SCOPE: AUDITABLE SCORE-ISOLATED AGENTIC ADAPTATION FOR VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models are increasingly used as simulators for planning and embodied reasoning, but inference-time adaptation creates an evaluation challenge: jointly evolving controls obscure causal attribution and can entangle adaptation with held-out feedback. We introduce SCOPE (*Score-Isolated Control Orchestration with Provenance-Bound Evidence*), which makes the inference harness an explicit typed state over text, sampling, verification, and selection. Development-supported edits update one axis at a time; the final policy and exact Base fallback are frozen before held-out scoring, with every registered transition bound to its evidence. On Physics-IQ, SCOPE achieves 34.94, improving on single-sample Base by $+14.24$ (95% CI $[+8.10, +21.23]$), including candidate expansion. A secondary comparison over the same candidate pool yields $+7.53$ over uniform selection (95% CI $[+4.04, +11.52]$). Matched ablations identify useful text, sampling, and learned-selection controls. Comparisons with the strongest agentic baseline and equal-budget random updates remain statistically unresolved, while prospective tests identify calibration under task and backbone shift as the next deployment challenge. Together, these results establish an auditable basis for developing inference controls and separating candidate quality from deployment choice.

1 INTRODUCTION

Video world models offer a flexible interface for planning and embodied reasoning: an observation and a prompt define a visual future that can be inspected before acting. For that future to serve as a simulator, motion and contact must remain physically meaningful. Inference-time controls provide a practical route to this goal without retraining the generator. A harness can clarify the scene description, adjust sampling, consult a tool, or select a better trajectory from several alternatives.

That flexibility also complicates evaluation. A better video may reflect a clearer description, a larger candidate pool, or a better choice among the same candidates. These are distinct improvements, even though a single benchmark score can blur them. When held-out outcomes also determine which intervention is retained, evaluation becomes part of adaptation itself. We call this ambiguity the *inference-control evaluation gap*.

Our key insight is to treat this harness as a persistent object of adaptation. SCOPE (*Score-Isolated Control Orchestration with Provenance-Bound Evidence*) represents text, sampling, verification, and selection in a typed state Ω_r. Each proposed edit identifies the control it changes, the development evidence that supports it, and the guards it must satisfy. Successful edits enter the next state; rejected edits leave the incumbent intact. After development, the state defines the complete generation and selection procedure, including an exact Base fallback. Freezing that procedure before held-out scoring turns an evolving workflow into a reproducible intervention.

The experiments connect this representation to measurable gains. On the 40-scene Physics-IQ common-base protocol, the full system improves over single-sample Base by $+14.24$ (95% CI $[+8.10, +21.23]$), including candidate expansion. Uniform selection from the same Base/card/rule pool gains $+6.71$; the secondary matched-pool comparison attributes a further $+7.53$ to selection (95% CI $[+4.04, +11.52]$). Component studies locate benefits in scene description, sampling, and learned selection. Cross-backbone and prospective evaluations then show which choices remain useful and which require recalibration. This combination of gains, matched controls, and deployment

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.

tests separates two research questions: how to produce better candidates and how to decide which interventions to deploy.

Contributions. Accordingly, our contributions are:

- **A typed adaptation framework:** an explicit control state with bounded development-only updates, a score-isolation invariant, and evidence-linked records for accepted and rejected attempts.
- **Matched evidence of effective controls:** system and same-pool gains on Physics-IQ, supported by separate text, sampling, and selection comparisons.
- **A deployment analysis:** cross-backbone and prospective tests that connect intervention quality to routing, uncertainty, and proposal coverage.

2 Related Work

Physics-aware generation and inference-time alignment. Training-time methods inject physical supervision or motion/geometry priors (Zhang et al., 2026a; Wang et al., 2026; Jiang et al., 2026; Li et al., 2026a); retrieval-based approaches add external knowledge (Cheng et al., 2026). PhyGround and WorldReasonBench test physical reasoning (Lin et al., 2026a; Wu et al., 2026). Rather than changing the generator during training, inference-time critics rank or steer samples (Yuan et al., 2026; Yin et al., 2026; Bucciarelli et al., 2026), and iterative refinement or repeated sampling allocates additional test-time compute (Kang et al., 2026; Brown et al., 2024; Ma et al., 2025). Repeated sampling changes candidate coverage, whereas a selector changes which sample is returned (Brown et al., 2024). SCOPE places these operators in a common typed control space and evaluates their effects through matched interventions. This connects inference-time optimization to the evidence required for retaining each change.

Agentic generation and tool orchestration. Agentic systems sharpen the attribution problem by interleaving generation and control. Qwen-Image-Agent uses tool-augmented feedback, NEWTON plans physically grounded videos, and LingBot-World structures interactive world generation (Zhang et al., 2026c; Feng et al., 2026; Gao et al., 2026). GenEvolve and COMFYCLAW further organize tool calls and reusable workflow experience for visual generation (Chen et al., 2026b; Li et al., 2026b). NEWTON trains a physics-tool planner with Flow-GRPO around a frozen generator and verifier; SCOPE instead governs which evidence may change an external state. These are complementary design choices: planner and tool feedback can generate proposals within SCOPE, while the development and evaluation interfaces determine how those proposals become persistent controls. Held-out benchmark outcomes remain reserved for evaluating the frozen procedure.

Persistent skills. Once controls are retained across tasks, attribution also becomes a question of memory. SkillOpt, SkillCAT, and SkillDAG evolve reusable skills (Yang et al., 2026; Chen et al., 2026a; Bai et al., 2026); other work diagnoses library drift and failures to translate updates into benefits (Zhang et al., 2026b; Lin et al., 2026b). SCOPE gives persistence an explicit evidence requirement: each retained change is bound to a development comparison, and the same state can then be evaluated on validation and confirmation sets.

Adaptive evaluation and safe updates. Reusable-holdout methods permit adaptive holdout queries using privacy-preserving mechanisms; high-confidence policy evaluation uses conservative deployment bounds (Dwork et al., 2015; Thomas et al., 2015). SCOPE pursues the same evaluation goal through structural score isolation: development updates and held-out evaluation have separate interfaces. Unlike privacy-preserving query mechanisms, it freezes the policy before score release. Task-disjoint prospective evaluation then measures how the development-selected policy transfers.

3 Method

3.1 Problem Setting and Typed Control State

Let $p_\theta(x_{1:T} \mid x_0, c)$ denote a pretrained video world model, where x_0 is the initial observation, c is the conditioning signal, and $x_{1:T}$ is the generated trajectory. The model parameters θ remain frozen

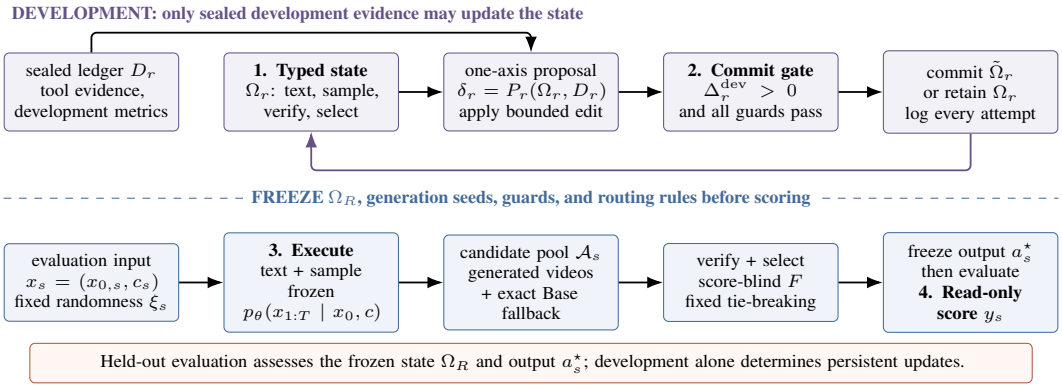

Figure 1: **Overview of SCOPE.** The four control axes form a generate–verify–select pipeline. Development evidence admits bounded edits; the resulting state and routing rules are frozen for held-out evaluation. Fixed-pool experiments share the generation stage across policies to isolate verification and selection.

throughout. Our object of adaptation is the inference harness, which SCOPE stores as the persistent typed state

$$\Omega_r = \left(\Omega_r^{\text{text}}, \Omega_r^{\text{sample}}, \Omega_r^{\text{verify}}, \Omega_r^{\text{select}}\right) \in \mathcal{C}, \tag{1}$$

where $\mathcal{C} = \mathcal{C}^{\text{text}} \times \mathcal{C}^{\text{sample}} \times \mathcal{C}^{\text{verify}} \times \mathcal{C}^{\text{select}}$ is the product control space. For an execution state Ω and input $x_s = (x_{0,s}, c_s)$, generation is

$$\begin{aligned} \tilde{c}_s^{(m)} &= \text{Text}(\Omega^{\text{text}}, x_s, m), \\ a_s^{(m)} &= \text{Sample}_\theta(x_{0,s}, \tilde{c}_s^{(m)}; \Omega^{\text{sample}}, \xi_s^{(m)}). \end{aligned} \tag{2}$$

Write Gen_θ for this construction: text supplies indexed conditionings, while sampling fixes the numerical procedure and budget around the unchanged model. The frozen predicate $V(\Omega^{\text{verify}}, x_s, a)$ determines candidate eligibility, after which selection applies proxy scores and deterministic tie-breaking. Exact Base is the default when no proposal qualifies. Both operations use score-blind features. This decomposition makes the source of a change explicit: text and sampling construct the pool, verification determines eligibility, and selection chooses the deployed output.

For example, a scene-description edit can be tested with the sampler and selector fixed, while a selector edit can reuse an unchanged pool of videos. In either case, the update names the axis being tested. The candidate budget belongs to the sampling state, so expanding a pool is recorded separately from choosing more effectively within it. This distinction links the executable harness to its experimental controls.

Having specified how a state produces an output, we next define how it may change. At development round r, a proposer produces a bounded edit

$$\delta_r = P_r(\Omega_r, D_r; \xi_r^{\text{dev}}), \tag{3}$$

where D_r is sealed development evidence. Each edit adds, deletes, or replaces content within one axis. VLM/tool workflows and fixed proposal families share this interface, allowing different proposal mechanisms to contribute to the same state. Learned components are fixed before deployment and use label-free features there. Appendix Table 7 instantiates SceneLang and Director cards, samplers, guards, OOF selectors, and task routers. These mechanisms can be replaced while preserving the same commitment contract.

