SAMPLING HEADROOM IS NOT SELECTION GAIN: A COMPUTE-VALUE AUDIT OF TEST-TIME SCALING FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time scaling (TTS) succeeds when additional computation improves the output a system returns. For video world models, however, a larger candidate pool can contain better videos while leaving selected quality nearly unchanged. We introduce the *Compute-Value Audit* (CVA), a sequential framework that traces this gap through four requirements: available improvement, informative inference-time signals, useful interventions, and gains that exceed matched uniform compute after generation and verification costs. On 192 Physics-IQ scenes, expanding the Wan2.2-5B pool from 4 to 16 candidates raises oracle quality by $+9.23$ IQ points (95% confidence interval $[+7.44, +11.14]$), while the estimated changes for denoising-stability, reconstruction-consistency, and learned-preference selectors range from -0.75 to $+0.26$ IQ. On VideoPhy2, perturb-and-refine sampling yields negative estimated changes in joint semantic and physical success across three fresh-seed replications at a matched budget of 60 denoiser evaluations; the intervals include zero. Reasoning controls expose the decisive distinction between predicting difficulty and identifying where additional sampling helps. An anchor–explorer policy beats matched uniform allocation on PRM800K, with gains concentrated in a small subset of problems, while privileged real-future references improve video selection. CVA thus shifts the scaling question from how much quality a pool contains to how much a realizable decision recovers at full cost.

1 INTRODUCTION

At inference time, a generative model can fail in two ways: it may never produce a good output, or it may produce one without recognizing it. Best-of-B sampling creates more opportunities by generating and ranking B candidates (Snell et al., 2024; Wang et al., 2023b; Hassan et al., 2026). Frozen video rollouts can look plausible while violating scene dynamics (Motamed et al., 2025).

A verifier must recognize better candidates, a confidence signal must identify where an intervention helps, and that intervention must beat a simpler use of the same compute. The best evaluation-scored candidate, or *oracle*, measures available quality; the selected output measures recovered quality.

We turn this distinction into four ordered requirements: *opportunity*, *state confidence*, *action value*, and *entry fee*. The last term is the improvement needed to beat uniform use of the same total budget. Every generated candidate is counted and scored, with smaller budgets using fixed prefixes of the same pool. Figure 1 traces where available quality stops becoming a better returned output.

CVA evaluates selection, allocation, and refinement under one contract. Each mechanism retains its own score and action; the first unmet requirement identifies the decision that needs improvement.

Figure 2 shows a $B = 4$ decision made by the reported history–motion selector, CVA-Select. The ball should land on sand and remain there. Denoising stability loses it, whereas the selected rollout retains the settled ball, with delayed release. The two methods rank the same four videos with a fixed model, prompt, and conditioning frame. Reconstruction consistency and global history also choose the CVA-Select output. This is a realizable decision rather than an oracle pick. CVA tests whether such choices recover headroom consistently and justify their full cost.

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

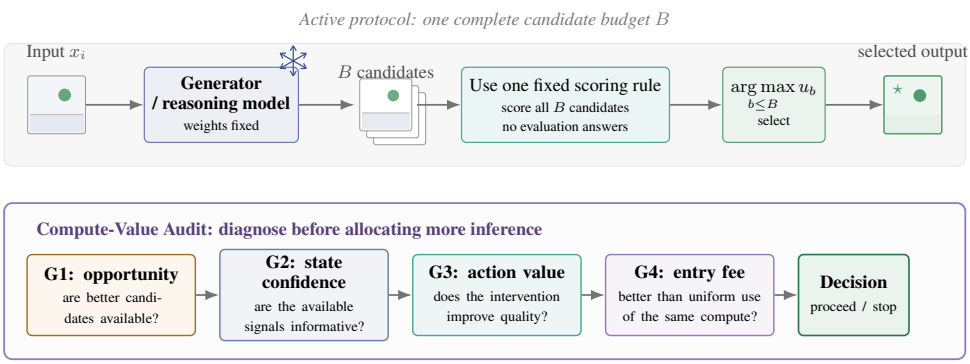


Figure 1: **Domain-general Compute-Value Audit.** *Top:* a fixed generator produces exactly B candidates, and one fixed rule scores all of them. *Bottom:* CVA audits opportunity, state confidence, action value, and the fee relative to matched uniform compute. The gates order the evidence: establish that an intervention improves the output, then compare its full cost with the alternative. Fixed-pool ranking tests a single selection decision; adaptive inference also makes intermediate allocation decisions. The same audit applies to consistency scoring, history-based ranking, and adaptive reasoning.

Our contributions are:

- **A theory-guided account of TTS value.** CVA decomposes scaling into four tests of information and compute use. Under a joint Gaussian model, verifier correlation determines the fraction of headroom recovered. The memory-value identity separates the benefit of adapting actions to inputs from the regret of choosing the wrong action.
- **A matched audit of video-world-model scaling.** Across three video generators and seven benchmarks, we distinguish oracle opportunity from selected quality. Wan2.2-5B exposes the central gap: oracle quality rises substantially from 4 to 16 candidates, while consistency and preference selectors recover little of that increase. Matched-budget refinement and target-score stopping diagnostics examine two further uses of inference compute.
- **Controls that explain when scaling pays.** Reasoning experiments distinguish signals of current accuracy from signals of additional-sampling value. Selective allocation succeeds when the latter guides the action. A privileged real future from the same video setup provides a complementary positive control for information-rich selection.

2 Related Work

We compare inference-time mechanisms under matched information and resource constraints.

Inference-time search and verifiers. Diffusion scaling separates generation from ranking. An inaccurate ranking score can improve under optimization while the true objective stagnates, a form of reward hacking (Ma et al., 2025; Gao et al., 2023). Outcome/process verification similarly relies on intermediate signals that predict terminal correctness (Cobbe et al., 2021; Lightman et al., 2024). Feynman–Kac steering uses sequential Monte Carlo (SMC) to propagate and resample partial generation trajectories, with gradient-free or gradient-guided proposals (Singhal et al., 2025). Adaptive reasoning methods also decide which inputs merit more samples (Zhai et al., 2026).

Changing intermediate denoising samples. Video-T1 combines noise search with Tree-of-Frames branching, while Stream-T1 propagates chunk noise and prunes beams with within-stream cache management (Liu et al., 2025a; Tu et al., 2026). Proprio instead ranks completed rollouts using the stability of the frozen model's predictions under small latent perturbations (Hassan et al., 2026). We also test Self-Refining Video Sampling's Predict-and-Perturb (P&P) procedure: repeatedly predict and re-noise an intermediate sample without retraining the generator (Jang et al., 2026). Off-Manifold Refinement (OMR) obtains guidance from a frozen external world model through a trained latent-to-

embedding adapter (Nguyen-Truong et al., 2026). Predict-and-re-noise updates, gradient guidance, particle resampling, and streaming search each incur distinct proposal, scoring, and training costs.

Physical and embodied benchmarks. Physics-IQ and VideoPhy2 probe observational dynamics and physical commonsense (Motamed et al., 2025; Bansal et al., 2025); PAI-Bench includes conditional robotics video generation (Zhou et al., 2025). RoboTwin and its randomized 2.0 extension introduce bimanual manipulation, contact, and task-conditioned dynamics (Mu et al., 2025; Chen et al., 2025). Our PAI/RoboTwin evaluations measure generated-video quality through perceptual and conditioning metrics. VideoReward supplies a generic preference ranking (Liu et al., 2025b), while physical-quality targets come from the evaluation benchmarks. Closed-loop robot-policy success is a separate evaluation setting.

Test-time memory. ComMem, Retrieve-then-Steer, and AdaMEM combine rapidly written experience with more slowly consolidated state (Sun et al., 2026; Zhao et al., 2026; Zhang et al., 2026). We isolate history-based reranking of candidates from a frozen generator. The two mechanisms are corrections fitted from past outcomes and retrieval of similar examples. Shuffled-history controls separate the value of relevant records from the benefit of fitting an additional model.

3 The Compute-Value Audit

We formalize how available candidate quality becomes selection gain, then separate the roles of verifier alignment, history, and compute.

3.1 One candidate budget

For fixed-pool best-of-B selection, B counts every generated candidate, and each selector must score the entire pool. The generated set and the decision set are therefore identical. For a frozen world model g_ϕ, scenario i produces

$$\mathcal{C}_i^{(B)} = \{z_i^{(b)}\}_{b=1}^{B}, \qquad z_i^{(b)} = g_\phi(x_i, \epsilon_i^{(b)}), \quad \epsilon_i^{(b)} \sim \mathcal{N}(0, I). \tag{1}$$

Minimizing an energy q_i is equivalent to maximizing the oriented score $S = -q_i$. Any history is declared separately, and no candidate outside the stated pool may be inspected. The single-candidate baseline uses $B = 1$ with replayable fixed seeds. Fixed-pool comparisons obey $G = V = B$; adaptive and within-generation policies additionally charge every preview, base step, and inner-loop network function evaluation (NFE) (Section 4.1).

3.2 CVA protocol: opportunity, state, action, and fee

With the comparison set fixed, we distinguish the available improvement from the amount recovered. For a declared reference Q_{ref} and positive headroom, define

$$H_B = Q_{\mathrm{oracle,B}} - Q_{\mathrm{ref}}, \qquad R_B = \frac{Q_{\mathrm{select,B}} - Q_{\mathrm{ref}}}{H_B}, \qquad A_B = H_B R_B. \tag{2}$$

Here H_B is the available headroom, R_B the recovered fraction, and A_B the quality gain before costs. Opportunity fixes the target, reference, and required headroom. State Confidence tests the permitted observations; Action Value tests whether acting on them improves the target. Adaptive allocation must then beat random assignment and uniform spending under matched resources. Predicting an error and identifying a useful response are separate requirements.

These checks order *evidence requirements*. In a static pool, G1 compares the oracle with $B = 1$, G2 checks score reliability, G3 measures selection recovery, and G4 compares cost-matched alternatives. Adaptive G2 also assesses the current result before allocating more compute. A dash denotes a bypassed or unreached check.

3.3 Ranking correlation and selection gain

Better candidates are useful only to the extent that the selector can identify them. We isolate this dependence with a minimal Gaussian model in which centered quality and score have correlation ρ:

$$Q_b = \rho S_b + \sqrt{1 - \rho^2}\, E_b, \qquad S_b, E_b \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, 1). \tag{3}$$

Proposition 1 (Verifier-quality law). *If the verifier selects $b^\star = \arg\max_{b \le B} S_b$, then*

$$\frac{\mathbb{E}[Q_{b^\star}]}{\mathbb{E}[\max_{b \le B} Q_b]} = \rho \tag{4}$$

for every $B > 1$.

The identity makes the two levers explicit: pool width increases the available maximum, while correlation determines the fraction recovered. It is a ratio of expectations under the i.i.d. joint-Gaussian model. Appendix E.1 gives the proof and shows how heavy tails and thresholded scores change the conclusion. Empirically, we report

$$R(v, B) = \frac{\overline{Q}_{\text{select}(v,B)} - \overline{Q}_{\text{random},B}}{\overline{Q}_{\text{oracle},B} - \overline{Q}_{\text{random},B}}, \quad \widehat{\rho}(v, B) = \frac{1}{n} \sum_i \widehat{r}_i. \tag{5}$$

Here $\widehat{r}_i$ is within-pool sample Pearson correlation, with degenerate pools zero-imputed (Appendix H). Ranking recovery $R(v, B)$ uses the within-pool random mean; deployment recovery R_B uses the incumbent. Their population ratio-of-expectations counterparts agree when reference expectations coincide, as for exchangeable c_{00}; finite-sample estimates can differ. Tracking both quantities separates the verifier's ranking ability from the gain over the output the system would otherwise return.

3.4 Frozen verifier signals

Denoising stability (Proprio (Hassan et al., 2026)) measures how much the generator's prediction changes after a small latent perturbation. *Reconstruction consistency* tests whether re-noising followed by denoising recovers the original sample; Cosmos3-Nano uses an analogous clean-sample reconstruction error. *Video preference* is the learned human-preference score supplied by VideoReward. Their alignment with physical correctness is measured against the benchmark target. Appendix E.4 defines the exact residuals. The "velocity" in these formulas is the model's latent denoising field, not object speed. Its raw magnitude is not a quality proxy: time reparameterization rescales it and linear-flow endpoint velocity need not vanish (Lipman et al., 2023; Ma et al., 2024).

3.5 Memory-based selector: Proprioceptive Memory (PM)

Proprioceptive Memory (PM) augments a fixed scoring rule with a correction learned from past outcomes and optional retrieval of similar examples. Neither component can access the current evaluation outcome or update the generator. Physics-IQ excludes history from the query's event family; PAI and RoboTwin use separate training folds with shuffled or wrong-task history as controls. Appendix E specifies the score, records, fitting, and tests; same-setup real futures remain separate additional-information references.

3.6 When does using history improve a decision?

For context x, fixed B, finite selector set $\mathcal{A}$, and candidate-generation noise ϵ, define

$$\begin{aligned} h(x) &= \mathbb{E}_\epsilon[Q_{\text{oracle,B}} - Q_{\text{random,B}} \mid x] > 0, \\ r_a(x) &= \mathbb{E}_\epsilon[Q_{\text{select}(a),B} - Q_{\text{random,B}} \mid x]/h(x). \end{aligned} \tag{6}$$

We average over generation noise before taking the ratio, so recovery describes the expected value of an action in context x. Write $\langle f \rangle_h = \mathbb{E}[hf]/\mathbb{E}[h]$ and $r^*(x) = \max_a r_a(x)$. A legal memory policy π uses only current observations and prior records.

Proposition 2 (Memory Value Law). *For any matched comparator $\pi_0 \in \mathcal{A}$,*

$$\underbrace{\langle r_\pi - r_{\pi_0} \rangle_h}_{\textit{memory value}} = \underbrace{\langle r^* - r_{\pi_0} \rangle_h}_{\textit{addressable value}} - \underbrace{\langle r^* - r_\pi \rangle_h}_{\textit{retrieval regret}}. \tag{7}$$

For the best fixed selector, the addressable term becomes the contextual heterogeneity gap $G = \langle r^* \rangle_h - \max_a \langle r_a \rangle_h \ge 0$; hence memory beats the best fixed selector iff regret is below G. Under the conditional Gaussian model, $r_a(x) = \rho_a(x)$. If all coordinates of $R = (r_a)_a$ lie in a common interval of length L, Bayes-optimal gross history value beyond context C is at most $L\sqrt{I_h(R; \mathcal{M} \mid C)/2}$. Here $\mathcal{M}$ is memory and information is in nats under the headroom-weighted distribution.

The heterogeneity gap measures the benefit available when different inputs favor different actions. Retrieval regret measures how much is lost when history leads to the wrong choice. History has positive net value only when the available benefit exceeds retrieval regret and the cost of using it.

Appendix Table 27 specifies the benchmark-specific history interfaces.

4 EXPERIMENTS

We compare selectors on fixed pools, use reasoning controls and richer video features to separate state prediction from action value, then test wider pools and matched-budget refinement. Target-score stopping diagnostics add a distinct allocation comparison. Each benchmark retains its own target.

4.1 EXPERIMENTAL PROTOCOL

We distinguish held-out tests, retrospective component comparisons, feedback-driven controls, and privileged references, and record when each study's selections were fixed. Physics-IQ begins with 198 scenarios; the headline analysis retains 192, compares $B = 4$, and extends the width study to $B = 16$. Uncertainty is clustered by scene or event family. PAI-Bench-robot uses the same eight candidates for every selector on 174 scenes and reports the official eight-dimensional mean (Zhou et al., 2025; Huang et al., 2024). VideoPhy2, OpenS2V, RoboTwin, PRM800K, and MMLU-Pro retain their own frozen populations, uncertainty units, and quality scales. Means, paired deltas, and decompositions are rounded independently after calculation from the full-precision results.

Matching the full cost. Reconstruction checks invoke the generator again, while a reward model adds a different type of computation. We therefore account separately for generation, verification by the same model, and external reward-model calls. DiT denotes the diffusion transformer inside the generator. Appendix Table 28 reports known per-scene calls and identifies the work that still requires hardware-specific pricing.

This accounting supports both fixed-pool ranking comparisons and quality comparisons at matched total cost. The first isolates the choice among outputs; the second tests whether that choice justifies verification. Denoising-stability and reconstruction scores add generator calls; the learned video-preference model adds one external call per candidate. History corrections run on the CPU alongside their charged base verifier. The same-setup real-future reference shares the retrieved-history rule's online calls and measures the value of additional predictive information.

4.2 VERIFIER RECOVERY FROM FIXED CANDIDATE POOLS

Once a pool contains better rollouts, can an available score identify them? Table 1 compares the single-candidate baseline with selectors based on denoising stability, learned video preference, reconstruction consistency, and retrieved history at $B = 4$. The selectors share the same four candidates; the baseline uses the first. The ‡ reference accesses a real future from the query's setup, providing an information-rich upper reference. Appendix Table 18 contains the corresponding controls.

The pool contains value that deployable scores leave unused. Reconstruction consistency leads among the three baseline scores. Video preference helps slightly on 5B, falls below a single candidate on 14B, and trails denoising stability on Cosmos. Retrieved history has mixed effects, while the same-setup real future improves selection on every generator. Figure 2 shows a verified $B = 4$ CVA-Select decision: the selected video retains the settled ball that denoising stability loses, reaching 75.39 IQ versus 56.12 at the same candidate budget. The delayed release remains visible in the full comparison.

Combining pooled history and motion. Table 2 reports all seven component combinations. We denote the $G + M_{\text{mot}}$ score by *CVA-Select*: G is pooled outcome history excluding the query family, and M_{mot} is the oriented frame-difference motion score. It averages 36.60 IQ, compared with 36.55 for G alone, a difference of $+0.05$; G remains higher on ABot-14B. Within these fixed pools, removing reconstruction from the three-component score raises the reported mean by $+0.59$ IQ.

Table 1: **Physics-IQ verifier audit ($B = 4$).** Mean IQ with 95% family-bootstrap CIs over 192 scenes/64 families. Bold highlights reconstruction consistency and the real-future reference. †History excludes the query's event family; ‡the reference accesses a real future from the same setup.

