M winners. All old M winners were therefore reproduced before scoring excluded actions. Discovery M also requires the original 30 batches of 300 candidates plus a singleton final mean. Replaying this partition reproduces all 9,001 archived scalar costs per decision and supplies predicted latents for the audited 256. The old batch-32 latents differ by at most 0.001945 in one component, but qualified-pool neutral winners remain identical in 80/80 decisions. Partial scoring versions stopped by the identity gate are archived separately, not pooled.

All 15,455 early-success actions were physically re-executed to step 25 with first-hit times matching their archives; one early action per nonempty decision was repeated exactly. Initial/goal encodings match the archives, and one qualified five-image terminal sequence was re-encoded per cohort. The maximum $|M - (L + E + q)|$ across the 240 decisions is 4.02×10^{-5}, 2.68×10^{-5}, and 6.58×10^{-5} for discovery, R1, and the second predictor, below the frozen 10^{-3} gate. Both new neutral winners are stable under repeat or alternative-batch scoring. All 100-step labels come from an observed first-plan success or verified cached continuation; no selected branch is missing, and one old long branch per cohort replayed with the same outcome.

At 25 steps, the source-paired neutral-minus-M_{all} contrasts under continued physics are +42.5 [31.25,53.75], +52.5 [37.5,66.25], and +30 [17.5,42.5] points; under absorption they are +40 [28.75,51.25], +47.5 [33.75,61.25], and +28.75 [17.5,41.25]. Table 4 reports 100-step counts. Intervals resample 40 sources 10,000 times after averaging their two optimizer seeds (seed 20260925). These already examined sources and the second predictor's reuse of R1 episodes define this as a development-cohort check.

L CALIBRATION PROTOCOL AND TEST RESULTS

Before training, we fixed 60 old source episodes for fitting (40 discovery and the first 20 sorted R1 sources), 20 other R1 sources for validation, and all 48 P2 original-model sources for one fixed method test. Each source's two optimizer seeds remain together. P2 outcomes were already known in the oracle study, so "method-unseen" here means only absent from calibrator fitting and configuration selection. The checkpoint and dataset-image chain are frozen. Each candidate supplies its 50 normalized action coordinates, the 192-dimensional initial-minus-goal embedding, all five 192-dimensional predicted embeddings minus the goal, and their five squared goal costs. No observed future enters the 1,207-dimensional feature vector.

The common MLP has widths 256 and 128, GELU and 0.1 dropout (342,402 parameters for two heads; 342,273 for one). A weighted SmoothL1 loss, AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}, batch size 1,024, 40 epochs, and three training seeds are shared across the energy, direct-cost, and decomposed heads. Each candidate in the predicted-score top 64 has five times the unit loss weight. Features and each target are standardized using training candidates only. Validation source-mean success, then lower true-D regret and earlier epoch, chooses one checkpoint from epochs 5, 10, ..., 40 and $\beta \in \{1, 2\}$ where relevant; the same choice rule averages all three seeds. Energy predictions are clipped to zero when constructing costs. The chosen epochs are 30/40/40 for energy/direct/decomposed, and both energy and decomposed choose $\beta = 1$.

The accepted feature extraction exactly reconstructs archived predicted terminal latents and M scores on all 80 discovery, 80 R1, and 96 P2 decisions (maximum score discrepancy 1.14×10^{-13}). It uses original 300-wide CEM batches for discovery, separate eligible/early 32-wide batches for R1, and sorted 32-wide batches for P2. An earlier R1 extraction stopped on a 0.0512 cost mismatch caused by mixing its two batch populations; none of its outputs is training input. On validation, M succeeds in 22/40 decisions (55%), whereas the three-seed decomposed mean is 35% despite lower

mean true-D regret. The predeclared online gate required at least a five-point success gain without increased regret and therefore closed before the P2 method test.

Table 16: Frozen P2 original-checkpoint 25-step fixed-pool test on the same 48 sources and 96 decisions. Counts are selected physical successes per training seed. Shuffling permutes the within-pool assignment of $\hat{p}$ while retaining its multiset.