3.2 DEVELOPMENT-ONLY STATE UPDATES

An edit must be evaluated before it can affect later decisions. Let $\tilde{\Omega}_r = \text{Apply}(\Omega_r, \delta_r)$ denote this trial state. Its first requirement is improvement in the predeclared development objective,

$$\Delta_r^{\text{dev}} = J_{\text{dev}}(\tilde{\Omega}_r; D_r) - J_{\text{dev}}(\Omega_r; D_r), \tag{4}$$

subject to the accompanying guard conditions,

$$G_r(\tilde{\Omega}_r; D_r) = \prod_{k=1}^{K} \mathbf{1}\left[G_{r,k}(\tilde{\Omega}_r; D_r) \le 0\right], \tag{5}$$

where K and the guard definitions are protocol-specific: motion/sharpness ratios, harm ceilings, or applicability checks (Appendix A.1). The objective measures the intended improvement, while the guards preserve the other predeclared requirements. Together they make acceptance a joint decision about benefit and admissibility. The persistent state follows the commit-or-retain rule

$$\Omega_{r+1} = \begin{cases} \tilde{\Omega}_r, & \Delta_r^{\text{dev}} > 0 \ \wedge \ G_r(\tilde{\Omega}_r; D_r) = 1, \\ \Omega_r, & \text{otherwise.} \end{cases} \tag{6}$$

Acceptance advances the persistent state; rejection preserves the exact incumbent. Later proposals therefore start from a well-defined parent rather than a mixture of accepted and abandoned edits. The same frozen endpoint can be evaluated across tasks and backbones, as in Section 4.4.

3.3 Frozen Deployment, Score Isolation, and Provenance

After R rounds, development yields a fixed state Ω_R independent of the held-out split H. For an evaluation instance s, this state and its input generate a candidate set that always includes exact Frozen Base,

$$\begin{aligned} \mathcal{A}_s &= \text{Gen}_\theta(\Omega_R^{\text{text}}, \Omega_R^{\text{sample}}, x_s, \xi_s) \cup \{a_s^{\text{Base}}\}, \\ a_s^\star &= F(\Omega_R^{\text{verify}}, \Omega_R^{\text{select}}, x_s, \mathcal{A}_s). \end{aligned} \tag{7}$$

and fixes the chosen route before y_s is revealed. In full deployment, the pool depends on the text and sampling controls that produced it. A candidate-fixed study holds that stage constant: every policy receives the same precomputed pool, so the comparison concerns routing rather than generation. Base is counted once and can be shared across selectors through the cache.

Proposition 1 (Score-isolation invariant)**.** *Fix proposal and execution randomness, sealed development ledgers and tool responses, the candidate-construction procedure, world model, evaluation inputs, and fallback rule. Replacing the complete validation, confirmation, held-out, or official-score ledger by arbitrary alternative values leaves both the final control state Ω_R and every frozen deployment route $a_s^\star$ unchanged before score release.*

The invariant, proved in Appendix A.2, makes the evaluation boundary explicit: the measured outputs belong to the policy selected by development evidence. It supplies the information-flow guarantee on which the matched and prospective comparisons below rely.

Candidate freezing alone does not provide this boundary. If the reported state is chosen as $\widehat{r}_H = \arg\max_r J_H(\Omega_r)$ using the held-out objective, evaluation outcomes still influence deployment after generation has finished. The same issue arises when confirmation results retune a router. SCOPE therefore separates development from deployment:

$$\begin{aligned} \Omega_R &= \Psi_{\text{dev}}(\Omega_0, D_{0:R-1}, \xi_{\text{dev}}), \\ a_s^\star &= F(\Omega_R^{\text{verify}}, \Omega_R^{\text{select}}, x_s, \mathcal{A}_s), \qquad s \in H. \end{aligned} \tag{8}$$

The first map depends entirely on development inputs. The second applies the resulting state to each evaluation instance using Equation 7 and fixed ξ_s. Instance-specific execution is therefore compatible with a globally frozen policy. Held-out outcomes assess that policy after its choices have been committed.

To make this separation auditable, each registered attempt records

$$\mathcal{T}_r = \big(h(\mathcal{T}_{r-1}), r, h(\Omega_r), h(\delta_r), h(D_r), b_r, \mathcal{R}_r, h(\Omega_{r+1})\big), \tag{9}$$

where h is canonical-JSON SHA-256, b_r is the commit decision, and $\mathcal{R}_r$ contains the guard results. Each record distinguishes an attempted edit from a state revision. Rejected edits keep the same parent and successor states but occupy distinct positions in the history. Binding the preceding record and advancing the counter after every attempt preserves this ordering. An independently retained initial-state digest, record count, and final digest anchor continuity and truncation checks. The audited history covers registered attempts; Appendix A.3 specifies the trust assumptions and executable checks.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Our evaluation addresses three complementary questions: how much the complete system improves generation, which controls contribute, and how the selected policies behave across backbones and tasks. Experiments use frozen Wan2.2 and CogVideoX backbones on Physics-IQ, P-AI (PAI-Bench-G), and OpenS2V-Eval. Primary Physics-IQ results cover 40 scenes. P-AI uses 96 robot split6 tasks and Qwen2.5-VL-72B; its Overall score is $(\mathrm{RO} + \mathrm{QS})/2$, where RO is official binary VQA accuracy and QS is the mean of eight quality dimensions (Appendix F). OpenS2V-Eval has 180 items per backbone. Prospective tests add PAI-Bench-G physics32, a 40-scene Physics-IQ confirmation, and PhyGround.

Each contrast remains within a backbone and protocol. The independent units are Physics-IQ scenes (with nested views), P-AI tasks, and category-stratified OpenS2V items. SceneLang averages three fixed seeds, the sampler study four, and reward selection uses scene-grouped OOF prediction. Unless noted, deltas pair against Frozen Base and bootstrap intervals resample the protocol's independent unit.

Generation, routing, and units are fixed before scoring. We distinguish full-system comparisons from candidate-budget-matched component studies and same-pool selection. Appendix Table 9 records each comparison's statistical role. Intervals are protocol-local; secondary analyses are descriptive, with no manuscript-wide multiplicity correction.

4.2 MAIN RESULTS ON PHYSICS-IQ

SCOPE improves Physics-IQ over both the single-sample Base and uniform selection from the matched candidate pool. Table 1 separates these two effects and places them alongside matched agentic alternatives.

The eleven methods share scenes, scorer, and Base within each backbone. Wan SCOPE reaches 34.94, gaining $+14.24$ over exact single-sample Base (95% CI $[+8.10, +21.23]$) through the complete generation-and-selection procedure. Selector rows then reuse the indexed Base/card/rule pool, including Base once: the secondary comparison gives $+7.53$ over uniform expectation (95% CI $[+4.04, +11.52]$). This isolates the value of selection once the candidates have been generated. The retrospective LingBot arm is a separately generated comparison. The strongest Qwen-style comparator scores 32.87, giving SCOPE a $+2.07$ point-estimate margin with uncertainty reported below.

The uniform reference is the expectation over the Base, physics-card, and rule-refinement arms. It measures the value of this heterogeneous pool; the secondary SCOPE–uniform contrast holds all its videos fixed and measures selection. Three views yield $40 \times 3 \times 3 = 360$ indexed videos, with Base included once. Thus the pool differs through conditioning as well as generation budget.

Paired SCOPE–Qwen scores differ on only $8/40$ scenes: $3/32/5$ wins/ties/harms. The median difference is zero, and the descriptive exact two-sided sign test on discordant scenes gives $p = 0.727$. The three wins total 93.20 points, whereas the five losses total -10.30; a single stable-blocks scene contributes 71.71. These large gains on a small set of scenes explain the positive mean alongside the zero median. Reporting both exposes the distribution of benefit behind the scene-equal primary estimate.

Figure 2 traces the endpoint through the Base, physics-card, rule-refinement, and selected states. The intermediate CogVideoX decrease shows why each stage is evaluated against its own control. R3 remains the final state after the later Wan candidates R4–R11 are considered.

4.3 CROSS-BACKBONE AND OPEN-DOMAIN GENERALIZATION

The common-base result characterizes one controlled setting. We next vary the backbone, benchmark, and metric to determine how far its ordering extends.

Table 1: **Common-base Physics-IQ comparison across two frozen video backbones.** All rows report means over 40 scene-level Physics-IQ composites. Within each backbone, methods share the same Base and evaluation protocol. The selector pool contains three candidates total, including Base once. The SCOPE–Qwen differences are $+2.07$ on Wan (95% CI $[-0.27, +6.14]$) and $+0.52$ on CogVideoX (95% CI $[-0.39, +1.65]$); both head-to-head intervals span zero. Comparisons are within backbone. The dagger denotes the separately generated fixed Director baseline; green indicates Base-relative intervals with positive lower bounds.

Method	Wan P-IQ	Wan Δ (95% CI)	CogVideoX P-IQ	CogVideoX Δ (95% CI)
Frozen Base	20.70	–	22.91	–
WMReward reward-only (Yuan et al., 2026)	21.81	+1.11 [−0.40, +3.17]	26.10	+3.19 [+0.85, +6.04]
VAE-θ reward	22.33	+1.63 [−0.67, +4.46]	25.14	+2.23 [+0.02, +4.83]
VIGOR reward-only (Yin et al., 2026)	25.15	+4.44 [+1.49, +7.96]	27.75	+4.83 [+1.71, +8.44]
uniform 3-candidate pool expectation (Brown et al., 2024; Ma et al., 2025)	27.41	+6.71 [+1.93, +11.45]	30.57	+7.65 [+4.51, +10.97]
LingBot-style Director→Pilot proxy† (Gao et al., 2026)	29.98	+9.27 [+2.37, +16.12]	34.91	+12.00 [+6.76, +17.62]
fixed GPT-5.4 physics card	30.04	+9.33 [+2.50, +15.93]	34.89	+11.98 [+7.05, +17.13]
fixed rule refinement	31.50	+10.80 [+2.67, +19.14]	33.89	+10.98 [+6.27, +15.93]
joint self-score ablation	32.07	+11.37 [+3.38, +19.69]	34.63	+11.71 [+6.83, +16.66]
Qwen-Image-Agent-style feedback (Zhang et al., 2026c)	32.87	+12.16 [+4.48, +20.06]	34.99	+12.08 [+7.36, +16.98]
SCOPE (ours)	**34.94**	**+14.24 [+8.10, +21.23]**	**35.51**	**+12.60 [+7.73, +17.60]**