Method	Extra info	Wan2.2-5B	ABot-14B	Cosmos-Nano
Baseline (no-TTS)	–	25.07 [21.26, 28.99]	40.52 [35.29, 45.91]	27.72 [23.51, 32.03]
Denoising stability	–	25.30 [21.19, 29.59]	41.07 [35.72, 46.66]	31.31 [26.60, 36.10]
Video preference	Learned score	25.76 [21.67, 29.97]	39.62 [34.16, 45.29]	29.89 [25.12, 34.88]
Reconstruction consistency	No history	**27.53** [23.51, 31.59]	**41.57** [36.05, 47.33]	**33.70** [28.91, 38.58]
Retrieved history†	Other families	29.21 [24.82, 33.67]	41.93 [36.44, 47.55]	32.82 [28.12, 37.66]
Real-future reference‡	Same-setup recording	**32.19** [27.71, 36.78]	**45.67** [39.81, 51.65]	**36.83** [31.73, 42.10]

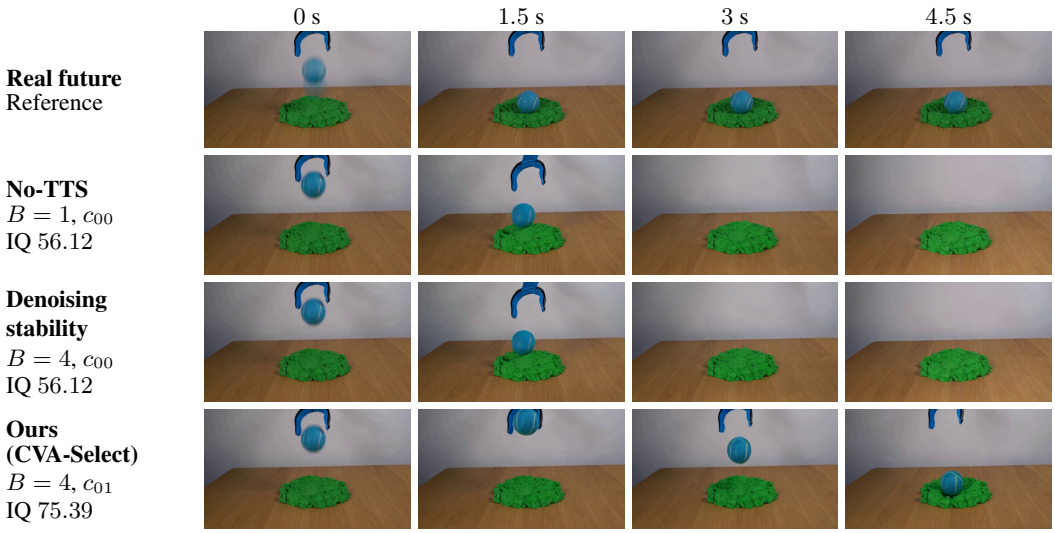


Figure 2: **Recovering the settled ball from the same four-candidate pool.** New $B = 4$ Wan2.2-TI2V-5B pool with one conditioning frame and the official description. No-TTS and denoising stability return c_{00}, which loses the ball; Ours returns c_{01}, retaining it with delayed release. G and reconstruction also select c_{01}. Both $B = 4$ rows use 160 generation DiT forwards; denoising stability adds 16 verification forwards (Appendix E.7).

The pooled-history score averages across records, whereas the rule in Table 1 retrieves individual examples. A larger ensemble can therefore dilute its stronger components. Appendix E.6 specifies the history bank, normalization, and tie rule.

Better ranking correlation accompanies greater recovery. Across held-out family–budget points, stronger correlation between the candidate score and evaluation quality accompanies greater recovery of available improvement ($R^2 = 0.629$). The association persists when calibration and evaluation use different families, but predictions transfer less accurately across generators. Appendix H reports the calibration results and their uncertainty.

Table 2: **Physics-IQ selector-component ablation ($B = 4$).** Post-hoc comparison of seven fixed combinations on 192 scenes/64 families, using the same candidate pools without query IQ. G, N, and M denote pooled history, reconstruction consistency, and low motion; CVA-Select is $G + M$. Δ compares with the full combination; single-component rows are references. The reconstruction average uses the displayed per-model means.

Selector	Wan2.2-5B	ABot-14B	Cosmos-Nano	Avg.	Δ
Global history (G)	30.59	44.13	34.94	36.55	–
Reconstruction (N)	27.53	41.57	33.70	34.27	–
Low motion (M)	30.33	43.69	35.10	36.38	–
Full fusion (G+N+M)	30.06	43.25	34.74	36.01	–
w/o global history (N+M)	29.14	42.61	34.58	35.44	−0.57
CVA-Select (G+M)	30.81	44.01	34.98	36.60	+0.59
w/o motion (G+N)	29.86	43.20	34.31	35.79	−0.23

Table 3: **Selective reasoning converts disagreement into allocation value.** The policy retains a reliable initial answer and samples more when answers disagree. On 128 PRM800K problems, the gain persists with a sampled initial answer and matched token cost. Appendix Table 12 reports uncertainty and the concentration of gains across problems.

Test	Estimate [95% CI]	Interpretation
Baseline (no-TTS; greedy)	Acc. 0.7344	Strong base
Sampled-anchor policy − uniform	+0.0358 [+.0119, +.0646] Accuracy gap +0.8627	Greedy unnecessary
Agreement signal	Extra-sampling value gap +0.1981	Predicts where sampling helps
Policy − token uniform/random	Uniform: +0.0159 Random: +0.0259	Gain after matching tokens

4.3 Where the chain breaks across benchmarks

The fixed-pool findings raise a second question: does a signal that identifies difficult inputs also reveal where extra compute will help? Appendix Table 8 organizes the corresponding tests. We begin with reasoning: answer disagreement can be measured before deciding whether to sample more.

Difficulty prediction and action value separate on MMLU-Pro. Agreement sharply separates easier from harder MMLU-Pro questions, satisfying the state-confidence test. Yet allocating more compute to questions with disagreeing answers does not beat matched uniform allocation. Appendix Table 9 thus places the failure at the choice of action, not at the availability of a predictive signal.

Selective sampling pays when disagreement predicts its value. PRM800K provides the contrasting case. Disagreement guides a policy, *anchor–explorer*, that retains a reliable initial answer and explores alternatives when answers differ. It beats uniform allocation matched by candidate count and by token cost. At the same mean candidate count, accuracy improves by $+2.64$ percentage points (95% CI $[+0.62, +5.14]$). The gains come from $7/128$ problems (5.5%), the top five account for 95.4% of positive mass, and trimming the top 10% removes the advantage. The policy earns its return through large improvements on a few problems rather than uniform gains across the population (Appendix Tables 11–12).

The video-state study tests candidate-set features, transfer, supervision, and partial-transformer previews. Across these alternatives, held-out action values fall short of the specified improvement criterion (Appendix Table 10). The next comparison asks whether combining inexpensive video signals improves selection, and how that combination depends on the generator.

Model-specific behavior on PAI. The no-history ensemble leads on 5B and 14B; visual reranking is competitive only on Cosmos and reduces some component scores on the other models. Appendix Table 29 reports all dimensions and uncertainty.

Table 4: **PAI-Bench-robot selector comparison ($B = 8$; 174 scenes).** Underline: best external score; bold: best reported benchmark-specific selection rule. no-TTS uses $B = 1$. †The ensemble and visual-reranking choices are model-specific retrospective comparisons. Appendix Table 29 reports paired uncertainty and all eight dimensions.

Method	Wan2.2-5B	ABot-14B	Cosmos3-Nano
Baseline (no-TTS; $B = 1$)	0.7594	0.7689	0.7730
Denoising stability	0.7630	<u>0.7704</u>	<u>0.7756</u>
Video preference	<u>0.7639</u>	0.7686	0.7731
Model-specific route†	**0.7698** (ensemble)	**0.7725** (ensemble)	**0.7766** (visual reranking)

Table 5: **Wider pools expose the gap between opportunity and recovery.** Physics-IQ $B : 4 \to 16$, Wan2.2-5B candidates only; 192 scenes/64 families, 50,000 family bootstraps. Cells show estimate and 95% CI. The last two columns give the algebraic contributions of added headroom and changed recovery under the symmetric identity in Eq. (11). †Access to a real future from the same setup.

Selection rule	IQ ($4 \to 16$)	ΔIQ	Added opportunity	Changed recovery
Baseline (no-TTS)	$25.07 \to 25.07$	0 (ref.)	–	–
		$+\mathbf{9.23}$		
Best candidate (oracle)	$35.09 \to 44.32$	$[+7.44, +11.14]$	–	–
		$+0.26$	$+1.78$	-1.52
Reconstruction	$27.53 \to 27.79$	$[-1.59, +2.10]$	$[+0.01, +3.45]$	$[-3.36, +0.51]$
		-0.61	$+2.75$	-3.36
Retrieved history	$29.21 \to 28.60$	$[-2.73, +1.62]$	$[+1.66, +3.74]$	$[-5.43, -1.24]$
		$+\mathbf{4.80}$	$+6.13$	-1.33
Real-future reference†	$32.19 \to 36.99$	$[+3.12, +6.57]$	$[+4.64, +7.70]$	$[-3.01, +0.47]$

Transfer and fresh-episode checks. Every admissible completion of four missing OpenS2V naturalness scores falls short of the shuffled-history criterion; Appendix Table 31 reports all seven dimensions. On fresh RoboTwin 2.0 episodes (Chen et al., 2025), the history correction matches the ordinary reward model on 5B and scores 0.000705 lower on 14B, with an interval spanning zero. This matched comparison isolates the history-specific contribution (Appendix Table 23).

Opportunity is model–pool specific. Cross-generator enrichment yields no multiplicity-corrected gain. We therefore measure opportunity in the target generator's pool before testing how selection changes with additional candidates.

4.4 Does additional compute pay for itself?

The selection comparisons ask which rule makes best use of a given pool. We next ask whether that rule benefits when the pool grows. We replay 192 scenes/64 families using *Wan2.2-5B-generated* candidates at fixed verifier depth. The ABot-14B row is a cross-model verifier of those same candidates.

Wider Wan2.2-5B pools expose a recovery gap. From $B = 4$ to 16, the best candidate gains $+9.23$ IQ, while reconstruction selection gains $+0.26$ and retrieved-history selection changes by -0.61. For retrieved-history selection, the additional available quality contributes $+2.75$ IQ to the accounting decomposition, while the decline in recovered fraction contributes -3.36 IQ. Table 5 uses the exact symmetric decomposition $\Delta(HR) = \Delta H(R_4 + R_{16})/2 + (H_4 + H_{16})\Delta R/2$ (Appendix E.2).

The retrieved-history selector's negative recovery term is consistent with proxy overoptimization: widening the pool also creates more ways to satisfy an imperfect score. In contrast, the real-future reference gains $+4.80$ IQ from the same expansion. Together with the fixed-pool comparison, this connects the scaling gap to the information used for selection (Appendix Tables 7 and 18).

Variable depth must overcome the uniform-allocation gap. A retrospective diagnostic gives stopping rules access to benchmark target scores for the generated prefix. At $B = 8$, the best of twelve

heuristics recovers 42–69% of the quality gap between random-depth and uniform allocation. The uniform control matches each rule's mean depth, so simply beating random assignment leaves part of the entry fee unpaid. This privileged diagnostic isolates the allocation comparison; deployment-time stopping additionally requires a realizable scoring interface and its cost (Appendix B).

Refinement reallocates computation within the sampler. The final experiment intervenes during generation on all 179 VideoPhy2-hard prompts. Perturb-and-refine sampling and the standard sampler share the initialization and use 60 denoising predictions. The three fresh-seed replications yield negative estimated changes in joint semantic and physical success, with intervals spanning zero. The primary run's unthresholded physical-commonsense rating changes by -0.140 on the ordinal 1–5 scale (95% CI $[-0.257, -0.022]$). A full Uniform40 follow-up on the same 179 primary prompts and recorded seeds separates changing the base grid from adding the refinement loop. All 537 archived and new videos are scored by one frozen evaluator. The six new grid/refinement contrast intervals include zero, leaving the relative contributions of integration and perturbation unresolved (Appendix C).

Both standard grids yield 9 successes among 179 primary prompts, while refinement yields 6. Joint success requires both ratings to clear their thresholds; ordinal scores also capture changes below them. With rare successes, equal totals do not establish grid equivalence. The paired component scores provide a finer comparison across the full prompt set. A resolved component-score difference can therefore coexist with an uncertain joint-success effect. Reporting both connects the pass-rate comparison to the behavior changed by additional computation.

Pool expansion, refinement, and variable-depth allocation change the candidate set, sampling trajectory, and distribution of compute across inputs, respectively. Each must improve the returned output over its matched alternative. The audit points to the corresponding next step: improve ranking when headroom is lost in selection, learn action value when difficulty alone is insufficient, and measure the full resource trade-off once the intervention improves quality.

5 Limitations

The width study at $B = 16$ uses Wan2.2-5B, and the refinement study uses one perturbation configuration. Extending these comparisons to other generators, gradient-guided proposals, particle steering, and streaming search would test how the identified bottlenecks change with the mechanism. The 40-step control leaves the integration-versus-perturbation explanation open, with precision constrained by rare joint successes. The privileged stopping and real-future comparisons define information boundaries; the OpenS2V score gaps and concentrated PRM800K gains define population boundaries. Cross-architecture reasoning confirmation remains an important next test.

Deployment cost also depends on hardware and execution strategy. Concurrent candidates increase peak memory and bandwidth demand, sequential generation adds latency, and reconstruction checks require successive solver calls. External reward models and offline training incur separate costs. Hardware-specific timing and memory measurements are therefore the next step from call-matched comparisons to an end-to-end efficiency frontier.

6 Conclusion

Sampling headroom is an opportunity, not a return on compute. CVA makes the conversion explicit by separating available quality, decision-relevant information, useful actions, and the full cost of executing them. Wan2.2-5B illustrates the gap: wider pools substantially improve oracle quality, while consistency and preference selectors recover little of the increase. The verifier and memory analyses identify score alignment and action relevance as the quantities connecting information to value. Selective reasoning and information-rich real-future references provide constructive examples of that connection. The resulting research target is clear: improve the output the system returns, and compare the complete procedure with simpler alternatives given the same resources. For video world models, this prioritizes signals tied to scene dynamics—including agreement across camera views and consistency of tracked 3D motion—and tests their value through held-out selection gains after charging generation and verification.

Reproducibility Statement

The appendices document candidate budgets, data filters, scoring rules, random seeds, paired uncertainty estimates, and the artifacts supporting each result. They distinguish recomputed estimates from archived summaries and record the available provenance. Code is available at https://anonymous.4open.science/r/sampling-headroom-is-not-selection-gain/.

AI Use Statement

Generative AI tools assisted research ideation, extraction and analysis code, experiment automation, literature retrieval, mathematical checks, and drafting. All experimental results are tied to recorded protocols and artifacts, with retrospective analyses labeled. The authors take responsibility for the claims, sources, and final manuscript. No AI system is an author.

References

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. *arXiv preprint arXiv:2503.06800*, 2025.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. RoboTwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025. URL https://arxiv.org/abs/2506.18088.

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 (NeurIPS)*, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, et al. Reinforced self-training (ReST) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Mariam Hassan, Kaouther Messaoud, Wuyang Li, and Alexandre Alahi. Proprio: Latent self-scoring and inference-time refinement for physically plausible video generation. *arXiv preprint arXiv:2605.28230*, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. *arXiv preprint arXiv:2406.15252*, 2024. URL https://arxiv.org/abs/2406.15252.

Yidong Huang, Zun Wang, Han Lin, Dong-Ki Kim, Shayegan Omidshafiei, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Planning with sketch-guided verification for physics-aware video generation. *arXiv preprint arXiv:2511.17450*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Sangwon Jang, Taekyung Ki, Jaehyeong Jo, Saining Xie, Jaehong Yoon, and Sung Ju Hwang. Self-refining video sampling. *arXiv preprint arXiv:2601.18577*, 2026. URL https://arxiv.org/abs/2601.18577.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 14165–14178, 2023. doi: 10.18653/v1/2023.acl-long.792.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2019.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *International Conference on Machine Learning (ICML)*, pp. 3652–3661, 2019.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Fangfu Liu, Hanyang Wang, Yimo Cai, Kaiyan Zhang, Xiaohang Zhan, and Yueqi Duan. Video-T1: Test-time scaling for video generation. *arXiv preprint arXiv:2503.18942*, 2025a. URL https://arxiv.org/abs/2503.18942.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. Improving video generation with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 82155–82192, 2025b. doi: 10.52202/085713-2750. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/76227feb18ea0ee40bd15cf02c33e18e-Abstract-Conference.html.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, 2024.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, Lunkai Lin, Zhiqiang Xie, Mingyu Ding, and Ping Luo. RoboTwin: Dual-arm robot benchmark with generative digital twins. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27649–27660, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Mu_RoboTwin_Dual-Arm_Robot_Benchmark_with_Generative_Digital_Twins_CVPR_2025_paper.html.

Hai Nguyen-Truong, Tuan-Anh Vu, and Dang Huynh. Off-manifold refinement: Guiding video generators with a frozen world model. *arXiv preprint arXiv:2608.29904*, 2026. URL https://arxiv.org/abs/2608.29904.

Alexander Pritzel, Benigno Uria, Sriram Srinivasan, Adrià Puigdomènech Badia, Oriol Vinyals, Demis Hassabis, Daan Wierstra, and Charles Blundell. Neural episodic control. In *International Conference on Machine Learning (ICML)*, 2017.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Phillip J. Schulte, Anastasios A. Tsiatis, Eric B. Laber, and Marie Davidian. Q- and A-learning methods for estimating optimal dynamic treatment regimes. *Statistical Science*, 29(4):640–661, 2014. doi: 10.1214/13-STS450.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. KNN-Diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. *arXiv preprint arXiv:2501.06848*, 2025. URL https://arxiv.org/abs/2501.06848.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Guanglong Sun, Shuang Cui, Bo Lei, Liyuan Wang, Zihan Zhai, Hongwei Yan, Hang Su, Jun Zhu, and Yi Zhong. ComMem: Complementary memory systems for test-time adaptation of vision-language models. *arXiv preprint arXiv:2606.28719*, 2026.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020.

Team Wan et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High confidence policy improvement. In *International Conference on Machine Learning (ICML)*, pp. 2380–2388, 2015.

Yijing Tu, Shaojin Wu, Mengqi Huang, Wenchuan Wang, Yuxin Wang, Chunxiao Liu, and Zhendong Mao. Stream-T1: Test-time scaling for streaming video generation. *arXiv preprint arXiv:2605.04461*, 2026. URL https://arxiv.org/abs/2605.04461.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023b.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977*, 2023.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, Jiayan Teng, Zhuoyi Yang, Wendi Zheng, Xiao Liu, Dan Zhang, Ming Ding, Xiaohan Zhang, Xiaotao Gu, Shiyu Huang, Minlie Huang, Jie Tang, and Yuxiao Dong. VisionReward: Fine-grained multi-dimensional human preference learning for image and video generation. *arXiv preprint arXiv:2412.21059*, 2024. URL https://arxiv.org/abs/2412.21059.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jianhao Yuan, Xiaofeng Zhang, Felix Friedrich, Nicolas Beltran-Velez, Melissa Hall, Reyhane Askari-Hemmat, Xiaochuang Han, Nicolas Ballas, et al. Inference-time physics alignment of video generative models with latent world models. *arXiv preprint arXiv:2601.10553*, 2026.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.01528.

