CANDIDATE-DEPENDENT MARGINS IN LATENT WORLD-MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Candidate actions change the measured margins between fixed latent world models, even when their top-ranked predictor remains unchanged. We quantify this dependence through paired candidate sets at fixed checkpoints, initial states, goals, horizons, and candidate counts. Among seven predictors sharing an encoder on 64 Push-T families, all six mean gain advantages over a post-hoc logged-data comparator contract by 36–65%. Their raw prediction-error advantages also contract by 36–71%, so this sensitivity extends beyond reference normalization. These six paired gaps share one comparator and do not constitute independent replications. Twelve of 21 post-hoc score-response contrasts survive a studentized sign-flip test with Holm correction; paired t tests retain the same findings. Separately, a public DINO-WM reproduction gains 0.100 in persistence-relative score despite an 11.6% increase in prediction error. Family-level accounting separates the opposing model-error and reference-error terms. A prospective PointMaze study identifies worse available task opportunity without resolving a change in within-set regret. We provide a reproducible audit of prediction error, reference difficulty, available opportunity, selected cost, regret, and cost discrimination. These measurements quantify candidate-dependent comparison margins and delimit their implications for action selection.

1 INTRODUCTION

Latent dynamics support planning from pixels and policy learning through imagined trajectories (Hafner et al., 2019; 2020). Recent visual world models score candidate futures in learned representations (Zhou et al., 2025; Assran et al., 2025; Terver et al., 2026). An evaluation therefore queries a trained model on an action distribution: logged trajectories, random proposals, or candidates concentrated by a planner (Chua et al., 2018). A model comparison is conditional on that distribution even when its summary is reported as a single score or improvement margin. The practical question is how much those margins change when the trained objects stay fixed.

Prediction error and control performance need not agree (Lambert et al., 2020; Lutter et al., 2021). Recent work measures the alignment between latent and task rankings during planning (Wang et al., 2026) and distinguishes data-averaged prediction from fidelity on planner-reachable states (You et al., 2026). These findings motivate a more specific measurement problem: when candidate composition changes, which part of an evaluation changes with it? A normalized prediction score can respond to both model error and reference difficulty. A selected action's cost can respond to both the opportunities supplied by the set and the choice made within it. The cost may also assign the same value to all candidates, leaving no resolution for comparing their selection. Each case calls for a different interpretation.

We study these distinctions through paired candidate interventions. Within each comparison, we fix the checkpoint, initial state, goal, horizon, and candidate count, then change the action sequences. We measure model error M, persistence error B, and relative gain $G = 1 - M/B$, together with best available task cost O, selected cost L, and within-set regret $R = L - O$. Gain is a reference-normalized diagnostic adopted for this study; it is not the headline metric of DINO-WM or JEPA-WMs. Reporting its components lets us test whether a change in relative performance also represents a change in unnormalized error. Reporting O and R separately locates a selected-cost change within the candidate set's possibilities.

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

The experiments support three contributions:

- **Candidate-dependent model-comparison margins.** Among seven predictors sharing an encoder, every advantage over a post-hoc logged-data comparator contracts in both normalized gain and raw prediction-error units. The analysis retains all pairs, with corrected inference, levels, and ranks; the top-ranked model remains unchanged (Section 4).
- **Score sensitivity accounted for at the family level.** A public DINO-WM reproduction has higher persistence-relative gain alongside higher prediction error. Exact accounting separates the opposing model-error and reference-error terms. A controlled study supports the same joint pattern under a narrower protocol (Section 5).
- **A reproducible audit of task opportunity and selection.** A prospective PointMaze test separates worsening available opportunity from an unresolved regret change. Simulator diagnostics identify sets on which zero regret cannot distinguish selectors. The released procedure makes these distinctions inspectable (Sections 6 and 7).

Why margins matter Model comparisons communicate how much a configuration improves a score and whether its advantage is resolved. Both can change while checkpoints remain fixed. Here, six gaps sharing a post-hoc logged-data comparator shrink substantially, the gap between the two leading predictors widens, and one previously resolved ordering becomes unresolved. Rankings need not change for these conclusions to depend on the evaluated candidates. Relatedly, reconstructed ImageNet tests shifted accuracy levels while gains on the original test set carried over (Recht et al., 2019).

Scope of claims Training data, objectives, and update counts vary jointly, so the measured gaps do not identify causal effects of individual training choices. Our claim concerns their conditional size and resolution on the specified banks. Prospective and released-planner tests leave regret or performance-margin changes unresolved; the observed score sensitivity therefore supplies no demonstrated harmful model-selection effect.

2 Related Work

Prediction and planning Static prediction diagnostics and downstream control measure different properties (Lambert et al., 2020; Lutter et al., 2021; Kégl et al., 2021). Value-aware model learning (Voelcker et al., 2022) and joint representation, model, and policy objectives (Ghugare et al., 2023) address this through training. DINO-WM and JEPA-style models plan in learned representations (Zhou et al., 2025; Assran et al., 2025; Terver et al., 2026). Our experiments hold the learned objects fixed and measure the effect of evaluation queries.

Candidate distributions and diagnostic resolution Concurrent work by Wang et al. (2026) measures latent–task rank agreement on random plans and across CEM stages, excluding constant-cost pairs when rank correlation is undefined. Concurrent work by You et al. (2026) distinguishes data-averaged error from fidelity on states a planner can reach. Zhang et al. (2026) evaluates world models in closed-loop environments and finds that visual quality alone does not ensure task success. We instead hold checkpoints and query budgets fixed to jointly measure prediction and reference errors, task opportunity, and within-set excess cost. Constant costs and prediction–control mismatch are established observations; their role here is to constrain interpretations of measured score and margin changes.

Empirical evaluation and reference scores Controlled re-evaluation has exposed the influence of implementation and training choices (Engstrom et al., 2020; Andrychowicz et al., 2021), learning objectives and baselines (Chen et al., 2019), and evaluation uncertainty (Agarwal et al., 2021). Reference-relative skill scores have a long history in forecasting (Murphy, 1988; Hamill & Juras, 2006; Mittermaier, 2008); their dependence on the reference and sample composition is established. We apply that accounting to latent world-model queries and quantify its relation to model-comparison margins and task opportunities. The algebra supplies the definitions; the empirical contribution is their measured magnitude and scope under matched candidate interventions.

3 PAIRED EVALUATION OF CANDIDATE DISTRIBUTIONS

Queries and independent units From simulator state s, candidate $a = (u_1, \ldots, u_H)$ produces endpoint $\Phi(s, a)$. A candidate bank $\mathcal{C}(s)$ contains K sequences. We use *candidate distribution* for the generator-induced empirical distribution over this finite bank; the contrasts concern these banks, rather than all possible action distributions. Each comparison uses two sets at the same checkpoint, start, goal, H, and K. In Push-T these are an ordinary mixture and a contact-enriched mixture; in PointMaze they are a reference set and its geometry-selected rotation. Every Δ denotes the second set minus the first. A trajectory family is the independent unit; its candidates, starts, and training-seed results remain together during aggregation and resampling.

Prediction and reference difficulty Let $z(s) = \phi(s)$ be the frozen representation (with its prescribed observation history included in s), $\hat{z}(s, a)$ the model prediction, and $z^\star(s, a) = \phi(\Phi(s, a))$ the realized endpoint representation. We compute squared errors

$$\varepsilon_\theta = d^{-1}\|\hat{z} - z^\star\|_2^2, \qquad \varepsilon_0 = d^{-1}\|z(s) - z^\star\|_2^2. \tag{1}$$

At each start, $M(s)$ and $B(s)$ are their means over candidates. Relative gain is

$$G(s) = 1 - M(s)/B(s), \qquad G(\theta, \mathcal{C}) = \frac{1}{N}\sum_{f=1}^{N}\frac{1}{|\mathcal{S}_f|}\sum_{s\in\mathcal{S}_f} G(s). \tag{2}$$

The ratio precedes state and family averaging. Pooled error ratios are additional descriptive readouts and do not determine the change in this mean of ratios. For $B_o, B_r > 0$, a single start satisfies

$$G_r > G_o \quad \Longleftrightarrow \quad M_r B_o < M_o B_r. \tag{3}$$

If both model errors are positive, this is equivalent to $M_r/M_o < B_r/B_o$. G is dimensionless; a shared numerical range does not make it a common quality scale across different encoders and references. A *comparison margin* is a score gap between models; an *equivalence tolerance* is a separately specified inferential threshold. Persistence supplies an interpretable unchanged-scene reference, familiar from video prediction (Lotter et al., 2017) and forecasting (Mittermaier, 2008). We use its skill score as a diagnostic and report M and B with it. Comparisons of raw errors require a common representation, including the encoder's input resolution and preprocessing.

Task opportunity and within-set regret The selector minimizes predicted latent goal cost, $\hat{a} = \arg\min_{a\in\mathcal{C}} d^{-1}\|\hat{z}(s, a) - z^g\|_2^2$. Exact predicted-cost ties choose the first candidate in the stored ordering. For realized task cost $J(s, a)$, lower being better, define

$$O(s) = \min_{a\in\mathcal{C}} J(s, a), \qquad L(s) = J(s, \hat{a}), \qquad R(s) = L(s) - O(s). \tag{4}$$

Thus $\Delta L = \Delta O + \Delta R$. This separates the best attainable task cost from the selected action's excess cost within that set. Both depend on candidate composition; R does not isolate an intrinsic property of a model. We also record cost ranges and ties. If every candidate has cost c, all selectors have $O = L = c$ and $R = 0$, regardless of their predictions.

3.1 EXPERIMENTAL PROTOCOLS AND INFERENCE

Table 1 identifies the main studies. In Push-T (Florence et al., 2022; Chi et al., 2023), task cost is one minus block–goal coverage. The ordinary set mixes anchor, directional, magnitude, temporal, and planner-proposal actions. The enriched set changes the non-anchor mixture using start geometry and proposed actions; the controlled bank retains the historical logged continuation as an explicit anchor. PointMaze uses an action-magnitude-matched rigid rotation selected from static maze geometry. Full constructions and goals are specified in Appendix C.

PA8 is the controlled predictor and one of seven configurations sharing a 15.0M-parameter transformer and fixed V-JEPA 2-derived features, with three seeds per configuration. Their labels denote logged data (LD), branched data (BD and BA), counterfactual augmentation (CF), planner-distribution augmentation (PA), and multiscale targets (MS); suffixes indicate thousands of updates. Data, objectives, and training duration vary together. These are comparisons of trained configurations, rather than isolated training ablations. Table 10 gives the model map.

Table 1: Main evidence and its sampling scope. H is physical controls and K is candidates per set. Checkpoint, start, goal, H, and K are matched within each pair.

Study / checkpoint	Families	H	K	Design and goal
Push-T / public DINO-WM reproduction	64	10	67	Fixed cohort/scoring; subsequent bank analysis; environment target.
Push-T / seven shared-encoder predictors	64	10	67	Same cohort; three seeds each; post-hoc gap analysis; environment target.
Controlled Push-T / PA8	16	16	68	Two starts and three seeds per family; specified reanalysis; logged endpoint goal.
PointMaze / public DINO-WM	32	25	64	Prospective after pilot; normalization-exposed public data; recorded target.
Push-T rotation / simulator	32	25	64	Descriptive cost discrimination; environment target.

Except in the controlled study, each family has one start. Five physical controls form one model step in the public-model protocols. Validation scope: Appendix E.2; supplementary studies: Appendix A.

The public reproduction is JEPA-WMs' released DINO-WM checkpoint evaluated through its native interface. Recomputed candidate-error records exactly reproduce its 64 family summaries. A separate input-level check explains why its persistence errors cannot be equated with those of the original DINO-WM implementation: the encoders receive 224×224 and 196×196 images, respectively. This resolves the interpretation of an earlier cross-pipeline reference check, but is not an independent end-to-end prediction parity test (Appendix E.2). Primary comparisons use one pipeline within each row.

For the main candidate-bank studies, contrasts are paired before family averaging. We report percentile family-bootstrap intervals; a joint gain–error increase requires both 95% intervals to exclude zero positively. The representative test for all 21 shared-encoder gap changes is a two-sided studentized sign-flip test (Winkler et al., 2014) with Holm correction (Holm, 1979) (99,999 common sign vectors; add-one Monte Carlo p values (Phipson & Smyth, 2010)). It assumes independent family contrasts symmetric about zero under the null; candidate assignment was not randomized. Paired-t sensitivity retains the same 12 findings. This reporting choice is post hoc; the original bootstrap zero-tail summaries remain historical diagnostics (Appendix G). PointMaze sample size, endpoints, margins, and decision rules were recorded internally after the pilot and before new outcomes; there is no external registration timestamp. Non-detection and equivalence have distinct decision rules.

4 Comparison Margins Change within a Shared Encoder

The seven predictors share their representation and persistence reference. This permits comparison in both gain and raw model-error units. Table 2 reports levels, ranks, and all six comparisons against the logged-data configuration LD6.

Effect sizes beyond the winner All six other configurations have higher mean gain than LD6 on both candidate sets. Every advantage contracts, by 35.6–64.8%, with a median contraction of 49.7% across the six pairs. These correlated contrasts share LD6 and are not six independent replications. The focus on this comparator was selected after inspection; all 21 pairs remain in the inferential family. Across their mean gaps, 15 contract and six expand, a descriptive count of $|\overline{G_i - G_j}|$, not a significance count. Taking absolute gaps before family averaging gives ten contractions and eleven expansions instead.

The top-ranked model, BA7, stays first. Its gap to MS8 increases from 0.0028 to 0.0086, with paired change $+0.0058$ $[+0.0022, +0.0091]$. We use the additive contrast because the ordinary-gap interval includes zero. For BD6 minus BA8, the ordinary gap is -0.0056 $[-0.0110, -0.0004]$ and the enriched gap is $+0.0005$ $[-0.0063, +0.0074]$. This ordering becomes unresolved; its point sign change does not establish a ranking reversal.

Table 2: Shared-encoder model comparison. Gains are family means; ranks order those means among the seven predictors. $\Delta G = G_{\text{rich}} - G_{\text{ord}}$ has a paired 95% family-bootstrap interval. Ranks are descriptive; the top predictor is unchanged.

Predictor	G_{ord}	Rank	G_{rich}	Rank	ΔG [95% CI]
BA7	0.5685	1	0.5706	1	+0.0022 [−0.0040, +0.0084]
MS8	0.5657	2	0.5621	2	−0.0036 [−0.0098, +0.0026]
PA8	0.5579	3	0.5616	3	+0.0037 [−0.0027, +0.0101]
BA8	0.5532	4	0.5482	5	−0.0050 [−0.0131, +0.0037]
BD6	0.5476	5	0.5487	4	+0.0011 [−0.0059, +0.0081]
CF6	0.5371	6	0.5461	6	+0.0090 [+0.0003, +0.0177]
LD6	0.5101	7	0.5330	7	**+0.0230 [+0.0160, +0.0300]**

Pair	Ordinary gap	Rich gap	Change [95% CI]	Shrinkage
BA7−LD6	0.0584	0.0376	**−0.0208 [−0.0270, −0.0150]**	35.6%
MS8−LD6	0.0556	0.0291	**−0.0265 [−0.0324, −0.0210]**	47.7%
PA8−LD6	0.0478	0.0286	**−0.0193 [−0.0244, −0.0142]**	40.3%
BA8−LD6	0.0431	0.0152	**−0.0279 [−0.0358, −0.0205]**	64.8%
BD6−LD6	0.0375	0.0157	**−0.0218 [−0.0276, −0.0163]**	58.1%
CF6−LD6	0.0270	0.0131	**−0.0140 [−0.0202, −0.0078]**	51.7%

The lower panel includes every comparison against LD6; focusing on this comparator is post hoc. Median shrinkage is **49.7%** across these six mean gaps, not across trajectory families. Bold highlights the logged-data comparator; underlining marks its gap shrinkages and the unchanged winner. Nominal intervals use shared family draws; all 21 tests remain in Appendix G, together with model-error gaps and reference accounting.

Contrasts are computed before rounding.