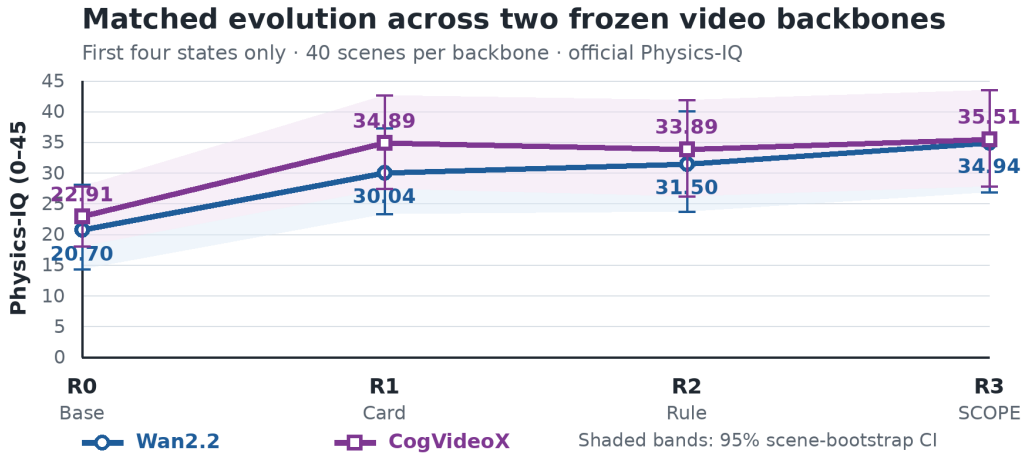

Figure 2: **Evolution of the matched Physics-IQ policy across two frozen backbones.** R0–R3 correspond to Base, the physics-card update, the rule-refinement update, and the final selected SCOPE state. Curves show the mean Physics-IQ score over 40 scenes, with shaded 95% scene-bootstrap confidence intervals. Scores are compared only within each backbone.

P-AI. Table 2 reports the matched robot split6 panel. We show both fixed SCOPE variants for each backbone, rather than retaining only their upper envelope.

P-AI reveals how the same control interacts with different backbones. Shared-global SCOPE gives $+0.80$ on Wan and -2.41 on CogVideoX, with both intervals spanning zero; without shared global, the estimates are $+0.12$ and $+0.38$. On CogVideoX, shared-global SCOPE raises QS by 0.85 points while VQA accuracy falls by 5.66. The metric breakdown therefore identifies the source of the Overall decrease: the visual-quality gain is outweighed by the task-answering loss. The cross-backbone differences remain descriptive because the intervals span zero. We show the fixed variants alongside the envelope so readers can distinguish the behavior of one policy from the best value across policies.

OpenS2V-Eval. OpenS2V-Eval extends this test beyond the P-AI task distribution. Within each backbone, methods share a frozen sampler and fixed routing. The matched LingBot-style proxy inserts a Qwen3.5-9B first-frame caption into a fixed Director→Pilot template, preserving the backbone and sampler. This prompt-only implementation provides a matched test of structured agentic conditioning.

Table 2: **Cross-backbone results on P-AI (PAI-Bench-G).** Deltas and paired 95% CIs use the same 96 tasks and each backbone's Base. † denotes a prompt-only LingBot-style proxy. * marks post-hoc diagnostics: fixed SCOPE variants and the per-backbone upper envelope. Overall combines official VQA accuracy and the eight-dimensional quality score; full profiles appear in Appendix F.

Matched method	Wan Overall	Wan Δ (95% CI)	CogVideoX Overall	CogVideoX Δ (95% CI)
Ranking-eligible matched methods				
Frozen Base	76.67	–	74.10	–
+ shared global	76.86	+0.19 CI [−2.13, +2.66]	74.41	+0.30 CI [−2.15, +2.77]
LingBot-style Director→Pilot + shared global†	76.77	+0.10 CI [−2.25, +2.54]	74.29	+0.18 CI [−2.30, +2.69]
Qwen-Image-Agent-style (SceneLang)	77.02	+0.35 CI [−1.50, +2.32]	74.72	+0.62 CI [−1.37, +2.68]
SceneLang + shared global	77.01	+0.34 CI [−1.86, +2.63]	71.57	−2.53 CI [−5.25, +0.12]
Post-hoc SCOPE diagnostics; no ranking or promotion credit				
SCOPE w/o shared global*	76.79	+0.12 CI [−1.80, +2.12]	74.48	+0.38 CI [−1.56, +2.37]
SCOPE + shared global*	77.47	+0.80 CI [−1.38, +3.12]	71.70	−2.41 CI [−5.04, +0.13]
SCOPE post-hoc upper envelope*	77.47	+0.80 CI [−1.38, +3.12]	74.48	+0.38 CI [−1.56, +2.37]

Table 3: **Cross-backbone results on OpenS2V-Eval (n=180 per backbone).** Scores and deltas are 100× normalized values; deltas precede display rounding. The Natural component uses the disclosed GPT-5.4/Qwen proxy; LingBot is prompt-only. Intervals are paired category-stratified bootstraps over frozen outputs. Secondary comparison roles and seeds are specified in Appendix E.

Backbone	Metric	Base	Proxy	SCOPE	Δ vs Base (95% CI)	Δ vs proxy (95% CI)
Wan	Mean	37.46	41.27	41.75	+4.29 [+2.80, +5.81]	+0.47 [−0.45, +1.43]
Wan	Weighted	36.60	42.84	43.17	+6.57 [+4.58, +8.60]	+0.33 [−0.93, +1.62]
CogVideoX	Mean	48.30	47.39	47.76	−0.54 [−1.85, +0.77]	+0.37 [−0.34, +1.07]
CogVideoX	Weighted	48.29	49.62	49.22	+0.93 [−0.71, +2.59]	−0.40 [−1.39, +0.52]

On Wan, SCOPE improves over Base on both Mean and Weighted in the secondary comparison. The matched proxy sets a stronger reference: all four SCOPE–proxy intervals include zero. CogVideoX's Base contrasts also span zero and change sign across metrics. Together with the P-AI breakdown, this shows that the backbone and evaluation objective jointly determine the benefit of a control. Reporting them separately preserves that information.

4.4 PROSPECTIVE DEPLOYMENT AND STOP TESTS

The preceding comparisons concern generalization across settings. Prospective tests pose the harder question of whether a frozen decision remains useful in deployment. They distinguish task-disjoint scoring, fixed-pool update confirmation, and pre-score applicability gates; none retunes on its evaluation pool.

The prospective tests identify the conditions that limit deployment. On fresh PAI physics32 tasks, the Base contrast is unresolved and the frozen-router contrast trends negative. The changed-task motion ratio, $0.842626 < 0.85$, triggers the guard despite the stronger all-task average. Typed-minus-random is -0.316 (95% CI $[-5.984, +5.145]$) and harms $6/40$ scenes against the incumbent, or $15\% > 10\%$. PhyGround likewise stops before outcome scoring: only $4/32$ proposals enter a protocol requiring at least $8/32$. The next experiment examines selective routing in more detail, including its gain over always-on assistance.

Selective routing under independent confirmation. For each task, every policy chooses between the same frozen no-tool output and a profiled tool-assisted candidate; only the route changes. This design isolates selection from candidate quality and generation randomness. Table 5 reports the resulting absolute DisMo cosine similarities.

The experiment separates card availability from deployment choice. Among 180 eligible tasks, the fit-only policy uses a card on 48. Random routing receives the same card budget within eligible strata, making the location of interventions the key difference between these policies.

Table 4: **Prospective evaluation and stop tests.** Intervals are paired 95% CIs. PAI physics32 uses Overall points; the typed Physics-IQ test uses P-IQ points. PhyGround reports proposal coverage, not an evaluated outcome.

Evaluation	Reference / Criterion	Outcome
PAI physics32 ($N = 32$)	Base	$+0.096$ $[-0.887, +1.077]$
	Frozen router	-1.019 $[-2.159, +0.146]$
Typed Physics-IQ ($N = 40$)	Random update	-0.316 $[-5.984, +5.145]$
PhyGround	Min. proposal coverage: $8/32$	$4/32$

Table 5: **Selective routing across target and confirmation tasks.** Policies share frozen no-tool/profiled candidates. Entries are DisMo cosine similarities; text contrasts are $100\times$ cosine differences. Of 288 tasks, 180 are eligible for a profiled card; the other 108 retain exact no-tool output. Cards count task-level routes, not additional videos. Seeds are averaged within task and the pooled mean weights tasks equally.

		DisMo Cosine Similarity		
		Target	Confirmation	Pooled
Policy	Cards	$N = 192$	$N = 96$	$N = 288$
Exact no-tool	0	0.4071	0.4425	0.4189
Always-on profiled card	180	0.4021	0.4352	0.4132
Fit-only group policy	48	0.4096	0.4401	0.4197
Same-budget stratified random	48	0.4084	0.4424	0.4197

Selective routing improves over always-on assistance in the pooled comparison, while its effect against exact no-tool varies across splits. On the target split, selective routing improves over exact no-tool by $+0.243$ DisMo points overall and $+1.460$ on the 32 switched tasks. Independent confirmation reverses both effects, to -0.237 overall and -1.420 over the 16 switched tasks, with a negative switched-task mean for every frozen seed. Pooling the splits leaves a resolved $+0.660$ advantage over always-on assistance (CI $[+0.016, +1.315]$), but not over exact no-tool ($+0.083$, CI $[-0.191, +0.381]$) or same-budget random routing ($+0.002$, CI $[-0.281, +0.299]$). The resolved benefit is therefore selective use relative to indiscriminate assistance; the confirmation reversal locates the remaining calibration problem. Because only 48 of 288 tasks change, the switched-task comparison is especially informative about the effect of each intervention.

4.5 Ablation Studies

Matched ablations identify three effective control axes and distinguish them from the update and tool mechanisms. Table 6 reports each effect with its comparator, benchmark, and independent unit, making the contribution of each tested intervention directly inspectable.

Component controls. SceneLang improves P-IQ by 5.62 points, the joint sampler by 1.26, and learned selection by 3.33 against their matched controls. These studies target different parts of the pipeline: scene specification, the sampling trajectory, and the final choice from a cached pool. SceneLang averages three fixed seeds; the reward model uses scene-grouped OOF supervision and label-free deployment features. Its advantage over self-score remains unresolved. The region-prior screen offers a contrasting case: local noise correlation passes its visual diagnostics yet loses 7.28 P-IQ points against the identity-noise control.