Shenghai Yuan, Xianyi He, Yufan Deng, Yang Ye, Jinfa Huang, Bin Lin, Jiebo Luo, and Li Yuan. OpenS2V-Nexus: A detailed benchmark and million-scale dataset for subject-to-video generation. *arXiv preprint arXiv:2505.20292*, 2025.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Zhiyuan Zhai, Bingcong Li, Bingnan Xiao, Ming Li, and Xin Wang. Adaptive test-time compute allocation for reasoning LLMs via constrained policy optimization. *arXiv preprint arXiv:2604.14853*, 2026.

Yunxiang Zhang, Yiheng Li, Ali Payani, and Lu Wang. AdaMEM: Test-time adaptive memory for language agents. *arXiv preprint arXiv:2606.05684*, 2026.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i17.29936.

Jianchao Zhao, Huoren Yang, Yusong Hu, Yuyang Gao, Qiguan Ou, Cong Wan, SongLin Dong, Zhiheng Ma, and Yihong Gong. Retrieve-then-steer: Online success memory for test-time adaptation of generative VLAs. *arXiv preprint arXiv:2605.10094*, 2026.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. PAI-Bench: A comprehensive benchmark for physical AI. *arXiv preprint arXiv:2512.01989*, 2025. URL https://arxiv.org/abs/2512.01989.

TERMS USED IN THE EXPERIMENTAL TABLES

The tables use concise labels for recurring comparisons. The definitions below specify score inputs, evaluation units, compute, and archived variant names.

Candidate scores and history. *Denoising stability* is the Proprio score: smaller changes in the generator's prediction under a controlled perturbation are preferred. *Reconstruction consistency* checks whether denoising a perturbed sample recovers the original; in the Physics-IQ archive, this score is called "Native" or "Cycle." *Video preference* is the learned VideoReward score. On PAI, "Native" instead denotes a no-history ensemble of reconstruction, motion, temporal-consistency, and condition-matching signals, rather than a single reconstruction score. *Retrieved history* uses selected past examples, whereas *pooled history* averages all eligible past templates. A *real-future reference* uses a recording from the query's physical setup and is marked as additional information.

Training and held-out evaluation. Leave-one-family-out (LOFO) excludes all views of the current event family. Leave-one-episode-out (LOEO) excludes the current episode. Out-of-fold (OOF) predictions are produced by models not trained on the corresponding held-out examples. FIT, TUNE, and CERT denote training, tuning, and validation partitions within the training data; these archived partition labels are separate from CVA's gates. A *pairwise reward model* (RM) learns to rank two candidates using past outcomes. Comparing it with a history-specific correction tests whether history adds value beyond ordinary supervised fitting.

What is counted as compute. B counts completed candidates. A network function evaluation (NFE) counts a denoising prediction. DiT stands for diffusion transformer, the neural network that implements this prediction. Classifier-free guidance (CFG) combines predictions with and without conditioning; in the refinement experiments, each guided prediction uses two transformer forward calls. The denoising-prediction count therefore differs from the raw transformer-call count. Wall-clock latency and memory use require separate measurements. In the trajectory experiment, Uniform40/Uniform60 denote standard sampling with 40/60 base predictions; P&P60 denotes 40 base predictions plus 20 refinement predictions. In allocation experiments, *uniform* instead refers to assigning each query the same expected candidate budget.

Quality scores and uncertainty. Semantic adherence (SA) and physical commonsense (PC) are the evaluator's integer ratings. *Combined success*, termed "Joint" in the original evaluation, requires both ratings to be at least four. "Official7" and "official8" denote the mean of the seven or eight specified benchmark dimensions; each mean is benchmark-specific. W/T/L counts paired wins, ties, and losses at the unit specified in each caption. A confidence interval containing zero retains uncertainty about the effect's direction at the stated confidence level; equivalence requires its own test.

Controls and historical variant names. *Wrong-task history* substitutes records from another task; *shuffled history* randomizes the relation between past records and current queries while preserving the amount of history. These controls test whether the relevant past experience helps. DEQM and Candidate-DEQM are archive identifiers for two earlier history-transfer variants; STOM denotes the set-based history-transfer experiment. These labels identify experiment variants; the control comparisons determine the evidence for each. "Slow" denotes a correction fitted across past examples; "fast" denotes query-specific retrieval.

Evaluation status and target access. A completed evaluation records a test of the method; a demonstrated improvement additionally meets the stated empirical criteria. *Superseded* indicates that a later evaluation replaces an earlier exploratory comparison. Results computed with evaluation targets are privileged-information diagnostics, distinct from evaluations using only deployment-time inputs.

The appendix follows the main audit: complete results come first, followed by analyses of where the requirements fail and how history is used. The remaining sections document provenance, developmental studies, and compute accounting.

Table 6: **Frozen PAI readouts** ($B = 8$)**.** Official 8D means over 174 scenes; 50,000 paired-scene bootstrap draws. The selected readout is the no-history ensemble for Wan2.2-5B and ABot-14B, and a reliability-gated historical selector for Cosmos3-Nano. Matched controls below separate selection gains from the effect of history. Components: Table 29.

(a) Official 8D mean

Model	no-TTS ($B = 1$)	Denoising stability	Video preference	Ensemble w/o condition match	**Selection**
Wan2.2-5B	0.7594	0.7630	0.7639	0.7685	**0.7698**
ABot-14B	0.7689	0.7704	0.7686	0.7721	**0.7725**
Cosmos3-Nano	0.7730	0.7756	0.7731	0.7754	**0.7768**

(b) Paired readout gains

Model	Δ no-TTS [95% CI]	Δ best baseline verifier [95% CI]
Wan2.2-5B	$+\mathbf{0.0104}$ $[+.0084, +.0124]$	$+\mathbf{0.0058}$ $[+.0041, +.0076]$
ABot-14B	$+\mathbf{0.0036}$ $[+.0022, +.0050]$	$+\mathbf{0.0021}$ $[+.0008, +.0034]$
Cosmos3-Nano	$+\mathbf{0.0038}$ $[+.0028, +.0049]$	$+\mathbf{0.0013}$ $[+.0006, +.0020]$

These model-specific readouts beat no-TTS and the best listed baseline verifier on all three models (all six CI lower bounds positive). The history-specific comparisons instead use matched controls. On Cosmos, removing condition matching also changes the history interface; Table 30 isolates the I2V component while holding that interface fixed. Separately trained pretraining gives $+0.000069/0/-0.000141$ on 5B/14B/Cosmos, with a scene-clustered cross-base delta of -0.000024 (CI $[-0.000208, +0.000155]$). These additional-fitting results are reported separately from the selected readouts.

Table 7: **Wan2.2-5B Physics-IQ scaling.** Complete 16-candidate coverage; 192 scenes/64 families. All candidates are generated by Wan2.2-5B; the ABot-14B row is only a cross-model verifier. Each column consumes exactly B candidates; oracle is evaluation-only. Paired 95% CIs use 50,000 family-bootstrap draws. Denoising-stability and video-preference intervals are recovered from the archived paired analysis; reconstruction-score and best-candidate intervals follow Table 5. Means and deltas are rounded independently; dashes denote point summaries without a reported interval.

Selector	$B = 4$	$B = 16$	Δ IQ [95% CI]
Baseline (no-TTS; $B = 1$)	25.07		0 (reference)
Denoising stability	25.30	24.55	-0.746 $[-2.290, +0.801]$
Video preference	25.76	25.16	-0.592 $[-2.448, +1.280]$
Reconstruction	27.53	27.79	$+0.26$ $[-1.59, +2.10]$
ABot-14B verifier of Wan5B	26.59	28.23	–
Low motion	30.33	31.90	–
True-IQ oracle	35.09	44.32	$+9.23$ $[+7.44, +11.14]$

PM audit: held-out B4/B16 29.21/28.60; paired-future B4/B16 32.19/36.99; $\Delta_{\text{paired}-\text{heldout}}$ B4/B16 $+2.98/+8.39$ IQ.

A Mechanism-level evidence for the audit

These comparisons connect each audit gate to a measured quantity. The cross-benchmark matrix locates the bottlenecks; the reasoning controls then isolate the signals and actions behind a successful allocation.

Table 8: **CVA gate matrix.** ✓/△/× denote pass/partial/stop. In each row, the first × marks the gate that stops the audit; rows with four ✓ entries meet all four requirements. A dash means not applicable or not reached; △ indicates partial evidence. Each row uses its own target and reference. The Physics-IQ width row concerns Wan2.2-5B candidates scored by denoising stability, reconstruction consistency, and learned video preference (Table 7). Table 15 separately reports refinement repeatability. The stopping sweep stops at G2 because it requires benchmark target scores; its privileged uniform-allocation comparison appears in Appendix B.

Eval.	G1 Opportunity	G2 State confidence	G3 Useful action	G4 Cost comparison
Video models				
Physics-IQ 192; $4 \to 16$	✓ $+9.23$ $[+7.44, +11.14]$	—	× 3/3 CIs cross 0	—
VideoPhy2 179; $B = 8$	✓ $+20.67$ oracle	△ rank .839 MAE .170	× best 0.00 $[-5.59, +5.59]$	—
Generation refinement 179; 60 NFE	△ fixed-budget test	△ uncertain regions	× -1.68 pp $[-4.47, +1.12]$	—
Target-score stopping	△ oracle > unif. (diagnostic)	× target-score access	—	—
Reasoning				
PRM800K 128	✓ oracle $+10.94$	✓ agreement	✓ $+3.93$ vs rand. $[+1.14, +6.78]$	✓ $+2.64$ vs unif. $[+0.62, +5.14]$ †
MMLU-Pro 448	✓ oracle $+19.87$	✓ $+15.20$ $[+10.16, +20.20]$	× -2.43 $[-5.27, +0.33]$	—

†Sparse: $7/128$ positive; top five contribute 95.4% of positive mass. The top-10%-trimmed advantage vanishes; no valid cross-architecture replication.

Table 8 synthesizes the frozen comparisons. The diffuse-gain attempt falls short on opportunity, MMLU-Pro stops at action value, and the stopping diagnostic requires target-score access. PRM anchor–explorer clears all four gates with gains on $7/128$ problems. Each row retains its own population, target, and matched comparator.

Table 9: **State confidence does not imply action value.** MMLU-Pro, Qwen3.5-2B, $n = 448$; effects are percentage-point differences with 95% bootstrap CIs. Green is pass (✓); red is stop (×).

Comparison	Result [95% CI]	Check
Preconditions		
no-TTS (greedy reference)	29.24% accuracy	–
Oracle@8	49.11%; $+19.87$ pp headroom	✓
Mechanism		
Initial accuracy: agree − disagree	$+\mathbf{15.20}$ $[+10.16, +20.20]$ pp	✓
Extra-sampling value: disagree − agree	$-\mathbf{2.43}$ $[-5.27, +0.33]$ pp	×
Difference between the above effects	$+\mathbf{17.63}$ $[+12.68, +22.50]$ pp	distinct
Policy control		
Policy − matched uniform	Candidate: -0.40 $[-1.10, +0.28]$ pp Token: -0.40 $[-1.13, +0.29]$ pp	×

Table 10: **Action-value tests for richer video features.** The same 179 VideoPhy2 prompts are used throughout. Target-score selection measures the available decision value; learned predictors are evaluated on data excluded from their fits. Appendix Section A reports the complete tests.

Available features / rule	Held-out result	Comparison	Check
Evaluation-score choice	+3.07 pp advantage	Uses unavailable target scores	✓
Candidate-set features	Negligible held-out gain	Matched random allocation	×
Transferred / supervised features	No reliable held-out gain	Separate refitting and transfer tests	×
Partial transformer preview[†]	Lower reported quality	Preview features and refitted predictor	×

[†] Compared with the learned video-preference score selected using training data.

Table 11: **Selective reasoning beats matched uniform sampling.** Qwen3-VL-4B on 128 disjoint PRM800K problems. Uniform matches the mean number of candidates; random allocation permutes the same candidate counts across problems. Accuracy differences are proportions, with 95% problem-bootstrap intervals. $B = 8$ consensus/best-candidate accuracy: $0.7891/0.8438$.

Policy	Mean candidates	Accuracy	Δ uniform	Δ random
Baseline (no-TTS; greedy)	1	0.7344	–	–
Matched random	3.6763	0.7498	−0.0129 [−.0299, +.0003]	0
Uniform	3.6763	0.7627	0	+0.0129 [−.0003, +.0299]
Anchor–explorer	3.6763	**0.7891**	**+0.0264** [+.0062, +.0514]	**+0.0393** [+.0114, +.0678]

Table 12: **Mechanism and concentration of selective-reasoning gains.** These fixed-output analyses share the original 128-problem set. Matched token costs include the prompt and generated answer. Accuracy differences are proportions, with 95% problem-bootstrap intervals.

Test	Estimate [95% CI]	Takeaway
Baseline (no-TTS; greedy anchor)	accuracy 0.7344	Strong base
Greedy − sampled (base)	baseline +0.0156 [−.0234, +.0547]	No anchor advantage
Greedy − sampled (policy)	policy −0.0078 [−.0234, .0000]	Greedy unnecessary
Sampled policy − uniform	+0.0358 [+.0119, +.0646]	Gain survives
Agree − disagree	anchor accuracy +0.8627 [+.7706, +.9349]	Confidence signal
Disagree − agree	deep-sampling value +0.1981 [+.0747, +.3347]	Compute-value signal
Policy − token uniform	+0.0159 [+.0027, +.0332]	Exact-cost gain
Policy − token random	+0.0259 [+.0061, +.0499]	Assignment matters
Problems contributing gains	7/128; top-5 mass 95.4%	Few problems drive the mean
Trim top 10%: policy − uniform	−0.0006 [−.0019, .0000]	Broad gain vanishes

B PRIVILEGED TARGET-SCORE STOPPING DIAGNOSTIC

This retrospective diagnostic studies allocation using benchmark target-score prefixes. The archived loader `posthoc_shape_aware_sizing.load_pools` reads Physics-IQ `official_iq_allcand_a14b.json` and `official_iq_allcand_cosmos.json`, plus the VideoPhy2 deep-pool `Overall` scores. Its six statistics therefore use privileged evaluation information. The analysis measures quality-side allocation value; a deployment test would also supply an inference-time scoring interface and charge its cost.

B.1 SIX STATISTICS AND TWELVE STOPPING RULES

Let $q_{i,1:K}^{(\omega)}$ be one random ordering of a fixed pool's *target* scores. For a prefix of length n, write $b_n = \max_{j\le n} q_j$, $\bar{q}_n = n^{-1}\sum_{j\le n} q_j$, and $s_n^2 = (n-1)^{-1}\sum_{j\le n}(q_j - \bar{q}_n)^2$. Each of the six statistics below is used in two independently tuned directions, yielding exactly twelve rules.

Table 13: **Complete archived stopping-rule family.** All inputs are prefix benchmark target scores. For each row, the low arm stops on $T_n < \tau$; the high arm applies the same strict first-crossing implementation to $-T_n$. Both arms are privileged-information diagnostics, using target scores in place of a deployment-time scoring model.

Statistic	T_n	Implementation detail
Gaussian marginal	$(a_{n+1} - a_n)s_n$	$a_n = \mathbb{E}\max_{j\le n} Z_j$, $Z_j \sim \mathcal{N}(0,1)$
Gap	$b_n - \bar{q}_n$	Running maximum minus prefix mean
Stall	$n - t_n$	t_n: last strict maximum improvement; $t_1 = 1$
Range	$b_n - \min_{j\le n} q_j$	Prefix range
Current increment	$b_n - b_{n-1}$	Zero after a non-improving draw
IQR	$Q_{.75}(q_{1:n}) - Q_{.25}(q_{1:n})$	NumPy linear percentile interpolation

For direction $\eta \in \{+1, -1\}$, the depth is

$$d_i^{(\omega)} = \min\big(\{n \in \{2,\dots,K-1\} : \eta T_{i,n}^{(\omega)} < \tau\} \cup \{K\}\big).$$

The minimum depth is therefore two; if no crossing occurs, the rule uses depth K. The terminal value is the target-score prefix oracle: $v_i^{(\omega)}(d) = \max_{j\le d} q_{i,j}^{(\omega)}$. It measures the best target value in the consumed prefix, rather than the output of a deployment-time verifier.

Threshold calibration and ordering. The analysis uses $J = 200$ random candidate orderings per pool, starting from study seed 20260802. For each dataset, statistic, direction, and ordering, it calibrates one threshold across all pools using 60 bisection steps to target a mean depth of eight. With $c = \eta T$ over all eligible pool–depth cells and $s = \max c - \min c + 10^{-12}$, the initial bracket is $[\min c - 0.001s, \max c + 0.001s]$. If the midpoint yields a mean depth above eight, it becomes the lower endpoint; otherwise, it becomes the upper endpoint. The final midpoint determines the depths, although discrete statistics need not attain the target exactly. Calibration is retrospective: each threshold uses the full cohort's criterion arrays. It defines a cohort-level diagnostic rather than an online or held-out threshold-fitting procedure. The recovered records store depths rather than thresholds; reproducing numerical thresholds for the twelve rules requires rerunning the cohort calibration for each ordering.

B.2 EXACT QUALITY-SIDE FEE AND UNCERTAINTY

Averaging each pool's adaptive value over orderings gives A_i. For each ordering, we randomly permute that rule's own depth multiset across pools; averaging the resulting prefix-oracle values gives R_i. Let $\bar{d}$ be the actual mean depth over pools and orderings, and let $\bar{v}_i(n)$ be the ordering-averaged uniform-depth curve. For $l = \lfloor \bar{d} \rfloor$ and $w = \bar{d} - l$, define $U_i(\bar{d}) = (1-w)\bar{v}_i(l) + w\bar{v}_i(\min(l+1, K))$. The same rule's fee and recovery are

$$\varphi = \overline{U(\bar{d})} - \bar{R}, \qquad \text{recovery} = \bar{A} - \bar{R}, \qquad \bar{A} - \overline{U(\bar{d})} = \text{recovery} - \varphi. \tag{8}$$

The fee φ can have either sign. A share is reported only when the archived fee interval has a positive lower endpoint. The displayed best rule maximizes recovery among nondegenerate rules whose actual mean depth lies within 0.25 of eight. A rule is classified as degenerate when the across-pool standard deviation of its ordering-averaged depth is below 10^{-9}. The display is a post-hoc best-in-family summary, with selection and reporting performed on the same cohort.

The original 5,000 bootstrap draws resample *pools* after averaging over orderings; the 200 orderings are within-pool repetitions. Bootstrap streams are keyed by the statistic/direction tuple using $20260802 + \mathrm{CRC32}(\mathrm{str}(\mathrm{key}))$. For Physics-IQ, the resampling unit is the pool, whereas the headline width audit clusters by event family. The archived intervals therefore describe pool-level uncertainty in this diagnostic.