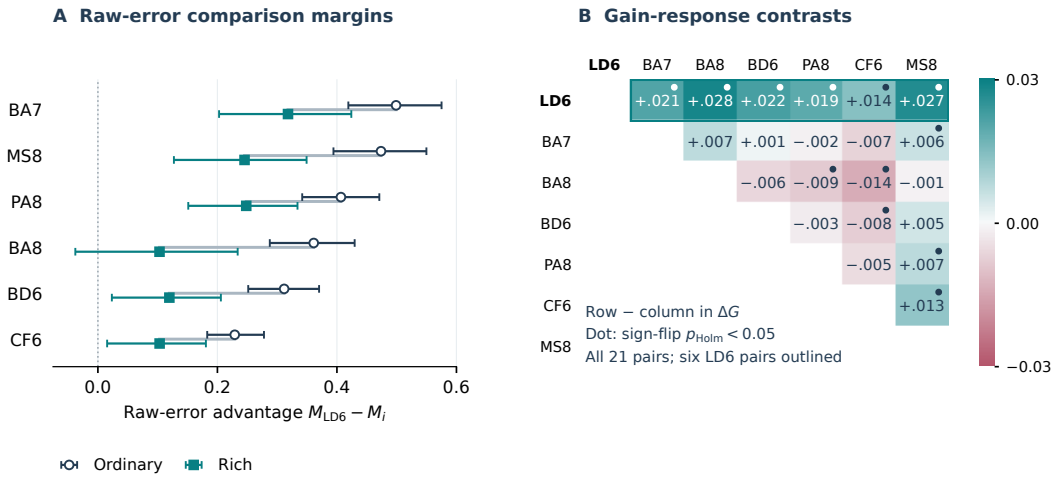

Figure 1: **Candidate sensitivity within a shared representation.** A: all six raw-error advantages over LD6 on each bank, with nominal paired 95% family-bootstrap intervals. BA8's rich-bank level includes zero; each paired *change* interval is negative. B: all 21 gain-response differences, row minus column. Dots mark studentized sign-flip $p_{\text{Holm}} < 0.05$ over all 21 pairs; the highlighted LD6 row shows the complete logged-data comparator set. Paired-t sensitivity retains all twelve findings. The focus on LD6 is post hoc; the correction covers all pairs.

Sensitivity beyond normalization Define a raw-error advantage over LD6 as $M_{\mathrm{LD6}} - M_i$. All six advantages also shrink, by 36.2–71.4% (pair median 51.5%), with negative nominal paired change intervals (Figure 1A). These raw-error intervals are not an additional multiplicity-corrected testing family. For BA7, the advantage falls from 0.4990 to 0.3181, a change of -0.1808 [-0.2455, -0.1213]. LD6's error increases by 0.0702 while the other increases range from 0.1959 to 0.3282. The effect is therefore present in the common representation's unnormalized errors. Exact family-level accounting assigns larger point magnitudes to the model-error-gap terms than to reference terms for all six comparisons, under both substitution orders (Appendix G).

For models i, j, the tested contrast is $D_{ij} = \Delta G_i - \Delta G_j$. Figure 1 displays the full set. Twelve of 21 differences survive the studentized sign-flip test with Holm correction: six involve LD6 and six compare the remaining models. Paired-t sensitivity, corrected over the same family, retains those twelve. The supported consequence is a change in the measured size and resolution of model comparisons at fixed checkpoints.

5 Prediction Scores Respond to Reference Difficulty

Table 3: Prediction-score sensitivity in Push-T. Rich minus ordinary changes have paired 95% family-bootstrap intervals. Each row holds its checkpoint and query conditions fixed; the rows span different representations and protocols and are not pooled. Ratios compare pooled mean errors within each row, not the mean of state-level gains. Bold pairs mark jointly positive gain and model-error changes.

Checkpoint / families	ΔG [95% CI]	ΔM [95% CI]	M_r/M_o	B_r/B_o
DINO-WM reproduction / 64	**+0.0995 [+0.0880, +0.1117]**	**+0.0066 [+0.0030, +0.0108]**	1.116	1.608
PA8 (controlled) / 16	**+0.0304 [+0.0069, +0.0544]**	**+0.2210 [+0.1054, +0.3529]**	1.065	1.104
PA8 (selection) / 64	+0.0037 [−0.0027, +0.0101]	+0.2286 [+0.1723, +0.2906]	1.060	1.066

Reproduction means JEPA-WMs' released DINO-WM checkpoint, not the original authors' release. M has each representation's squared-error units; cross-representation magnitudes are not comparable. The controlled row uses $H = 16$ physical controls; both 64-family rows use $H = 10$. PA8/64 uses the same shared resampling as Table 2. Protocol and validation scope: Table 1 and Appendix E.2. Full controlled endpoints: Table 17.

Contrasts are computed before rounding.

A public checkpoint on matched queries On 64 Push-T families, the public reproduction's mean gain rises from 0.6813 to 0.7807. Its model error rises from 0.0574 to 0.0640 and persistence error from 0.1830 to 0.2943: pooled increases of 11.6% and 60.8%, respectively. Table 3 gives paired uncertainty. Gain increases in 63 families, and both gain and model error increase in 38. The higher relative score therefore accompanies worse unnormalized prediction error on average.

Accounting at the correct aggregation level To distinguish the terms without relying on pooled ratios, we decompose each start's gain change and then apply Equation (2). Averaging the two sequential substitution orders gives

$$\Delta G(s) = \underbrace{-\tfrac{1}{2}(M_r - M_o)(B_o^{-1} + B_r^{-1})}_{A_M(s)} + \underbrace{\tfrac{1}{2}(M_o + M_r)(B_o^{-1} - B_r^{-1})}_{A_B(s)}. \tag{5}$$

For the reproduction, the model-error term is -0.0250 [-0.0415, -0.0108] and the reference-error term is $+0.1244$ [$+0.1125$, $+0.1385$]; they sum to $+0.0995$. Both substitution orders retain these signs. This post-hoc accounting is descriptive, not a causal intervention on either error term (Appendix F.3).

A controlled counterpart and its boundary The 16-family PA8 study holds two starts and three seeds per family fixed. At $H = 16$, model error rises by $+0.2210$, reference error by $+0.6391$, and

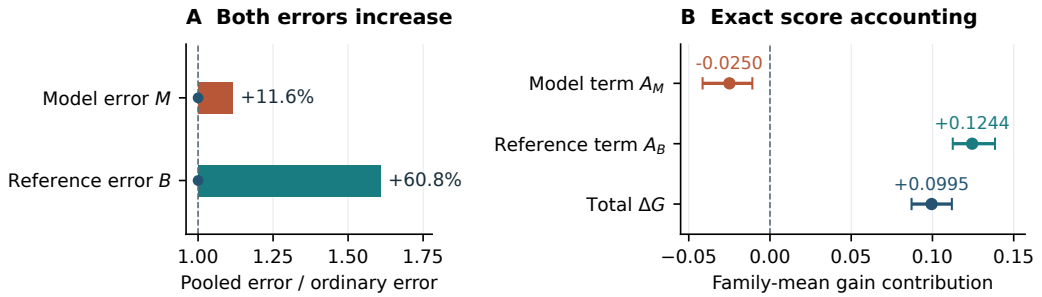

Figure 2: **Why gain can rise with error in the public Push-T reproduction.** A: pooled error ratios describe the two error increases. B: exact accounting is performed within each family and then averaged, with 95% family-bootstrap intervals. The reference term outweighs the negative model term. This is a descriptive decomposition, not causal attribution; its intervals use 20,000 shared draws (Appendix F.3).

gain by $+0.0304$ (Table 3). The seed-average joint criterion holds, as does one of three individual-seed criteria; all three gain changes have positive signs. Among horizons $H = 4, 8, 16$, only $H = 16$ meets the joint criterion. On the separate 64-family $H = 10$ cohort, PA8's gain change is unresolved despite higher model error; its 32-family $H = 16$ test-stream change is also unresolved. These cohorts differ in starts and candidate counts as well as horizon, so they do not isolate a horizon effect.

A secondary endpoint detects reduced latent–task ranking agreement, which does not establish top-1 selection harm. Complete endpoints, seed and horizon contrasts, and separate-cohort boundaries are in Appendix F.

6 Task Opportunity and the Resolution of Selection

A prediction-score change does not specify how the available actions or the selected action change. We apply $L = O + R$ to continuous task costs and test where the cost can distinguish choices.

A prospective PointMaze boundary test After an eight-family pilot, we fixed the sample size, primary regret hypothesis, equivalence margins, and decision rules for 32 new families. The geometry set rigidly rotates each reference sequence using the initial position and map, preserving action magnitudes, $K = 64$, and five model steps (25 controls). Selection minimizes predicted terminal latent goal cost. Task cost is obstacle-aware shortest-path distance under the specified point-agent geometry, measured at control 25 and averaged over states after controls 1–25, excluding the initial state. Distances use the environment's native planar coordinate units. These readouts need not agree because the selector optimizes a terminal objective.

Table 4: Prospective PointMaze evaluation ($n = 32$, separate from the eight-family pilot). Changes are geometry minus reference. Task intervals are paired family 95% intervals; model-metric equivalence uses 90% intervals. Regret was not tested for equivalence. Bold contrasts separate worse available opportunity from unresolved selection regret.

Task quantity	Reference	Geometry	Change [95% CI]	Test
Time-mean regret R	0.352	0.313	$\mathbf{-0.039}$ $\mathbf{[-0.191, +0.117]}$	**Not supported**
Terminal regret R_H	0.235	0.264	$+0.030$ $[-0.190, +0.216]$	Holm: $p = 1.000$
Time-mean optimum O	0.584	0.666	$\mathbf{+0.082}$ $\mathbf{[+0.046, +0.119]}$	Holm: $p < .001$
Model metric		Margin	Change [90% CI]	Criterion
Relative gain G		± 0.05	$+0.0131$ $[+0.0051, +0.0204]$	Met
Log model error $\log M$		± 0.10	$+0.0040$ $[-0.0052, +0.0133]$	Met

Contrasts are computed before rounding.

The primary increase in time-mean regret does not replicate: $\Delta R = -0.039$ [-0.191, $+0.117$]. Terminal regret is also unresolved. By contrast, the best available time-mean cost worsens by $+0.0818$ [$+0.0460$, $+0.1187$], about 14.0% of its reference-set mean. This is the only secondary outcome surviving Holm correction over six endpoints. The selected-cost change, $\Delta L = +0.043$, combines that worse opportunity with the unresolved regret change. The set's worse optimum is a manipulation check on what actions it supplies, not evidence of model-attributable harm.

Both model metrics meet the prespecified equivalence criteria (Table 4). The gain tolerance ± 0.05 exceeds the controlled Push-T estimate; the actual 90% upper bound, 0.0204, is more informative about the observed scale. The log-error interval and tolerance are reported alongside gain; the tolerances were pilot-informed and have no task-performance calibration (Appendix H). Regret was not tested for equivalence. The families were unseen by this analysis, but their public split contributed to checkpoint normalization; they are not a clean held-out sample.

When zero regret carries no discrimination The same rotation construction in Push-T yields constant coverage cost in 48 of 64 candidate sets, paired within 24 of 32 families. Every selector has $R = 0$ within these sets. Among their candidates, 364 record contact events and block motion: interaction alone does not ensure that coverage distinguishes actions. The rotation uses arena-wall clearance and does not explicitly encourage block contact. This simulator-only diagnostic complements the exclusion of constant-cost pairs from rank correlation because that correlation is undefined (Wang et al., 2026); it is not an additional model-quality result. Wall and Drawer pilots use different candidate constructions and remain supplementary scope probes (Appendix I.3).

Extensions to released planner proposals Two additional Push-T studies use released JEPA-WM and DINO-WM checkpoints with CEM proposals. A scale study over 21 episodes finds that halving the initial CEM proposal scale lowers success from 45.2% to 26.2% for JEPA-WM and from 50.0% to 28.6% for DINO-WM, without resolving the prespecified change in their between-model success margin. A crossed-proposal study over 24 starts from 15 episodes also leaves both margin contrasts unresolved under its specified start-weighted estimand. These extensions connect the question to an operational planner but do not establish harmful model selection. Their designs, all primary outcomes, and the aggregation correction for the latter are in Appendices J.2 and J.3.

7 A Reproducible Audit of Candidate-Conditioned Results

The audit takes per-candidate records and returns paired family summaries, uncertainty, denominator diagnostics, and task-cost discrimination checks. Its value is to make the interpretation accompanying a score explicit. Table 5 shows the distinctions supported by the experiments.

Table 6 specifies the executable procedure. Metadata matching belongs to the experiment runner; the scalar API checks numerical records. Shared-encoder sign-flip inference and episode-clustered planner analyses are separate from the generic audit's historical bootstrap-tail output.

The primary gain needs bank-mean $B > 0$. Because the current API also computes candidate ratios, it requires every candidate reference to be positive; zero references and missing costs are rejected, rather than replaced. Positive references are never clipped. Each set can have a constant cost while the two constants differ: then $\Delta R = 0$ but $\Delta O = \Delta L$ can be nonzero. These cases are distinguished explicitly. Appendix E maps each result to its records, commands, dependencies, and validation scope.

The earlier planner records illustrate why denominator flags matter: $M = 7.826$ and $B = 2.63 \times 10^{-8}$ yield a gain near -2.98×10^{8}, whose scale is dominated by the reference. Those records also change query states, so they diagnose the scoring rule rather than a controlled action-only intervention (Appendix J.1).

Table 5: **What joint reporting changes in the interpretation.** Each row combines quantities from the same matched comparison. Unresolved contrasts remain unresolved.

Observed pattern	Supported interpretation
Push-T: $\Delta G = +0.100$, M rises 11.6%	Higher reference-relative gain accompanies worse unnormalized prediction error.
Shared encoder: all six advantages over LD6 contract	Candidate composition changes measured improvement margins even with an unchanged winner.
PointMaze: $\Delta L = +0.043$, $\Delta O = +0.082$, $\Delta R = -0.039$	The resolved change is worse available opportunity; selected-cost and regret intervals include zero.
Push-T rotation: 48/64 sets have constant costs	Zero regret on these sets provides no discrimination between selectors.

Table 6: **Paired audit: inputs, operations, and checks.** The released scalar implementation starts from stored measurements; simulator and predictor execution remain external.

Step	Operation
1	Caller fixes checkpoint, representation/history, goal, horizon and K. Supply family, set, candidate index, model/reference errors, and predicted/realized costs. API checks pairing, equal counts and finite values; it does not verify checkpoint or goal metadata.
2	Average candidate errors to M, B; form $G = 1 - M/B$. Keep the candidate-mean ratio as a separately named sensitivity.
3	Select the first predicted-cost minimum in index order; return O, L, and $R = L - O$. A selected rollout alone supplies L, with O, R unavailable.
4	Pair banks within families; preserve prescribed seed/start aggregation and resample the independent unit. Use episode clusters when starts share episodes.
5	Report paired intervals, declared equivalence tolerances, and the full inferential family. Apply the stated study-specific test and multiplicity correction.
6	Report denominator minima/quantiles and small-positive counts, cost ties and constant sets, and the residual of $\Delta L = \Delta O + \Delta R$. Keep flagged observations in the estimates.

8 Limitations and Conclusion

Scope The strongest score and margin results are from Push-T. The controlled predictor uses frozen V-JEPA 2 features and is not official V-JEPA 2-AC; its joint detection depends on horizon and seed aggregation. Shared-encoder training configurations differ in several factors, and their gap analysis is post hoc. The cross-environment constructions address different questions and do not constitute repeated tests of one universal intervention. PointMaze does not reproduce the pilot regret effect. The original 32-family closed-loop cohort was stopped after inspecting results, before its planned 128 families, and supports no utility claim. Released-planner extensions also leave model-margin changes unresolved. Input-resolution checks explain the cross-pipeline reference discrepancy but do not replace independent full-inference validation.