Training seed	M	Energy only	Direct scalar	Decomposed	Shuffled $\hat{p}$
1101	41	33	41	39	33
2202	41	37	49	48	37
3303	41	38	42	48	37
Mean success (%)	42.71	37.50	45.83	46.88	37.15

The decomposed-minus-M source-paired success contrast is +4.17 points [-8.33,17.01]; decomposed-minus-direct is +1.04 [-3.47,6.25], decomposed-minus-energy is +9.38 [0.69,19.10], and decomposed-minus-shuffled-projection is +9.72 [-1.04,21.18]. These intervals bootstrap 48 sources after averaging two optimizer seeds and all three training seeds. Decomposition's true-D regret difference from M is +2.62 [-8.54,12.74], so neither test endpoint establishes a benefit over the original score. Test candidate-level R^2 ranges across seeds are 0.19–0.32 for predicted E, 0.06–0.25 for predicted p, and 0.14–0.28 for direct cost correction. Warmed H200 forward time is about 0.087–0.088 ms per 256 candidates, excluding world-model rollout and feature construction. No online branches were launched after the failed validation gate.

M Additional related work

Planning with learned worlds. Model-based planning spans Dyna and probabilistic policy search through image-based imagination, latent model-predictive control, and large-scale learned dynamics (Moerland et al., 2023; Sutton, 1991; Deisenroth and Rasmussen, 2011; Ha and Schmidhuber, 2018; Kaiser et al., 2020; Hafner et al., 2019; 2020; Hansen et al., 2022; 2024; Hafner et al., 2025). Joint-embedding prediction supplied a separate route to visual representations (Assran et al., 2023; Bardes et al., 2024; Balestriero and LeCun, 2025); V-JEPA 2 and DINO-WM connect predictive visual features to planning (Assran et al., 2025; Zhou et al., 2025). LeWorldModel (LeWM), our frozen backbone, trains its encoder and predictor jointly from pixels (Maes et al., 2026). Architectural, objective, and efficiency comparisons of this family address different levers from our fixed-model score audit (Terver et al., 2026; Gao and Xu, 2026).

Prediction error and decision quality. Teacher forcing and free-running prediction can induce different state distributions (Ross et al., 2011; Bengio et al., 2015); model-based RL studies multi-step amplification, self-correction, and how much imagined experience to trust (Talvitie, 2017; Asadi et al., 2018; Pan and Duraisamy, 2018; Janner et al., 2019; Song and Cai, 2026; Liu et al., 2026b). Task-aware representations and model losses ask which information must survive for control (Ferns et al., 2004; Gelada et al., 2019; Zhang et al., 2021; Farahmand et al., 2017; Grimm et al., 2020; Lambert et al., 2020; Voelcker et al., 2025). More broadly, decision-focused learning optimizes downstream choice (Wilder et al., 2019; Elmachtoub and Grigas, 2022). Recent theory analyzes planner-reachable cost gaps and harmful one-sided underestimation (You et al., 2026); our controlled comparison measures the contribution of a realized signed error term to action selection.

Closest finite-pool diagnostics. Decision-Metric Alignment measures rank agreement and candidate margins through CEM search (Wang et al., 2026). Trajectory Reachability Metrics trains a new terminal cost with a frozen model; temporal-distance and geometry-aligned approaches modify learned costs or representations (Li et al., 2026b; Bai and Xiong, 2026; Hu et al., 2026; Chahe et al., 2026). ASAR separates candidate coverage from selection, ARC-Bench studies how replanning affects first-plan rankability, and D-JEPA measures decision-local inversions before learning a relational correction (Li et al., 2026a; Zhang and Li, 2026; Liu et al., 2026a). Our oracle contrast keeps the original squared-goal metric and each candidate's realized error energy fixed, isolating the goal-related projection's contribution to ranking and physical outcomes.

Selection and evaluation. Cross-entropy optimization repeatedly concentrates on low estimated costs (Rubinstein, 1999; de Boer et al., 2005), so post-selection optimism and model exploitation are relevant prior concerns (Smith and Winkler, 2006; Chua et al., 2018; Yu et al., 2020; Kidambi et al., 2020). Our physical audits distinguish an available successful action from one selected by a score. Paired source-level uncertainty and exact replay gates matter because small evaluation cohorts and implementation details can change apparent control gains (Henderson et al., 2018; Agarwal et al., 2021; Engstrom et al., 2020).