Distinct source populations. The 14B archive label "Wan2.1-I2V-14B" names the ABot-PhysWorld pool based on that backbone, distinguishing it from a vanilla-Wan baseline. The ABot and Cosmos loaders require at least 16 target scores per pool, and their declared diagnostics contain 198 pools each. These complete caches are separate from the common Wan2.2-5B 192-scene/64-family width intersection. The VideoPhy2 loader retains deep-pool entries with at least 100 scored candidates. Its archived population consists of 20 prompts with $K = 100$, distinct from the 179-prompt P&P cohort. Availability determines inclusion, making this a selected deep-pool diagnostic rather than a representative replication of that cohort. Here K denotes the stored pool capacity, while eight is the target mean stopping budget; within each dataset, the code uses the minimum available pool length.

Table 14: **Privileged target-score stopping (retrospective).** For each pool collection, the best eligible member of 12 target-score heuristics is shown at nominal $B = 8$; uniform matches its mean depth. Prefix benchmark target scores supply the privileged inputs; fee φ measures the quality-side control gap. Deployment costs are outside this retrospective comparison. Red is the diagnostic deficit; orange is the evaluation-only allocation oracle. Historical CIs use 5,000 pool-level bootstrap draws. The VideoPhy2 row uses Overall-score units; the Physics-IQ rows use IQ units.

		vs. random		vs. uniform	
Generator	Pools	Fee φ	Best	Best	Oracle
VideoPhy2, 5B ($K = 100$)	20	+0.114	+0.079 (69%)	−0.035	**+0.099** [+.032, +.167]
ABot-14B (PIQ; $K = 16$)	198	+1.621	+0.675 (42%)	−0.946	**+2.554** [+2.24, +2.89]
Cosmos-Nano (PIQ; $K = 16$)	198	+2.020	+0.985 (49%)	−1.035	**+3.340** [+2.91, +3.79]

Every target-score heuristic is compared with uniform evaluated at its *own* mean depth, so Eq. (8) is applied per rule. The archived $B = 8$ summary reports that reversing five of six statistics beats random on each pool collection, yet no eligible rule clears uniform. The signal–remaining-headroom Spearman range runs from -0.165 to $+0.901$ across the eighteen statistic–pool pairs (sealed preregistration `76a721b3`). The quality-side gap and the pool-curve curvature are empirical quantities. Bonferroni over all 132 rule–budget–generator comparisons leaves no win for $B \geq 4$. Separately, the historical finite-ordering curvature diagnostic classifies 261 of 7,504 second differences as negative and 3 as positive (largest $+0.183$), with the remaining 96% unresolved under its recorded bootstrap correction. This pool-ordering diagnostic is separate from the 132 rule comparisons and uses its own resampling procedure. At the $B = 2$ boundary, three rules survive uncorrected, two survive the within-cell correction and one survives the global one. The raw sign count says 43% of second differences are positive on the deep pool (sealed preregistration `5b4d5d75`). The resulting deficit is specific to the twelve privileged heuristics: it quantifies how much of the matched-uniform gap their allocation decisions recover.

C Refinement during generation and matched-compute controls

To test an intervention within the sampler, we translate the uncertainty-masked P&P loop of Self-Refining Video Sampling (Jang et al., 2026) to Wan2.2-TI2V-5B sampling. The standard sampler uses 60 denoising levels; P&P uses 40 base levels plus 20 fixed inner-loop predictions. Both arms use the same prompt seed and initial latent, exactly 60 denoiser evaluations and 120 CFG transformer forwards, and no target feedback. All 358 videos are hash-bound before the fixed VideoPhy2 evaluator is run. Semantic adherence (SA) and physical commonsense (PC) are its two ratings. Combined success, labeled "Joint" in the archive, denotes $\mathbf{1}[\mathrm{SA} \geq 4]\mathbf{1}[\mathrm{PC} \geq 4]$. Classifier-free guidance (CFG) combines conditioned and unconditioned predictions; each guided denoising prediction here requires two transformer calls. Thus 60 denoiser evaluations correspond to 120 such calls.

The end-to-end contrast at 60 NFEs changes both the integration grid and the inner-loop action relative to uniform 60. The added Uniform40 experiment in Appendix C.1 compares the grids separately from adding P&P to the 40-step grid. Adding the loop costs 20 additional NFEs, while P&P versus Uniform60 holds total compute fixed.

Table 15: **Original VideoPhy2 P&P audit.** Each run contains all 179 hard prompts; effects are paired percentage-point differences with 50,000 prompt-bootstrap draws. Replicas change only the prospectively frozen seed salt. P_+ is the fraction of bootstrap draws with a positive effect; sign p is the one-sided test for improvement.

Run	Standard success	Refined success	Δ success [95% CI]	P_+	sign p
Primary	5.03%	3.35%	-1.68 $[-4.47, +1.12]$	.083	.938
Replica 1	6.70%	4.47%	-2.23 $[-5.59, +1.12]$	.073	.945
Replica 2	6.15%	4.47%	-1.68 $[-4.47, +0.56]$	.047	.969
Replica 3	5.59%	5.03%	-0.56 $[-4.47, +2.79]$	.324	.726

Table 16: **Primary P&P component scores and distributions.** Both arms use 60 NFEs on the same 179 prompts. Panel (a) retains unthresholded ordinal evaluator-score units; higher is better. Paired 95% prompt-bootstrap intervals use 50,000 draws and seed 20260903, as in the original primary result. These component intervals are nominal, with no multiple-comparison adjustment. Panel (b) gives counts from the hash-matched original score CSVs; every row sums to 179.

(a) Raw-score means and paired differences

Component	Uniform60	P&P60	P&P − Uniform [95% CI]
Physical commonsense (PC)	3.592	3.453	-0.140 $[-0.257, -0.022]$
Semantic adherence (SA)	2.698	2.659	-0.039 $[-0.101, +0.022]$

(b) Unthresholded rating counts

Component	Arm	Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
PC	Uniform60	0	6	79	76	18
PC	P&P60	0	5	97	68	9
SA	Uniform60	0	69	95	15	0
SA	P&P60	0	73	95	10	1

The low primary Joint success rate is largely driven by semantic adherence: only $15/179$ Uniform60 and $11/179$ P&P samples reach $\mathrm{SA} \geq 4$, whereas $94/179$ and $77/179$ reach $\mathrm{PC} \geq 4$. In this primary paired comparison, the raw PC interval lies below zero, whereas the raw SA interval includes zero. The component scores therefore reveal a PC deficit alongside uncertainty in the Joint contrast. These outcomes are ordinal evaluator judgments, rather than continuous physical measurements. The treatment combines perturbation and an integration-grid change. The result describes P&P60 versus Uniform60 on this primary seed; replication and other trajectory-steering methods require their own evidence. The additional grid control is reported separately in Table 17.

The recovered PC/SA CSV hashes, `b7afa05c...5bfbd3` and `3d4ba0f0...ca6a5e`, match the primary result receipt. We verified the identities, paired means, and the 9 versus 6 Joint-success counts before reporting these distributions.

The preregistered success rule requires every primary criterion to pass. The primary run falls short of the action-value, uncertainty, sign, PC, SA-noninferiority, and groupwise thresholds. The standard sampler has 9–12 successes per run, P&P has 6–9, and every paired 95% interval includes zero. The robustness criterion requires at least two nonnegative replica effects and a positive mean; instead, $0/3$ are nonnegative and the mean is -1.49 pp. The runs share prompts, so uncertainty is evaluated within each run rather than treating 3×179 outcomes as independent. Compute and basic safety checks pass: runtime ratios are 0.998–1.000, motion ratios are 1.327–1.390, sharpness ratios are 0.964–1.057, and black-frame deltas are zero. The primary result SHA-256 is `bb57a1de...0a65e0d`; the three replica results are `0bdcfdcf...55050f6`, `bfacd0ee...bd7e1`, and `3c263e99...4a4740a`.

C.1 Feedback-driven three-arm re-evaluation

We re-evaluated the archived Uniform60/P&P60 videos together with new Uniform40 videos on all 179 primary prompts to distinguish the change in base grid from the addition of the P&P inner loop.

The control retained the original generator checkpoint, recorded prompt seeds, and conditioning procedure, with 33 frames at 480×832, 16 fps, CFG scale 5, and sigma shift 5. Only the P&P inner predictions were disabled on its 40-level base grid. This feedback-driven follow-up used a protocol frozen before scoring; it supplements the original preregistration with a targeted control, rather than independently confirming a selected method.

All controls matched the frozen original initializer replay byte for byte in initialization and conditioning. The replay used the recorded checkpoint and seeds; historical latent tensors were not saved. The fixed first-prompt smoke test was reused within the original cohort, rather than added as a 180th sample. All 179 new videos were sealed before evaluation, and all 537 videos were scored with one newly frozen common evaluator. All nine contrasts use these newly rated CSVs. The evaluator's 25 checkpoint, tokenizer, preprocessing, and code assets were hash-bound before scoring; it used 32 sampled frames, BF16 evaluation without gradients, and greedy decoding (`do_sample=False`, `top_k=1`, temperature 0.001, maximum length 256). All 358 reference SA ratings and all 358 reference PC ratings were reproduced exactly. The exact rating matches establish functional agreement on the reference videos. The recovered historical asset receipt binds the generator, whereas verifying historical grader byte identity would require the complete historical grader assets. Using the shared re-evaluation avoids mixing historical and new labels.

Table 17: **Uniform40 base-grid and added-loop control.** All rows concern the same 179 primary prompts. Panel (a) reports raw evaluator-score means and actual denoiser/CFG counts per video. Panel (b) uses 50,000 paired-prompt bootstrap draws, seed 20260903; the six new 95% intervals are nominal, exploratory, and all include zero. Joint differences are percentage points; SA/PC differences are raw score units. Means and deltas are rounded independently.

(a) Matched prompt/seed population

Arm	NFE / CFG	Joint	Joint rate	SA	PC
Uniform60	60/120	9/179	5.03%	2.698	3.592
Uniform40	40/80	9/179	5.03%	2.670	3.559
P&P60 (40 + 20)	60/120	6/179	3.35%	2.659	3.453

(b) Separate grid and added-loop contrasts

Contrast	Δ Joint, pp [CI]	Δ SA [CI]	Δ PC [CI]
Uniform40 − Uniform60	0.00 [−3.91, +3.91]	−0.028 [−0.089, +0.034]	−0.034 [−0.117, +0.050]
P&P60 − Uniform40	−1.68 [−5.59, +1.68]	−0.011 [−0.084, +0.061]	−0.106 [−0.218, +0.006]

The Uniform40 minus Uniform60 contrast measures base-grid sensitivity, distinct from error against an exact ODE solution. P&P60 minus Uniform40 fixes that grid and adds perturbation, inner-loop predictions, and 20 NFEs, comparing the added loop at higher compute. P&P60 minus Uniform60 remains the original compute-matched contrast. Its raw PC difference remains -0.140 $[-0.257, -0.022]$. The separate contrasts share Uniform40 and are correlated; a total-effect interval must preserve that covariance rather than add the marginal interval endpoints.

Matching 9/179 Joint means summarize aggregate success; establishing per-prompt agreement, equivalence, or noninferiority of 40 and 60 steps requires separate evidence. The added-loop comparison with Uniform40 and the base-grid comparison both have intervals spanning benefit and harm, limiting attribution of the original deficit to either component.

Generation used four idle A800 GPUs and 7,160 denoiser NFEs (14,320 CFG forwards), plus initialization audits and evaluator work recorded separately. The generation/scoring campaign took 23.72 minutes. These figures describe the generation/scoring campaign, including parallel execution, rather than a deployment-latency benchmark. The protocol, video-seal, and result hashes are `8a601d58...2830537b`, `d138f627...fd595802`, and `c5d55607...7d28341d`. An independent CPU verifier reproduces all nine paired contrasts and verifies the reference-score identities and smoke reuse.

D DEVELOPMENTAL COMPARISONS AND SENSITIVITY CHECKS

The developmental analyses test which parts of the main result persist under changes in data, representation, and allocation. Earlier auxiliary logs remain in the source history; the comparisons here retain explicit metrics, populations, and controls.

The width study separates available quality from recovered quality through Eq. (11). The target-score stopping calculation recovers 41.6–69.0% of its matched-uniform gap, locating a further loss at allocation. The constructive reasoning comparison starts from a 73.44% one-answer greedy baseline on PRM800K. The Qwen3.5 transfer attempt and a separate chain-of-thought smoke test fall short of their validity criteria. MMLU-Pro supplies the complementary case in which confidence predicts current accuracy more effectively than the value of further sampling.

Support concentration in a broader reasoning test. An attempt to spread gains across more problems yields $+5.86$ points of best-candidate headroom, $12/256$ positive contributions, and a top-10%-trimmed effect of -0.42 points. It fails the stated validity criteria, reinforcing the need to measure where an average allocation gain originates.

Held-out action value from richer video features. A predictor using candidate-set structure obtains an advantage of $+0.27$ points with $p = .572$, whereas a choice based on the evaluation target would yield $+3.07$ points. Better choices are therefore available in this analysis, but the predictor does not reliably identify them. Attention-based transfer between source sets reaches $+1.67$ points ($p = .160$), cross-target transfer has $p = .385$, and explicit source-window action supervision has $p = .825$. None meets its specified improvement criterion. A preview using the first 8 of 30 transformer blocks is also below its improvement criterion after fitting on candidates c_{00}–c_{07} and testing once on c_{08}–c_{15}. The gap between target-score and predicted actions identifies the remaining challenge: extracting information that predicts the return on additional generation.

E HISTORY-BASED SELECTION: IMPLEMENTATION AND EVIDENCE

This section isolates what prior outcomes add to a fixed verifier. We define the available information and matched controls, then describe the benchmark-specific adapters and their measured contributions.

E.1 VERIFIER-QUALITY LAW: PROOF AND SCOPE

Let $S_{1:B}$ and $E_{1:B}$ be independent standard-normal vectors with independent coordinates, and define $Q_b = \rho S_b + \sqrt{1-\rho^2}E_b$. The selected index $b^\star = \arg\max_b S_b$ is measurable with respect to $S_{1:B}$. It follows that

$$\mathbb{E}[E_{b^\star} \mid S_{1:B}] = \sum_{b=1}^{B} \mathbf{1}[b^\star = b]\mathbb{E}[E_b \mid S_{1:B}] = 0, \tag{9}$$

$$\mathbb{E}[Q_{b^\star} \mid S_{1:B}] = \rho \max_{b \le B} S_b. \tag{10}$$

Taking the total expectation gives $\mathbb{E}[Q_{b^\star}] = \rho m_B$, where $m_B = \mathbb{E}[\max_{b \le B} S_b]$. Marginally, the Q_b are also independent standard normals, so $\mathbb{E}[\max_b Q_b] = m_B > 0$ for $B > 1$. Dividing these expectations proves Proposition 1, including negative ρ. At $B = 1$, both centered expectations vanish and the ratio is undefined. Conditioning on x and allowing a common location and a positive quality scale yields Eq. (6) with $r_a(x) = \rho_a(x)$. The result concerns a ratio of expectations; the expectation of a random finite-pool recovery ratio is a distinct quantity.

Why non-Gaussian tails change the result. The identity relies on a linear conditional mean and matched marginal order statistics. For example, let $S \sim \mathcal{N}(0,1)$, $Q = \beta S + E$ with $\beta > 0$, and let E be independent centered Pareto noise with tail exponent $\alpha > 2$. The Pearson correlation remains fixed and positive, yet score selection has expected gain $\beta m_B = O(\sqrt{\log B})$. The oracle's expectation is at least that of the largest E_b, which is $\Theta(B^{1/\alpha})$: at the maximum-noise index, the independent S_b

still has zero mean. Thus random-relative recovery tends to zero. This mathematical counterexample uses unbounded noise, whereas IQ scores are bounded.

Thresholding can have the same effect without heavy tails. With $Q \sim \mathcal{N}(0,1)$ and score $\mathbf{1}[Q > t]$ at fixed finite t, uniform tie-breaking selects a candidate whose mean converges to the finite $\mathbb{E}[Q \mid Q > t]$, whereas the oracle grows as $\sqrt{2\log B}$. Recovery again tends to zero. The outcome depends on the score–quality relation: $S = Q$ yields recovery one for any integrable quality distribution with positive headroom; for bounded $Q \sim \mathrm{Uniform}[0,1]$ and the same threshold rule with $0 < t < 1$, recovery tends to t, not zero. These examples show that width can amplify proxy error or reward hacking, while its effect on recovery depends jointly on the quality distribution and scoring rule. A finite-pool audit therefore evaluates the relevant generator, metric, score orientation, and width together.

E.2 Exact opportunity–alignment accounting

For two widths B_0, B_1 with the same reference, let $\Delta H = H_{B_1} - H_{B_0}$, $\Delta R = R_{B_1} - R_{B_0}$, $\bar{H} = (H_{B_1} + H_{B_0})/2$, and $\bar{R} = (R_{B_1} + R_{B_0})/2$. We use the symmetric midpoint product decomposition

$$\Delta Q_{\text{select}} = \Delta(HR) = \underbrace{\Delta H\, \bar{R}}_{\text{opportunity}} + \underbrace{\bar{H}\, \Delta R}_{\text{alignment}}. \tag{11}$$

Expanding both products gives the identity. This symmetric convention averages the two possible update orders and attributes the observed change algebraically; causal identification would require additional assumptions. The opportunity term differs across selectors sharing a pool because it is weighted by each selector's $\bar{R}$. If the reference changes with width, the change must be added explicitly. All quantities and their joint bootstrap draws in Table 5 use the same c_{00} reference.

E.3 Memory-value law: proof and information boundary

Add and subtract $\langle r^* \rangle_h$:

$$\langle r_\pi - r_{\pi_0} \rangle_h = \langle r^* - r_{\pi_0} \rangle_h - \langle r^* - r_\pi \rangle_h. \tag{12}$$

Taking π_0 to be $a^*_{\text{fix}} = \arg\max_a \langle r_a \rangle_h$ gives G; nonnegativity follows from $\langle \max_a r_a \rangle_h \geq \max_a \langle r_a \rangle_h$. Conditionally applying Proposition 1 at each x gives $r_a(x) = \rho_a(x)$ and therefore

$$V_{\text{mem}} = \langle \max_a \rho_a \rangle_h - \max_a \langle \rho_a \rangle_h - \langle \max_a \rho_a - \rho_\pi \rangle_h. \tag{13}$$

The headroom weights are necessary when candidate quality scales differ. For the information bound let P_h be the distribution with density $h/\mathbb{E}[h]$ relative to the context distribution, and suppose every R_a lies in a common interval of length L. For fixed c, write $P_m = P_h(R \mid c, m)$, $P_c = P_h(R \mid c)$, and let a_m, a_c maximize expected reward under these respective distributions. The tower property gives $\mathbb{E}_{m|c}\mathbb{E}_{P_m} R_{a_c} = \mathbb{E}_{P_c} R_{a_c}$. Thus the conditional *average* decision advantage satisfies

$$V^*_c = \mathbb{E}_{m|c}[\mathbb{E}_{P_m} R_{a_m} - \mathbb{E}_{P_c} R_{a_c}] \tag{14}$$

$$\leq \mathbb{E}_{m|c}[\mathbb{E}_{P_m} R_{a_m} - \mathbb{E}_{P_c} R_{a_m}] \leq L\, \mathbb{E}_{m|c}\, \mathrm{TV}(P_m, P_c). \tag{15}$$

The first inequality uses optimality of a_c under P_c; the second bounds a single coordinate with oscillation L after averaging over memory. That averaging step is essential: the realized conditional gain $\mathbb{E}_{P_m}(R_{a_m} - R_{a_c})$ can exceed $L\,\mathrm{TV}$ at fixed m. Averaging over C, then applying Pinsker and Jensen, yields

$$0 \leq V^*_h(\mathcal{M} \mid C) \leq L\, \mathbb{E}_h \sqrt{\tfrac{1}{2} D_{\mathrm{KL}}(P_m \| P_c)} \leq L \sqrt{I_h(R; \mathcal{M} \mid C)/2}. \tag{16}$$

KL divergence and mutual information use natural logarithms (nats). For example, three equiprobable one-hot reward states and perfect memory give $V^*_h = 2/3 \leq \sqrt{\log(3)/2} \simeq 0.7412$ when $L = 1$. Conditional independence $R \perp \mathcal{M} \mid C$ implies zero gross Bayes value. Learned retrieval policies incur regret, and their observation and retrieval costs further reduce net value.