Conclusion Candidate composition changes both the prediction problem and the task opportunities against which a world model is evaluated. In paired Push-T comparisons, model-comparison margins change within a shared representation, and relative gain increases despite higher prediction error. The raw-error contrasts show that comparison sensitivity extends beyond the selected normalization. Prospective and simulator diagnostics delimit the implications for selection: worse available actions, unresolved regret changes, and non-discriminating costs require different conclusions. Reporting model error, reference difficulty, task opportunity, and within-set excess cost together makes those distinctions inspectable. The audit turns this conditional interpretation into a reproducible evaluation procedure.

Reproducibility Statement

The supplement provides manuscript source, scalar records, audit code and tests, figure builders, and exact accounting and planner reanalysis. Its coverage map distinguishes recomputable quantities, frozen summaries, and external dependencies. Scalar checks verify aggregation, not simulator or full-predictor outputs. Large observations and checkpoints are external. The appendices retain protocol status, input-trace scope, and unresolved outcomes.

LLM Use Statement

LLM assistance was used for literature retrieval, for grammar and writing polish, and for code implementation and figure visualization. All other contributions, including the experiment design, the interpretation of results and the research questions, are our own.

References

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html.

Marcin Andrychowicz, Anton Raichuk, Piotr Stańczyk, Manu Orsini, Sertan Girgin, Raphaël Marinier, Léonard Hussenot, Matthieu Geist, Olivier Pietquin, Marcin Michalski, Sylvain Gelly, and Olivier Bachem. What matters for on-policy deep actor-critic methods? a large-scale study. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nIAxjsniDzg.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985. arXiv preprint.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkxLXnAcFQ.

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. URL https://www.roboticsproceedings.org/rss19/p026.html.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

John Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019. URL https://www.jmlr.org/papers/v20/17-750.html.

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 *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1etN1rtPB.

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

Proceedings of the 5th Conference on Robot Learning, volume 164 of *Proceedings of Machine Learning Research*, pp. 158–168, 2022. URL https://proceedings.mlr.press/v164/florence22a.html.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning, 2020. URL https://arxiv.org/abs/2004.07219.

Raj Ghugare, Homanga Bharadhwaj, Benjamin Eysenbach, Sergey Levine, and Ruslan Salakhutdinov. Simplifying model-based RL: Learning representations, latent-space models, and policies with one objective. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=MQcmfgRxf7a.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1lOTC4tDS.

Thomas M. Hamill and Josip Juras. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, 132(621C): 2905–2923, 2006. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/qj.06.25.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Balázs Kégl, Gabriel Hurtado, and Albert Thomas. Model-based micro-data reinforcement learning: What are the crucial model properties and which model to choose? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=p5uylG94S68.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 761–770, 2020. URL https://proceedings.mlr.press/v120/lambert20a.html.

William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/pdf/277fb50ed09a36b54842b8129c0f9027fd7edc53.pdf.

Michael Lutter, Leonard Hasenclever, Arunkumar Byravan, Gabriel Dulac-Arnold, Piotr Trochim, Nicolas Heess, Josh Merel, and Yuval Tassa. Learning dynamics models for model predictive agents, 2021. URL https://arxiv.org/abs/2109.14311. arXiv preprint arXiv:2109.14311.

Marion P. Mittermaier. The potential impact of using persistence as a reference forecast on perceived forecast skill. *Weather and Forecasting*, 23(5):1022–1031, 2008. URL https://journals.ametsoc.org/view/journals/wefo/23/5/2008waf2007037_1.xml.

Allan H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 116(12):2417–2424, 1988. doi: 10.1175/1520-0493(1988)116⟨2417:SSBOTM⟩2.0.CO;2. URL https://journals.ametsoc.org/view/journals/mwre/116/12/1520-0493_1988_116_2417_ssbotm_2_0_co_2.xml.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

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. URL https://openreview.net/forum?id=cHZn5Gdh8e. arXiv:2512.24497.

Claas A. Voelcker, Victor Liao, Animesh Garg, and Amir-massoud Farahmand. Value gradient weighted model-based reinforcement learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=4-D6CZkRXxI.

Jiawei Wang, Ke Rui, Yushen Zuo, Yichun Feng, and Minglei Li. Decision-metric alignment in latent world models: Diagnostics and action-conditioned objectives for MPC planning, 2026. URL https://arxiv.org/abs/2608.18746. arXiv preprint.

Anderson M. Winkler, Gerard R. Ridgway, Matthew A. Webster, Stephen M. Smith, and Thomas E. Nichols. Permutation inference for the general linear model. *NeuroImage*, 92:381–397, 2014. doi: 10.1016/j.neuroimage.2014.01.060.

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, 2026. URL https://arxiv.org/abs/2607.10362. arXiv preprint.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pp. 1094–1100, 2020. URL https://proceedings.mlr.press/v100/yu20a.html.

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

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

A EVIDENCE, COHORTS, AND PROTOCOL STATUS

The supplementary analyses document construction details, conditional results, and validation limits for the main claims. Prediction–control mismatch and constant-cost diagnostics are established in the related work (Section 2). The additions here concern matched candidate interventions and the joint interpretation of prediction, reference, and task quantities.

Table 7 groups analyses by inferential status. Detailed outcomes are reported once in the relevant results tables; negative and unresolved outcomes retain their original status.

Table 7: Evidence inventory. Specifying a reanalysis of inspected data in advance differs from fixing a new cohort before its outcomes exist; a selector is fixed before its test stream. "Preregistered" here means the analysis plan was written down and fixed before the outcomes were computed, in records kept with the project. These are not deposits in an external registry and carry no third-party timestamp.

Study	Inferential status	Boundary of the evidence
Controlled PA8, 16 families	Reanalysis of previously inspected development data under a protocol fixed in advance; task ranking is secondary.	Joint detection at $H = 16$ for the seed average and one seed; shorter-horizon and separate forecasting checks limit generality.
Reserved 29 families	Preregistered, disjoint from development data, evaluated once.	The gain and ranking-gap claims pass; this does not test an independent candidate-bank swap.
Selection cohort, 64 families	Fixed cohort and scoring; between-set model comparisons and the LD6 focus are post hoc.	All 21 shared-encoder contrasts remain the Holm family.
Test stream, 32 of 128 planned	Outcome-informed stopping; final-cost comparison and continuous replay are exploratory.	No resolved final-cost effect; replay metrics are post-outcome diagnostics.
Cross-environment pilots	Eight families per environment; generators fixed in advance but different; exploratory, on public data.	Drawer actions are demonstration-conditioned. Pilot aggregation checks are post hoc.
PointMaze confirmation	Preregistered with fixed $n = 32$; margins chosen with the pilot visible; one pre-outcome correction to the render-acceptance rule.	Primary regret claim fails; model-metric equivalence holds within the margins; only time-mean optimum survives secondary Holm.
Wall / Push-T rotations	Cohorts and generators fixed in advance; no registered hypothesis or equivalence margin.	Wall opportunity worsens; Push-T cost discrimination is limited. Push-T's original-checkpoint model endpoints remain provisional.
Released-planner scale study, 42 windows	Endpoint plan fixed before the main outcomes; smoke exposure and a wrapper amendment after a viewed failed run are recorded (Appendix J.2); not a reserved cohort: the windows are the authors' public validation split.	Primary margin-change endpoint fails; level change at $\sigma_0 = 0.5$ detected for both models; released-scale margin sign not reproduced; 21 episodes cannot resolve a margin of the published size; sample and aggregation differ from the published evaluation.
Original DINO-WM rescoring	Post-hoc scoring in a different input/representation pipeline.	Input tracing explains the reference discrepancy; no independent full-prediction parity. These conditional records supply no primary evidence.
Crossed CEM populations	Two checkpoints on three fixed sets, 24 starts from 15 episodes.	Both primary contrasts unresolved under the specified start-weighted estimand; original analysis weighting corrected after results; some secondary endpoints unavailable.

Cohorts and exclusions The controlled study uses 16 families, two starts each and 2,176 candidates per bank; the reserved study uses 29 disjoint families, 58 starts and 3,944 candidates on the rich bank. Their estimates are not pooled. Model inference on the reserved cohort was executed once. Its frozen outputs are reused only for explicitly labelled post-hoc sensitivity analyses, not as new confirmatory evaluations. Ordinary error magnitudes use the PA8 checkpoint at 8,000 updates ($M = 3.3793$); an earlier table at 6,000 updates gives 3.5980 because it uses different weights, and a comparison across those weights is excluded from the model-dependent evidence, as is an earlier cross-architecture run whose terminal states did not match the simulator.

Selection and test lineage Selection uses 64 roots, 67 candidates per bank, and 96 proposal groups of four candidates (two proposers, three repeats, 16 replanning states). The ordinary, rich and proposal selectors were fixed as DINO-WM, DINO-WM and MS8 before the test stream was opened. The stream was stopped after outcomes from the first 32 of its 128 predetermined families had been seen, without selecting families by outcome. Its closed-loop and replay results are retained in Table 32.

Chronology The inventory distinguishes analyses specified before their outcomes from post-hoc analyses. The gain accounting and corrected CEM aggregation in this revision use existing outcomes and are explicitly post hoc. PointMaze's one pre-outcome correction is described in Appendix H; Drawer Open's horizon followed a model-free diagnostic.

B Models and Representations

The controlled goal is the logged continuation's final block pose; the selection and test cohorts use the environment target. Table 8 collects notation.

Table 8: Notation for lookup. Core metrics are defined in Section 3; the reserved cohort's ranking decomposition is defined in Appendix F.4.

Symbol	Meaning
s, u, a, H	Start state, pusher control, control sequence and horizon. A Push-T control lasts 0.1 seconds; pilot model steps contain five controls.
$\Phi(s, a), g(s)$	Realized endpoint and goal: logged endpoint in the controlled study, environment target in the model comparison.
$\mathcal{C}(s), K, N$	Candidate bank, candidates per bank and independent trajectory-family count (Table 1).
$\phi, z, \hat{z}, z^\star$	Frozen encoder, current representation, predicted future and encoded realized future.
$\varepsilon_\theta, \varepsilon_0; M, B, G$	Candidate model/reference squared errors; bank-mean errors and persistence-relative gain.
$\hat{J}, J^\star, J_{\text{task}}$	Predicted latent goal cost, realized latent goal cost and realized task cost.
$O, L, R = L - O$	Best available cost, selected cost and within-set excess cost.
$\rho_{\text{dyn}}, \rho_{\text{task}}$	Within-bank Spearman agreement of predicted costs with realized latent and task costs.
$A, B, C; \Gamma_{\text{contract}}, \Gamma_{\text{pred}}$	In Appendix F.4 only: task-ranking correlations for pose, realized-latent and predicted-latent costs; their differences $A - B$ and $B - C$.

The internal predictor packs two physical controls per model step. The main protocol table reports horizons in physical controls; predictor lead h counts model steps. Table 9 identifies the fixed PA8 predictor; it is not a complete training recipe.

Checkpoints bind data, configuration, optimizer and RNG states. Bitwise reproduction also requires the corresponding runtime and pipeline; evaluation requires weights, cached inputs and scoring code. Table 10 lists the training configurations.

All seven configurations share the architecture and frozen representation. Their data, loss and duration vary as specified in Table 10: CF6 adds a 0.3-weighted paired-difference loss, while MS8 follows PA8's BD6 starting weights and input stream with a fixed low-frequency auxiliary loss on every batch. DINO-WM reproduction denotes the JEPA-WMs release, evaluated with three paired

Table 9: PA8, the predictor in the controlled and reserved-cohort studies, identified by configuration and checkpoint; Section 3 describes its use.

field	value
architecture	continuous transformer predictor in direct mode with causal attention masks: a future query at lead h attends context and action tokens of steps $\leq h$ only, so no information from later actions or leads can reach it
parameters	15,047,040 per seed
latent	$d = 64 \times 384$: last tubelet of a frozen V-JEPA 2 ViT-L/16 encoder, FP32, TF32 off, past-only 16-frame left-padded clip, resized and centre-cropped by the official preprocessor from the 224×224 render to the encoder's 256×256 input, so the last tubelet carries $16 \times 16 = 256$ tokens; pooled to an 8×8 grid, then a frozen PCA $1024 \rightarrow 384$ fitted once and never refitted
objective	absolute latent prediction error only, a masked squared error against the future representation
optimizer	AdamW, weight decay 0.01, constant learning rate 3×10^{-4}
updates	8,000, continued for 2,000 updates from BD6 at U6000
training batches	1,000 logged-window batches and 1,000 same-state branch-pair batches; the latter use a predeclared 50/50 old/expanded-palette sampling rule, realized as 513/487. Expanded branches use predefined planner-distribution action templates
seeds	0, 1, 2, all three evaluated, none selected

Table 10: Model identity across studies. All seven internal labels denote training configurations of one predictor architecture on the same fixed encoder, three seeds each; the label gives the training data or objective and the suffix the number of updates in thousands. PA8 is the controlled-study model. The two DINO-WM checkpoints are distinct public releases.

Label	Study use	Training configuration
PA8	**Controlled 16; reserved 29**; shared-encoder 64	Branched training with predefined planner-distribution action templates, 8,000 updates.
LD6	Shared-encoder 64	Logged-window MSE training, 6,000 updates.
BD6	Shared-encoder 64	Logged and same-state branch MSE, 6,000 updates.
CF6	Shared-encoder 64	BD6 data stream plus a paired-difference loss, 6,000 updates.
BA7 / BA8	Shared-encoder 64	Continuation of BD6 with the original branch palette, 7,000 and 8,000 updates.
MS8	Shared-encoder 64; test-stream selector	Same BD6 start and stream as PA8, with an auxiliary spatial-scale loss, 8,000 updates.
DINO-WM reproduction	Public Push-T 64; test stream	The DINO-WM Push-T checkpoint released with JEPA-WMs, run through that release's interface.
DINO-WM original	Supplementary Push-T 64; rotation 32	The original authors' Push-T release; its model-dependent results are provisional (Appendix E.2).

Public navigation and Drawer checkpoints are environment-specific (Tables 1 and 30).

planning schedules rather than training seeds; the original authors' checkpoint remains separately identified.

History and representation fitting For the internal models, z_t abbreviates a history-dependent embedding: the encoder receives frames $\max(0, t-15), \ldots, t$, repeating frame 0 on the left until the clip contains 16 frames. No frame after t enters that clip. We retain the last temporal tubelet's 16×16 spatial grid, pool it to 8×8 tokens, and apply a fixed mean-centered PCA projection from 1024 to 384 dimensions. The mean was fitted from 446,464 pooled tokens from the public Push-T training episodes `train/000`–`train/063`. The PCA basis used a seed-0 subsample of 120,000 of those tokens and `torch.pca_lowrank` with 384 components and four iterations; its recorded retained variance fraction is 0.9871. Calibration and evaluation episodes did not fit the PCA. This basis was subsequently reused for simulator-rendered histories without refitting. Image normalization uses the upstream evaluation transform's fixed ImageNet statistics; it is distinct from fitting the PCA mean. The historical extraction runtime was Python 3.12, PyTorch 2.7.0+cu128, FP32 with TF32 disabled, on an RTX 4090.

`manifests/representation_manifest.json` records the fitting split, source hashes, and runtime.

C Candidate construction and task coordinates

C.1 Push-T: ordinary and geometry-enriched banks

The controlled study uses $K = 68$ action sequences at each start, each of length 16 physical controls. The independent selection and test streams use $K = 67$. The selection cohort scores the first ten controls of these sequences; the matched test-bank contrast uses sixteen. The latter remove exactly the logged-continuation entry from each 68-entry recipe, without a replacement or a change to any other quota. In both banks, the first entry is hold. Entries retain their construction order and multiplicity: there is no deduplication, including when the two toward-landmark magnitudes coincide.

Coordinates and physical controls. Let $u_k \in [-1, 1]^2$ denote the predictor's normalized action. The environment receives $a_k = \sigma \odot u_k + \mu$, where $\sigma = (0.2019, 0.2002)$ and $\mu = (-0.0087, 0.0068)$, and sets the pusher target to $p_k + 100a_k$. The target stays fixed for ten physics steps of length 0.01 seconds:

$$v \leftarrow v + 0.01[100(p_k + 100a_k - p) - 20v], \qquad p \leftarrow p + 0.01v.$$

Thus one physical control lasts 0.1 seconds; $H = 16$ physical controls last 1.6 seconds and correspond to eight steps of the internal predictor. A target displacement d in pixels is encoded as $u(d) = \mathrm{clip}_{[-1,1]^2}((d/100 - \mu) \oslash \sigma)$. Hold is $u(0)$, not normalized zero. Generated arrays are rounded to six decimal places and stored as float64; the simulator receives the affine transform above. The environment itself does not add normalized-action clipping.