Updates, fallback, and tools. Random and typed edits both improve over frozen Ω_0; their direct contrast does not resolve a typed-update advantage. On PAI physics32, retirement returns 17 additional tasks to Base and costs 0.773 Overall points (95% CI $[-1.471, -0.203]$). The Base candidate remains available throughout, so this contrast measures additional conservative retention. The retrieval and SAM comparisons have intervals spanning zero. Taken together, the ablations concentrate the demonstrated benefits in text, sampling, and learned selection, while making update quality and tool applicability clear targets for further development. Appendix G provides complementary qualitative illustrations.

Table 6: **Ablation studies of SCOPE.** Effects are protocol-local score-point differences. Sc. denotes scenes; views and seeds are nested within the declared unit. The region prior is a 12-scene development screen; all tool contrasts use 66 scenes with three views and one fixed seed. Paired confidence intervals resample scenes or tasks; effects from different protocols are interpreted separately.

Variant	Comparator	Protocol	Unit / N	Metric	Δ (95% CI)
Text (SceneLang)	Base	P-IQ s100	66 sc.; 3 seeds	P-IQ	$+5.62$ $[+0.55, +10.55]$
Joint sampler	fixed sampler	P-IQ s110	66 sc.; 4 seeds	P-IQ	$+1.26$ $[+0.52, +2.01]$
Reward	uniform Bo16	P-IQ OOF	64 sc.; 179 views	P-IQ	$+3.33$ $[+1.49, +5.30]$
Reward	self-score	same OOF pool	same 64 sc.	P-IQ	$+0.66$ $[-1.52, +2.92]$
Region prior	identity noise	P-IQ development	12 sc.; 1 seed	P-IQ	-7.28 $[-13.33, -2.53]$
Random update	Ω_0	P-IQ s115	40 sc.	P-IQ	$+4.53$ $[+1.14, +9.17]$
Typed update	Ω_0	same s115	same 40 sc.	P-IQ	$+4.22$ $[+0.81, +8.30]$
Typed update	random update	same s115	same 40 sc.	P-IQ	-0.316 $[-5.984, +5.145]$
Retirement routing	commit+fallback	PAI physics32	32 tasks	Overall	-0.773 $[-1.471, -0.203]$
Retrieval	no tool	P-IQ tool axis	66 sc.; 1 seed	P-IQ	$+0.562$ $[-0.700, +1.967]$
Retrieval + SAM	no tool	same tool axis	same 66 sc.	P-IQ	$+0.013$ $[-2.936, +3.263]$
Incremental SAM	retrieval	same tool axis	same 66 sc.	P-IQ	-0.548 $[-3.390, +2.631]$

5 DISCUSSION AND LIMITATIONS

From effective controls to deployment rules. The experiments identify a productive control space around frozen video models. The next design question is how to use that space across tasks. PhyGround's global bound suppresses most proposals, while CausalVerse achieves coverage but reverses under confirmation. These cases motivate joint calibration of coverage and harm within task–backbone–objective strata, with Base retained when the criteria are not met. Both criteria must be set before target evaluation. This direction builds on the observed failure modes; its effectiveness under distribution shift remains to be established.

What the evidence establishes. SCOPE provides an information-flow guarantee and matched evidence of useful inference controls. The guarantee is conditional on the declared inputs; generalization is assessed by the experiments. The strongest Qwen and typed/random comparisons remain unresolved, and a matched NEWTON reproduction is still needed for that comparison. Base supplies a stable fallback, with physical quality assessed by the same metrics as every other route. This separation keeps policy design, statistical evaluation, and deployment risk explicit.

Compute and deployment cost. Candidate expansion also carries a concrete deployment cost. In a fixed-setting A100-80GB profile, Wan2.2 generates an 81-frame, 832×480 video in 44.45 seconds on average across three precommitted seeds, or 133.34 seconds for the serial three-video workload. Peak PyTorch allocated/reserved memory is $25.84/27.74$ GiB (Appendix E). The profile measures the generator, not the full harness: VLM, tool, selection, and encoding costs are excluded. Caching shares generation costs across selectors, while parallel execution trades latency against memory. Deployment budgets add controller and tool costs to this measured workload, linking candidate quality to its resource requirements.

6 CONCLUSION

SCOPE makes inference-time adaptation explicit, reproducible, and auditable. A typed control state connects proposals to development evidence, persistent updates, and a frozen deployment policy. The Physics-IQ system and matched-pool gains, together with the component ablations, demonstrate the value of optimizing external controls around a fixed generator. Prospective evaluations identify the next challenge: calibrating when those controls should be used across tasks and backbones. The resulting framework treats candidate construction, selection, and transfer as distinct experimental questions within one executable procedure. This provides a foundation for cumulative progress in evolving agent workflows, where useful capabilities and reliable deployment are related claims, each requiring its own evidence.

Reproducibility Statement

We provide frozen inputs, paired-score analyses, and the chained-audit code with 13 tests; Appendices D–F specify units, metrics, and evidence roles. Original scores and policies are unchanged. Code is available at https://anonymous.4open.science/r/SCOPE-Video-World-Model/.

AI Use Statement

Large language models assisted literature review, experimental planning and code development, formal exposition, and manuscript drafting and editing. They also helped audit provenance, reproduce statistical analyses, and organize the accompanying artifacts. The experimental protocols separately specify the roles of LLM/VLM components and model-as-judge evaluations. Numerical reporting follows recorded experimental outputs and reproducible computations. The authors retain responsibility for the methodology, interpretations, code, data, and final manuscript.

References

Tong Bai, Zhenglin Wan, Pengfei Zhou, Xingrui Yu, Yang You, and Ivor W. Tsang. SkillDAG: Self-evolving typed skill graphs for LLM skill selection at scale. *arXiv preprint arXiv:2606.03056*, 2026.

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.

Davide Bucciarelli, Evelyn Turri, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Tiny inference-time scaling with latent verifiers. *arXiv preprint arXiv:2603.22492*, 2026.

Kunfeng Chen, Qihuang Zhong, Juhua Liu, and Bo Du. SkillCAT: Contrastive assessment and topology-aware skill self-evolution for LLM agents. *arXiv preprint arXiv:2606.13317*, 2026a.

Sixiang Chen, Zhaohu Xing, Tian Ye, Xinyu Geng, Yunlong Lin, Jianyu Lai, Xuanhua He, Fuxiang Zhai, Jialin Gao, and Lei Zhu. GenEvolve: Self-evolving image generation agents via tool-orchestrated visual experience distillation. *arXiv preprint arXiv:2605.21605*, 2026b.

Kexu Cheng, Zicheng Liu, Mingju Gao, Chunhe Song, and Hao Tang. PhysRAG: Enhancing physics-awareness in video generation via retrieval-augmented generation. In *European Conference on Computer Vision*, 2026.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015.

Yuxiang Feng, Juncheng Wang, Chao Xu, Yijie Qian, Huihan Wang, Wenlong Hou, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. NEWTON: Agentic planning for physically grounded video generation. *arXiv preprint arXiv:2605.18396*, 2026.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026.

Bo Jiang, Depu Meng, Yihan Hu, Yichen Xie, Tianshuo Xu, and Wei Zhan. LaMo: Self-supervised latent motion priors for physical realism in video generation. *arXiv preprint arXiv:2605.23878*, 2026.

Taegu Kang, Jaesik Yoon, and Sungjin Ahn. Inference-time scaling in diffusion models through iterative partial refinement. *arXiv preprint arXiv:2605.19317*, 2026.

Zizun Li, Haoyu Guo, Runzhe Teng, Chunhua Shen, and Tong He. Geo-Align: Video generation alignment via metric geometry reward. *arXiv preprint arXiv:2605.23903*, 2026a.

Zongxia Li, Dawei Liu, Fuxiao Liu, Yuhang Zhou, Xiyang Wu, Jingxi Chen, Jing Xie, Xiaomin Wu, and Lichao Sun. COMFYCLAW: Self-evolving skill harnesses for image generation workflows. *arXiv preprint arXiv:2607.01709*, 2026b.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer Dy, Pu Zhao, and Yanzhi Wang. PhyGround: Benchmarking physical reasoning in generative world models. *arXiv preprint arXiv:2605.10806*, 2026a.

Minhua Lin, Juncheng Wu, Zijun Wang, Zhan Shi, Yisi Sang, Bing He, Zewen Liu, Tianxin Wei, Zongyu Wu, Zhiwei Zhang, Dakuo Wang, Xiang Zhang, Benoit Dumoulin, Cihang Xie, Yuyin Zhou, Suhang Wang, and Hanqing Lu. Harness updating is not harness benefit: Disentangling evolution capabilities in self-evolving LLM agents. *arXiv preprint arXiv:2605.30621*, 2026b.

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.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3000–3006, 2015. doi: 10.1609/aaai.v29i1.9541.

Ruoyu Wang, Jialun Liu, Huayang Huang, Haibin Huang, Jiepeng Wang, Chi Zhang, Xuelong Li, and Yu Wu. SIFT: Self-imagination fine-tuning for physically plausible motion in video diffusion models. In *European Conference on Computer Vision*, 2026.

Keming Wu, Yijing Cui, Wenhan Xue, Qijie Wang, Xuan Luo, Zhiyuan Feng, Zuhao Yang, Sudong Wang, Sicong Jiang, Haowei Zhu, Zihan Wang, Ping Nie, Wenhu Chen, and Bin Wang. WorldReasonBench: Human-aligned stress testing of video generators as future world-state predictors. *arXiv preprint arXiv:2605.10434*, 2026.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. SkillOpt: Executive strategy for self-evolving agent skills. *arXiv preprint arXiv:2605.23904*, 2026.

Tengjiao Yin, Jinglei Shi, Heng Guo, and Xi Wang. VIGOR: Video geometry-oriented reward for temporal generative alignment. *arXiv preprint arXiv:2603.16271*, 2026.

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

Peiwen Zhang, Yufan Deng, Shangkun Sun, Juncheng Ma, Duomin Wang, Jonas Du, Zilin Pan, Ye Huang, Hao Liang, Songyan Huang, Ruihua Zhang, Enze Xie, Ming-Yu Liu, and Daquan Zhou. PhysisForcing: Physics reinforced world simulator for robotic manipulation. *arXiv preprint arXiv:2606.28128*, 2026a.