Decision value and measured action-space checks. For arbitrary action utilities U_a, Bayes-optimal memory has gross value

$$V^*_{\mathcal{M}|C} = \mathbb{E}\Big[\max_a \mathbb{E}[U_a \mid C, \mathcal{M}] - \max_a \mathbb{E}[U_a \mid C]\Big] \geq 0. \tag{17}$$

Conditional Jensen proves nonnegativity; equality holds whenever one C-measurable action is posterior-optimal almost surely. Thus memory may predict outcomes without changing the optimal action, and observation, retrieval, or compute costs may make net value negative.

On label-sealed episode 29, mean random-to-oracle region headroom is 0.1386; normalized addressable value 0.4855 minus retrieval regret 0.0609 gives 0.4246 over Ordinary and reconstructs the raw $+0.0588$ region gain. Its simultaneous official7 loss shows that the regional gain trades off against the aggregate target. In the retrospective cross-model LOEO audit, history routes between frozen 5B and 14B no-history selection rules. Model choice improves over the pooled-history rule by $+0.002655$ on the seven-metric mean (task CI $[+0.000522, +0.005227]$), cyclic wrong-task routing (CI $[+0.002501, +0.015403]$), and 256 task derangements ($p = 0.003891$). Its law is 0.004024 addressable value minus 0.001369 regret. All four e20–23→e24–27 temporal replications fail, including the strict e16–19/e20–23/e24–27 train/validation/test chain, placing the initial gain within the developmental analysis.

Applying the law to frozen cross-model picks. We apply Eq. (7) to the frozen cross-world-model picks, including the no-history baseline, the fitted history correction, and a pairwise reward model trained on the same data in the comparison envelope and using RM as π_0. This finite-pool plug-in check evaluates the accounting identity directly; the conditional-Gaussian relation $r_a = \rho_a$ and information bound concern separate population quantities. The archived 5B envelope reports zero addressable value and regret, and its Semantic/RM picks coincide. On 14B, the normalized terms are approximately 0.068 addressable value, 0.109 retrieval regret, and -0.041 memory value (14-task cluster CI $[-0.144, 0.039]$). We use this precision because the full-precision decomposition receipt is unavailable and the archival entries were rounded separately. The raw PM–RM difference is -0.000705 and mean headroom approximately 0.0171; both are reported at their recorded precision.

Appendix E.4 defines score normalization and Eq. (21) combines fitted and retrieved history. For a label-free candidate descriptor $x(z)$, the reward-tier slow state fits prior within-scene outcome residuals,

$$\begin{aligned}\beta_t = \arg\min_\beta \sum_{j<t}\sum_{a<b} w_{jab}\big[(y_{ja} - y_{jb}) - (u_v(z_{ja}) - u_v(z_{jb})) \\ - (x(z_{ja}) - x(z_{jb}))^\top \beta\big]^2 + \gamma\|\beta\|_2^2.\end{aligned} \tag{18}$$

This compact outcome-trained reranker leaves the world model fixed. The ordinary pairwise-RM control is trained on the same outcomes. In the final frozen episodes 16–19 audit, the two rules are identical on 5B and Semantic Slow PM is -0.000705 below ordinary RM on 14B (CI $[-0.002508, +0.000639]$); thus both 95% upper bounds exclude a $+0.001$ gain.

E.4 Frozen verifier implementation

Denoising stability (the Proprio flow residual). The score measures the generator's residual against the known rectified-flow target after re-noising a clean latent video (Hassan et al., 2026):

$$\begin{aligned} z_{t,p} &= (1-t)z + t\epsilon_{t,p}, \qquad \epsilon_{t,p} \sim \mathcal{N}(0, I), \\ \ell_{t,p} &= \mathrm{MSE}_{m(z)}\big(v_\phi(z_{t,p}, t),\, \epsilon_{t,p} - z\big), \\ q_{\mathrm{flow}}(z) &= \sum_{t\in\mathcal{T}} w_t \bar{\ell}_t, \qquad w_t = \frac{(\widehat{v}_t + \varepsilon_{\mathrm{v}})^{-1}}{\sum_{s\in\mathcal{T}}(\widehat{v}_s + \varepsilon_{\mathrm{v}})^{-1}}. \end{aligned} \tag{19}$$

Here $\bar{\ell}_t$ and $\widehat{v}_t$ are the mean and sample-variance statistic over the perturbation outcomes. The mask $m(z)$ emphasizes moving regions, and the masked MSE is normalized by its broadcast mask mass. The archived scorer uses antithetic noise and inverse-variance weighting, with the conditioning frame clamped in the model input. The replay settings are specified in Appendix E.7. Lower residuals are preferred, so the audit uses the negative residual as the oriented utility.

Reconstruction consistency after re-noising. We re-noise z to $z_{t_0} = (1-t_0)z + t_0\epsilon$, then integrate $\dot{z} = v_\phi(z,t)$ back to zero using K Euler steps to obtain $\hat{z}_0$. The cycle energy is

$$q_{\mathrm{cyc}}(z) = \|\hat{z}_0 - z\|/\|z\|. \tag{20}$$

The first frame remains clamped. Wan2.2-5B and ABot-PhysWorld-14B use native cycle scores; Cosmos3-Nano uses analogous denoising-x_0 error. All scores are frozen, with lower values preferred. The benchmark-specific verification forwards are accounted for in Table 28.

E.5 Normalizing scores and storing permitted history

For a scalar score s_v, choose $\eta_v \in \{-1, +1\}$ so that higher values are preferred. The history-correction interface uses $u_v(z) = \eta_v(s_v(z) - \mu_{i,v})/(\sigma_{i,v} + \varepsilon)$, with mean and standard deviation computed only over the consumed candidate pool $\mathcal{C}_i^{(B)}$. This standardization preserves the base score's preferred ordering. At query t, $\mathcal{D}_t = \{(k_j, p_j)\}_{j<t}$ contains only earlier observable keys and stored outcomes. The fitted correction $x(z)^\top \beta_t$ forms the archived "slow" component, while query-specific retrieval $m_{d,i}(z)$ forms the "fast" component. Both components use history available before the current evaluation outcome. For benchmark d and base score v, the resulting score and selection are

$$u_{\mathrm{PM}}^{(d,v)}(z) = \alpha_{d,v}\big(u_v(z) + x(z)^\top \beta_t\big) + g_d(c_i)\lambda_{d,v} m_{d,i}(z), \tag{21a}$$

$$b_i^\star = \arg\max_{b \le B} u_{\mathrm{PM}}^{(d,v)}(z_i^{(b)}). \tag{21b}$$

The features $x(z)$ are evaluation-label-free; $g_d(c_i)$ gates retrieval, and the weights are fixed before evaluation. Setting $\beta_t = \lambda_{d,v} = 0$ with $\alpha_{d,v} > 0$ recovers the base selections. The correction acts on the scoring rule while keeping generator weights frozen.

E.6 Exact post-hoc history–motion readout

For Physics-IQ candidate b, let G_b be its score against the arithmetic mean of the 195 history templates in the 65 non-query families of the 198-view/66-family take-2 bank. Excluding the query family removes all three of its views. Each template contains 24 uniformly sampled 64×64 grayscale frames, stored in float16 and read in float32 for scoring. This bank is distinct from the 192-scene/64-family evaluation intersection and from PGCM's source bank. This bank definition also applies to any prospective wider-pool evaluation: using 63 history families at $B = 16$ would change the protocol.

For a consumed-pool vector $a \in \mathbb{R}^B$, let $Z_B(a) = (a - \bar{a})/\,\mathrm{SD}_{\mathrm{population}}(a)$ when that standard deviation exceeds 10^{-12}, and the zero vector otherwise. The frozen component readout is

$$b^\star = \arg\max_{b \le B} \{[Z_B(G)]_b + [Z_B(-\mathrm{motion})]_b\},$$

with unit coefficients and ties broken by the first candidate index. The archived cross-model $B = 4$ means are 36.6028437416 for this readout and 36.5526402832 for G, giving a descriptive difference of 0.0502034583 IQ. The combination was highlighted retrospectively, so the contrast is a descriptive component comparison rather than an independent test of synergy.

Testing the rule at $B = 16$ requires its picks or the complete motion/template caches, both absent from the recovered archive. Any extension must retain the same bank, family exclusions, coefficients, normalization, and tie rule. The historical PM width curve concerns a separate readout. Likewise, the archived OpenS2V template adapters and RoboTwin residual PM controls test their own interfaces; transfer of this exact $G + M$ rule remains untested.

E.7 Verified qualitative CVA-Select replay

Figure 2 uses a newly generated pool and the exact $G + M$ rule in Appendix E.6. The historical four-candidate videos and corresponding selection stage were unavailable, so we fixed 16 center-view conditions and regenerated four candidates per condition. The generator was frozen Wan2.2-TI2V-5B, with 81 frames at 832×480, 16 FPS, 20 denoising steps, CFG 5, sigma shift 5, and seeds $0, 1, 2, 3$. Each video used 20 scheduler evaluations and 40 observed DiT forwards. Both candidate identities and effective prompts are bound to generation receipts.

We first used the recovered runner's generic prompt and then ran one separately frozen input variant with the unedited official scene descriptions. Both variants used the same ordered conditions, model, seeds, sampler settings, history bank, and selection functions. The text-conditioned variant supplies the displayed example; neither variant replaces the historical benchmark populations or table estimates. The 198 take-2 history videos match the archived input hashes. Their 24-frame, 64×64 grayscale templates were reconstructed with the original float16 storage convention; each query excludes all three views of its event family. The template score, pixel-motion statistic, normalization, and tie rule match the original functions.

The original G, low-motion, and CVA-Select choices were hash-sealed before target scoring. We evaluated every candidate with the original per-view Physics-IQ composite and SHA-identical official mask and metric implementations, using the first 80 frames and both real takes. Original c_{00} scores reproduced exactly in the calibration checks. The generic-prompt campaign contains 64 videos, frozen in batches of 8, 6, and 2 pools as GPU capacity became available. The official-description campaign contains another 64 videos with all 16 pools sealed together before scoring. Each campaign costs 1,280 scheduler evaluations and 2,560 observed DiT forwards; combined generation-container time is 1.42 GPU-hours.

In the official-description cohort, CVA-Select improves on no-TTS in 9 conditions, ties in 5, and loses in 2. It ties global history in 13 conditions and loses in 3; its choices coincide with low motion throughout. These controls distinguish a recovered selection gain from a benefit uniquely attributable to combining the components. The displayed ball-in-sand case retains the original CVA-Select candidate c_{01} (IQ 75.39). Denoising stability selects c_{00} (56.12), a difference of $+19.26$ IQ computed from the unrounded scores. No-TTS also returns c_{00}; global history, low motion, and reconstruction consistency return c_{01}. We inspected all 81 frames of each displayed generated video. The denoising selection repeatedly resets or loses the ball and does not retain it on the sand. Ours releases the ball late, near 3 seconds, but preserves the settled ball through the end. Early transient artifacts remain; all displayed frames use the original 0, 1.5, 3, and 4.5 s timestamps.

Matched-pool denoising-stability baseline. We recovered the original `ProprioScorer.score` implementation used by the five-billion-parameter generation runner. Its source and the runner's encoding and model-call adapters are unchanged. The scorer VAE-encodes the first 17 decoded frames of each frozen MP4, uses noise levels 0.35 and 0.5, and takes two antithetic perturbations per level with seed 0. The motion-mask floor is 0.05; inverse-variance weighting uses $\varepsilon_{\mathrm{v}} = 10^{-8}$. All operations follow the original BF16 scoring path. The first latent frame is clamped in the model input, with the original residual mask retained. No pruning, refinement, or generation is invoked. The scores for c_{00}–c_{03} are 0.417969, 0.484375, 0.464844, and 0.476562, respectively. Their actual minimum is c_{00}. Each candidate costs four DiT forwards, giving 16 verification calls on top of the shared pool's 160 generation calls. The displayed baseline therefore costs 176 DiT forwards versus 160 for CVA-Select, at the same candidate budget. The baseline was added after the original IQ evaluation, without IQ inputs to its scoring or selection; all four scores and the pick were sealed before joining the existing IQ values. The source and result SHA256 prefixes are `88e29df12ba9` and `8103f58b7d1f`. Full-precision scores and configuration are retained in `fig_denoising_b4_record.json`.

Retained reconstruction controls. We score all 16 official-description pools with the native reconstruction-consistency verifier, reusing the completed ball-in-sand result and evaluating the other 15 pools without new generation. The candidates, prompts, Ours choices, and verifier settings remain fixed. The recovered generated-pool scorer VAE-encodes the first 17 frames, re-noises at $t = 0.6$, and averages relative reconstruction drift over two noise draws from seed 1234, each with five Euler steps. The first latent frame is clamped in each model input and excluded from the drift. For ball-in-sand, the scores for c_{00}–c_{03} are 0.380225, 0.354058, 0.372304, and 0.361776, respectively. The minimum selects c_{01} at 75.39 IQ, tying Ours. We retain this tie as a separate matched-pool control. The scoring functions match the recovered implementation, and all model and recorded runtime identities match the original A800 generation environment. This baseline was added after the original IQ evaluation, with no IQ inputs to its scoring or selection; its new score/pick record was sealed before joining the existing candidate-IQ values. Its 10 verification forwards per candidate add 40 calls to the shared pool's 160 generation forwards. This reconstruction control matches candidate budget, while its full procedure costs 200 DiT forwards for reconstruction versus 160 for G or CVA-Select. The 15

new pools cost 600 verifier forwards; all 16 pools total 640. The ball-in-sand reconstruction result SHA256 prefix is `efcf5fd85b69`.

Across all 16 pools, Ours wins in 6, ties in 5, and loses in 5 against reconstruction. Mean IQ is 34.65 versus 35.67, a difference of -1.01. The six distinct higher-IQ selections were visually screened in the preceding figure audit. The full comparison, including every tie and loss, is retained in `reconstruction_cohort_records.json`. These full-cohort measurements remain separate from the single-pool denoising comparison.

A native H200 portability check used another 40 verifier forwards. It preserved the calibration pool's selected index and full ranking, but its maximum score difference was 0.003651, exceeding the predeclared 0.001 tolerance. We retained this failed check and used the original A800 runtime for every reported baseline value, without changing the tolerance or verifier settings. Total reconstruction work, including this portability check, was 680 DiT forwards. The additional denoising check brings supplementary verification work to 696 DiT forwards, with no new generated videos.

The official-description selection stage and result have SHA256 prefixes `7d3a1a5e58f6` and `bf0b33f803ce`; its generation software identity is `01c0ec3d5987`. Figure frames are taken at 0, 1.5, 3, and 4.5 seconds from the sealed videos. The only publication transform is full-frame, aspect-preserving resizing; the frame trace records source-video identities, selected indices, and decoded-pixel hashes.

Physics-IQ: observed-outcome memory. For same-stem repeated setup, a previously observed take-2 real outcome $\hat{T}_i$ enters through

$$m_{\mathrm{PIQ},i}(z) = Z[\mathrm{IoU}(M(z), M(\hat{T}_i))] - Z[d_{\mathrm{app}} + d_{\mathrm{temp}} + d_{\mathrm{motion}}]. \tag{22}$$

Physics-IQ take-2 supplies $\hat{T}_i$ and take-1 remains the evaluation realization. Both the history key and query key are computed from the same canonical c_{00} conditioning file for each stem; the resulting $198/198$ match checks that key construction is consistent. It does not compare the take-1/take-2 first frames. The paired-future reference measures the selection value of additional predictive information from the same setup. Family-held-out retrieval is worse than a train-global history control.

The real-future reference exceeds reconstruction consistency by $+4.68/+4.12/+3.25$ IQ, with paired 95% CIs $[+2.94, +6.53]/[+2.38, +6.15]/[+1.82, +4.71]$. A preregistered combination of denoising stability and history lowers the real-future-only reference by $1.71/2.24/2.15$ IQ. The exactly paired future supplies more useful information than retrieval from other families: family-held-out retrieval has a cross-world-model delta of $+0.44$ with CI $[-0.77, 1.57]$ and is worse than the train-global outcome control by -1.87 IQ with CI $[-3.28, -0.64]$.

PAI selection without history. PAI-Bench-robot evaluates perceptual consistency and conditioning faithfulness alongside plausible motion. The baseline without history ($\beta = 0, \lambda = 0$) provides the reference for testing any history correction. For the same $B = 8$ candidates, its score s_v sums four signals standardized within the scene:

$$q_{\mathrm{PAI}}(z) = Z[-q_{\mathrm{cyc}}(z)] + Z[q_{\mathrm{motion}}(z)] + Z[q_{\mathrm{consist}}(z)] + Z[q_{\mathrm{i2v}}(z)]. \tag{23}$$

The last three terms are inexpensive video and conditioning statistics that use neither an external VLM nor PAI labels. The ablation removes q_{i2v}. The no-history ensemble rows in Table 29 select directly with this score. History-based corrections are compared against this same baseline and reported separately. The cheap selector feature q_{i2v} is not the official `i2v_background` outcome withheld from seven-dimensional fitting.

PAI/RoboTwin: learning corrections from past outcomes. Here the selector uses past candidate outcomes rather than a same-setup real future. Equation (18) fits a correction from those permitted outcomes. Shuffled outcomes and wrong-task records test whether the relevant history matters. The studied cross-fitted corrections, transfer variants, fresh-episode adaptation, and event-based models do not meet all their prespecified improvement requirements.