Geometry-enriched recipe. Let b be the initial block position, p the pusher position and $\beta = (b - p)/\max(\|b - p\|, 10^{-12})$. For a direction d, a constant normalized sequence is $md/\max(\|d\|, 10^{-12})$, clipped componentwise to $[-1, 1]$. The initial block footprint consists of the local quadrilaterals $[(-60, 30), (60, 30), (60, 0), (-60, 0)]$ and $[(-15, 30), (-15, 120), (15, 120), (15, 30)]$, rotated by the initial block angle and translated by b. Its six local landmarks are $(0, 30), (-45, 15), (45, 15), (0, 105), (0, 15), (0, 75)$. Nominal contact means that the free-space PD path comes within the 15-pixel pusher radius of this fixed footprint. It uses no collision-resolved future.

Only the toward-landmark component searches a magnitude ladder; it takes the first passing value. The other geometry components use the enumerated constants. No candidate is rejected on predicted or realized quality, and geometry failures do not trigger resampling. Nominal engagement, penetration and dwell are recorded diagnostics rather than extra acceptance filters. The five intended no-contact controls are design labels, not a fixed realized no-contact count. The controlled rollout protocol permits at most two retries of an execution exception with identical inputs, then records the

Table 11: Ordinary Push-T recipe. Counts are 68 with, and 67 without, the logged continuation. $D = 19.03876248493026$ pixels is a target displacement per physical control, rather than realized pusher motion.

Component	Count	Exact rule
Anchor palette	6 / 5	Hold; constant $\pm x, \pm y$ commands of target magnitude D; plus the root-specific logged 16-control continuation only in the 68-entry bank.
Diagonal	4	Constant target vectors $(\pm D/\sqrt{2}, \pm D/\sqrt{2})$, all four signs.
Magnitude	8	Each cardinal target at $0.5D$ and $1.5D$, with componentwise normalized clipping.
Temporal	8	Four $8 + 8$ turns $(+x, +y), (+y, -x), (-x, -y), (-y, +x)$; two sequences alternating $(+x, +y)$ or $(-x, -y)$ every control; two $+x$ or $+y$ ramps with magnitude $D \sin[\pi(k + 1/2)/16]$, $k = 0, \ldots, 15$.
Initial proposals	42	Independent 16×2 standard-normal arrays, componentwise clipped to $[-1, 1]$; NumPy `default_rng` seed 4990, drawn once and reused at every start.

Table 12: Geometry-enriched Push-T recipe. These are construction rules, not guarantees of realized contact. Magnitudes below are in normalized action units.

Component	Count	Exact rule
Anchor palette	6 / 5	The ordinary anchor entries, in the same order.
Toward-landmark	12	For each landmark, choose the first nominal-contact magnitude in $(0.25, 0.40, 0.55, 0.70, 0.85, 1.00)$, using 1 if none passes; emit that magnitude and $\min(1, m + 0.30)$.
Tangential	10	Bearing offsets $+60, -60, +75, -75, +90$ degrees from β, each at magnitudes $0.60, 1.00$. No dwell optimization is performed.
Cross-object	8	Aim at $b + 140\beta + c(-\beta_y, \beta_x)$ for $c \in \{-60, -25, 25, 60\}$ pixels, at magnitudes $0.80, 1.00$.
Magnitude sweep	10	Aim at b at magnitudes $0.10, 0.20, \ldots, 1.00$.
Direction sweep	12	Magnitude 0.80 and bearing offsets $0, 30, \ldots, 330$ degrees.
Temporal	6	Approach b at magnitude 0.80 for eight controls then rotate the bearing by $+90, -90, 180$ degrees for eight; also three curves rotating uniformly over sixteen controls by $+45, -45, +90$ degrees at magnitude 0.80.
Initial proposals	4	The first four already frozen ordinary Gaussian proposal arrays.

candidate as failed without replacement; this technical failure rule is separate from the construction rule.

Inputs, goals and pairing. The geometry recipe reads initial pusher position and velocity and block pose; it does not read a goal, model output or candidate outcome. The controlled bank retains the pre-existing logged continuation, which also defines the goal observation and block pose at its endpoint. This anchor is an explicit exception to the outcome-free geometry construction, not a generated proposal. For the independent streams it is removed and the goal is the environment's fixed target pose. Their starts are obtained by seeded reset followed by four physical zero controls, and the stored five-frame history is shared across banks. The internal predictor uses context times $t-4, t-2, t$; the public reproduction uses its native context. All candidates at a start use the same checkpoint, history, goal and physical horizon; the bank need not pair individual candidates. Fresh replay verifies exact equality of the root snapshot and stored history. Predicted-cost selection uses NumPy's first-index argmin for exact ties, so hold wins an exact tie involving its first entry. This is distinct from tolerance-based ties in pairwise task-ranking statistics.

C.2 Matched rigid rotations across environments

The PointMaze pilot and prospective cohort, the Wall rotation cohort, and the Push-T rotation cohort use the same $K = 64$, five-macro-action generator. A scrambled 10-dimensional Sobol sequence supplies pairs (q_1, q_2); a macro action is $\sqrt{q_1}(\cos 2\pi q_2, \sin 2\pi q_2)$ and is repeated for five physical controls. For each candidate, the alternative applies one rigid rotation to the whole sequence, chosen from angles $2\pi j/16$, $j = 0, \ldots, 15$. For each rotation, a nominal path starts at the initial agent position and accumulates macro actions multiplied by the environment scale in Table 13. Each segment is probed at 33 equally spaced locations, including its endpoints. The chosen rotation maximizes the minimum signed clearance to the static rectangular obstacles over all probes; the smallest rotation index wins a tie. No candidate is discarded or deduplicated. This rule uses position, static geometry and the frozen seed; it reads no goal, history, model output or candidate outcome.

Table 13: Environment-specific parameters of the matched rotation generator. All use 25 physical controls. Distance units and Push-T's affine action mapping are retained explicitly rather than treated as a shared physical scale.

	PointMaze	Wall rotation	Push-T rotation
Nominal scale per macro	0.25 qpos units	10 image-coordinate units	40.4931 pixels, from stored free-space calibration
Obstacle geometry	Official U-maze walls, radius-0 center-point convention	Slabs $[29, 4, 35, 26]$ and $[29, 34, 35, 60]$	Arena segments at $5, 506$, expanded by wall radius 2 plus pusher radius 15
Goal	Public episode state after its first 25 controls, accepted by replay gate	$(\mathrm{clip}(64 - x, 8, 56), \mathrm{clip}(y, 8, 56))$	Fixed environment target pose
Bank seed	$560001+i$ (pilot); $570001+i$ (prospective)	$580001 + i$	$620001 + i$
Task cost	Radius-0 geodesic distance in qpos units	Door-aware center-point geodesic distance in image-coordinate units	Terminal $1 - \mathrm{coverage}$, dimensionless

The two navigation studies evaluate $J_{\mathrm{mean}} = 25^{-1} \sum_{t=1}^{25} d(s_t, g)$, at every post-control state, excluding $t = 0$; terminal cost uses $t = 25$. The selector minimizes predicted terminal latent-goal cost and is then evaluated with either realized cost. Thus time-mean evaluation does not change the selector's objective. The Push-T rotation study uses terminal coverage cost, not a time-mean cost. Rotation preserves action-vector norms before native preprocessing. These vectors are raw environment actions in PointMaze and Wall but normalized u in Push-T; the affine, anisotropic Push-T transform above does not preserve physical-action norms exactly. In Push-T only arena boundaries affect clearance: the movable T block is not an obstacle, so the rule does not favor block contact or goal progress.

C.3 Environment-specific pilots

The earlier Wall pilot uses a different intervention from rigid rotation. For each of its 64 candidates and 25 controls, paired magnitudes are drawn from $N(1, 0.4^2)$ truncated by rejection to $[0.2, 1.8]$. Ordinary directions start uniformly and accumulate von Mises increments with concentration 5. Geometry directions steer an analytic position proxy through a door lane $y \sim U[27, 33]$ to a point just beyond the opposite wall face, then toward a uniform point in the opposite room. The proxy updates by twice the action and clips its position to $[8, 56]^2$. Only directions change; paired raw-action magnitudes agree. The frozen generator seed and exact draw order are in the released recipe. Its door-aware distance is averaged over the 25 post-control states, with $t = 0$ excluded.

Drawer Open uses 16 candidates over 100 physical controls (20 model steps). At each seeded root, the official expert produces a fixed 100-control prefix. The ordinary set randomizes each expert XYZ action direction while preserving its norm and gripper control. The alternative includes the exact expert as candidate 0; candidate $b > 0$ perturbs the expert unit direction by isotropic Gaussian noise of standard deviation $0.03 + 0.02b$, renormalizes to the expert XYZ norm, and retains its gripper control. Directions outside the action box are rejected and redrawn; the frozen RNG is `default_rng`(root seed $+400000$). This is demonstration-conditioned construction: the expert prefix contains task information even though candidate scoring and model outputs are unavailable at construction. The latent goal is the fixed expert endpoint; realized drawer cost is $|q_{\text{drawer}} - q_{\text{goal}}|$ in simulator length units, reported at control 100 and averaged over post-control times $1, \ldots, 100$. The model selects by terminal latent-goal cost, with first-index tie-breaking. These pilot constructions are not replications of one common intervention.

D Metrics, Aggregation, and Inference

For the controlled PA8 study, ε_θ and ε_0 are means over the $d = 64 \times 384$ coordinates of the squared difference between predicted and realized representation at the horizon, per candidate. They are computed separately for each of the three seeds and then averaged over seeds; rank statistics are computed once on the seed-averaged goal cost $\hat{J}$, so they are not an average of per-seed correlations. Task pairs differing by at most 10^{-12} are discarded as ties; representation-cost ties are exact-equality only. The bootstrap resamples trajectory families, not candidates, with 2000 draws, a fixed seed and one shared index vector across all contrasts, so paired quantities share resampling noise. Public-checkpoint dimensions follow their encoder token grids; the prospective and planner studies use their separately specified resampling schemes. The selection cohort and test stream use the same estimators in each predictor's own representation space, with the per-family gain formed as in (2) at the family's single start state, and family bootstraps with a fixed seed; the replayed closed-loop endpoints use a second fixed seed.

Selection and ties Predicted-cost minimization uses the first exact minimum in stored candidate order. This rule is separate from excluding task-cost pairs within the stated tolerance in ranking diagnostics. For the shared-encoder comparisons, the representative inferential family and sign-flip assumptions are specified in Appendix G; bootstrap intervals remain nominal paired intervals. Other studies retain their own recorded procedures and status.

Selection ties The scalar audit and prospective PointMaze reaggregation sort candidates by integer index and apply `numpy.argmin`; predicted-cost ties therefore select the lowest-index candidate. Where a bank places hold first, hold wins an exact predicted-cost tie. This is separate from the average ranks used for Spearman correlation and from the omission of tied task pairs in pairwise agreement. The crossed-population runner likewise applies `numpy.argmin` in stored population order. Native CEM refinement uses the upstream `torch.topk` elite operation, whose tied-elite ordering is not overridden by this audit.

E ARTIFACTS AND VALIDATION

E.1 ARTIFACT COVERAGE

The supplied coverage inventory distinguishes scalar records that can be recomputed, frozen summary outputs, display-only figures, and external simulation or inference dependencies. The principal paired estimates and the new accounting and planner reanalyses are reproducible from the included records. These checks establish consistency of numerical aggregation; they do not independently validate the simulation or learned inference that produced the records.

E.2 ARTIFACT ENTRY POINTS AND VALIDATION SCOPE

The anonymous archive's entry points are `verify_results.py` and `verify_revision4.py`; `README.md` and `RESULTS_COVERAGE.md` specify the supported computations. The distinction between measurement generation and reaggregation is explicit in Table 14. Original experiment source is supplied under `source/` for inspection, with neutral external-asset paths. Running that source requires the external checkpoint, dataset, upstream repository, and simulator identified by its protocol. The portable checks use author-derived records and execute none of those components.

Table 14: Validation layers and their scope. Hash identity establishes provenance, not numerical inference parity.

Layer	Evidence checked	Scope boundary
Stored scalar aggregation	Controlled means, shared-encoder gaps/intervals, public reproduction accounting, PointMaze endpoints, and selected planner contrasts recompute from supplied records.	This does not validate the predictions and simulator outputs that originally produced the records.
Encoder input/reference trace	One family, four candidate endpoints: matching raw-frame hashes, native $196/224$ inputs, and persistence values within 1.1×10^{-6} of recorded values.	No full-predictor history, action-packing, or rollout parity comparison; no assertion of coverage over all families.
Simulator-derived diagnostics	Supplied rotation candidate costs reproduce 48/64 constant banks; saved planner records satisfy $J = 1 - \text{coverage}$.	The portable commands do not replay the simulator. Contact callbacks do not independently identify all body pairs.
Native full inference	Loading and inference follow each checkpoint's native path; the inspected conventions and trace are reported separately.	An independent end-to-end output comparison remains incomplete. Neither scalar checks nor the limited trace substitute for it.

Commands and expected outputs Run from the extracted archive with Python 3.10 or later and NumPy; SciPy is needed for the shared-encoder inference verifier, and Matplotlib is needed only for figures. The recorded scalar run used Python 3.12.4 and NumPy 2.2.6. The following commands write only verification products; use an external output directory when checking an unchanged archive's hash manifest.

```
python -m unittest discover -s tests -v
python verify_results.py --output scalars.json
python verify_revision4.py --output v4.json
python verify_shared_inference.py --output inference.json
```

The recorded execution returned nine passing boundary tests, 1,025 scalar checks, and 32 revision-4 numerical checks plus stored-trace/candidate-shape checks, respectively. The shared-encoder verifier reproduces all 21 pairs with maximum difference zero from the stored sign-flip and t results. The JSON outputs include per-check residuals. `validation/RUN_LOG.json` records the actual commands and exit codes; `requirements.txt` names dependencies. These recorded check counts describe the archive version, not an independent replication claim.

The optional commands `pythonfigures/build_paired_figure.py--outputpaired_results.pdf` and `pythonfigures/build_score_`

Table 15: Display-to-record map for numerical checks. All listed paths begin at the archive root. V1 denotes `verify_results.py`; V2 denotes `verify_revision4.py`. The commands above run all entries for the selected script.

Display / result	Supplied record and aggregation entry	Expected output and dependency boundary
Controlled contrasts, Table 3; complete endpoints	`data/controlled/family_statistics.csv`, paired differences, selection/regret summaries; V1 `controlled()`	Mean differences and $\Delta L = \Delta O + \Delta R$; original bootstrap draw order is supplied as a result, not regenerated.
Shared-encoder levels, gaps, raw-error panel, all 21 comparisons	`data/shared_encoder/family_G_M_B.tsv`, `gap_summary.json`, `statistical_sensitivity.json`; V1 `shared_encoder()`	All score-gap intervals and six raw-error change intervals; `verify_shared_inference.py` additionally reproduces all sign-flip/t statistics and raw/adjusted p values.
Public reproduction and Figure 2	`data/reproduction/family_scores.csv`, `gain_accounting.json`; V1 `reproduction()`, V2 `reproduction_accounting()`	64-family levels and two-order/symmetric accounting intervals. Figure builder consumes the distributed summaries.
Prospective PointMaze, Tables 4 and 27	`data/pointmaze/candidates.csv`, `registered_results.json`; V1 `pointmaze()`	Eleven contrasts, two equivalence intervals, recorded six-endpoint adjustment; primary regret criterion remains unmet.
Push-T constant-cost diagnostic	`data/pusht_rotation/candidates.csv`, recorded audit; V1 `pusht_rotation()`	Six means and 48/64 constant banks; model-facing endpoints retain their validation qualification.
Encoder trace, Table 16	`data/pipeline_trace/original.npz`, `repro.npz`, stored trace check; V2 `pipeline_trace()`	Raw-input digest agreement, native input shapes, and four persistence errors per implementation. No new encoder call.
Proposal-scale study, Table 34	`data/proposal_scale/episode_results.jsonl`, protocol, original analysis; V2 `proposal_scale()`	252 rows, 42 starts/21 episodes, two primary contrasts and recorded adjustment. Upstream evaluation aggregation is not replicated.
Crossed CEM populations, Table 35	`data/planner_cross/rows.jsonl`, protocol, corrected analysis, `sets/*.npz`; V2 `planner_cross()`	Both weighting conventions, clustered intervals, exact cluster sign-flip sensitivity, 24 array checks. Full candidate costs and O, R are absent.
Remaining robustness, pilots, closed-loop displays	Final `paper_source/` tables/figures; selected summaries in `data/`	Supplied display evidence only where no record/script entry is listed. The archive does not regenerate every table from raw observations.