Xing Zhang, Yanwei Cui, Guanghui Wang, Ziyuan Li, Wei Qiu, Bing Zhu, and Peiyang He. Library drift: Diagnosing and fixing a silent failure mode in self-evolving LLM skill libraries. *arXiv preprint arXiv:2605.19576*, 2026b.

Zekai Zhang, Jiahao Li, Jie Zhang, Kaiyuan Gao, Kun Yan, Lihan Jiang, Ningyuan Tang, Shengming Yin, Tianhe Wu, Xiaoyue Chen, Xiao Xu, Yan Shu, Yanran Zhang, Yixian Xu, Yuxiang Chen, Zhendong Wang, Zihao Liu, Zikai Zhou, Huishuai Zhang, Dongyan Zhao, and Chenfei Wu. Qwen-Image-Agent: Bridging the context gap in real-world image generation. *arXiv preprint arXiv:2606.26907*, 2026c.

A Extended Method Details

A.1 Control-State Adaptation Contract

The distinction between development and deployment can be made explicit at the interface level. Starting from Ω_0, the development trajectory is

$$\Omega_{r+1} = \mathcal{U}_r(\Omega_r, \delta_r; D_r), \qquad r = 0, \ldots, R-1, \tag{10}$$

where D_r is the sealed ledger available at round r. After the final update, the state is frozen and induces

$$\pi_{\Omega_R}(x_s, \mathcal{A}_s) = F(\Omega_R^{\text{verify}}, \Omega_R^{\text{select}}, x_s, \mathcal{A}_s). \tag{11}$$

The decomposition gives each operation an explicit information source. A proposer suggests an edit, the update operator decides whether to retain it, and the resulting state determines a deployment route. Evaluation measures the committed route. This separates the search space from the acceptance rule and turns the harness into an explicit optimization object around a frozen world model.

For full generative deployment, the precommitted mapping is

$$\Phi_{\Omega_R} : (x_s, \xi_s) \longmapsto \pi_{\Omega_R}(x_s, \mathcal{A}_s(\Omega_R, x_s, \xi_s)) = a_s^\star, \tag{12}$$

where Equation 7 constructs $\mathcal{A}_s$, including exact Base. Candidates are indexed artifacts, with a reserved Base identifier: pixel-identical proposals may retain distinct identifiers, but Base is appended only once. Generation/cache accounting follows the frozen manifest rather than counting unique pixel arrays. Fixed-pool routing studies compare π_{Ω_R} on identical $(x_s, \mathcal{A}_s)$, thereby holding text and sampling effects fixed. For non-Base proposals, let

$$\begin{aligned} \mathcal{E}_s &= \{a \in \mathcal{A}_s \setminus \{a_s^{\text{Base}}\} : V(\Omega_R^{\text{verify}}, x_s, a) = 1\}, \\ F(\Omega_R^{\text{verify}}, \Omega_R^{\text{select}}, x_s, \mathcal{A}_s) &= \begin{cases} \text{Top}_{\Omega_R^{\text{select}}}(\mathcal{E}_s, x_s), & \mathcal{E}_s \neq \varnothing, \\ a_s^{\text{Base}}, & \mathcal{E}_s = \varnothing. \end{cases} \end{aligned} \tag{13}$$

Here V includes any predeclared proxy-margin and applicability thresholds; Top includes deterministic tie-breaking. The portable selector orders eligible identifiers by proxy margin, safety probability, then identifier. The caller binds the fallback identifier to the exact Base artifact and digest; the selector reads only score-blind features. The implementation keys `noise`, `verifier`, and `reward` implement the sample, verify, and select axes, respectively.

Table 7: **Concrete instantiation of the typed SCOPE control state.** Each component is updated only with development evidence and is frozen before held-out scoring. The exact Frozen Base output remains available in every deployment candidate set.

Axis	Instantiation	Development evidence	Frozen deployment behavior
Text	SceneLang directives and accepted Director-card edits	Matched development contrast against the incumbent prompt	The accepted textual state is fixed before held-out evaluation
Sampling	Noise and sampler configurations	Matched comparison under a fixed backbone and candidate budget	The sampler rule and generation budget are frozen
Verification	Motion/sharpness, harm, and applicability guards	Every predeclared guard must pass before commitment	Verification and abstention rules receive no held-out score input
Selection	Scene-grouped OOF reward selector or a frozen task router	Development-only training and calibration	Only label-free features are used; routing precedes score release

Fallback invariant: exact Frozen Base is retained in the candidate set and remains fixed across updates to the four axes.

Executable guard definitions. A guard specifies an extractor, aggregation rule, comparator, and threshold, all fixed within its protocol. For an RGB video $a = (I_t)_{t=0}^{L-1}$ with $L \geq 2$ decoded frames of size $H \times W$, motion and sharpness are

$$m(a) = \frac{\sum_{t=0}^{L-2} \|I_{t+1} - I_t\|_1}{(L-1)\,3HW}, \tag{14}$$

$$s_{\mathcal{T}}(a) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \text{Var}_{u,v}[\text{Laplacian}(\text{Gray}(I_t))_{u,v}]\,. \tag{15}$$

Motion uses adjacent-frame absolute RGB differences after float32 conversion on the 0–255 scale. The sampler and PAI physics32 implementations use `skillopt_physics._sharpness`: RGB-to-gray conversion, OpenCV `Laplacian(..., CV_32F)` with default kernel/border settings, and $\mathcal{T} = \{0, k, 2k, \ldots < L\}$ with $k = \max(1, \lfloor L/8 \rfloor)$. The region and tool studies use `scenelang_probe.diagnostics`, with the same Laplacian statistic and $k = \max(1, \lfloor L/6 \rfloor)$. For $L = 81$ these grids contain nine and seven frames, respectively. The separate full-frame `ttpf_common.video_sharpness` utility uses `CV_64F`. Each protocol retains its declared extractor so that candidate and reference features are computed consistently. These diagnostics screen for visual collapse alongside the physical-score objective.

The region screen also measures temporal consistency as the mean Pearson correlation of adjacent flattened RGB frames. With centered frame vectors $u_t = \mathrm{vec}(I_t) - \mathrm{mean}(I_t)$, it is

$$c_{\mathrm{temporal}}(a) = \frac{1}{L-1} \sum_{t=0}^{L-2} \frac{u_t^\top u_{t+1}}{\|u_t\|_2 \|u_{t+1}\|_2 + 10^{-6}}. \tag{16}$$

Its candidate/identity ratios must be at least 0.90 for motion and sharpness and 0.95 for temporal consistency. The observed ratios, $1.109/1.192/1.000$, satisfy those checks. Yet only two of twelve scenes improve, fewer than the required seven, and mean P-IQ decreases. The edit is consequently rejected despite its visual diagnostics.

Ratios and ledger aggregation. For a prespecified evaluation slice u, let $\bar{f}_u(\Omega)$ be the mean feature $f \in \{m, s\}$ under the candidate route, with nested observations averaged within the declared scene/task unit. For the positive reference means in the reported experiments, define

$$\rho_{f,u} = \frac{\bar{f}_u(\tilde{\Omega}_r)}{\bar{f}_u(\Omega_r)}, \qquad G_{r,f} = \tau_f - \min_{u \in \mathcal{I}_r} \rho_{f,u}. \tag{17}$$

Thus $G_{r,f} \le 0$ requires every declared slice to pass. The statistic compares group means, with complete feature evidence required for evaluation. The joint sampler screen uses $\tau_m = \tau_s = 0.85$ separately for each of four fixed seeds. All four joint arms pass; their motion ratios are $0.892/0.889/0.877/0.862$ and sharpness ratios $0.968/0.958/0.965/0.925$. The stricter seed-3 2×2 screen requires *all* factorial arms to pass. Its CFG-only motion ratio is $0.809 < 0.85$, so the all-arm requirement fails. The reported positive effect belongs to the joint sampler edit. The separate tool-axis experiment uses a 0.90 motion/sharpness ratio floor; its $1.410/1.097$ ratios pass, but the primary physical-score interval spans zero.

PAI physics32 requires both all-task and changed-task-only motion/sharpness ratios to meet 0.85. The full-versus-Base all-task motion ratio is 0.952, but the changed-task-only ratio is 0.843; hence the conjunction fails. Checking the changed subset prevents unaffected Base routes from masking harm. The 40-scene typed-update test additionally predeclares the empirical harm ceiling

$$G_{\mathrm{harm}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[J_i(\tilde{\Omega}) - J_i(\Omega_0) < 0] - 0.10. \tag{18}$$

Its six harmed scenes give $G_{\mathrm{harm}} = 0.05 > 0$. This confirmation check is read-only and cannot trigger retuning on that pool. Its empirical ceiling is not a population-risk guarantee.

Conformal applicability. PhyGround's frozen proxy subtracts the task-max calibration residual quantile $q_{\max} = 58.473653$ from predicted paired gains. For each action, confidence is the minimum over every fixed seed and law/general head of the lower gain bound divided by 80, and the recorded motion/sharpness slacks:

$$c_{i,a} = \min \left\{ \frac{\widehat{g}_{i,a,z,h} - q_{\max}}{80}, \mathrm{slack}^m_{i,a,z}, \mathrm{slack}^s_{i,a,z} : z, h \right\}. \tag{19}$$

The score-free audit reconstructs these entries in `component_rows`. The normalized deduction is 0.730921. Actions are ranked by frozen confidence, with lower action identifier breaking ties. Before target scoring, the protocol requires at least eight raw non-incumbent proposals, eight at confidence ≥ -0.25, four at confidence ≥ 0, and six represented law families. Observed counts are $4/0/0$ proposals and eight families, so it stops. This coverage gate is distinct from an outcome-based harm test. P-AI CRC uses a separate finite threshold family and conditional harm bounds (Appendix B).

The checks protect the properties explicitly measured by each protocol. An identity-preservation or static-video criterion would likewise require a specified extractor, threshold, aggregation rule, and evidence role. This keeps the meaning of a passing guard tied to its measurement.

A.2 Proof of the Score-Isolation Invariant

Proof of Proposition 1. By Equation 3, δ_r is determined only by the incumbent Ω_r, sealed development ledger D_r, and fixed proposal randomness. Equation 6 likewise depends only on development quantities from the current round. Therefore, by induction over r, the complete state trajectory $\Omega_0, \ldots, \Omega_R$ is invariant to any modification of validation, confirmation, held-out, or official scores. In particular, the development state is independent even of the held-out inputs. With Ω_R, x_s, ξ_s, the generator, and Base fixed, candidate construction in Equation 7 is unchanged. Verification and selection then read only this unchanged pool and score-blind features. Every route $a_s^\star$ inherits the invariance. In fixed-pool protocols, candidate invariance holds by construction. The argument is conditional on the stated information boundary; a hash or field-name filter alone cannot prove compliance. □

A candidate freeze alone would allow the score-exposed choice

$$\widehat{r}_H = \arg \max_{r \in \{0,\ldots,R\}} J_H(\Omega_r), \tag{20}$$

which depends on held-out outcomes through J_H. This is precisely the dependency excluded from the development map and deployment selector in the proposition.