A preregistered episodes 8–10 $\rightarrow$ 28 study improves the held-out transition-region score over ordinary RM by $+0.075982$ (CI $[+0.028359, +0.129885]$). It nevertheless misses the wrong-task, global-history, strict-win, and official7-noninferiority criteria. A subsequent integrity check found that

72/112 candidate MP4 hashes matched an earlier pool whose official7 scores had been available before these selections were fixed. We therefore classify episode 28 as a developmental diagnostic.

The history-calibrated stopping study, originally called Conformal Memory Stopping, proceeds through three variants. The numeric-feature version still uses all eight candidates. Adding a bound on selection risk reduces the mean count from 5.964 to 5.571 (global 5.982), meets the quality noninferiority test against B8, and beats wrong-task memory with a positive interval. Its gain over global history is nevertheless $+0.000732$ with CI $[-0.001756, +0.003533]$, below the shuffle distribution's 97.5th percentile. Worst-case quality and best-candidate-hit checks also fail, and a separately calibrated reliability filter further attenuates the benefit. An evaluation-time oracle choice between two safe options saves 0.518 candidates, and choosing a safe stopping prefix with target-score access saves 2.964. These oracle comparisons quantify the opportunity remaining above the realizable stopping rules.

Adding prefix-content features—verification scores, latent features, and temporal coverage—reduces the global rule's mean count from 5.982 to 5.429 candidates, raises estimated quality by $+0.001746$, halves the violation rate from 16.1% to 8.9%, and increases oracle hits from 12 to 18. However, the quality and compute-aware CIs narrowly cross zero, discrimination measured by area under the receiver-operating characteristic curve rises only $+0.0045$, and the original two-episode task-history residual worsens the content-aware variant. A revised correction uses six leave-one-episode-out residuals per task, retains only positive corrections with a stable sign, and shrinks estimates toward a pooled mean. This reduces the mean candidate count from 6.321 to 5.964 and changes the compute-aware development estimate to $+0.001314$ (CI $[-0.000230, +0.003191]$), outperforming 256 task derangements ($p = .007782$). Its interval against wrong-task history still includes zero ($+0.000874$, CI $[-0.000753, +0.002456]$).

A check conducted before opening test outcomes favors the task average over three-neighbor content retrieval: the latter's out-of-fold Brier score is 0.167102, compared with 0.164665 for the task average, where lower values indicate more accurate probability predictions. The revised correction improves the development point estimate, but falls short of the representation and task-specificity criteria for proceeding to the planned episode-30 test.

The replacement episode-29 evaluation uses a fresh, outcome-blind pool: 112 unique candidate hashes and 14 source hashes with zero overlap against both previously evaluated episode-28 pools. The score focused on the task-relevant region improves over the pairwise reward model by $+0.058848$ (CI $[+0.016109, +0.103457]$), beats wrong-task memory (CI $[+0.006397, +0.096932]$), and beats 256 shuffled task assignments ($p = 0.003891$). Yet it remains indistinguishable from pooled success memory ($+0.000866$, CI $[-0.019748, +0.019357]$) and reduces official7 versus the pairwise reward model by -0.007112 (CI $[-0.011187, -0.003448]$). The regional gain comes with an overall-quality loss and no resolved advantage over pooled success memory.

The final fixed-selection audit covers all 14 tasks, episodes 16–19, and 56 scenes per generator. On 5B, both Semantic Slow PM and ordinary RM improve on the no-history baseline by $+0.005805$, but return exactly the same selections and means (history correction minus reward model 0, CI $[0, 0]$). On 14B, the history correction minus no-history baseline is $+0.001086$ (CI $[-0.001728, +0.004211]$), whereas PM–RM is -0.000705 (CI $[-0.002508, +0.000639]$). An additional constrained correction, archived as "Joint-Null Cone," is inconclusive on 5B ($+0.000106$ vs the baseline, CI $[-0.000994, +0.001209]$). Its positive 14B result belongs to a separate exploratory arm; the aligned and cross-world-model criteria still fail. The sealed result SHA256 values are `fa7495f0...dcc33d` (matched-RM control) and `5c2f0528...1813bd` (Joint-Null Cone).

Choosing a generator from the current input. Training on episodes 16–19 and validating on episodes 20–23 tests six representations: first-frame image embeddings (GME); 1370 robot, scene, and image features available at generation time; a principal-component compression fitted on training data; 107 foreground-geometry features; responses from each model's first candidate; and a model ranking constrained to the same compute allocation. None meets the criterion after accounting for the number of alternatives tested. Choosing the better model with evaluation-score access could correct nine cases across seven tasks, but the available mean improvement is only 0.001275. A target gain of $+0.001$ would require recovering 78.4% of it. If exactly 16 of 56 queries use 5B, choosing those queries with evaluation scores improves on task-based routing by $+0.001093$ (CI

$[+0.000090, +0.002504]$), with 21.77% savings in the recorded compute proxy. The tested history model changes none of the assignments.

For matched positions within a task, agreement between past and current candidate rankings is 0.607 ($p = .0817$). After removing task-average effects, the best of seven feature mappings has correlation 0.081, below the 97.5th percentile of the multiple-comparison null, 0.306 ($p = .704$). Even that mapping loses 0.001528 to using the task alone ($p = .611$): it corrects five model choices but makes four previously correct choices wrong. Geometry corrects five and harms seven; each tested model preview corrects three and harms seven. The feature comparisons use training and validation data only; episodes 24–27 remain sealed for this analysis.

Table 19 is the complete main-text claims ledger; Appendix L reports the earlier diagnostics.

Table 23: **RoboTwin task-disjoint diagnostics ($B = 8$).** Official7 over 112 scenes/14 tasks; task-bootstrap CIs; oracle is evaluation-only. Slow − RM: $+0.000445$ $[-0.000209, +0.001371]$.

Method	B	Off.7	Δ zero	95% CI	Regret	W/T/L
Baseline (no-TTS)	1	0.653234	−0.001080	–	0.024388	–
Baseline without history	8	0.654314	0	–	0.023308	ref.
Pairwise RM[†]	8	0.660807	+0.006493	[+0.003252, +0.010368]	0.016815	21/91/0
Fitted history correction[‡]	8	**0.661251**	**+0.006937**	[+0.003430, +0.011082]	**0.016371**	21/91/0
PM + pretrain	–	–	–	not measured	–	–
Oracle (eval. only)	8	0.677622	+0.023308	–	0	–

[†]Post-outcome control; relocation outcomes excluded from fit/picks. [‡]Earlier outcome-blind diagnostic, superseded by the fresh-episode audit. Per-task: Table 33.

PM audit: held-out PM–RM $0/-0.000705$ (5B/14B); paired-future not measured; $\Delta_{\text{paired}-\text{heldout}}$ not measured.

OpenS2V-Eval: changing samples or their scores using history. The six-condition pilot tests two uses of earlier samples: shifting a sample's latent representation toward a previous high-scoring example, and correcting its selection score from outcomes generated with other seeds. Both fall short of the shuffled-history criterion. Table 31 reports the complete pilot comparisons.

F Prospective pretrained global capability memory

The next design learns one task-agnostic memory on a public source distribution, freezes it, and transfer it without target outcomes. We preregister *Pretrained Global Capability Memory* (PGCM). For each frozen model, two seeded first-step, no-decode probes yield 27 features; concatenating the 5B and 14B vectors and their difference gives 81. Appendix F.1 enumerates all features and the eight actual model calls per paired query. Source-trained ridge regression predicts the difference between the 14B and 5B no-history selections. We use this pooled predictor as the comparison model, isolating the additional value of retrieving examples. The memory alternative uses one shared public bank

$$\mathcal{M}_{\text{pub}} = \{(k_i, v_i)\}_{i=1}^{192}, \qquad k_i = P_d \,\text{Std}(z_i), \quad v_i = Q_i^{14\text{B}} - Q_i^{5\text{B}}, \tag{24}$$

where $z_i \in \mathbb{R}^{81}$ is the paired probe. A query retrieves at most one view per source family and uses a Gaussian-weighted mean of the nearest stored values. PCA dimension $d \in \{8, 16, 32, 64, 81\}$, distinct-family neighbor count $K \in \{4, 8, 16, 32\}$, and temperature $\tau \in \{0.25, 0.5, 1, 2\}$ are chosen by nested leave-one-source-family-out prediction MSE. The protocol calls for refitting the transform and bank on all 64 public families before target use. The selected tuples and action maps are absent from the recovered records. The stored memory includes no task identifier, prompt-text feature, target history, target example, or target label.

F.1 Exact early-denoise feature construction

The frozen extractor uses seeds 314159 and 271828 independently for each model. For one seed, we capture the initial latent z, the conditional and unconditional model outputs p, u, and the updated latent z' after the first scheduler step. Using the recorded CFG scale γ, we set $v = u + \gamma(p - u)$

and $\Delta z = z' - z$. The hook stops before the next model evaluation or video decode; the two model evaluations, conditional and unconditional, have already run. Per-run receipts specify the actual scheduler timestep, nominal schedule, dimensions, CFG scale, and conditioning hashes.

For a $[N, C, T, H, W]$ latent, $D(X)$ omits the conditioning frame when $T > 1$. Let $R(X) = \sqrt{\text{mean}(X^2)}$, $E(X) = \max(R(X), 10^{-8})$, and let $\cos(X, Y)$ be the cosine of the flattened tensors, returning zero if the product of their Euclidean norms is at most 10^{-12}. Define temporal roughness $T(X) = R(\Delta_t D(X))/E(D(X))$ and spatial roughness $S(X) = \text{mean}_{a\in\{H,W\}} R(\Delta_a D(X))/E(D(X))$, using only available axes. Temporal roughness is zero when fewer than two frames remain; spatial roughness is zero when no axis admits a difference. For per-channel RMS values $r_c = R(D(X)_{:,c})$, let $C(X) = \text{SD}_{c,\text{sample}}(r_c)/\max(\text{mean}_c\, r_c, 10^{-8})$.

Table 24: **The thirteen per-seed PGCM probe statistics.** The ordering is the extractor's ordering. t is its recorded model timestep, and $\Delta z_{:,:,0}$ is the first-frame update. Tensor reductions use float32 before conversion to the float64 feature vector.

Index	Statistic	Definition
1	Normalized timestep	$t/1000$
2	Dynamic latent RMS	$R(D(z))$
3	Latent temporal roughness	$T(z)$
4	Guided velocity RMS	$R(D(v))$
5	Guided velocity absolute mean	$\text{mean}\,\lvert D(v)\rvert$
6	Guided temporal roughness	$T(v)$
7	Guided spatial roughness	$S(v)$
8	Guided channel variation	$C(v)$
9	CFG gap ratio	$R(D(p-u))/E(D(v))$
10	Conditional/unconditional cosine	$\cos(D(p), D(u))$
11	Relative update	$R(D(\Delta z))/E(D(z))$
12	Update/velocity cosine	$\cos(D(\Delta z), D(v))$
13	First-frame update ratio	$R(\Delta z_{:,:,0})/E(D(\Delta z))$

If the two 13-vectors are $f^{(1)}, f^{(2)}$, the model's 27 features are

$$F = \left[(f^{(1)} + f^{(2)})/2, \quad |f^{(1)} - f^{(2)}|, \quad \cos(f^{(1)}, f^{(2)})\right].$$

The features are thirteen seed means, thirteen absolute seed differences, and one seed-feature cosine, in that order. The paired 81-vector is $[F_{5B}, F_{14B}, F_{14B} - F_{5B}]$. The query probes for both models therefore require $2\ \text{seeds} \times 2\ \text{CFG evaluations} \times 2\ \text{models} = 8$ DiT evaluations: four 5B and four 14B calls. Conditioning encoders and memory processing add inference work beyond this decode-free DiT count.

Hyperparameter search and action-map provenance. The recovered analyzer searches $5 \times 4 \times 4 = 80$ (d, K, τ) settings using nested source-family-held-out prediction MSE. It fits normalization and PCA on the corresponding training split, keeps at most one view per retrieved family, and assigns the K nearest families weights proportional to $\exp[-\|k - k_j\|_2^2/(2\tau^2)]$, then normalizes them. Near-minimum losses are identified with NumPy `isclose(loss,best,atol=1e-15)`; ties favor smaller d, then larger K, then larger τ. Each outer family's tuple is selected by inner cross-validation; the full-source refit selects its own tuple.

The primary source action rule selects 5B exactly when the predicted 14B-minus-5B advantage is negative; otherwise, it selects 14B, including ties. The 5B allocation follows the predicted signs rather than a fixed quota and differs from the clipped $1/192$ allocation used only in the later post-stop diagnostic.

We report the search specification. The per-budget `SOURCE_RESULT.json` and `SOURCE_RESULT_B8_TRANSFER.json` records are needed to verify selected (d^*, K^*, τ^*) values and the memory/ridge out-of-fold action maps; these records are absent from the recovered material. Equal average quality is compatible with different actions. The zero-effect source entries are historical negative diagnostics whose mechanism has yet to be independently reproduced.

On nested source OOF predictions, the memory rule must improve on ridge regression by at least $+0.25$ IQ, with a family-bootstrap lower bound above zero. It must also exceed the 97.5th percentile of 256 family-value shuffles and 256 family key/value mismatches. These controls distinguish the shared bank's contribution from a transferable regression fit.

PAI is selected before quality scoring because it changes event semantics (natural interactions to robot manipulation), scene/camera composition, evaluation target (IQ to official8 perceptual and conditioning quality), and visibility (public videos to private condition images). RoboTwin is public and is therefore ineligible for the private-target claim. "Far OOD" additionally requires zero image/task overlap and two independent outcome-free audits: both a frozen DINOv2 ViT-L/14 condition-image encoder (fixed 518 center crop, 1024-D CLS token, frozen revision/weight hash) and the 81-D denoise probe must obtain group-held-out domain AUC ≥ 0.80 and group-mean linear-MMD permutation $p \leq 0.01$. Failure of either audit removes the far-OOD claim regardless of quality.

The primary target controls are global ridge, source-majority routing, equal-quota random routing, both memory shuffles, always-5B/14B, and an equal-quota oracle. Global memory must improve both ridge and source-majority by at least $+0.0005$ official8 with positive paired uncertainty, pass the matched nulls, remain noninferior to always-14B, increase oracle hits, and save compute after charging both probes. The full frozen gates are in `eval/PREREG_pretrained_global_capability_memory_20260720.md` and its two amendments, including `AMENDMENT_pgcm_shared_memory_ood_ladder_20260720.md`.

Outcome-free probes cover all 192 public source scenes. Both source-data tests yield zero estimated memory-minus-ridge gain. In the separately specified $B = 4$ analysis matching Table 1, the memory-minus-ridge difference is $+0.00$ IQ (95% CI $[0.00, 0.00]$), with value-shuffle $p = 0.9455$ and key/value-mismatch $p = 0.9728$. In the separate $B = 8$ source-data test, memory minus ridge is again $+0.00$ IQ (95% CI $[0.00, 0.00]$), with value-shuffle $p = 0.8911$ and key/value-mismatch $p = 0.8444$.

Both primary results use the frozen analyzer's summed squared Euclidean distance in the selected PCA space; its code hash is bound in the analysis input receipt. The B4 source comparison stops at the memory-specific improvement criterion. A separate exploratory follow-up with a clipped 1/192 5B quota scores 41.56 and differs from held-out 14B PM by -0.36 IQ (95% CI $[-1.80, 1.06]$). For OOF 5B rate r, this diagnostic sets $q = \mathrm{clip}(\lfloor 192r + 0.5 \rfloor, 1, 191)$ and selects the lowest q OOF predictions, with scene-ID tie breaking. The quota belongs to this follow-up alone. The source zero-effect results report quality, not action identities: without the original route maps, a $[0, 0]$ interval cannot determine whether memory and ridge chose identical actions. The completed tests use held-out families within the public source domain. Since private PAI labels were previously inspected, the target confirmation protocol requires a new owner-held private split of at least 100 scenarios.

G Memory-conditioned seed screening and velocity diagnostics

The preceding experiments rank completed candidates. This section asks whether history can save generation compute earlier, while preserving natural-noise sampling and guarding against off-manifold optimization.

Select natural noise; do not rewrite it. For deterministic sampling $z^{(b)} = F_\phi(x, \epsilon^{(b)})$, selecting a completed candidate also selects its initial seed, but only after paying for the rollout. We preregister the compute-saving variant

$$\epsilon^{(1:M)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I), \quad h_\tau^{(b)} = F_{\phi,0:\tau}(x, \epsilon^{(b)}), \quad \mathcal{S} = \mathrm{Top}_B\, \widehat{R}_\tau(h_\tau^{(1:M)}, \mathcal{D}_t), \tag{25}$$

and complete only $b \in \mathcal{S}$. Memory predicts a residual over a matched history-blind preview ranker; random-screen, wrong-memory, and 256 shuffled-history controls receive identical proposals, checkpoints, and NFE. Every $\epsilon^{(b)}$ remains an independent production-prior draw. The fresh matched-seed comparison motivates this restriction: continuous shaping has an estimated official7 difference of -0.000454 (CI $[-0.002209, +0.001481]$) from the unshaped control, while direct noise-affine optimization and averaging exhibit off-manifold Goodhart behavior. Equation (25) defines the prospective experiment; its frozen gates are in `eval/PREREG_memory_conditioned_seed_screening_20260720.md`.

Verifier target. An ideal intermediate verifier predicts terminal utility, e.g., the Feynman–Kac potential

$$R^*_\tau(h, c) = \log \mathbb{E}[\exp\{\lambda R(z_0, c)\} \mid h_\tau = h, c], \tag{26}$$

whose small-λ limit ranks conditional expected reward (Singhal et al., 2025). This is the diffusion analogue of a process verifier, evaluated by held-out pairwise ranking of final outcomes. Search amplifies proxy error (Gao et al., 2023; Ma et al., 2025), so the protocol requires a matched ordinary-preview ranker, positive held-out alignment, history shuffles, and motion/sharpness/diversity diagnostics.

Why smaller late velocity is not generally better. For an interpolant

$$x_t = \alpha_t x^* + \sigma_t \epsilon, \qquad v(x, t) = \mathbb{E}[\dot\alpha_t x^* + \dot\sigma_t \epsilon \mid x_t = x], \tag{27}$$

the field depends on the path and schedule (Lipman et al., 2023; Ma et al., 2024). Under a differentiable time change $s = h(t)$ with $h'(t) > 0$, the same trajectory has $\widetilde v(x, s) = v(x, h^{-1}(s))/h'(h^{-1}(s))$, so its norm can change without changing the sample. For $\alpha_t = 1-t, \sigma_t = t$, the conditional target is $\epsilon - x^*$ and need not vanish at the data endpoint. Small raw velocity can therefore mean a slow schedule, a static mode, or local convergence.