`accounting.py--outputscore_accounting.pdf` redraw figures from structured summaries. They require Matplotlib and do not estimate the underlying statistics. The manuscript's final display files are included separately.

Identity, exposure, and external assets `manifests/model_manifest.json` identifies each internal configuration, all three seed checkpoints, and both public DINO-WM releases by SHA-256. The reproduction checkpoint begins `8ec9cb05f228`; the original release begins `e909f0cec958`; the complete hashes are in the manifest. `manifests/seed_manifest.json` preserves the historical integer seeds, and `manifests/registration_manifest.json` records protocol order and hashes. `manifests/source_provenance.json` distinguishes original records from the anonymized copies; `MANIFEST.json` hashes the distributed files. Checkpoint weights, raw videos, and upstream runtime installations are external dependencies. PointMaze's public validation episodes contributed to checkpoint normalization statistics. This exposure is independent of whether our paired analysis was fixed before its outcomes were inspected.

Public reproduction: validation record and limits The main result uses the JEPA-WMs reproduction of DINO-WM, a different checkpoint from the original authors' release. The reproduction was run through its own loading and inference interface, and its conventions were read from the source: image mean and standard deviation 0.5, the last of five root observations as context, two successive five-control blocks, the final unroll output, and target frame 14 (zero-based). These conventions are mutually consistent on inspection; the later trace below checks encoder inputs and persistence values, while an independent full-predictor output comparison remains unavailable.

Recomputing all 8,576 ordinary and rich candidate rows reproduces every family score with zero residual. The family means are $G : 0.6812589 \to 0.7807226$, $M : 0.0573847 \to 0.0640176$ and $B : 0.1830384 \to 0.2943479$, with the intervals reported in the main text. These checks are conditional on the stored predictions. No stored independent official end-to-end comparison verifies input tensors, action packing, predicted endpoints and aligned simulator/encoded target indices; this validation remains needed. The cohort and native-gain selector were protocol-defined, but the bank-effect interpretation is a subsequent selection-cohort analysis.

Encoder-input verification An earlier control compared persistence errors between the original DINO-WM pipeline and the JEPA-WMs reproduction, requiring family-wise correlation at least 0.90 and a mean ratio within 25%. Four of 64 families missed the correlation criterion. Recomputed differences were present throughout the cohort: median relative disagreement across 8,576 candidates was 13.4%, with family medians 6.7–27.5%. This result did not identify a defective release; it showed that the assumed equivalence of our two reference computations required inspection.

On byte-identical stored frames, a forward pre-hook at the encoder patch embedding reveals different actual inputs (Table 16). The original world-model class resizes according to $(224/16) \times 14 = 196$, combining decoder scale 16 with encoder patch size 14. The reproduction encodes at 224. Both configuration files declare image size 224, so matching that field alone does not match the tensor entering the backbone. The traces explain why persistence distances computed over different token

Table 16: Stored input trace on one family and four candidate endpoints. The raw frames are identical; effective encoder input and token grids differ. This is an encoder/reference check.

Quantity	Original DINO-WM	JEPA-WMs reproduction
Raw frame size	224×224	224×224
Encoder input size	196×196	224×224
Tokens / dimension	$196/384$	$256/384$
Maximum absolute persistence discrepancy to stored values	8.64×10^{-7}	1.07×10^{-6}

grids need not be numerically equal. The comparison should therefore not serve as a cross-pipeline parity test. The check covers the encoder input and four persistence values, not full predictor histories, action packing, multi-step outputs, or every family. Original-checkpoint score records remain conditional supplementary results. The main public reproduction and shared-encoder contrasts com-

pare candidates within their own unchanged pipelines. Their end-to-end validation limit remains as stated in Appendix E.2.

E.3 Scalar API boundaries

The record-level procedure accepts environment-specific task costs through a common schema. Its input is one record per scored candidate, carrying the family, the query set, the model and reference errors, the realized task cost and the predicted cost; simulation and inference stay outside it. It reports the two error levels, the gain of (2) with the mean-candidate-ratio form as a labelled sensitivity, and paired $\Delta O, \Delta R, \Delta L$ under one shared family-bootstrap index. It rejects non-positive references and unequal candidate counts, verifies $\Delta L = \Delta O + \Delta R$ family by family, and keeps constant-cost sets in the means while naming them in the table notes. Equivalence is available only through declared margins and 90% interval containment.

Small positive references are never clipped or excluded: candidate-level and family–set denominators are summarized by their minima, quantiles and counts below descriptive thresholds ($10^{-12}, 10^{-9}, 10^{-6}$ in input units). An all-zero contrast yields a point interval and a two-sided tail summary of one; a nonzero constant contrast reaches the $1/B$ reporting floor, and this tail summary is not an exact randomization test. Recomputing the PointMaze primary, gain-equivalence and optimum estimates, the Push-T cost identity, and the rotation values and counts from the released records reproduces the printed values. That check does not validate simulation, encoding or inference. The portable PointMaze check also reaggregates log-error equivalence and the recorded six-secondary-endpoint Holm analysis. The Wall and Push-T rotation studies were analysed with this procedure.

Numerical and missing-data contract The ratio-of-means gain requires a positive bank mean B. The current scalar API is stricter: it rejects any nonpositive candidate reference error because it also computes candidate-level ratios. A zero candidate reference with positive bank mean therefore has a mathematically defined primary gain but is outside this API's accepted input; it is not silently clipped or discarded. Small positive values remain unchanged, with diagnostic thresholds $10^{-12}, 10^{-9}, 10^{-6}$ in that representation's error units. Missing realized costs are not accepted as zero. Experiments without all-candidate costs report unavailable O, R outside this complete-record API, as in Appendix J.3.

F Score Sensitivity and Robustness

These analyses use the released controlled-study and reserved-cohort records. They are post hoc unless a protocol-specified endpoint is explicitly identified. The separate prospective PointMaze study retains its own status (Appendix H). Statistical intervals below resample trajectory families; candidates and the three fixed predictor seeds are not independent samples.

F.1 Controlled comparison and verification

The complete controlled result is in Table 17. Checkpoint bytes are identical across query sets; model-independent quantities reproduce to 8.9×10^{-16} against a 10^{-6} tolerance. Removing the logged-continuation candidate changes no sign or interval-exclusion verdict. For the 288 candidates shared by the sets, action-matched predictions agree exactly across all three seeds: maximum latent and goal-cost differences are both 0.0. Thus the historical ordinary-set predictor discrepancy in Appendix A does not occur between the two sides of this comparison. The horizon replay also reproduces the $H = 16$ arm bitwise before reading shorter prefixes.

Seeds and readout horizon The joint criterion requires both the model-error and gain contrast intervals to have positive lower bounds. Both changes have the same sign at all three seeds, but only one individual seed meets this criterion; the seed average meets it. Per-seed family contrasts correlate 0.70–0.94, and the variance of their mean is 0.86 of the mean per-seed variance, rather than the 0.33 independent-seed value. Inference is conditional on these checkpoints. Across the five specified primary endpoints, the original Holm-adjusted bootstrap-tail summaries are 0.0025,

Table 17: Changing only the candidate set: each seed checkpoint is byte-identical on the two sides (SHA-256), start states, goal, horizon and candidate count are the same, and only the action content differs. Δ is intervention-rich minus ordinary, formed per start state, averaged within family, then bootstrapped over the 16 families (2000 resamples, percentile 95% interval). Bold marks an interval excluding zero. The primary and secondary blocks are the split fixed before these numbers were read.

Quantity	Ordinary	Rich	Δ	95% CI	Fam.>0
Primary, squared-error units					
model error $\mathbb{E}[\varepsilon_\theta]$	3.3793	3.6003	**+0.2210**	[+0.1054, +0.3529]	12/16
baseline error $\mathbb{E}[\varepsilon_0]$	6.1643	6.8034	**+0.6391**	[+0.3376, +0.9471]	13/16
Primary, dimensionless					
gain G	0.4369	0.4673	**+0.0304**	[+0.0069, +0.0544]	12/16
$\rho_{\mathrm{dyn}} = \rho(\hat{J}, J^\star)$	0.6506	0.6134	-0.0372	[-0.1026, +0.0309]	6/16
pair accuracy vs $J^\star$	0.7426	0.7342	-0.0084	[-0.0332, +0.0163]	8/16
Secondary, dimensionless					
$\rho(J^\star, J_{\mathrm{task}})$	0.4366	0.4655	+0.0289	[-0.0263, +0.0812]	10/16
$\rho_{\mathrm{task}} = \rho(\hat{J}, J_{\mathrm{task}})$	0.2785	0.1689	**-0.1096**	[-0.1861, -0.0362]	4/16
pair accuracy vs J_{task}	0.6111	0.5610	**-0.0500**	[-0.0852, -0.0160]	5/16

The first block is in representation-space squared-error units; the others are dimensionless. Fam.>0 counts trajectory families whose paired difference is positive. An interval containing zero records a non-detection, not equivalence.

0.0025, 0.0390, 0.5880, 0.5880: the two errors and gain retain their findings. These are approximate percentile-bootstrap tail summaries with a $1/2000$ floor, not exact null tests.

Table 18 uses deterministic prefixes of the same sixteen-control candidates, with states, goals, weights and renders fixed. The joint criterion was fixed before this follow-up computation, which remains post hoc relative to the study. It holds only at $H = 16$; $H = 8$ fails on gain, and $H = 4$ detects none of the three contrasts. This is a boundary within the tested protocol, not a universal horizon threshold or causal attribution to persistence.

Table 18: Matched readout horizons on the controlled cohort. Rich minus ordinary, with nominal paired 95% family-bootstrap intervals. Joint detection requires positive lower bounds for both ΔM and ΔG.

H	ΔM [95% CI]	ΔB [95% CI]	ΔG [95% CI]	Joint
16	$+0.2210$ [$+0.1054$, $+0.3529$]	$+0.6391$ [$+0.3376$, $+0.9471$]	$+0.0304$ [$+0.0069$, $+0.0544$]	yes
8	$+0.1239$ [$+0.0616$, $+0.1929$]	$+0.3018$ [$+0.0245$, $+0.5871$]	$+0.0176$ [-0.0085, $+0.0442$]	no
4	-0.0044 [-0.0380, $+0.0302$]	-0.0341 [-0.1871, $+0.1276$]	-0.0037 [-0.0199, $+0.0138$]	no

Forecasting boundary A separate multi-lead latent forecasting evaluation compares short and long episode regimes using the same checkpoint within each seed. It is not another physical environment. Its pre-specified criterion required co-increasing error and gain at a majority of leads in all three seeds. It failed: only lead 1 meets it, in 3 of 96 seed–lead cells, with error $5.85 \to 7.27$ and gain $-0.713 \to -0.146$. All other 93 cells have higher error and lower gain. Agreement with the reference-ratio condition is algebraic, not an independent validation of that condition. This check limits generalization across forecasting leads.

F.2 Reference weighting and alternative baselines

Using the relative-error definition of skill (Murphy, 1988) and its reference-dependent interpretation (Hamill & Juras, 2006), let $m_k, b_k > 0$, $g_k = 1 - m_k/b_k$ and w_k be stratum weights. Substitution of $m_k = b_k(1 - g_k)$ gives

$$G(s; w) = \frac{\sum_k w_k b_k g_k}{\sum_k w_k b_k}. \tag{6}$$

It is a convex combination of local gains. Moving weight between two strata raises both error and gain exactly when $b_2/b_1 > m_2/m_1 > 1$. Verification on the eight construction strata matches this condition for all 1,792 within-state ordered pairs, but only 54 of 56 pairs after pooling states first. The identity holds to 10^{-15}; it neither replaces the specified aggregation nor supplies evidence independent of the observed ratios.

The specified mean-of-state-ratios gain is $0.4369 \to 0.4673$; the pooled ratio gives $0.4518 \to 0.4708$. Both accompany the same $+0.2210$ model-error change. Alternative references show that the joint pattern is not universal (Table 19). The terminal-mean reference reads realized futures and is an oracle, not a deployable forecast; the predictor loses to it on both sets.

Table 19: Alternative action-independent references on the development cohort. Model error is unchanged by the choice of reference; these gain contrasts are post hoc.

Reference	ΔG [95% CI]	Joint increase
Mean realized terminal latent (oracle; uses realized futures)	$+0.1523$ [$+0.0974$, $+0.2026$]	detected
One constant mean-context latent	-0.0553 [-0.0769, -0.0357]	absent
Linear context extrapolation	$+0.0049$ [-0.0069, $+0.0164$]	unresolved

Oracle-reference gain changes from -0.2292 to -0.0769.

Fixed-support reweighting The reweighting calculation uses the standard χ^2 variance bound (Duchi & Namkoong, 2019), applied to $h_t(c) = m_c - (1-t)b_c$: gain endpoints solve $\mathbb{E}[h_t] \pm \sqrt{\delta \operatorname{Var}(h_t)} = 0$ separately at each start state. Roots are unique for $\delta < \mathbb{E}[b]^2 / \operatorname{Var}(b)$; the smallest such threshold is 8.44. Attainment additionally requires nonnegative normalized tilts. With $K = 68$ and uniform base weights, $\delta \le 1/(K-1) = 0.0149$ is sufficient; beyond it feasibility is checked directly. The constraint is a product across states on their existing candidate supports, not a restriction on arbitrary new actions.

At $\delta = 0.01$, the attained band is $[0.4493, 0.4849]$, width 0.0356 and maximum displacement 0.0181. At $\delta = 0.0278$, both witnesses are feasible at all 32 states: they move gain by -0.0304 or $+0.0289$, and model error by $+0.1090$ or -0.1147, respectively. Neither endpoint reproduces the co-increase; the upper gain endpoint is also below the observed $+0.0304$ gain rise. Larger-radius results at 0.05, 0.25 and 1 are outer bounds because at least one state has an infeasible tilt. This band has no sampling-confidence interpretation. It does not rule out co-increases under larger reweightings or changed support.

Construction-defined action generators Table 20 holds model, states, goals and horizon fixed, but the generators have different candidate quotas. They are subsets of the rich set, distinct from the ordinary-set groups. The observed gain span is 0.2318 [0.1812, 0.2841]; the extrema were selected from the data and the interval is not selection-adjusted. Among 56 ordered generator pairs, two meet the two-interval joint criterion and eleven co-increase at their point estimates. The selected planner-proposal-like to cross-object contrast has $\Delta M = +0.1936$ [$+0.0013$, $+0.4119$] and $\Delta G = +0.1888$ [$+0.1311$, $+0.2488$]; its error lower bound only narrowly exceeds zero. These are conditional comparisons, not an exhaustive action-space test.

F.3 Within-Pipeline Accounting of the Public Reproduction

The 64-family reproduction has one start per family. We recompute the identity from its 128 family–set scalar records, check $G = 1 - M/B$ in each row, and pair each family's sets before averaging. Both sequential substitutions exactly sum to the recorded gain change:

$$\Delta G = -(M_r - M_o)/B_o + M_r(B_o^{-1} - B_r^{-1}), \tag{7}$$

$$\Delta G = -(M_r - M_o)/B_r + M_o(B_o^{-1} - B_r^{-1}). \tag{8}$$

Their symmetric average is Equation (5). We use 20,000 shared family bootstrap draws for the newly computed intervals. This post-hoc calculation is an exact description of the score components, not a causal decomposition. It does not use the provisional original-checkpoint or cross-encoder results. The total-change interval from these draws is $[+0.0871, +0.1119]$; the main result table retains its original bootstrap interval. Small endpoint differences reflect the new resampling stream.