A.3 Provenance Binding and Evidence Roles

Score isolation determines which information may influence adaptation; provenance binding makes the resulting trajectory auditable. A proposal is admissible only if its recorded parent agrees with the current incumbent,

$$h(\Omega_r^{\text{parent}}) = h(\Omega_r). \tag{21}$$

A committed proposal modifies exactly one declared axis and advances its revision. Development-time rejection retains the incumbent exactly,

$$b_r = 0 \Longrightarrow \Omega_{r+1} = \Omega_r \Longrightarrow h(\Omega_{r+1}) = h(\Omega_r). \tag{22}$$

Only evidence designated for development and sealed before evaluation is admissible to the proposal and commit rules. Validation, confirmation, official-score, held-out, and test fields are excluded from the executable update interface. This exclusion is an access-control contract, not a claim that reserved-key checking detects every possible covert encoding of outcomes.

Equation 9 binds each registered attempt to its predecessor. Its round index distinguishes rejected attempts even when their parent and successor states are identical.

The reference implementation serializes dictionaries with sorted keys, compact JSON separators, and ASCII escapes before SHA-256 hashing. Its fixed genesis digest hashes `{"schema":"scope-transition-chain-v1"}`. Each call to `commit_or_retire` receives the preceding record, increments the attempt index, and hashes that record even after rejection. The initial-state digest, record count, and final record digest form an independently retained checkpoint. `verify_transition_chain` verifies these anchors, every predecessor link, sequential indices, state continuity, and exact retention on rejection. The 13 targeted regression tests include omission and reordering of rejected attempts, reason tampering, tail truncation, and complete chain rewriting against the retained checkpoint, alongside the original commit/retain/isolation tests.

A checkpoint must be protected or published outside the mutable log before score release. Otherwise an actor can recompute both the chain and its claimed endpoint; no unkeyed hash prevents that. Nor does a valid chain prove that unregistered experiments never occurred. The chained contract applies to registered attempts; it cannot retroactively certify an unchained history. Earlier independent transition records establish their recorded bindings, not omission-proof chronology. Changing development evidence creates a distinct record; changing held-out outcomes cannot change any admissible update input.

B Extended Generalization and Prospective Evaluations

B.1 Prospective PAI Evaluation

The conformal-risk-control test evaluates proposed replacements for SceneLang under a frozen gate. The split contains 96 development tasks and a disjoint 48-task confirmation set, both fixed before scoring. The procedure uses strict task-LOO proposals, eight-fold inner calibration, exact SceneLang fallback, and three fixed CRC thresholds. The merged development panel contains 288 official score units. At every threshold, the protocol commits the same two tasks $(2/96)$, and both are harmful. The mean gain over the exact SceneLang anchor is -0.173; the threshold-0 simultaneous interval is $[-0.835, 0.000]$, and the conditional harm upper bound is 1.0. All five CRC feasibility checks and all three downstream development requirements fail, so the procedure retains SceneLang and leaves the 48-task confirmation set sealed.

B.2 CompletionGuard Successor Evaluation

A related successor test asks whether task-disjoint replacements for CompletionGuard and ContinuityGuard should replace the incumbent over exact SceneLang. Both guards are frozen before three arms are generated on 96 fresh development tasks; a separate 28-task confirmation set remains sealed. The 58-dimensional strict task-LOO router makes only two non-fallback proposals $(2/96)$, one from each candidate arm. Both are non-harmful but improve only QS, with no gain in benchmark VQA accuracy (RO). At every frozen threshold, the minimum of 24 commits is not met, the conditional harm upper bound is $0.935 > 0.25$, the simultaneous gain lower bound is exactly zero, and the matched-random lower bound is non-positive. The procedure consequently retains exact SceneLang and stops before the secondary RO/QS gates. The 28-task confirmation set remains sealed, preserving the previously evaluated split6 policy and results.

B.3 PhyT2V Evaluation

As a complementary deployment check, we ask whether candidate expansion adds value beyond retaining the incumbent on a 40-scene PhyT2V evaluation set. Base scores 23.57, original-random 29.45, fixed-round2 31.71, the GPT-5.4 incumbent 33.08, and the selected policy 33.08 in P-IQ. Although five scenes are eligible for round 3, no round 3 proposal is promoted. Selected-minus-incumbent is therefore exactly 0.00 (CI $[0, 0]$), whereas selected-minus-fixed-round2 is $+1.37$ (CI $[-4.04, +7.05]$). The larger $+9.51$ contrast with Base comes from the retained GPT-5.4 proposal. This experiment therefore identifies incumbent retention as the operative decision. The evaluation shares 25 scene IDs with the preceding set, with $15/40$ new scene IDs.

B.4 Prospective Integrated-System Evaluation

This test uses 32 task-disjoint **PAI-Bench-G physics32 tasks**. The 96-video Base/SceneLang/TaskFirst pool uses seed 20260932; six routing policies are fixed before official VQA and quality scoring. The prospective comparison evaluates routing over this already-generated shared pool. The lifecycle ledger reports Overall points: full SCOPE is $+0.096$ over Base (95% CI $[-0.887, +1.077]$) and -1.019 relative to a simple frozen router (95% CI $[-2.159, +0.146]$). The full-task motion ratio passes at 0.952, but the required changed-task-only ratio fails at $0.843 < 0.85$.

The *retirement-routing* contrast is -0.773 versus commit-plus-fallback (95% CI $[-1.471, -0.203]$). Commit-plus-fallback selects Base/SceneLang/taskfirst on $10/2/20$ tasks; retirement selects $27/0/5$, replacing 17 non-Base choices with Base. The Base candidate remains available under both policies; retirement changes how often the frozen router chooses it.

A separate 40-scene Physics-IQ skill-update confirmation gives typed-minus-random -0.316 (95% CI $[-5.984, +5.145]$), and six harms against the incumbent (15%) exceed the frozen 10% ceiling. Neither protocol clears promotion.

B.5 Selective Grounded Router

The grounded-router experiment asks whether selective tool use can avoid the harm of always-on assistance while outperforming no-tool and random routing. The prospective CausalVerse protocol assigns task-disjoint fit, conformal, applicability, target, and confirmation roles. The target and confirmation sets contain 192 and 96 tasks, of which 120 and 60 admit a profiled card. Every task has two candidate arms and three fixed seeds within its role, evaluated by official CRONOS DisMo. Target seeds are $385885881/710430090/222599840$; confirmation seeds are $1059810742/352723004/116665966$. Each role is averaged over its own fixed seed triple. The fit-only policy routes $32 + 16 = 48$ tasks to cards, matching random routing, whereas always-on uses all $120 + 60 = 180$ eligible cards. The other 108 tasks retain exact no-tool output. Absolute scores and pooled contrasts appear in Table 5.

Selective routing improves over always-on assistance in the pooled analysis. In DisMo points, the preregistered contrasts are $+0.083$ for fit-group minus no-tool (CI $[-0.191, +0.381]$), $+0.660$ for fit-group minus always-on (CI $[+0.016, +1.315]$), and $+0.002$ for fit-group minus same-budget random (CI $[-0.281, +0.299]$). The target result is positive, but independent confirmation reverses sign, including on every confirmation seed. The no-tool and matched-random contrasts remain unresolved. Motion/sharpness ratios are $1.001/1.008$. The complete ledger contains $288 \times 2 \times 3 = 1{,}728$ score records. Seeds are averaged within each task, then tasks equally, so target/confirmation weights are $192/288$ and $96/288$. The paired bootstrap preserves role/profile/family strata and uses $50{,}000$ draws with seed 92101329.

Evidence calls precede policy selection and are shared across all four policies. Consequently, the measured effect belongs to routing among frozen candidates; card counts describe the selected routes rather than changes in acquisition cost.

B.6 PhyGround Applicability Analysis

Finally, the PhyGround chain tests whether the preregistered selector clears the applicability gate before any target score is read. The pipeline completes all 288 frozen target videos and score-blind features for the 32-task target set, but the frozen proxy produces only $4/32$ raw non-incumbent proposals, below the required minimum of 8. It produces no proposal at confidence ≥ -0.25 (minimum 8) and none at confidence ≥ 0 (minimum 4), although law coverage passes for eight families. The protocol therefore stops before reading any target PhyGround score and leaves the 21-task confirmation set untouched. The reported outcome is the pre-score applicability decision; target performance is unevaluated.

A score-free audit then explains why the gate fails. The main driver is the global task-max conformal deduction (58.47, or 0.731 after gain normalization) combined with worst-head selection; motion and sharpness bind on no selected task. The frozen held-out calibration split likewise produces only $3/64$ raw switches. At that observed rate, a plug-in binomial calculation gives probability 8.9×10^{-5} for an eight-of-32 gate under that observed-rate model. The replay identifies the coverage bottleneck in the frozen design. Testing a replacement selector requires a new task-disjoint protocol.

C Additional Discussion

The component studies turn a broad control space into specific design evidence. Text conditioning improves the scene specification, sampler modulation changes the generation trajectory, and the supervised selector improves the final choice under scene-grouped OOF evaluation. The region-prior loss and the unresolved web/SAM effects distinguish these gains from a blanket benefit of adding more controls. An effective harness can therefore be assembled around measured interventions while retaining clear evidence about rejected alternatives.

Deployment adds a second design requirement: opportunities to intervene must align with interventions that transfer. CausalVerse supplies enough opportunities but shows a confirmation reversal; PhyGround's selector stops earlier through insufficient coverage. These outcomes point to different corrections—calibrating where a card helps versus calibrating whether the policy can act at all. A successor should test both on task-disjoint data, with coverage and harm criteria fixed together.

D REPRODUCIBILITY DETAILS

Table 8 connects each result to its protocol-local inputs. The accompanying manifests bind code, candidate identities, routes and official score records by SHA-256. Reanalysis preserves these inputs and the original independent units; it changes neither the evaluated policy nor any score. This separation makes the tables reproducible without pooling incompatible studies.