The proposed diagnostic compares raw velocity with the scheduler-scaled update $\|\Delta x\|/(\|x\| + \varepsilon)$ and a one-step/two-half-step defect. To make the numerical operator explicit, for forward Euler define $\Phi_{t,h}(x) = x + h\, v_\phi(x, t)$, with signed scheduler step $h = t_{k+1} - t_k$. Its original displacement-normalized form is

$$d_{\text{step}}(t_k) = \frac{\|\Phi_{t_k,h}(x) - \Phi_{t_k+h/2,h/2}(\Phi_{t_k,h/2}(x))\|}{\|\Phi_{t_k,h}(x) - x\| + \varepsilon}. \tag{28}$$

This ratio is unstable when displacement is near zero. A fully specified *prospective* sensitivity check replaces its denominator by $\max(\|\Phi_{t_k,h}(x) - x\|, 10^{-3}\|x\|) + 10^{-8}\sqrt{D}$ for a D-dimensional latent, evaluates every scheduler interval, and reports the median, maximum, and fraction of intervals activating the state-magnitude floor. The full and first half steps share the base velocity; the second half step adds one NFE per interval, which must be charged (including CFG). This prospective diagnostic would be compared with the scheduler-scaled update using a predeclared, multiplicity-corrected test of held-out terminal-quality value. The frozen experiments retain their original operators.

H Reproducibility Details

This section links the headline comparisons to their candidate sets, selection files, uncertainty procedures, and result hashes. We organize the records by experiment so each claim can be traced from its protocol to the reported number.

Broader positioning. Reasoning-time search includes repeated sampling, rank aggregation, tree search, and world-model planning (Brown et al., 2024; Jiang et al., 2023; Yao et al., 2023; Hao et al., 2023; Schrittwieser et al., 2020; Hafner et al., 2019; Chua et al., 2018). Video generation and evaluation span Wan, VideoPhy, learned rewards, physics-aware steering, and sketch verification (Team Wan et al., 2025; Bansal et al., 2024; Xu et al., 2023; He et al., 2024; Xu et al., 2024; Liu et al., 2025b; Yuan et al., 2026; Huang et al., 2025). Persistent-state alternatives include test-time adaptation, retrieval, episodic memory, and self-training (Sun et al., 2020; Wang et al., 2021; 2022; Yuan et al., 2023; Iwasawa & Matsuo, 2021; Sheynin et al., 2022; Blattmann et al., 2022; Pritzel et al., 2017; Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2023a; Zelikman et al., 2022; Gulcehre et al., 2023). Baseline-relative decision rules are related to A-learning, residual control, safe policy improvement, and reward mixtures (Schulte et al., 2014; Johannink et al., 2019; Thomas et al., 2015; Laroche et al., 2019; Ramé et al., 2023); CVA differs by auditing a frozen finite-pool compute intervention rather than updating a policy or model.

The Physics-IQ comparisons use deterministic candidate seeds. A result at $B = 4$ consumes candidates c_{00}–c_{03}; a result at $B = 16$ consumes c_{00}–c_{15}. Every consumed candidate is scored by the selector. PM picks are materialized by `eval/table1_flow_local_template.py`: the 198-entry take-2 bank and retrieval gate are hash-bound, picks are frozen without official IQ, and take-1 IQ is opened only in the analyze stage. The correlation-to-headroom analysis is implemented in `eval/verifier_law_heldout.py`. Five fixed folds cover the union of 66 event families from

the complete per-model populations. Each held-out fold compares empirical alignment and recovery with the Gaussian prediction. The script specifies its slope, intercept, R^2, and MAE criteria before computing the result. Uncertainty uses 5,000 seeded bootstrap draws that resample event families within the frozen folds and retain all world-model–verifier–budget points for each sampled family. This population differs from the headline complete-16 intersection: requiring complete coverage retains 192 scenes/64 families from the original 198 scenes/66 families. The calibration statistics use the original family population; the 64-family headline subset has the additional complete-coverage restriction. The recovered Wan2.2-5B target cache contains 3,153/3,168 scores: 15 missing late-candidate scores affect all six views of `weight-on-pillow` and `weight-protects-duck`. Dropping those six incomplete views yields the 192×16 headline matrix.

Empirical correlation convention. For centered pool vectors s_i, q_i, the archived implementation sets $\widehat{r}_i = s_i^\top q_i / (\|s_i\|_2 \|q_i\|_2)$ when $\|s_i\|_2 \|q_i\|_2 > 10^{-12}$, and zero otherwise. The average in Eq. (5) retains these zero-imputed pools. Zero is an implementation convention for an undefined sample correlation. Even for nondegenerate Gaussian pools, sample correlation is attenuated toward zero at small B:

$$\mathbb{E}[\widehat{r}_i] \simeq \rho \left(1 - \frac{1-\rho^2}{2(B-1)}\right).$$

At $B = 4$ the leading relative attenuation is $(1-\rho^2)/6$. We retain the recorded sample statistic without a post-hoc bias correction; Proposition 1 uses the population ρ. Heterogeneity adds another distinction: under the conditional Gaussian model, population aggregate recovery is $\langle \rho(x) \rangle_h$, not the unweighted mean of pool correlations. The archived unweighted $\widehat{\rho}$ describes average within-pool alignment, while calibration measures how well it predicts aggregate recovery.

The motion-strata result is materialized by `eval/physiq_recurrent_pm_motion_strata.py`: the median split is over ground-truth event-family motion and is independent of method picks, and all CIs bootstrap the 33 held-out families rather than individual views. The post-hoc disjoint-family calibration audit is materialized by `eval/verifier_law_calibration_split.py`: a SHA256 rule assigns 27 and 39 of the 66 event families to calibration and evaluation, respectively; all families had entered the earlier full-law analysis, while evaluation labels do not estimate ρ in this check. It gives MAE 0.0697 and Pearson 0.802, while leave-one-world-model-out MAE is 0.146. The more-than-twofold increase in error favors calibrating alignment on held-out data from the generator being used. The result SHA256 is `1662c6c9...259d0`. The fixed-compute frontier is materialized by `eval/verifier_compute_frontier.py` on the 192-scene/64-family complete-16 intersection. It reports strict DiT NFE and separately scoped shared-pipeline GPU-seconds with 5,000 family bootstrap draws; result SHA256 is `956e1f0b...d9f51`. The headline width deltas and PAI comparisons are materialized by `eval/paper_paired_uncertainty.py`. It first reproduces every displayed mean from frozen candidate identities, then uses 50,000 paired event-family draws for the 192-scene Physics-IQ complete-16 intersection and paired scene draws for all 174 PAI scenarios; result SHA256 is `06f894ee...b31f4`. The previously omitted denoising-stability and video-preference width intervals in Table 7 are documented in `paper/CLAIM_EVIDENCE_LEDGER_20260730.json`: respectively -0.746 IQ $[-2.290, +0.801]$ and -0.592 $[-2.448, +1.280]$. Their original 50,000-draw paired-family bootstrap seeds are 20260718 and 20260722. These are recovered archival estimates; the oracle interval is independently recomputed from the target cache. The joint opportunity/alignment decomposition uses common seed 20260722, whereas the original standalone reconstruction-score contrast uses seed 20260719. The small interval difference reflects Monte Carlo variation. We report the joint run's total and component intervals together. The PAI result is materialized by `eval/pai_table2_n8.py`; its frozen artifact records $B = 8$, 174 scenes, all picks, input hashes, and complete official eight-dimensional scores. Its label-free table picks precede official scores. The displayed PAI PM+history row uses preregistered five-fold cross-fitting over the same 174 PAI scenes. In each fold, only the other four folds' PAI seven-dimensional outcomes (all official dimensions except I2V-B, `i2v_background`) may write memory; held-out picks freeze before their labels are opened, and no Physics-IQ inputs are loaded. The official eighth-dimension I2V-B audit reuses exact candidate-identity scores for 347/349 required candidates and evaluates only the two remaining frozen picks; `eval/pai_fulln8_memory_crossfit.py` then verifies that the zero-memory control exactly reproduces its Table 6 row before computing the paired historical comparison. The developmental selector repair is materialized by `eval/pai_pairwise_pm_crossfit.py`: all 1,392 candidate videos receive temporal descriptors without evaluation labels; nested training, tuning,

and validation partitions fix four diagnostic pick sets before outer outcomes; and only the preregistered `pairwise_rich` arm is eligible as primary. The representation amendment belongs to the same-benchmark developmental analysis. For metric-matched auditing, frozen repaired policies reuse exact candidate-identity eighth-dimension scores and evaluate only 20 missing candidates; the zero-memory control must reproduce the complete Table 6 value before any repaired 8D comparison is accepted. The history-specificity experiment fixes the two archived variants' (DEQM and Candidate-DEQM) selections before new-task outcomes and compares them with matched shuffled-history controls. Their raw gains ($+0.006485$ and $+0.006211$) do not pass memory- specificity gates. The subsequent relocation study froze 112 RoboTwin scenes without outcomes, but the provenance record `eval/PROVENANCE_RESOLUTION_ROBOTWIN_V3_V4_20260716.md` marks it as superseded diagnostic evidence because an authoritative follow-up was selected and then stopped before execution. The post-outcome ordinary pairwise RM control is materialized by `eval/robotwin_pairwise_rm_control.py`; it uses no relocation labels for fit or picks and is a retrospective control. Task-catalog and archived-code checks identify the RoboTwin 2.0 release (Chen et al., 2025); the historical video-extraction commit remains unavailable. Statistical inference for the PAI result uses the six changed interventions; the 174-scene mean and five ID folds are descriptive summaries. Two implementation amendments made before scoring fix the intended four-arm schema, including the diagnostic ungated proposer, and validate the complete 174×16 official cache before extracting the fixed c_{00}–c_{07} deployment subset. The active model-forward and timing ledger is materialized by `eval/active_compute_ledger.py`; it hashes all eight generation and cycle logs and reports external reward-model calls separately from DiT NFE. The fixed task-local fresh-episode confirmation uses episodes 0–3 as history and previously unscored episodes 8–11 for evaluation on the same 14 tasks. All six arms and 256 wrong-task assignments, each excluding the correct task, are fixed before computing evaluation official7 scores. The authoritative analyzer is hash-bound at SHA256 `5475401f...4366525`; its result and the runner result are byte- identical (SHA256 `ce1d6c41...7c043f`). Uncertainty is bootstrapped over tasks rather than individual scenes. The untouched set-transport experiment freezes 48 evaluation picks before official7; its independently audited result SHA256 is `c84f3f77...21d9c2d`. The event-transition experiment likewise freezes 56 QUERY picks and both wrong-family and outcome-shuffle nulls before official7; its independently audited result SHA256 is `6331db5e...18fcb`. Both analyses pass the preregistered checks of the no-TTS baseline and available improvement above the no-history selector, and both resample tasks rather than individual scenes. The replacement episode-29 success-transition confirmation freezes all six observed arms and 256 task derangements before materializing source futures or official7. Its freezer additionally requires zero SHA256 overlap against both opened episode-28 candidate pools and source futures. A pre-outcome extractor ABI amendment changes only episode validation and is included in the 27-file code manifest. Frozen-pick, result, and independent-audit SHA256 values are `06146daf...dd3fc`, `a0ac3527...7c0e0`, and `813dc137...537e`, respectively. The Memory Value Law implementation is materialized by `eval/memory_value_law.py`; its final adapter reads only frozen picks and subsequently opened official7 outcomes. The 5,000-draw task-cluster artifact reconstructs both raw PM–RM deltas exactly and has SHA256 `f3d482c9...3862bf8a`.

I Detailed interfaces and compute accounting

These ledgers collect the interfaces and costs referenced throughout the main text; they are separated from the result tables so that accounting choices do not interrupt the empirical narrative.

Table 27: **How history is used in each benchmark.** The generator and base scoring rule stay fixed; matched controls measure the contribution of the history interface.

Benchmark	Available history	Base score	Use of history / result
Physics-IQ	Prior real videos	Stability or reconstruction	Compare with past outcomes; benefit depends on extra information
PAI-Bench-robot	Separate training outcomes	No-history baseline	Fitted or retrieved corrections; intervals span zero
OpenS2V-Eval	Best past sample from other seeds	Denoising stability	Correct sample or score; assess against shuffled history

Information and controls: PIQ held-out PM 34.65 vs. privileged paired future 38.23 IQ (cross-model means). PAI5B pretraining changes its zero-memory comparator by $+0.000069$ (n.s.) using training-fold outcomes.

Table 28: **Per-scene compute.** Gen./ver. count generator/verification calls to the diffusion transformer (DiT); external reward-model calls are separate. Reconstruction checking uses 10 ver./candidate on PIQ and 12 on PAI/RT; history distances and frame statistics are CPU-only. External RM, conditioning encoders, and offline costs are separate from the known DiT totals. PGCM uses a two-model query probe.

Setting	Method	B	Gen.	Ver./probe	Ext.	Known DiT
PIQ	no-TTS	1	20	0	0	20
PIQ	Denoising stability	4	80	16	0	96
PIQ	Video preference	4	80	0	4	80
PIQ	Reconstruction	4	80	40	0	120
PIQ	PM / CVA-Select	4	80	0	0	80
PIQ	Real-future reference[†]	4	80	0	0	80
PAI/RT	no-TTS	1	20	0	0	20
PAI/RT	Denoising stability	8	160	32	0	192
PAI/RT	Video preference	8	160	0	8	160
PAI/RT	Baseline / fitted scores	8	160	96	0	256
PAI	Zero-memory / PM	8	160	96	0	256
PAI	Extra fitting[‡]	8	160	96	0	≥ 256
PGCM	Two-model query probe	–	0	$4_{5B} + 4_{14B}$	0	$8^{\star}$

PIQ history/motion readouts require no new DiT passes; the PAI readout includes cycle and therefore pays $8 \times 12 = 96$ verification passes. Comparing VideoReward's $4/8$ external calls with reconstruction requires model- and hardware-specific FLOPs or timing.

[†]Privileged paired-future reference. [‡]PAI fitting consumes cached candidate features, not the PGCM query probe. The DiT entry is a lower bound: producing cached features, offline feature compression, regression, and validation add work whose original cost is only partially recorded. *Two seeds and two CFG evaluations per seed give four calls per model. The model-indexed 5B and 14B ledger preserves their different per-call costs for FLOP/runtime comparisons; eight heterogeneous calls require this distinction (Appendix F.1).

J COMPLETE PAI DIMENSIONS

The main table reports aggregate official means. The following tables resolve those means into all eight dimensions and isolate each signal's contribution through the cross-base ablation.

Table 29: **PAI-Bench-robot dimensions ($B = 8$).** Official means over 174 scenes. AQ: aesthetic quality; BG: background consistency; IQ: imaging quality; MS: motion smoothness; OC: overall consistency; SC: subject consistency; I2V-B/S: background/subject faithfulness to the conditioning image. no-TTS uses c_{00}; search uses c_{00}–c_{07}. I2V-B (`i2v_background`) is the eighth dimension withheld from 7D memory fitting and added only after picks freeze. Visual reranking uses the historical Top-3 candidate shortlist. History corrections are reported as separate controls.

World model	Method	AQ	BG	IQ	MS	OC	SC	I2V-B	I2V-S	Mean
Wan2.2-5B	Baseline (no-TTS)	0.4642	0.9347	0.6821	0.9827	**0.1934**	0.9216	0.9652	0.9310	0.7594
	Denoising stability	0.4708	0.9363	0.6860	0.9857	0.1927	0.9267	0.9692	0.9368	0.7630
	Video preference	**0.4738**	0.9333	**0.6972**	0.9861	0.1922	0.9244	0.9691	0.9356	0.7639
	Ensemble w/o condition match	0.4715	0.9391	0.6946	0.9895	0.1913	0.9368	0.9777	0.9476	0.7685
	Baseline without history	0.4693	**0.9425**	0.6947	**0.9897**	0.1916	**0.9414**	**0.9791**	**0.9499**	**0.7698**
	Visual reranking (Top-3)	0.4678	0.9396	0.6943	0.9880	0.1915	0.9377	0.9750	0.9440	0.7672
	Added history (5-fold)	0.4692	0.9423	0.6947	0.9896	0.1913	0.9407	0.9791	0.9497	0.7696
ABot-14B	Baseline (no-TTS)	0.4770	0.9318	0.6955	0.9911	0.1932	0.9338	0.9786	0.9502	0.7689
	Denoising stability	0.4791	0.9322	**0.7036**	0.9912	**0.1940**	0.9346	0.9783	0.9500	0.7704
	Video preference	0.4790	0.9300	0.7021	0.9899	0.1936	0.9290	0.9776	0.9479	0.7686
	Ensemble w/o condition match	**0.4803**	0.9350	0.7014	**0.9923**	0.1930	0.9398	0.9816	**0.9538**	0.7721
	Baseline without history	0.4792	**0.9364**	0.6993	**0.9923**	0.1939	**0.9430**	**0.9819**	**0.9538**	**0.7725**
	Visual reranking (Top-3)	0.4807	0.9349	0.7017	0.9915	0.1940	0.9395	0.9803	0.9522	0.7719
Cosmos3-Nano	Baseline (no-TTS)	0.4658	0.9391	0.7218	0.9930	0.1931	0.9342	0.9838	0.9532	0.7730
	Denoising stability	0.4649	0.9425	0.7259	0.9930	0.1932	0.9440	0.9861	0.9552	0.7756
	Video preference	**0.4667**	0.9387	0.7238	0.9915	0.1936	0.9351	0.9835	0.9522	0.7731
	Ensemble w/o condition match	0.4646	0.9425	0.7216	**0.9935**	0.1944	0.9423	**0.9871**	**0.9570**	0.7754
	Baseline without history	0.4648	0.9437	0.7209	0.9934	0.1959	0.9437	0.9871	0.9569	0.7758
	Visual reranking (Top-3)	0.4658	**0.9449**	**0.7257**	0.9929	0.1948	**0.9453**	0.9866	0.9565	**0.7766**
	Gated reranking (held out)	0.4664	**0.9447**	**0.7263**	0.9929	**0.1953**	**0.9454**	0.9869	0.9568	**0.7768**

Additional-fitting check: baseline means (5B/14B/Cosmos) $0.769758/0.772483/0.776845$; after extra fitting, $0.769828/0.772483/0.776704$. Differences are $+0.000069/0/-0.000141$.

Table 30: **Removing signals from the no-history ensemble.** Retrospective leave-one-component audit at $B = 8$; 50,000 scene-bootstrap draws. Picks are frozen without official labels; no-TTS ($B = 1$) anchors the score scale. Appendix Table 29 reports all eight dimensions.