Table 20: Controlled predictor evaluated separately on each construction-defined generator. Counts differ by design. Planner-proposal-like means the planner's initial proposal distribution, not converged proposals (Appendix C).

Generator	n	M	G
planner-proposal-like	128	3.2470	0.3489
magnitude sweep	320	4.1342	0.3534
toward-object	384	4.1865	0.3604
tangential	320	3.4070	0.4649
mixed temporal	192	3.3794	0.4933
direction sweep	384	3.2170	0.5218
cross-object	256	3.4406	0.5377
anchor palette	192	3.2962	0.5806

Table 21: Public reproduction gain accounting, with paired 95% family-bootstrap intervals. The mean gain change is $+0.09946$. Both orders assign opposing signs to the two terms.

Substitution order	Model-error term	Reference-error term
Model first	-0.0320 [-0.0532, -0.0138]	$+0.1314$ [$+0.1164$, $+0.1494$]
Reference first	-0.0180 [-0.0298, -0.0078]	$+0.1174$ [$+0.1077$, $+0.1285$]
Symmetric average	-0.0250 [-0.0415, -0.0108]	$+0.1244$ [$+0.1125$, $+0.1385$]

F.4 Reserved-cohort ranking diagnostics

The rank-correlation decomposition below is exact on common support. One of 58 start states has constant privileged pose cost and undefined rank A; its task-cost range is only 2.2×10^{-16} across 68 candidates. Following the specified rule, its family contributes through its other state. Consequently the reported gap sum 0.5594 differs from the stage-mean subtraction $0.7474 - 0.1905 = 0.5569$ (Table 22); these summaries have different supports. Realized-displacement strata are descriptive: reference error is 5.1549 without translation versus 7.5004 in the largest-translation stratum, but the moving strata are not monotone. Neither this stratification nor small block translation establishes task-cost equality.

Applying the 10^{-12} task-tie tolerance of Appendix D to rank correlations changes the reserved gaps from $(0.2872, 0.2722)$ to $(0.2884, 0.2759)$, with both intervals still positive; dropping the constant state instead gives $(0.2872, 0.2735)$. The controlled task-rank contrast becomes -0.1219 [-0.1996, -0.0423], versus -0.1096 under the specified convention. On the development cohort, changing the pose orientation weight from 0 to 100 px/rad leaves the contract-gap interval positive, with point values 0.2587–0.3006; at 200 px/rad it reaches zero. The privileged comparator is therefore weight-dependent, not a ceiling.

Top-1 choice and its limits For each start state, regret is the selected candidate's realized task cost minus that set's best available task cost. Excluding the logged continuation that reaches the goal exactly, ordinary regret is 0.1659 [0.1116, 0.2234] and rich regret 0.2131 [0.1644, 0.2646]. Their paired difference is $+0.0472$ [-0.0302, $+0.1275$]; normalizing each by its own set's cost spread gives $+0.0045$ [-0.1350, $+0.1338$]. Both include zero. The sets offer different actions and opportunities, so these within-set comparisons do not establish performance loss on common independent control tasks.

Ranking the rich set by realized latent cost instead of prediction still gives regret 0.0863 [0.0549, 0.1232], against a task-cost spread of 0.6772. This shows residual misalignment under that latent cost, not a causal allocation of loss between representation and prediction. Pairwise accuracy need not track top-1 regret: task ties occupy 0.265 and 0.087 of ordinary and rich pairs. Envelopes computed from the realized tied costs span almost the full regret range at the observed accuracies (both ends attainable at all 32 rich states and 30 of 32 ordinary states). They limit what the statistics imply about each other, rather than demonstrate model or planning harm.

Realized-latent and pose diagnostics The privileged comparator ranks candidates using simulator block position and orientation, with orientation weighted at 25 px/rad relative to translation. Let A, B and C be the task-ranking correlations of this pose cost, realized latent cost and predicted latent cost, respectively. The descriptive identity is

$$\underbrace{A - C}_{\text{total}} = \underbrace{A - B}_{\Gamma_{\text{contract}}} + \underbrace{B - C}_{\Gamma_{\text{pred}}}. \tag{9}$$

On 29 reserved families, the registered contrasts are 0.2872 $[+0.1417, +0.4311]$ and 0.2722 $[+0.2088, +0.3378]$ (Table 22). They depend on the pose comparator and are not causal attributions. The support difference described above also prevents treating the registered sum as a difference of stage means.

Table 22: The reserved cohort: 29 reserved trajectory families, 58 start states, 3944 candidates on the intervention-rich candidate set, evaluated exactly once. Each claim passes when the lower bound of its paired interval over families exceeds zero, a rule registered before the cohort was opened; the evaluation was designed before the candidate-set comparison of Table 3 existed and does not test it. The development column is context only and is never pooled with the reserved column.

Claim	Quantity	Value	95% CI	Fam.>0	Development
1	baseline-relative gain	0.4664	[+0.4513, +0.4801]	29/29	0.4673
2	ρ(predicted, realized)	0.6181	–	29/29	0.6134
2	ρ(predicted, task)	0.1905	–	21/29	0.1689
2	their paired difference	0.4276	[+0.3130, +0.5435]	27/29	0.4445
3	A: privileged pose comparator	0.7474	–	29/29	0.7554
3	B: realized repr. vs task	0.4626	–	27/29	0.4655
3	C: predicted repr. vs task	0.1905	–	21/29	0.1689
3	contract gap $A - B$	0.2872	[+0.1417, +0.4311]	23/29	0.2899
3	prediction gap $B - C$	0.2722	[+0.2088, +0.3378]	29/29	0.2965

Stage A is a privileged comparator, never a deployable cost and not a proven upper bound. The gaps are paired per-start-state differences, not differences of the stage values above: stage A is undefined at one start state, so the supports differ, and subtracting unrounded stage values gives 0.2847 instead of 0.2872. Which gap is larger is not tested.

G Shared-Encoder Comparisons

The testing family contains all 21 pairs of seven predictors on 64 families, with $d_f = (G_{i,r,f} - G_{j,r,f}) - (G_{i,o,f} - G_{j,o,f})$ and three-seed aggregation inside each family. For the final inferential display, we uniformly use Holm-adjusted studentized sign-flip tests (Winkler et al., 2014) across these 21 contrasts and retain paired t tests as a sensitivity analysis. This reporting change is post hoc; neither the pairs nor the test are selected separately for favorable results. Both procedures identify the same 12 contrasts (Table 25).

The sign-flip statistic is $|\bar{d}/(s_d/\sqrt{64})|$. The stored analysis uses 99,999 shared sign vectors and add-one Monte Carlo p values (Phipson & Smyth, 2010), $(1 + \#\{|t^*| \geq |t|\})/100000$. It assumes independent family contrasts symmetric about zero under the null. Candidate assignment was not randomized, so its calibration follows this symmetry assumption rather than treatment randomization. The paired t test has 63 degrees of freedom and is exact for independent Gaussian contrasts, otherwise approximate. Agreement is a same-cohort sensitivity result, not an independent confirmation.

The original 2,000-draw bootstrap doubled the smaller zero-tail fraction and floored it at $1/2000$. Those adjusted tail summaries are retained in Table 25 as historical diagnostics, not null-calibrated p values. Percentile intervals remain nominal paired family-bootstrap intervals; their presentation does not imply simultaneous coverage.

Complete comparison groups and raw errors The six LD6 comparisons were emphasized after inspection. All six mean gain advantages contract by 35.6–64.8%; their pair-level median contraction is 49.7% [33.6%, 71.5%] under shared family draws, not a median of family-level percentages. All ordinary denominators of these six contraction ratios remain positive in those draws. Raw-error

advantages also contract by 36.2–71.4% (median 51.5%; Table 23A). BA8's rich raw-error advantage includes zero, and its contraction interval extends above 100%.

Across all 21 pairs, $|\overline{G_i - G_j}|$ contracts in 15 and expands in six; these are point counts, not significance counts. With absolute values taken before family averaging, the counts are ten and eleven. The BA7–MS8 gap changes from 0.00281 [−0.00098, +0.00666] to 0.00856 [+0.00329, +0.01356], a paired widening of +0.00575 [+0.00220, +0.00907]. Its point ratio is 3.05, but 150 of 2,000 ordinary-gap bootstrap draws are nonpositive; inference uses the additive contrast.

Accounting before family aggregation Cached reference errors are identical across models, and $G = 1 - M/B$ holds within family to numerical precision. Define $A_f = M_{\mathrm{LD6},f} - M_{i,f}$ and $d_f = A_f/B_f$. Averaging the two sequential orders gives

$$\Delta d_f = \tfrac{1}{2}(A_{f,r} - A_{f,o})(B_{f,o}^{-1} + B_{f,r}^{-1}) + \tfrac{1}{2}(A_{f,o} + A_{f,r})(B_{f,r}^{-1} - B_{f,o}^{-1}).$$

Table 23B averages these descriptive terms across families; both orders are supplied separately. The pooled reference ratio 1.06565 implies 6.16% contraction only under hypothetical uniform rescaling with fixed model errors; it does not apportion the observed mean-of-ratios change. For every LD6 comparison, the model-error-gap term is negative and larger in point magnitude than the reference term in both orders and their symmetric mean.

Table 23: All six comparisons with the logged-data comparator LD6, in unnormalized error units and symmetric gain-gap accounting. Positive raw-error advantage $A_i = M_{\mathrm{LD6}} - M_i$ favors model i. Every interval is a nominal paired 95% family-bootstrap interval, with the same shared draws as Table 2. This is a post-hoc descriptive analysis.

A: Raw model-error advantages

Model i	$A_{i,o}$	$A_{i,r}$	Change [95% CI]
BA7	0.4990	0.3181	−0.1808 [−0.2455, −0.1213]
MS8	0.4737	0.2454	−0.2283 [−0.2912, −0.1716]
PA8	0.4067	0.2482	−0.1584 [−0.2119, −0.1081]
BA8	0.3612	0.1031	−0.2581 [−0.3445, −0.1783]
BD6	0.3117	0.1196	−0.1921 [−0.2529, −0.1354]
CF6	0.2290	0.1033	−0.1257 [−0.1888, −0.0677]

B: Gain-gap change, symmetric arithmetic accounting

Model i	Model-error-gap term [95% CI]	Reference-error term [95% CI]
BA7	−0.01776 [−0.02420, −0.01174]	−0.00303 [−0.00404, −0.00202]
MS8	−0.02388 [−0.03025, −0.01815]	−0.00267 [−0.00369, −0.00162]
PA8	−0.01682 [−0.02220, −0.01146]	−0.00245 [−0.00335, −0.00157]
BA8	−0.02613 [−0.03464, −0.01826]	−0.00180 [−0.00275, −0.00082]
BD6	−0.02002 [−0.02602, −0.01444]	−0.00179 [−0.00261, −0.00101]
CF6	−0.01265 [−0.01911, −0.00631]	−0.00132 [−0.00196, −0.00071]

Panel B averages both algebraic orders and then families; its two terms sum to the corresponding gain-gap change. It is accounting, not causal attribution. Both sequential orders and all family values accompany the analysis code. No new model-error or median-endpoint tests are introduced by this table. BA8's rich advantage is 0.1031 with level interval [−0.0378, +0.2339]; this differs from its paired change interval above. Contrasts are computed before rounding.

Supplementary predictor comparisons Table 26 reports supplementary raw-error and task-ranking responses, excluding the conditional original-checkpoint row. Gain levels and intervals use the canonical shared analysis in Table 2; the historical bootstrap stream remains in the supplied records. Table 2 uses the later shared 2,000-draw analysis; small interval-endpoint differences reflect resampling. The complete 21 paired tests and sensitivity analyses appear once in Appendix G (Table 24); the main Holm family remains those seven shared-encoder predictors. The original-checkpoint rescoring is preserved in the historical records but supplies no additional model-comparison claim in this revision. Pooled reference ratios are 1.066 internally and 1.608 for the reproduction. The family-level ratio identity was verified for all $8 \times 64 = 512$ pairs of the original seven internal models plus reproduction, excluding the later original release. This is an arithmetic

Table 24: All shared-encoder gain comparisons on 64 families. For pair $i - j$, the level columns report $G_i - G_j$ and the change is rich minus ordinary. Intervals are nominal paired 95% family-bootstrap intervals. The final inferential display uses studentized sign-flip tests with Holm adjustment across all 21 changes; 12 meet $p < 0.05$. This uniform reporting choice and the comparisons are post hoc. Paired-t sensitivity and historical bootstrap tails are in Table 25. Contrasts are computed before rounding.

Pair $i - j$	Ordinary gap	Rich gap	Change [95% CI]	Sign-flip p_{Holm}
LD6−BA7	−0.0584	−0.0376	+0.0208 [+0.0150, +0.0270]	2.10×10^{-4}
LD6−BA8	−0.0431	−0.0152	+0.0279 [+0.0205, +0.0358]	2.10×10^{-4}
LD6−BD6	−0.0375	−0.0157	+0.0218 [+0.0163, +0.0276]	2.10×10^{-4}
LD6−PA8	−0.0478	−0.0286	+0.0193 [+0.0142, +0.0244]	2.10×10^{-4}
LD6−CF6	−0.0270	−0.0131	+0.0140 [+0.0078, +0.0202]	2.60×10^{-4}
LD6−MS8	−0.0556	−0.0291	+0.0265 [+0.0210, +0.0324]	2.10×10^{-4}
BA7−BA8	+0.0153	+0.0224	+0.0071 [+0.0006, +0.0130]	0.2503
BA7−BD6	+0.0209	+0.0219	+0.0010 [−0.0039, +0.0057]	1.0000
BA7−PA8	+0.0106	+0.0091	−0.0015 [−0.0058, +0.0026]	1.0000
BA7−CF6	+0.0314	+0.0246	−0.0068 [−0.0130, −0.0010]	0.2503
BA7−MS8	+0.0028	+0.0086	+0.0058 [+0.0022, +0.0091]	0.0267
BA8−BD6	+0.0056	−0.0005	−0.0061 [−0.0112, −0.0010]	0.2503
BA8−PA8	−0.0047	−0.0134	−0.0087 [−0.0140, −0.0031]	0.0308
BA8−CF6	+0.0161	+0.0021	−0.0140 [−0.0199, −0.0082]	4.80×10^{-4}
BA8−MS8	−0.0125	−0.0139	−0.0014 [−0.0055, +0.0030]	1.0000
BD6−PA8	−0.0103	−0.0128	−0.0025 [−0.0060, +0.0011]	0.7183
BD6−CF6	+0.0105	+0.0026	−0.0078 [−0.0108, −0.0050]	2.10×10^{-4}
BD6−MS8	−0.0181	−0.0134	+0.0047 [+0.0004, +0.0088]	0.2503
PA8−CF6	+0.0208	+0.0155	−0.0053 [−0.0100, −0.0008]	0.2503
PA8−MS8	−0.0078	−0.0005	+0.0073 [+0.0048, +0.0098]	2.10×10^{-4}
CF6−MS8	−0.0286	−0.0160	+0.0126 [+0.0074, +0.0177]	2.10×10^{-4}

Table 25: Inference and historical summaries for all 21 shared-encoder gain-gap changes. Sign-flip tests supply the final display and paired-t tests provide sensitivity; each adjusts the same 21 endpoints with Holm and identifies the same 12 pairs at 0.05. The last column preserves the original adjusted bootstrap zero-tail summaries for provenance; these are not null-calibrated p values. All analyses use the same families and are post hoc.