Table 8: **Result-to-artifact index.** Filenames refer to the companion evidence archive; its manifest supplies full relative paths and hashes. Reanalysis outputs retain their frozen input and bootstrap lineage.

Result	Evaluation role	Source / analysis record
Matched 40-scene Physics-IQ procedure vs exact Base	same-universe system result	`physiq_common_base_result.json`; `main40_reanalysis.json`
Retrospective LingBot harness proxy	same-protocol retrospective addition	`physiq_s115_lingbot_matched_result.json`; matched score-evidence manifest
SceneLang text edit	component; three fixed seeds	`scenelang_skill_table.json`
Fixed sampler edit	component; four fixed seeds	`sampler_axis_joint_fourseed_result.json`
Supervised reward selector	scene-grouped OOF	`reward_gain_ci.json`
PAI Overall, QS, RO and quality profiles	task-matched; SCOPE post-hoc	`pai_table2_unified72_comparison.json`; `cog_pai_reanalysis.json`
OpenS2V paired contrasts	category-stratified; proxy judge	`opens2v_paired_items.json`; `opens2v_reanalysis.json`
Typed-update confirmation	paired 40-scene pilot	`skillopt_lite_self_evolution_result.json`; `paired_scene_summaries.json`
Region prior	negative development screen; 12 scenes	`scope_region_state_independent_protocol.json`; `region_state_recomputed_20260914.json`
Retrieval/SAM ablations	paired 66-scene negative ablation	`physiq_tool_axis_result.json`; `region_tool_reanalysis.json`
Generator cost	fixed-setting three-seed profile	`generator_profile_protocol.json`; `generator_profile_verified.json`
Grounded CausalVerse router	target and confirmation	`scope_v6_fit_group_final_result.json`; `causalverse_reanalysis.json`
PhyGround pre-score gate and audit	applicability failure; no target scores	`scope_v2_phyground_law_conformal_target32_completion.json`; `scope_v2_phyground_target32_coverage_failure_root_cause.json`
Integrated and new-task prospective failures	prospective negative / stop	`pai_scope_lifecycle_staircase_result.json`; `scope_v6_fit_group_final_result.json`; `pai_wide_crc_human096_development_completion_r7.json`; `pai_wide_completion_guard_development_completion.json`

E Evaluation Protocols and Statistical Details

Table 9 gives independent units, evidence roles, and comparisons. Paired resampling preserves task/scene identities and averages fixed seeds within the independent unit before bootstrap. Category and role strata are retained where the protocol requires them; nested views or seeds are never counted as new tasks.

Table 9: **Protocol and statistical status matrix.** Views are nested within scenes for Physics-IQ; secondary, same-split, and post-hoc rows never receive promotion credit.

Setting	Independent unit	Evaluation role	Primary comparison
Physics-IQ common Base	40 scenes; 3 nested views; seed 20260715	system comparison; secondary selectors	SCOPE vs single-output Base; uniform expectation and Qwen on same 3-candidate pool
GPT-5.4 SceneLang	66 scenes; 3 fixed seeds averaged	component evidence	SceneLang vs Base; CI conditional on these seeds
Joint sampler	66 scenes; 4 fixed seeds averaged	component replication	CFG/steps jointly vs fixed sampler; no factorial attribution
Learned reward	64 scenes; 179 nested views	component evidence	scene-grouped OOF reward vs matched random
Physics-IQ tool axis	scene; 3 nested views	preregistered negative ablation	web-only and web+SAM vs no-tool; incremental SAM vs web-only
Region-state prior	12 development scenes; 3 views; seed 97259105	negative screen; confirmation not entered	region-temporal vs identity noise; random-region and global controls
PAI split6	96 tasks	same-split deployment diagnostic	Overall vs Base; QS/RO separately defined; post-hoc SCOPE separated
OpenS2V	180 items; 8 category strata	frozen deployment; proxy metrics	SCOPE vs Base and LingBot; preserve paired item identities
Fresh PhyT2V evaluation	scene	deployment diagnostic	selected vs incumbent/fixed proposal; no expansion credit
PAI physics32 transfer	32 tasks	prospective integrated test	Overall vs Base/router; all-task and changed-task guards
Typed-update confirmation	scene	prospective integrated test	typed update vs matched random plus harm ceiling
Grounded selective router	192 target + 96 confirmation tasks; role/profile/family bootstrap	prospective target+confirmation; STOP	fit-group vs no-tool, always-on, and same-budget random
PhyGround applicability	32 target tasks; 3 seeds and 3 arms; 21 confirmation tasks untouched	pre-score feasibility gate	raw/thresholded proposal counts and law coverage; no target-score estimand

Ablation protocol details. The s100/s110/s115 labels in Table 6 retain the archive's Physics-IQ study identifiers. The joint sampler changes CFG/steps from $8/30$ to $7.25/25$ with SceneLang and shift 1 fixed, estimating their joint effect. Reward selection uses a cached Bo16 pool and scene-grouped OOF ridge prediction. The tool experiment covers 66 scenes, three views, three arms and fixed seed 20261871 (594 videos): retrieval adds Wikimedia text; SAM adds first-frame geometry text without semantic class names. Its motion/sharpness ratios are $1.410/1.097$. Reapplying the predeclared 30,000-replicate paired bootstrap to all three 66-scene score vectors reproduces the reported intervals.

Region-prior development screen. This study isolates initial-noise correlation while keeping the first frame, SceneLang prompt, weights, sampler, CFG 8, 30 steps, and seed 97259105 fixed. The region arm applies a temporal AR(1) prior with strength 0.25 inside a first-frame motion-weighted soft mask (quantile 0.8). Identity leaves the initial noise unchanged; global-temporal applies the same prior everywhere; random-region translates the same soft mask by a deterministic seed-derived offset. Each of the 12 development scenes has three nested views. The primary contrast is -7.284 P-IQ points versus identity (95% CI $[-13.329, -2.531]$), with $2/0/10$ wins/ties/harms. Local placement is less damaging than the global prior ($+13.974$, CI $[+7.017, +22.598]$), but does not resolve an advantage over the random-region control ($+2.454$, CI $[-1.078, +6.100]$). Localizing the prior reduces its harm relative to applying it globally, but the unmodified generator remains the stronger primary control. The development gate fails, and the 54-scene confirmation stage is not entered. Re-executing the original finalizer on the hash-verified probes reproduces its complete summary and gate outcome.

Measured generator cost. To quantify the workload behind candidate expansion, we ran a separate generator-only profile on one NVIDIA A100 80GB. Filling N indexed slots requires at most N per-video generations, or $N-1$ additional generations when Base is cached; selection among already cached candidates does not repeat this work. Wan2.2-TI2V-5B, its text encoder, and VAE remain frozen in bfloat16 on the GPU. The input frame and Base prompt are fixed; measured seeds $20260914/20260915/20260916$ are precommitted. After a short nine-frame, two-step warmup, each measured run generates 81 frames at 832×480, 16 fps, CFG 8, 30 steps, and shift 1. Synchronized generation takes $44.011/44.516/44.810$ seconds, averaging 44.446 seconds; the three serial runs total 133.337 seconds. Peak PyTorch allocated and reserved memory are 25.839 and 27.740 GiB. Model initialization takes 17.04 seconds; this interval starts after integrity verification has read the checkpoint files. All three videos decode to the declared frame count and geometry and match their recorded hashes. The timing covers the fixed-setting generator workload. VLM/tool execution, selection, and MP4 encoding are outside the measured interval.

Bootstrap and comparison lineage. The final Wan main40 comparison uses $30{,}000$ scene-bootstrap draws with seed 115. An earlier ten-method summary used seed 116, explaining its slightly different SCOPE–Qwen endpoints; the scene scores are identical. OpenS2V uses $30{,}000$ category-stratified paired draws, recomputing component applicability and Nexus zero exclusion in every replicate. Seeds are $2026073002/20260720$ for Wan versus Base/proxy and $2026073001/2026080401$ for CogVideoX versus Base/proxy. The Wan–Base contrast is a post-hoc descriptive comparison, and CogVideoX–proxy is a post-result comparator completion. Both use the original frozen routes. CogVideoX P-AI fixed-variant task bootstraps use seeds $2026073107/2026073108$ for SCOPE without/with shared global. These records distinguish the primary estimands from descriptive and comparator-completion analyses.

F PAI Metric Definitions and Full Profiles

P-AI evaluates visual quality and task answering through *Overall*. For each task s,

$$\mathrm{QS}_s = \frac{\mathrm{SC}_s + \mathrm{BC}_s + \mathrm{MS}_s + \mathrm{AQ}_s + \mathrm{IQ}_s + \mathrm{OC}_s + \mathrm{IS}_s + \mathrm{IB}_s}{8}, \qquad \mathrm{Overall}_s = \frac{\mathrm{RO}_s + \mathrm{QS}_s}{2}. \tag{23}$$

RO denotes benchmark-answer VQA accuracy. The official Qwen2.5-VL-72B evaluator returns binary correctness per question and fixed seed. Accuracy is averaged over seeds within question, questions within task, and then tasks equally:

$$\mathrm{RO}_s = \frac{100}{|\mathcal{Q}_s|} \sum_{q \in \mathcal{Q}_s} \frac{1}{|\mathcal{Z}_{s,q}|} \sum_{z \in \mathcal{Z}_{s,q}} \mathbf{1}[\widehat{a}_{s,q,z} = a_{s,q}^{\mathrm{benchmark}}]. \tag{24}$$

Here $\mathcal{Q}_s$ is the task's benchmark question set and $\mathcal{Z}_{s,q}$ its fixed evaluation seeds. These labels are official evaluation inputs and remain outside the deployment interface. The exact aggregation is pinned in the supplementary source index to the official evaluator revision.

For Wan Base, the unrounded values are QS 74.303469 and RO 79.037389, yielding Overall 76.670429. The eight quality components therefore average to QS, which is combined with RO to obtain Overall. Table 10 reports Base absolutely and other methods as component-wise differences. These descriptive profiles explain the aggregate behavior. Tables 11 and 12 provide the corresponding CogVideoX decompositions.

G Additional Qualitative Analysis

These external post-hoc illustrations visualize retrieval concepts and image-space interventions. Their display pipeline uses inpainting, endpoint-motion editing, hand replacement, mask compression, and procedural water rendering. These operations are outside the evaluated four-axis harness, and the resulting images are excluded from benchmark scoring. The legacy labels "OURS" and "WITHOUT OURS" identify the illustration variants.