Control	Wan2.2-5B	ABot-14B	Cosmos3-Nano
no-TTS mean ($B = 1$)	0.7594	0.7689	0.7730
Full ensemble mean ($B = 8$)	0.7698	0.7725	0.7758
Reconstruction Δ	-0.0002 $[-0.0011, +0.0006]$ $+0.0006$	$+0.0005$ $[-0.0003, +0.0013]$ -0.0006	$\mathbf{+0.0009}$ $[\mathbf{+0.0003}, \mathbf{+0.0015}]$ -0.0004
Motion Δ	$[-0.0003, +0.0016]$ $\mathbf{+0.0016}$	$[-0.0014, +0.0001]$ $\mathbf{+0.0006}$	$[-0.0009, +0.0001]$ $+0.0000$
Consistency Δ	$[\mathbf{+0.0007}, \mathbf{+0.0026}]$ $\mathbf{+0.0012}$	$[\mathbf{+0.0001}, \mathbf{+0.0011}]$ $+0.0003$	$[-0.0004, +0.0004]$ $+0.0004$
I2V Δ	$[\mathbf{+0.0003}, \mathbf{+0.0022}]$	$[-0.0002, +0.0009]$	$[-0.0001, +0.0010]$

K OpenS2V-Eval six-condition mechanism probe

The Single-Domain pilot evaluates selector behavior within one frozen candidate pool on OpenS2V-Eval (Yuan et al., 2025). Table 31 reports all seven preregistered components and the range of results allowed by missing scores. This is an exploratory mechanism comparison: it fails the preregistered precondition and uses a different population from the public Open-Domain leaderboard.

Table 31: **Complete OpenS2V-Eval selector comparison ($B = 4$).** Panel (a) reports aesthetic quality, smoothness, motion, face similarity, embedding similarity (GME), the Nexus benchmark score, and naturalness, with their equal-weight mean; bold marks the best non-oracle entry per column. Columns identify evaluation measures. Panel (b) retains the 2,925-completion identified-set sweep. Four NaturalScore cells are missing and the untouched 48-condition test remains sealed.

(a) Existing selectors and controls

Method	Aes.	Smooth.	Motion	FaceSim	GME	Nexus	Natural	Mean
Baseline (no-TTS; $B = 1$)	0.3721	0.7942	0.3147	0.7253	0.4504	0.7335	0.5972	0.5696
Denoising stability	0.3828	0.8148	0.0522	0.7607	0.4475	0.7596	0.5139	0.5331
Latent-history correction	0.3766	0.7984	0.0575	0.7607	0.4547	0.7596	0.5417	0.5356
History selection	0.4148	**0.8601**	0.2079	0.7720	0.4342	0.7562	0.5972	0.5775
Low motion	**0.4334**	0.8169	0.0580	**0.8114**	0.4309	**0.7624**	**0.6250**	0.5626
Pooled history	0.4036	0.8313	**0.3274**	0.7652	**0.4639**	0.7479	0.6111	**0.5929**
Oracle (eval. only)	0.4167	0.8004	0.3058	0.8091	0.4281	0.7562	0.6944	0.6015

(b) Results allowed by the missing scores

Quantity	Pooled history	Shrunken correction	Conclusion
Possible mean scores	$[.\mathbf{5443}, .\mathbf{5588}]$	$[.\mathbf{5443}, .\mathbf{5643}]$	**Result remains an interval**
Memory delta	$[-.0065, +.0124]$; $868/2925$ positive		sign unidentified
Shuffle gate	maximum margin -0.0011		$0/2925$ pass

Across the six conditions, individual component advantages do not translate into a gain that clears the shuffled-history criterion. The identified-set calculation reaches the same conclusion for every admissible completion.

L Additional PAI history fitting: paired comparisons

These analyses isolate the effect of additional fitting relative to the no-history baseline in Table 29. The feature-rich pairwise correction adds four principal components of generator-derived features and four of temporal video descriptors. Feature compression is fitted only on the training partition. Each generator has its own compression, regression model, intervention threshold, and validation check, with no parameters or outcomes shared across generators.

The three model files have distinct hashes. The scene-clustered mean effect is -0.000024 (CI $[-0.000208, +0.000155]$; W/T/L $12/150/12$). The paired estimates center near zero, with no model clearing all its improvement thresholds.

Cross-fitted score corrections. In five-fold cross-fitting over 174 scenes, each held-out fold may use history only from the other four folds' seven-dimensional outcomes, excluding I2V-B (`i2v_background`). Only one of five training checks passes. The baseline ensemble scores 0.739852, compared with 0.739656 for the gated history correction ($\Delta = -0.000197$, CI $[-0.000493, +0.000021]$); oracle top-1 drops from $61/174$ to $58/174$. Completing I2V-B after picks freeze gives baseline 0.769758 and history correction 0.769590 ($\Delta = -0.000169$, CI $[-0.000446, +0.000038]$). Four fold gains are zero and the fifth is -0.000863.

Pairwise fitting increases intervention coverage. The original fold-wide filter leaves 170 selections unchanged. A correction trained on within-scene outcome differences retains the baseline ensemble as an option and uses a threshold chosen on training data to decide when to intervene. It changes $43/174$ selections and moves the seven-dimensional mean from 0.739852 to 0.740252 ($\Delta = +0.000399$, CI $[-0.000272, +0.001126]$), with $28/131/15$ wins/ties/losses. Only $1/5$ certification folds pass and $3/5$ outer folds are positive. A later train-only risk diagnostic reaches 0.770343 ($\Delta = +0.000584$, CI $[+0.000042, +0.001237]$) in a diagnostic proposed after the outer result. The fixed protocol above supersedes that diagnostic for model-specific inference; its 5B delta is $+0.000069$ with CI $[-0.000275, +0.000461]$.

Matched-history specificity and the relocation boundary. On 12 task-disjoint RoboTwin tasks, DEQM changes nine scenes and obtains $+0.006485$ with 9 wins/0 losses; Candidate-DEQM changes six and obtains $+0.006211$ with 6 wins/0 losses. Neither beats its matched shuffled-history criterion, separating an outcome-fitting gain from a history-specific contribution. The superseded residual diagnostic on the 14 remaining relocation tasks reaches $+0.006937$ (CI $[+0.003387, +0.011135]$), while the post-outcome ordinary pairwise RM reaches $+0.006493$ (CI $[+0.003252, +0.010368]$); their difference is inconclusive. The fresh episodes 16–19 evaluation supersedes this diagnostic for memory-specific inference: PM and ordinary RM are identical on 5B, and PM does not beat ordinary RM on 14B. The residual rule's six-intervention outer-fold PAI evaluation also fails: deployment -0.000222, changed-only -0.006433 (CI $[-0.015561, +0.001295]$), 2 wins/4 losses, and exact sign $p = 0.6875$. Together, the controls locate the gains in outcome-trained reranking, with the history-specific correction adding no resolved advantage.

Fresh-condition task-local confirmation. We next keep task identity fixed while evaluating entirely new episodes. History comes from episodes 0–3, and evaluation uses previously unscored episodes 8–11 on the same 14 tasks (56 scenes). The always-on task-specific correction changes $28/56$ selections relative to a global model fitted on the same history and raises the estimated mean by $+0.001514$, but the task-bootstrap CI $[-0.002031, +0.005258]$ crosses zero; scene W/T/L is $16/28/12$, only $6/14$ task means are positive, and the worst scene is -0.025177. Correct task identity beats all 256 wrong-task derangements (empirical $p = 0.003891$) and raises oracle hits by four, showing a real task-specific signal. It nevertheless scores -0.001191 versus the parent pairwise proposer without current history (CI $[-0.006448, +0.003100]$). Four prior outcomes reveal a task-specific signal, while the quality and reliability intervals remain too wide to meet the prespecified deployment criteria.

M Complete-budget accounting checklist

The experimental comparisons follow a common accounting rule. Every fixed-pool result falls into one of three categories:

- no-TTS generates and scores one candidate ($B = 1$);
- a best-of-B row generates and scores exactly B candidates;
- an oracle uses the same B candidates but selects with the official metric and is explicitly evaluation-only.

No fixed-pool result generates more candidates than its displayed budget. Adaptive rules additionally declare the full prefix-depth distribution and charge generation and scoring at every consumed prefix. P&P declares base and inner-loop NFEs at $B = 1$; the prospective seed-screening extension charges every partial proposal, not just the completed candidates.

N Additional PAI held-out results

The earlier branch-development split also provides a check on scenes not used to develop that branch. On the last 94 sorted PAI scenes, the unified-$B = 8$ means are $0.7530/0.7551/0.7566/0.7607/0.7615$ for the single-candidate baseline, denoising stability, video preference, the ensemble without condition matching, and the full no-history ensemble. These scenes are disjoint from the first 80 used for branch development. The ordering agrees with the all-scene

result, with the full ensemble above the listed baselines. This split-level comparison reproduces the selector ordering within the same generator.

O WHY CANDIDATE SAVINGS REQUIRE COMPLETE ACCOUNTING

The single-budget convention ties sample counts to end-to-end cost. Reducing verification saves only the verification component when generation remains unchanged. For a world-model generation cost C_{gen} and verifier cost C_{ver}, a procedure that generates G candidates but scores only V costs

$$GC_{\mathrm{gen}} + VC_{\mathrm{ver}}.$$

When video generation dominates, the contribution of reducing V depends on the unchanged cost of G. Fixed-pool reranking therefore uses $G = V = B$, with generation and verification calls reported separately. A diagnostic that generates a larger collection and scores a subset must retain both counts; the headline comparisons use the complete declared pool.

Table 18: **Physics-IQ memory ablation ($B = 4$).** Same 192 scenes/64 families as Table 1; LOFO excludes the query family. ‡Privileged same-stem outcome.

Variant	Information	Wan2.2-5B	ABot-14B	Cosmos-Nano	Avg.
Baseline (no-TTS; $B = 1$)	None	25.07	40.52	27.72	31.10
Low motion (zero memory)	None	30.33	43.69	35.10	36.38
History selection	LOFO outcomes	29.21	41.93	32.82	34.65
Train-global prior	LOFO outcomes	30.59	44.13	34.94	36.55
Reconstruction + motion	LOFO outcomes	29.14	42.61	34.58	35.44
History + reconstruction	LOFO outcomes	29.86	43.20	34.31	35.79
Pooled history + motion	LOFO outcomes	30.81	44.01	34.98	36.60
Real-future reference‡	Same-stem outcome	**32.19**	**45.67**	**36.83**	**38.23**

PM audit: held-out 34.65 IQ (complete-192 avg.); paired-future 38.23 IQ (complete-192 avg.); $\Delta_{\text{paired}-\text{heldout}}$ +2.98/+3.74/+4.01 IQ (avg. +3.58).

Table 19: **Memory-independence audit.** Attribution requires valid information boundaries and matched-history controls. PIQ: Physics-IQ; RT: RoboTwin; 95% CIs.

Setting	Effect	Control	Verdict
PIQ held-out	Across WMs $+0.44$ $[-0.77, 1.57]$ -0.000197	LOFO; train-global $+1.87$ vs retrieval	No confirmed transfer
PAI five-fold	$[-0.000493, +0.000021]$	Train-only writes; frozen outer folds	No confirmed history benefit
RT DEQM	$+0.006485$; 9W/0L	History shuffle	Specificity fails
RT Candidate-DEQM	$+0.006211$; 6W/0L $+0.001514$	Fixed-memory shuffle	Specificity fails
RT fresh local	$[-0.002031, +0.005258]$ Region vs RM $+0.058848$ $[0.016109, 0.103457]$ Off.7 -0.007112	Eps. 8–11; 256 wrong-task nulls	Correction unconfirmed
RT transition (e29)	$[-0.011187, -0.003448]$ Task–global -0.003324 $[-0.008816, +0.002076]$	No candidate/source alias; wrong-task/shuffle pass; global CI crosses 0	Region improves; overall score drops
RT extended (dev.)	3W/5T/6L Off.7 vs global $+0.002655$ $[0.000522, 0.005227]$	Eps. 12–15→16–19; wrong-task/shuffle pass	Task labels add no clear gain
RT reliability (dev.)	4W/10T/0L	No-history, episode-held-out; wrong-task CI > 0; shuffle $p = .003891$	Initial gain; later tests fail
RT stopping family	Mean B: initial/predictive/barrier/reliability $5.964/8.000/5.571/5.607$; content-global 5.429; shrunk task/global $5.964/6.321$ 5B PM–RM 0 $[0, 0]$ 14B -0.000705	Split calibration; global/wrong-task/256 shuffles; write-stage audit	Content risk helps; shrunk-task CIs cross 0; nearest-neighbor retrieval fails
RT cross-WM	$[-0.002508, +0.000639]$	Frozen eps. 16–19; matched RM	No gain over reward model
RT STOM	-0.001046 $[-0.003138, 0]$; $0/46/2$	48 sealed scenes; wrong-task/outcome shuffles	Transfer unconfirmed
RT event transition	0 $[0, 0]$; $0/56$ changed	56 sealed queries; global/shuffle	No improvement

Information and controls: PIQ held-out PM 34.65 vs. privileged paired future 38.23 IQ (cross-model means). PAI5B pretraining changes its zero-memory comparator by $+0.000069$ (n.s.) using training-fold outcomes.

Table 20: **Complete qualitative-replay cohorts.** Descriptive mean per-view IQ over all 16 conditions in each input variant. These are supplemental cohorts, separate from the historical 192-scene comparison. Reconstruction is evaluated on the official-description cohort.

Generation text	No-TTS	Global history	Low motion	Reconstruction	CVA-Select
Generic prompt	28.03	24.62	24.41	–	24.67
Official description	28.54	35.26	34.65	35.67	34.65

Table 21: **Privileged same-stem Physics-IQ reference ($B = 4$).** Unpruned 198 scenes/66 families, rather than the complete-16 192-scene/64-family intersection in Table 1; this population difference explains the offset in means. 95% family-bootstrap CIs; paired take-2 is privileged.

Selector	Wan2.2-5B	ABot-14B	Cosmos-Nano
Reconstruction	27.48 [23.65, 31.45]	41.16 [35.71, 46.76]	33.62 [28.97, 38.40]
History selection	29.11 [24.79, 33.47]	41.67 [36.28, 47.25]	32.80 [28.17, 37.46]
Real-future reference (priv.)	32.16 [27.80, 36.62]	45.27 [39.58, 51.09]	36.87 [31.92, 41.99]
Δ: *pretrain − held-out*	+3.05	+3.60	+4.07

PM audit: held-out 34.53 IQ (all-198 avg.); paired-future 38.10 IQ (all-198 avg.); $\Delta_{\text{paired}-\text{heldout}}$ +3.57 IQ.

Table 22: **Privileged Physics-IQ motion strata ($B = 4$).** 198 unpruned scenes: 33 calm/33 dynamic families (99 scenes each), not the headline 192-scene subset; 95% CIs use 50,000 family bootstraps. "Future" uses the real second recording of the same setup; "Low motion" selects the candidate with the least measured motion. $\Delta = \text{Future} - \text{Low motion}$ on dynamic families.

	Calm		Dynamic				
Model	Low motion	Future	no-TTS	Recon.	Low motion	Future	Δ [95% CI]
Wan2.2-5B	31.54	32.09	25.97	25.57	28.93	32.24	+3.31 [1.47, 5.21]
ABot-14B	46.93	46.82	40.80	39.99	39.86	43.72	+3.86 [2.13, 5.77]
Cosmos3-Nano	33.23	33.78	32.48	35.78	36.87	39.96	+3.08 [1.37, 4.92]

PM audit: held-out 34.53 IQ (all-198 avg.); paired-future 38.10 IQ (all-198 avg.); $\Delta_{\text{paired}-\text{heldout}}$ +3.57 IQ.

Table 25: **PGCM attribution.** Source arms share probes, labels, folds, candidates, and the sign-based routing rule; their realized model counts need not coincide. Memory must beat ridge and both matched nulls.

Arm	Source object	Target labels	Role
Source-majority	Mean advantage	None	Feature-free null
Global ridge	81-D parametric fit	None	Utility control
Global memory	Shared key/value bank	None	Nonparametric memory route
Value/key shuffle	Deranged bank	None	Specificity nulls
Target-history ridge	Private fit	Used	Supervised upper comp.

PGCM diagnostic: held-out 14B PM 41.93 IQ; post-stop PGCM 41.56 IQ; $\Delta_{\text{PGCM}-\text{heldout}} = -0.36$ IQ. The action is a choice between generators.

Table 26: **Outcome-free transfer ladder.** One frozen memory throughout. PAI outcomes are open; confirmation requires a new owner-held split.

Tier	Data	Shift	Role
Pretrain	Physics-IQ, 192/64	Natural events	Source fit
Near	Held-out PIQ families	New families	Public diagnostic
Mid	OpenS2V-Eval	New scenes/interface	Public diagnostic
Far-dev.	Private PAI, 174	Robot I2V/official8	Retrospective
Far-confirm.	New robot, ≥ 100	Owner-held assets	Claim required

PGCM diagnostic: held-out 14B PM 41.93 IQ; post-stop PGCM 41.56 IQ; $\Delta_{\text{PGCM}-\text{heldout}} = -0.36$ IQ. The action is a choice between generators.

Table 32: **Additional model-specific fitting on PAI (eight-metric mean).** Each model uses its own 174×8 labels and five-fold state. The frozen comparator is the no-history ensemble for Wan2.2-5B and ABot-14B, but the historical reliability-gated selector for Cosmos3-Nano. Deltas are relative to these historical, model-specific comparators. Each model is evaluated against its own prespecified improvement requirements.

Model	no-TTS	Comparator	Extra fit	Δ [95% CI]	Valid. pass	Test +
Wan2.2-5B	0.759376	0.769758	0.769828	$+.000069$ $[-.000275, +.000461]$	$1/5$	$1/5$
ABot-14B	0.768894	0.772483	0.772483	0 $[0, 0]$	$0/5$	$0/5$
Cosmos3-Nano	0.773021	0.776845	0.776704	$-.000141$ $[-.000566, +.000214]$	$1/5$	$0/5$

Table 33: **Superseded exploratory RoboTwin per-task diagnostic.** Eight scenes/task, 112 scenes in total; RM and slow use matched outcomes, frozen 12-D features, $\alpha = 100$, and top-21 coverage. $\Delta = \text{slow} - \text{RM}$. The earlier mean difference is $+0.000445$; the later fresh episodes 16–19 audit supersedes it with $0/-0.000705$ on 5B/14B.

Task	RM	Slow	Δ
move card	$+0.006631$	$+0.006631$	0.000000
move stapler pad	$+0.004432$	$+0.002843$	-0.001589
pick bottles	0.000000	0.000000	0.000000
place bread/skillet	$+0.010142$	$+0.010388$	$+0.000246$
place burger/fries	0.000000	0.000000	0.000000
place cans/box	0.000000	0.000000	0.000000
place cup	$+0.025067$	$+0.027042$	$+0.001975$
place fan	$+0.004785$	$+0.004785$	0.000000
place mouse/pad	$+0.012089$	$+0.012089$	0.000000
place object/basket	$+0.002392$	$+0.002392$	0.000000
place object/stand	0.000000	0.000000	0.000000
stack 3 blocks	$+0.004597$	$+0.004597$	0.000000
stack 2 blocks	$+0.014666$	$+0.014666$	0.000000
stack 2 bowls	$+0.006097$	$+0.011689$	$+0.005592$
Mean (14 tasks)	$+0.006493$	$+0.006937$	$+0.000445$