Pair	Sign-flip p_{Holm}	Paired-t p_{Holm}	Historical adjusted tail
LD6−BA7	2.10×10^{-4}	2.18×10^{-7}	0.0105
LD6−BA8	2.10×10^{-4}	3.52×10^{-8}	0.0105
LD6−BD6	2.10×10^{-4}	6.67×10^{-9}	0.0105
LD6−PA8	2.10×10^{-4}	1.46×10^{-8}	0.0105
LD6−CF6	2.60×10^{-4}	6.82×10^{-4}	0.0105
LD6−MS8	2.10×10^{-4}	8.26×10^{-12}	0.0105
BA7−BA8	0.2503	0.2489	0.1840
BA7−BD6	1.0000	1.0000	1.0000
BA7−PA8	1.0000	1.0000	1.0000
BA7−CF6	0.2503	0.2489	0.1840
BA7−MS8	0.0267	0.0276	0.0330
BA8−BD6	0.2503	0.2489	0.1170
BA8−PA8	0.0308	0.0327	0.0400
BA8−CF6	4.80×10^{-4}	3.04×10^{-4}	0.0105
BA8−MS8	1.0000	1.0000	1.0000
BD6−PA8	0.7183	0.7125	0.6480
BD6−CF6	2.10×10^{-4}	2.05×10^{-5}	0.0105
BD6−MS8	0.2503	0.2489	0.1840
PA8−CF6	0.2503	0.2489	0.1840
PA8−MS8	2.10×10^{-4}	1.11×10^{-5}	0.0105
CF6−MS8	2.10×10^{-4}	1.79×10^{-4}	0.0105

Table 26: Predictor responses on 64 paired families. Seven internal configurations and the public DINO-WM reproduction are reported. Error contrasts are rich minus ordinary with nominal 95% family-bootstrap intervals; bold marks an interval excluding zero. Ratios are pooled diagnostics; the last columns count positive-gain and joint-positive families. Labels: Section 3.1.

Predictor	$\Delta\mathbb{E}[\varepsilon_\theta]$ [95% CI]	ε_θ ratio	ε_0 ratio	$G\uparrow$	joint
Internal predictors on the fixed V-JEPA 2 representation, three seeds each					
LD6	**+0.0702** [+0.0207, +0.1245]	1.017	1.066	54/64	30/64
BD6	**+0.2623** [+0.2041, +0.3250]	1.067	1.066	30/64	21/64
CF6	**+0.1959** [+0.1269, +0.2677]	1.049	1.066	37/64	20/64
BA7	**+0.2510** [+0.1841, +0.3183]	1.068	1.066	32/64	21/64
BA8	**+0.3282** [+0.2479, +0.4111]	1.085	1.066	24/64	16/64
PA8	**+0.2286** [+0.1722, +0.2888]	1.060	1.066	34/64	25/64
MS8	**+0.2985** [+0.2361, +0.3655]	1.080	1.066	24/64	22/64
Public checkpoint in its own representation space					
DINO-WM (reproduction)	**+0.0066** [+0.0030, +0.0108]	1.116	1.608	63/64	38/64

Gain levels and intervals appear in Table 2 and Table 3; historical gain intervals are retained in the records. Contrasts are computed before rounding. Error columns are in the squared-error units of each representation space and are never compared across blocks. The ratio columns are descriptive diagnostics (Appendix F). Horizon 10 controls, environment target pose (Section 3.1).

check, not independent evidence or a test of which aggregate intervals exclude zero; pooled and family-mean gain need not agree.

H PROSPECTIVE POINTMAZE: COMPLETE OUTCOMES

Protocol and cohort After the eight-family pilot, we fixed the primary claim, the equivalence margins, the sample size, the bootstrap seed and replicate count, the six-endpoint Holm family, the substitution rule and the outcome labels before any further outcome was computed. That plan was recorded with the project rather than deposited in an external registry, so it carries no third-party timestamp; what it supports is that the rule was not chosen after the numbers were seen. We selected the first 32 already-extracted, non-pilot public validation episodes reaching source index 25, in ascending episode order. Selection rule, horizon, candidate count, and task costs follow the pilot. These episodes contributed to the checkpoint's normalization statistics, so the evaluation is on data we had not inspected rather than on a clean held-out split. Before any outcome was computed, the render-acceptance rule was corrected to the one the pilot had used: the simulator state is authoritative to 10^{-5}, and a rendered-frame mean absolute pixel difference of at most 1 is a non-gating diagnostic. All starts match exactly and goal-state errors are at most 1.85×10^{-7}.

Estimands and outcomes The registered gain is $G_f = 1 - \overline{M}_f / \overline{B}_f$, followed by averaging paired family differences, as in Equation (2) with one start per family. This differs from the pilot's mean of per-candidate gains. With pilot estimates visible, the protocol set margins of ± 0.05 for gain (about one fifth of pilot ordinary gain) and ± 0.10 for log error; these tolerances have no task-performance calibration. The two one-sided equivalence tests require the full 90% interval within its margin; a zero-containing 95% interval alone does not establish it. The primary claim is unsupported; both equivalence tests pass, and set optimum is the only secondary endpoint surviving Holm. Tables 27 and 28 retain all endpoints and intervals. Terminal and time-mean regret evaluate the same terminal-latent selector. Time-mean distance averages controls 1–25, excluding $t = 0$, in native planar units. The log-error tolerance ± 0.10 permits error ratios from 0.905 to 1.105; the observed 90% interval implies 0.995–1.013. Equivalence remains relative to the stated tolerances.

Table 27: All registered task endpoints in the prospective PointMaze cohort ($n = 32$). Reference and geometry columns are means over families. Changes and unadjusted 95% intervals are paired within family. The primary endpoint is tested separately; Holm adjustment covers exactly the six secondary endpoints below it. All costs use the model's single selection based on predicted terminal latent distance. Terminal costs evaluate the same selected action. The earlier eight-family pilot is not pooled into this table.

Endpoint	Reference	Geometry	Change [95% CI]	p_{raw}	p_{Holm}
Mean geodesic regret (primary)	0.3520	0.3130	-0.0390 $[-0.1914, +0.1166]$	0.6220	—
Mean geodesic optimum	0.5838	0.6657	$+0.0818$ $[+0.0460, +0.1187]$	0.0001	0.0006
Mean geodesic selected cost	0.9358	0.9786	$+0.0428$ $[-0.1189, +0.2110]$	0.6196	1.0000
Terminal geodesic regret	0.2347	0.2644	$+0.0297$ $[-0.1901, +0.2159]$	0.7274	1.0000
Terminal geodesic optimum	0.1333	0.2009	$+0.0676$ $[+0.0086, +0.1323]$	0.0236	0.1180
Terminal geodesic selected cost	0.3680	0.4653	$+0.0973$ $[-0.1501, +0.3241]$	0.4122	1.0000
Mean Euclidean regret	0.3492	0.3125	-0.0367 $[-0.1871, +0.1185]$	0.6400	1.0000

Table 28: Registered model-metric equivalence criteria in the prospective PointMaze cohort ($n = 32$). Both entire paired 90% family-bootstrap intervals are inside the prespecified margins. Meeting these operational criteria does not establish zero change or negligible task impact. The gain margin exceeds the controlled Push-T effect; the observed interval gives a tighter bound. The log-error margin permits error ratios from 0.9048 to 1.1052.

Contrast	Estimate	90% CI	Margin	Criterion
ΔG	$+0.0131$	$[+0.0051, +0.0204]$	± 0.05	Met
$\Delta \log M$	$+0.0040$	$[-0.0052, +0.0133]$	± 0.10	Met

I Exploratory Pilots and Cost Discrimination

I.1 Distinct eight-family pilots

Drawer Open uses Meta-World (Yu et al., 2020); PointMaze uses the D4RL maze task (Fu et al., 2020) as implemented in DINO-WM (Zhou et al., 2025). The three eight-family pilots use different second-set constructions, not repeated instances of one intervention (Table 30). All use public training- or normalization-exposed pipelines; none is a clean held-out confirmation or a model-selection test. Candidate generation reads no candidate outcome, cost, goal field or model output; Drawer Open additionally uses an executed expert prefix. Gain levels are family means; pooled ratios are descriptive and do not determine detection. Table 29 preserves each pilot's native aggregation and the candidate-mean ratio of (2), plus terminal and time-mean opportunity decompositions. Under (2), Drawer, Wall and PointMaze have gain changes $+0.731$, -0.135 and $+0.010$, versus native $+0.953$, -0.135 and $+0.011$; positive-family counts remain 8/8, 0/8 and 6/8. Drawer and PointMaze's native means of candidate ratios can be unbounded below near zero reference error. The separate prospective PointMaze study uses 32 families and is reported in Appendix H; its results must not be substituted for this pilot.

The paired model-error changes are -0.057 $[-0.168, +0.085]$ in Drawer and $+0.109$ $[+0.067, +0.154]$ in Wall (nominal 95% family intervals). Drawer's ordinary gain is negative: the model loses to persistence on that set.

Table 29: Pilot sensitivity, recomputed from the stored per-family records. Top: the gain change under the aggregation of (2) (candidate means, then the ratio), beside the artifact-native form each pilot froze (Drawer Open and PointMaze average per-candidate gains; Wall/Two-Room already uses (2)); both are family means with a paired percentile bootstrap over families (seed `0x135283C`, $B = 10{,}000$), so the native intervals here are a recomputation and may differ from the frozen records in the last digit. Bottom: the opportunity decomposition $\Delta L = \Delta O + \Delta R$ of the selected action's realized cost, where O is the best realized cost in the set and $R = L - O$ the within-set regret, for the time-mean cost the pilots report and for the terminal cost that matches the selection rule; each of the three is bootstrapped on its own per-family values, never by subtracting interval ends. Push-T has only a terminal cost (seed 49, $B = 2000$, 16 families). Bold marks an interval excluding zero; units are each environment's own.

	Push-T	Drawer Open	Wall/Two-Room	PointMaze pilot ($n = 8$)
Gain change ΔG				
(2) form	*as Table 3*	**+0.731 [+0.618, +0.825]**	**−0.135 [−0.198, −0.077]**	+0.010 [−0.010, +0.028]
native form	*same*	**+0.953 [+0.847, +1.071]**	**−0.135 [−0.198, −0.077]**	+0.011 [−0.009, +0.030]
families $\Delta G > 0$, (2) form	12/16	8/8	0/8	6/8
Time-mean cost				
ΔL selected	*n/a*	**−0.033 [−0.038, −0.023]**	**−14.1 [−18.9, −8.7]**	**+0.515 [+0.187, +0.842]**
ΔO optimum	*n/a*	**−0.039 [−0.041, −0.038]**	**−4.6 [−6.4, −2.4]**	**+0.118 [+0.028, +0.232]**
ΔR regret	*n/a*	**+0.007 [+0.001, +0.017]**	**−9.5 [−14.9, −3.7]**	**+0.397 [+0.101, +0.714]**
Terminal cost				
ΔL selected	+0.001 [−0.082, +0.083]	**−0.157 [−0.185, −0.110]**	**−21.7 [−31.3, −11.9]**	+0.488 [−0.042, +1.241]
ΔO optimum	**−0.046 [−0.090, −0.007]**	**−0.189 [−0.190, −0.189]**	**−9.5 [−15.2, −2.8]**	+0.023 [−0.132, +0.222]
ΔR regret	+0.047 [−0.030, +0.128]	**+0.032 [+0.004, +0.079]**	**−12.2 [−19.6, −4.0]**	+0.465 [−0.021, +1.277]

I.2 Shared-generator Wall extension

The byte-identical PointMaze generator supplies Sobol macro-actions in the unit disk and rotates each sequence using the maximum map-only clearance among sixteen rotations, with a fixed tie break. The 32 public Wall episodes exclude pilot episodes; goals reflect starts across the wall into free space. Only geometry and displacement scale change between environments. The generator reads neither goal nor outcomes; paired action norms agree to 10^{-7}. Gain rises $+0.012$ $[+0.002, +0.021]$, model error falls -0.005 $[-0.009, -0.001]$, and reference change is unresolved, $+0.008$ $[-0.008, +0.023]$. Selected cost rises $+6.86$ $[+4.37, +9.44]$, with $\Delta O = +8.96$ $[+7.44, +10.43]$ and $\Delta R = -2.10$ $[-4.37, +0.10]$. Thus Wall and prospective PointMaze both show worse opportunity but unresolved regret changes, not a replicated null. The cohort and generator were fixed before any outcome, without a registered hypothesis or margin; PointMaze's equivalence verdict does not transfer to it.

Table 30: Pilot protocol cards. Each uses eight families, one fixed public checkpoint and two sets fixed before candidate rollouts. All model steps pack five physical controls. Intervals use paired family percentile bootstraps (10,000 draws). Selection minimizes predicted terminal-goal latent distance; the reported primary pilot cost is time-mean, with terminal-cost sensitivity in Table 29.

	Drawer Open	Wall/Two-Room	PointMaze
Environment / checkpoint	Meta-World drawer-open-v3 / JEPA-WMs Meta-World	Single-layout DINO-WM Wall / public Wall checkpoint	DINO-WM PointMaze / public PointMaze checkpoint
H steps; K/set	20; 16	5; 64	5; 64
Ordinary set	Isotropic directions, expert-prefix norms and gripper channel	Uniform initial heading, von Mises increments	Broad headings and magnitudes
Second set / inputs	Perturbed executed expert-prefix directions at matched norms; demonstration-conditioned	Matched magnitudes steered through the door; start and static map	Rigid rotation of each ordinary sequence; start and static maze
Native gain	Mean candidate ratios	Ratio of candidate means	Mean candidate ratios
Realized cost	$\lvert q_{\mathrm{drawer}} - q_{\mathrm{goal}} \rvert$; fixed expert terminal goal	Door-aware geodesic goal distance	Radius-0 geodesic goal distance
Exposure	Public in-task episodes	Public single-layout archive	Public normalization-exposed episodes

Model error is final visual-latent MSE; regret is selected realized cost minus the within-set minimum. Drawer uses the expert's simulator-executed prefix; the other generators use only start and static geometry. All retain seeded randomness and read no candidate outcomes or model outputs.

Table 31: One rotation generator across environments. Each cohort has 32 families, five model steps and 64 candidates per bank. Navigation entries are paired changes with 95% family intervals; bold marks unadjusted interval exclusion of zero. Underlined counts mark exact task-cost constancy. Costs use environment-specific units. Push-T reports simulator-only diagnostics here; the validation scope of its original-checkpoint scoring is described in Appendix I.3.

	Wall (descriptive)	PointMaze (preregistered)
gain ΔG	**+0.012 [+0.002, +0.021]**	**+0.013 [+0.004, +0.022]**
error $\Delta\mathbb{E}[\varepsilon_\theta]$	**–0.005 [–0.009, –0.001]**	+0.001 [–0.002, +0.005]
selection regret ΔR	–2.1 [–4.4, +0.1]	–0.039 [–0.191, +0.117]
selected cost ΔL	**+6.9 [+4.4, +9.4]**	+0.043 [–0.119, +0.211]
set optimum ΔO	**+9.0 [+7.4, +10.4]**	**+0.082 [+0.046, +0.119]**
families with $\Delta G > 0$	19/32	23/32
sets with a constant task cost	0/64	0/64
Push-T: task-cost discrimination (descriptive)		
Banks with exactly constant realized cost		48/64
Families with both banks constant		24/32
Candidates with logged contact and block motion		1107/4096
Such candidates within exactly constant banks		364

Within a constant bank, every selector has $R = 0$; this does not identify a superior selector. If two banks have different constant levels, $\Delta O = \Delta L$ need not be zero even though $\Delta R = 0$. Two additional banks have range 2.22×10^{-16}; exact constancy is a conservative count. No family is excluded.