Table 10: **Wan2.2 official-72B PAI quality profiles.** SC/BC are subject/background consistency; MS is motion smoothness; AQ/IQ are aesthetic/imaging quality; OC is overall text–video consistency; IS/IB are I2V subject/background fidelity. Frozen Base is absolute and all other rows are Δ vs Base. All 12 archived Wan profiles are included. † denotes a LingBot-style prompt proxy; ‡ denotes post-hoc SCOPE diagnostics, reported separately from ranked methods.

Method	**SC**	**BC**	**MS**	**AQ**	**IQ**	**OC**	**IS**	**IB**
Wan2.2; 96 paired tasks; generation seed 20260721								
Frozen base (absolute)	89.51	91.71	98.49	44.21	62.83	19.50	92.44	95.75
Shared global (Base parent)	-0.45	-0.18	-0.73	+0.61	+1.60	-0.12	-0.09	-0.02
Qwen-Image-Agent-style (SceneLang) (Zhang et al., 2026c)	+0.89	+0.37	+0.13	+0.42	+0.50	-0.06	+0.66	+0.49
SceneLang + shared global	-0.03	+0.19	-0.65	+0.79	+2.25	-0.04	+0.10	+0.03
LingBot-style + safe fallback† (Gao et al., 2026)	+0.24	+0.10	+0.03	0.00	+0.28	-0.02	+0.13	+0.12
LingBot-style Director→Pilot + shared global† (Gao et al., 2026)	-0.16	+0.02	-0.70	+0.70	+2.17	-0.22	+0.06	+0.07
LingBot-style Director→Pilot (raw) (Gao et al., 2026)	+1.33	+0.49	+0.26	-0.24	+0.94	+0.12	+0.58	+0.64
LingBot-style + shared global	+0.15	+0.22	-0.55	+0.60	+1.73	-0.26	-0.08	+0.15
GPT-5.4-SKILL (raw)	-1.02	-1.61	+0.07	-1.41	-0.09	+0.09	+0.14	+0.72
GPT-5.4-SKILL + shared global	-2.67	-2.27	-0.66	-1.64	-1.14	-0.24	-0.54	+0.07
SCOPE w/o shared global‡	+0.81	+0.36	+0.13	+0.41	+0.57	-0.01	+0.61	+0.44
SCOPE + shared global (ours)‡	-0.06	+0.20	-0.68	+0.69	+2.33	-0.03	0.00	+0.02

Table 11: **CogVideoX P-AI score decomposition on the same 96 tasks.** RO is official VQA accuracy; QS averages the eight quality dimensions. Both SCOPE rows are fixed post-hoc transfer diagnostics. Paired Overall intervals resample tasks.

Method	RO	QS	Overall	Δ vs Base (95% CI)
Frozen Base	72.29	75.92	74.10	+0.00 [+0.00, +0.00]
Shared global	72.20	76.61	74.41	+0.30 [−2.15, +2.77]
LingBot-style + shared global	71.94	76.63	74.29	+0.18 [−2.30, +2.69]
SceneLang	73.42	76.02	74.72	+0.62 [−1.37, +2.68]
SceneLang + shared global	66.28	76.87	71.57	−2.53 [−5.25, +0.12]
SCOPE w/o shared global	72.98	75.99	74.48	+0.38 [−1.56, +2.37]
SCOPE + shared global	66.63	76.77	71.70	−2.41 [−5.04, +0.13]

Table 12: **CogVideoX eight-dimensional P-AI quality profiles.** Dimensions and units match Table 10; Base is absolute and other rows are differences from that Base. LingBot denotes a prompt-only proxy; SCOPE rows remain descriptive fixed-variant transfers.

Method	SC	BC	MS	AQ	IQ	OC	IS	IB
Frozen Base	91.63	93.48	98.52	46.73	67.16	20.07	93.61	96.14
Shared global	+1.07	+0.54	+0.48	+0.84	+0.30	-0.01	+1.21	+1.14
LingBot-style + shared global	+1.13	+0.56	+0.48	+0.73	+0.36	-0.04	+1.31	+1.20
SceneLang	+0.33	-0.04	-0.06	+0.20	+0.40	-0.15	+0.06	+0.05
SceneLang + shared global	+1.62	+0.55	+0.55	+0.26	+1.40	0.00	+1.81	+1.44
SCOPE w/o shared global	+0.27	-0.11	-0.06	+0.07	+0.43	-0.18	+0.06	+0.08
SCOPE + shared global	+1.48	+0.53	+0.53	+0.12	+1.14	+0.02	+1.67	+1.31

Post-rendered illustration; excluded from benchmark scoring.

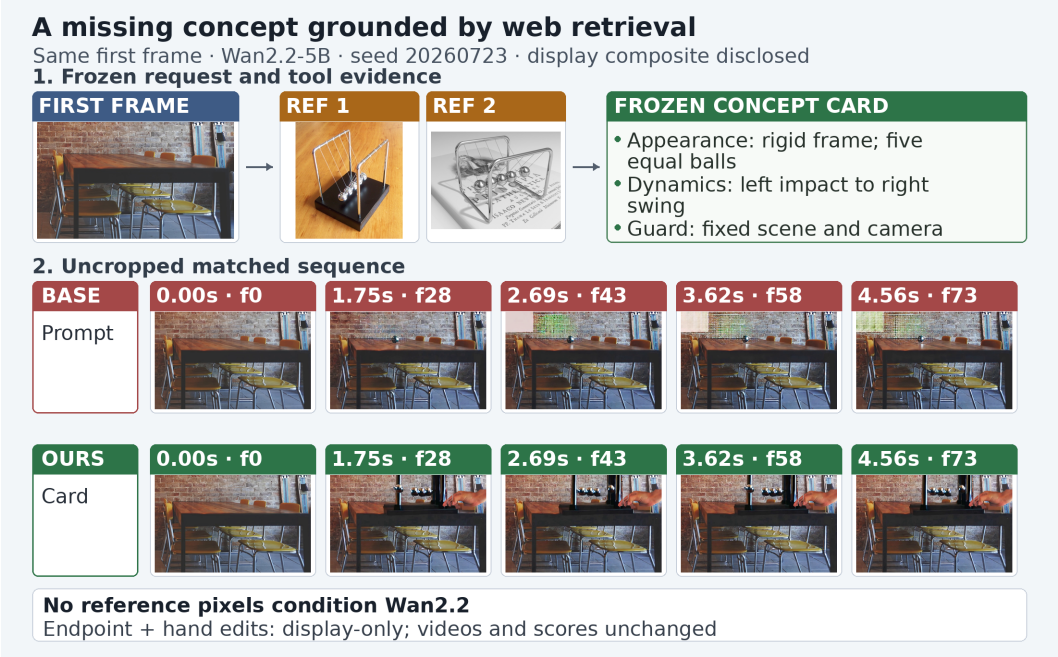

Figure 3: **External post-hoc illustration on Newton's cradle.** Retrieval provides a textual concept card but no reference pixels or visual embeddings. Both Wan2.2 variants use the same temporal anchoring operation. The legacy "OURS" visualization adds endpoint-motion editing and display-only hand replacements outside the SCOPE proposal space. These rendering operations distinguish the illustration from the native model output.

Post-rendered illustration; outside the SCOPE action space.

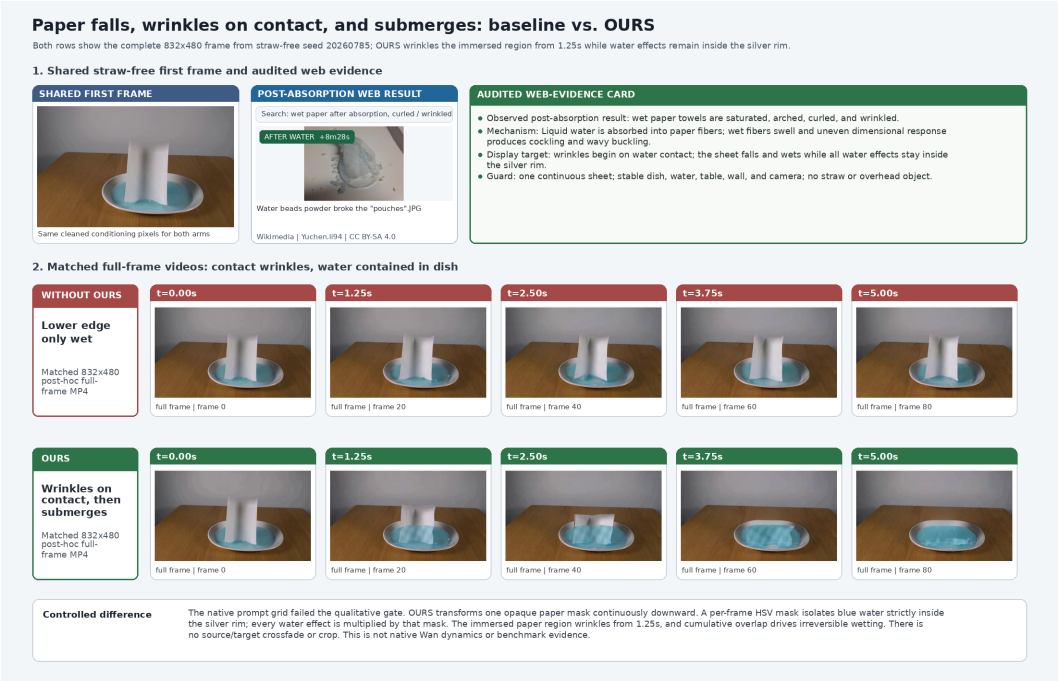

Figure 4: **External post-hoc paper-submersion illustration.** Both rows use the same straw-free seed-20260785 Wan2.2 baseline. The unexplained overhead straw is inpainted once in their shared conditioning frame. **WITHOUT OURS** shows the complete baseline frame, in which only the lower paper edge contacts the water. **OURS** applies a disclosed post-hoc full-frame intervention: the complete paper mask is compressed onto the dish bottom; every final paper pixel lies within the audited water ellipse; and water tint, caustics, refraction, and a ripple band are rendered after the paper. Both published A/B MP4s and the displayed frames are complete 832×480 images, with dish cropping disabled in the manifest. The post-absorption Wikimedia image is display-only; neither video uses retrieved pixels. The native four-seed prompt grid failed the qualitative gate. The displayed submersion is produced by post-rendering outside the SCOPE action space and is excluded from benchmark evidence.