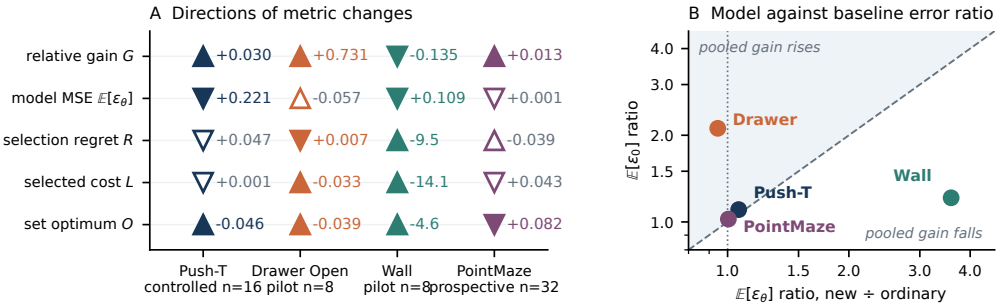

Figure 3: Cross-environment diagnostics, with environment-specific costs. Push-T is the controlled 16-family PA8 study. The Wall and Drawer columns are eight-family pilots; PointMaze is the 32-family prospective cohort. The additional 32-family Wall and Push-T rotation studies are separately reported in Table 31. In A, up means higher G or lower remaining quantities; labels give mean changes and filled markers denote unadjusted 95% intervals excluding zero. B uses pooled error ratios, which do not determine the interval for family-mean gain. Neither panel establishes generalization across interventions.

I.3 Push-T rotation: scope and validation

The same generator is applied to 32 fresh simulator seeds outside the declared cohorts, at five model steps of five controls. It uses the original public checkpoint through the official loader, preprocessor and rollout interface, without independent output parity. This separate run neither inherits nor resolves the cross-pipeline disagreement described in Appendix E.2: all model-dependent quantities, including selected actions, remain provisional. Simulator cost ranges, minima and contact counts do not depend on those predictions.

Rotation maximizes clearance from the four radius-inflated arena walls; the T block is the manipulandum, not an obstacle. It therefore does not favor block approach or goal progress. The macro displacement is 40.5 pixels, measured along sixteen free-space headings and retaining those whose block endpoint matches a zero-action control; an instant-settling assumption would give 100.5 pixels, $2.48\times$ too large. Because a scripted-reset block settles about sixteen pixels without any action, displacement alone cannot certify the absence of contact.

Model-facing outputs from this run remain conditional on the original checkpoint scoring path. They are not used to support a model-quality effect in this revision. The simulator-derived cost discrimination below is independent of that validation status.

Coverage cost is exactly constant in both banks of 24/32 families: 48/64 banks, with two additional ranges of 2.22×10^{-16}. The paired constant levels also agree. Every selector then has $O = L$ and $R = 0$, and these values remain in the means. Of 4,096 candidates, 1,107 have logged contact events and block motion, including 364 in exactly constant banks. The collision callback does not identify body pairs, so these are not independently verified pusher–block contacts. The cost can remain unchanged despite logged contact and motion, a scope limitation consistent with known ranking diagnostics (Wang et al., 2026). These constant-cost counts are simulator diagnostics and do not rely on a model-score contrast. We do not use conditional original-checkpoint gain variation as additional evidence.

J Operational Planner Extensions

J.1 Earlier proposal groups and fixed selectors

A separate analysis uses LD6, BD6 and CF6, three seeds each, with the earlier predictor lineage in Appendix A; these levels are not the controlled-study checkpoints' levels. Comparing size-matched ordinary-set construction and planner-proposal groups gives differences between

ranking-axis changes of $+0.1774$ [$+0.0207$, $+0.3283$], $+0.2125$ [$+0.0879$, $+0.3420$] and $+0.2644$ [$+0.1287$, $+0.4134$]. Here realized-latent alignment falls more, unlike the controlled between-set comparison. Thus axis separation recurs, but its direction does not generalize. The ordering of these three models remains unchanged. On all 42 planner proposals, LD6–CF6's reported gap is $+0.0375$ [-0.0640, $+0.1389$], versus $+0.3556$ over the whole set; this is unresolved ordering, not a demonstrated switch. The size-matched top-1 task-regret changes are $+0.0390$ [-0.0598, $+0.1295$], $+0.0063$ [-0.1034, $+0.1052$] and $+0.0000$ [-0.1104, $+0.0964$]. None resolves a worsening of the selected action despite changed rank statistics.

Table 32: The first 32 families of the predetermined test stream: exploratory, because the run was stopped at 32 of the 128 planned families after partial outcomes had been seen. Query-set contrasts are intervention-rich minus ordinary for a single frozen predictor at a time; closed-loop contrasts are MS8 minus DINO-WM, the two sealed selections, with three repeats averaged within family. Intervals are 95% family bootstraps (2000 resamples; seed `0x135283A` for the query-set contrasts and the registered final cost, `0x135283C` for the replayed endpoints). Bold marks an interval excluding zero.

Block	Quantity	N	Mean [95% CI]
Query-set contrast, DINO-WM, $H = 10$ controls			
	ΔG	32	**+0.0933** [+0.0727, +0.1128]
	$\Delta\mathbb{E}[\varepsilon_\theta]$	32	**+0.0078** [+0.0021, +0.0148]
	$\Delta\mathbb{E}[\varepsilon_0]$	32	**+0.1136** [+0.0990, +0.1293]
Query-set contrast, PA8 (internal), two readouts			
$H = 10$	ΔG	32	-0.0041 [-0.0141, +0.0053]
	$\Delta\mathbb{E}[\varepsilon_\theta]$	32	**+0.2595** [+0.1920, +0.3286]
	$\Delta\mathbb{E}[\varepsilon_0]$	32	**+0.5970** [+0.4192, +0.7852]
$H = 16$	ΔG	32	+0.0091 [-0.0058, +0.0227]
	$\Delta\mathbb{E}[\varepsilon_\theta]$	32	**+0.3532** [+0.2581, +0.4519]
	$\Delta\mathbb{E}[\varepsilon_0]$	32	**+0.9758** [+0.7554, +1.2177]
Closed loop, MS8 minus DINO-WM, 32 controls, replanning every two			
registered	final task cost $1 - \mathrm{cov}$	32	+0.0264 [-0.0015, +0.0662]
registered	success rate ($\mathrm{cov} > 0.95$)	32	0 of 768 episodes succeed for any model
post hoc	maximum coverage, controls 0–32	32	$-$**0.0320** [-0.0702, -0.0051]
post hoc	final block–goal distance (px)	32	+1.56 [-2.63, +6.87]
post hoc	mean distance, controls 2–32 (px)	32	+1.36 [-2.30, +6.15]
post hoc	final wrapped angle error ($^\circ$)	32	-0.76 [-4.35, +2.36]
post hoc	mean angle error, controls 2–32 ($^\circ$)	32	+0.01 [-2.95, +2.73]
reference	DINO-WM minus hold, final cost	32	-0.0222 [-0.0617, +0.0057]
reference	MS8 minus hold, final cost	32	+0.0042 [0.0000, +0.0125]

Positive cost and pose error and negative coverage favour DINO-WM. Of the 768 closed-loop episodes over all eight predictors, 723 end at exactly the hold cost, 20 below it and 25 above it (tolerance 10^{-9}). Replayed final frames match the frozen episode images exactly and replayed final coverage agrees with the recorded cost within 10^{-10}.

Reduced test stream and continuous replay The test run stopped after outcomes from 32 of 128 planned families had been viewed; all results in Table 32 are therefore exploratory. Closed-loop comparisons hold start, goal and planning budget fixed: 32 controls, replanning every two, with three repeats averaged within family. The selections fixed before the test stream are MS8 for planner proposals and the public reproduction for ordinary/rich queries. Their final-cost contrast is $+0.0264$ [-0.0015, $+0.0662$]. No episode among 768 across eight predictors meets the coverage threshold; 723 end at the hold cost. In the two-model contrast, 168/192 do so (MS8 93/96; reproduction 75/96, tolerance 10^{-9}). This largely uninformative deployment supplies no utility conclusion, and its interval establishes neither equivalence nor absence of a selection effect. Block–goal position, angle and maximum coverage were added by replay after outcomes were seen; they are post-hoc endpoints, with replay checks recorded in the table. The separate query-set contrasts show increased reproduction gain and error, whereas PA8's error rises at both $H = 10$ and $H = 16$ and both gain intervals include zero.

Planner references and the fixed selector Each planner score forms $1 - \bar{M}/\bar{B}$ over four candidates, averages across 96 query states (two proposing predictors, three repeats, sixteen replanning states) and internal training seeds, then across 64 families. The seven internal predictors share 6,144 distinct family–query denominators; their 18,432 query–seed rows are repeated references, not independent observations. The reproduction has 6,144 denominators. All 1,536 family scores recompute exactly from the stored records.

Table 33: Reference-error diagnostics on recorded planner proposals. Each reference error is the mean over four candidates at one query state. Both representations have 6,144 distinct family–query groups; internal values repeat exactly across three seeds and seven predictors. The threshold is a descriptive, scale-dependent diagnostic. These queries also change the evaluation states and do not constitute a controlled candidate-only intervention.

	Shared internal representation	Public reproduction
Minimum B	2.01×10^{-8}	2.68×10^{-11}
First percentile B	3.38×10^{-8}	3.07×10^{-11}
Fifth percentile B	1.7624	5.45×10^{-11}
Median B	3.041	0.0930
Groups with $B < 10^{-6}$	90/6144 (1.46%)	343/6144 (5.58%)

All denominators are positive, yet the most negative query/seed gains for the two registered selector choices are -2.98×10^8 for MS8 and -3.05×10^9 for the public reproduction. Their corresponding (M, B) values are $(7.826, 2.63 \times 10^{-8})$ and $(0.08675, 2.85 \times 10^{-11})$. These are group-level ratios, whereas the fixed selector averages them within each family and then across families, giving -2.90×10^6 for MS8 and -9.02×10^7 for the reproduction. The diagnostics retain the registered scores without clipping, excluding families or replacing the selector.

These strictly positive but tiny denominators make the fixed selector's mean-of-ratios score numerically unusable, and the fixed selector chose MS8 as the least negative score. A post-hoc, unregistered alternative averages within-family median group gains: reproduction 0.330, PA8 0.038, MS8 0.034, agreeing in first choice with ordinary/rich selectors. Neither ordering is validated model quality. Planner proposals change query states as well as actions; this diagnoses the recorded scoring rule, not a controlled action-only effect. Initial CEM populations were not retained for this earlier dataset. The separate crossed-population study in Appendix J.3 stores common initial and refined sets, but evaluates only selected-action costs. Thresholds in the diagnostic table are representation-scale dependent, not exclusion rules.

J.2 Sensitivity to CEM Proposal Scale

We run the released JEPA-WM and DINO-WM reproduction checkpoints through the JEPA-WMs planner (Terver et al., 2026). Each of 21 public validation episodes supplies two fixed windows. Six arms cross two models with initial proposal scales $\sigma_0 \in \{0.5, 1, 2\}$, for 252 plans. The released `var_scale` parameter multiplies the sampled standard deviation, so σ_0 denotes initial standard deviation and the isotropic covariance is $\sigma_0^2 I$. Population size is 300, with 30 CEM iterations, ten elites, six model steps, and five controls per step. The success rule is the released pose threshold: position within 20 pixels and orientation within $\pi/9$. Arms share windows and fixed planning seeds.

The primary contrast is the change in the JEPA-minus-DINO success margin relative to $\sigma_0 = 1$. The analysis uses 10,000 episode-clustered bootstrap draws and Holm over the two alternatives. The plan was recorded internally before outcomes, with smoke exposure and execution amendments documented in the supplement. A technical change to the rollout wrapper followed a failed run whose intermediate results had been viewed; the final 252-row run is the reported dataset. The primary endpoint and final sample did not change.

Table 34 preserves all outcomes. Neither primary margin change is resolved: $+2.4$ [-9.5, $+14.3$] and $+7.1$ [-11.9, $+23.8$] percentage points, with Holm tail summaries of one. Halving the scale lowers each model's success level as a secondary result, without establishing a changed relative margin. The point margin changes sign at scale two, but its interval includes zero. The released-scale sign check is also unsuccessful.

Table 34: Released-planner proposal-scale sensitivity. Two checkpoints use the same CEM implementation ($N = 300$, 30 iterations, ten elites, six model steps), dataset-segment goals, and pose-success rule on 42 windows from 21 episodes. Only initial scale changes. This is not an exact replication of the published sample or aggregation. Success is in percentage points; intervals cluster by episode. Bold marks a nominal interval excluding zero; the last row is primary and remains unresolved after Holm correction.

σ_0	0.5	1 (released)	2
JEPA-WM success	26.2	45.2	54.8
DINO-WM success	28.6	50.0	52.4
JEPA-WM, change from released	$\mathbf{-19.0}$ [−31.0, −7.1]	–	+9.5 [−4.8, +23.8]
DINO-WM, change from released	$\mathbf{-21.4}$ [−31.0, −11.9]	–	+2.4 [−9.5, +14.3]
margin, JEPA-WM − DINO-WM	−2.4 [−7.1, 0.0]	−4.8 [−16.7, +7.1]	+2.4 [−11.9, +16.7]
margin change (primary)	+2.4 [−9.5, +14.3]	–	+7.1 [−11.9, +23.8]

This is not a reproduction of the original published success comparison. That evaluation samples 96 episodes per training epoch and aggregates late epochs and seed folders; here one released checkpoint per model is evaluated on 42 windows from 21 public validation episodes. Some source episodes were used in earlier validation checks. The published values are not contradicted by these estimates. We retain the failed primary test as a limit on extending the score-margin claim to operational planner success. We do not interpret a noisy sign change as a model-ranking reversal.

J.3 Crossed Evaluation of Initial and Refined CEM Populations

Design and completed outcomes Twenty-four fixed starts from 15 of the same public validation episodes supply three candidate sets of 300 sequences each: the initial CEM population $\mathcal{C}_0$, the final population generated by JEPA-WM $\mathcal{C}_1$, and that generated by the DINO-WM reproduction $\mathcal{C}_2$. The initial populations from the two generation runs are bitwise identical. Both models score the same stored arrays. Each selected sequence is executed for 30 controls; task cost is one minus final block–goal coverage. All 144 primary cells are complete. The windows are distinct from the scale study's windows but the source episodes overlap, so this is not an independent confirmation cohort.

The primary quantity is the mean across starts of JEPA-selected cost minus DINO-selected cost. Two contrasts compare that gap on refined populations with its value on the initial population. The recorded protocol specifies a mean over starts, whereas the first analysis averaged within episodes and then weighted all episodes equally. Episodes contain one to three starts, so those estimands differ. We correct this mismatch using the existing outcomes and explicitly label the correction as post-result analysis. Resampling remains at episode level: each draw samples episodes and divides their total outcomes by their total number of starts.

Table 35: Crossed CEM evaluation under the specified start-weighted estimand. Intervals resample 15 episode clusters; negative gaps favor JEPA-WM. All primary changes are unresolved.

Quantity	Mean [95% interval]	Holm tail summary
Gap on initial $\mathcal{C}_0$	−0.02710 [−0.06125, +0.00127]	—
Gap on JEPA-refined $\mathcal{C}_1$	−0.00192 [−0.01212, +0.00669]	—
Gap on DINO-refined $\mathcal{C}_2$	+0.00223 [−0.00873, +0.01484]	—
Gap change, $\mathcal{C}_1 - \mathcal{C}_0$	+0.02518 [−0.00039, +0.05584]	0.1072
Gap change, $\mathcal{C}_2 - \mathcal{C}_0$	+0.02934 [−0.00335, +0.06423]	0.1072

The 10,000-draw paired bootstrap uses the same episode indices for both contrasts. Bootstrap tail summaries are supplemented with all 2^{15} cluster sign flips; the two Holm-adjusted values are 0.2020. This sensitivity analysis assumes independent episode clusters and null symmetry of their contrasts. Under either analysis, the specified primary claim is unsupported. The initial-set gap is also unresolved under the specified estimand, so the experiment does not show a resolved advantage disappearing after refinement.

Scope of completion The populations do change: median refined-to-initial dispersion ratios are 0.0307 and 0.0311. This manipulation check is not evidence of a changed model margin. Only selected sequences were executed; full-population task costs, opportunity O, and regret R were not obtained. The planned third-checkpoint secondary analysis was not executed. The runner recorded the released pose-success rule whereas the protocol specified a coverage threshold, so we do not present success as a protocol-conforming endpoint. These limitations leave the primary continuous-cost calculation intact but prevent claiming a completed joint audit of planner populations. Original and corrected aggregations are supplied with the scalar records.