NEGATIVESPACE-WM: LEARNING TO RULE OUT IMPOSSIBLE FUTURES FOR WORLD-MODEL PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

World models help agents choose actions by predicting their consequences, but generating physically impossible futures wastes computation. Rejecting these futures before generation requires identifying physical conflicts while preserving feasible alternatives. We propose NegativeSpace-WM, a framework that learns minimal impossibility certificates to guide this decision. Each certificate describes a physical conflict, a repair with no unnecessary edits, and the futures that the conflict rules out. A shared model trained on valid simulated events and their violating and repaired variants learns to recognize conflicts in videos and predict them from observations, actions, and brief descriptions of proposed futures. These predictions guide independent physical checks; generation is skipped only when the checks rule out every future allowed by the candidate request. Experiments show that minimal repairs yield a 2.2-point gain in balanced accuracy on unseen physical settings over redundant repairs of equal length. Certificate guidance also helps the verifier reach definite conclusions for more candidates than unguided search at the same checking budget. Across four planning task families, NegativeSpace-WM uses 33.20% less total GPU time per episode than generating every candidate, with task success rates of 81.6% versus 82.5% under identical planning conditions.

1 INTRODUCTION

World models support planning by predicting how an environment will respond to an agent's actions. Advances in learned dynamics and video prediction have expanded their use in model-based control and embodied decision-making (Chua et al., 2018; Hafner et al., 2019; 2025a; Assran et al., 2025). As candidate pools and planning horizons grow, repeatedly generating and evaluating futures can consume a substantial share of the planning budget. With limited decision time, computation spent on implausible futures leaves less time to compare useful alternatives. A planner therefore needs to decide which proposed futures warrant detailed prediction and which can already be ruled out. Physical constraints offer an opportunity to make this decision before paying the cost of a full rollout.

Recognition benchmarks test whether models distinguish physically valid events from violations, while other approaches assess the consistency of generated videos (Riochet et al., 2018; Bordes et al., 2025; Yuan et al., 2026). For planning, a plausibility score can prioritize candidates and independent physical checks can establish a conflict. The challenge is to connect these capabilities before full generation: the learned model must help find useful evidence, and the verifier must determine which futures that evidence actually excludes. For example, a wall may block a proposed crossing while leaving stopping or rebound motions feasible under the same action; a useful rejection should preserve those alternatives for the planner to compare. Otherwise, screening can either discard feasible alternatives or spend so much on checking that it loses the computational benefit of early rejection.

This motivates our central question: *can learning why a future is physically impossible improve both physical reasoning and the efficiency of world-model planning?* We propose NegativeSpace-WM, a framework that learns the physical reasons that make candidate futures impossible and uses them to guide selective generation. It represents these reasons as *minimal impossibility certificates*, connecting a physical conflict to the changes needed to resolve it and the alternatives it rules out. A shared Critic and Gate learn this representation from counterfactual events, linking video recognition

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.

to screening before rollout. Predicted certificates then guide independent physical checks, so that learning identifies useful evidence and verification establishes exclusion before full generation.

We evaluate physical reasoning and planning efficiency across five categories of physical constraints and four planning task families. Controlled comparisons test whether minimal repairs improve representations, whether certificate content guides verification, and whether these gains translate into lower complete planning cost. The results show stronger transfer and more effective use of a fixed verification budget. Certificate-guided planning also reduces GPU time relative to Full Rollout, with task success reported alongside cost. These findings connect the learning value of physical explanations to decisions about where a world model should spend computation.

Our contributions are threefold:

1. **Method.** We propose NegativeSpace-WM, which learns minimal impossibility certificates and combines them with independent verification to guide selective world-model generation.
2. **Results.** At equal target length, minimal certificates improve transfer balanced accuracy by 2.2 percentage points over redundant targets. Planning uses 33.20% less GPU time per episode, with 81.6% success versus 82.5% for Full Rollout under identical branch contracts.
3. **Community resources.** We provide a reusable certificate specification and controlled evaluation protocols that separate representation quality, verification, and complete planning cost. Code will be released to support reproducible comparisons and further development.

2 Related Work

World models and planning. DINO-WM and V-JEPA 2 plan in learned feature spaces (Zhou et al., 2025; Assran et al., 2025), Dreamer 4 trains policies in imagination (Hafner et al., 2025b), and PAN and DreamZero model action-conditioned futures (PAN Team, 2025; Ye et al., 2026). GVP-WM grounds generated plans in feasible trajectories (Ziakas et al., 2026). Diffusion Policy and Cosmos provide action modeling, reasoning, and generation (Chi et al., 2023; NVIDIA, 2025; 2026). Our approach complements these models by screening proposed futures before detailed prediction. We keep the world model and planner fixed, so the comparison isolates how learned physical evidence directs checking effort and determines which candidates warrant the cost of generation.

Physical reasoning and supervision. PhysBench broadens physical understanding evaluation (Chow et al., 2025), complementing event benchmarks (Weihs et al., 2022; Jassim et al., 2024). Physics-IQ, Morpheus, VideoPhy-2, WorldModelBench, and WorldReasonBench assess physical fidelity and future-state consistency (Motamed et al., 2026; Tragoudaras et al., 2026; Bansal et al., 2026; Li et al., 2025; Wu et al., 2026). PLATO, CLEVRER, and IPV-Bench cover object prediction, counterfactual reasoning, and impossible scenarios (Piloto et al., 2022; Yi et al., 2020; Bai et al., 2025). WISA conditions generation on physical attributes (Wang et al., 2025), TRAVL trains plausibility judges (Motamed et al., 2025), and LikePhys uses likelihood preference (Yuan et al., 2026). Our supervision couples physical judgments with executable inclusion-minimal repairs and exclusion scope.

Diagnosis and verification. Diagnosis and QuickXplain formalize consistency restoration and minimal conflicts (Reiter, 1987; Junker, 2004). World Action Verifier uses cycle consistency to check state plausibility and action reachability (Liu et al., 2026); World-Coherent Decoding ranks generated futures using internal signals and execution feedback (Zhang et al., 2026). Our certificates guide independent checks before full generation; cancellation requires exclusion of every future admitted by the request. A common check inventory and fixed search order isolate learned guidance.

3 Methodology

NegativeSpace-WM learns minimal impossibility certificates to screen candidate futures before full world-model rollout (Figure 1). Replay-verified legal, violating, and repaired events supervise a shared Critic and Gate: Critic recognizes physical conflicts in candidate videos, while Gate predicts certificates from observed history, actions, and cheap commitments. These predictions guide independent checks that cancel a generation request only when verification excludes every future it accepts. The planner generates and evaluates the retained candidates, executes the first step of the selected action sequence, and replans from the updated history.

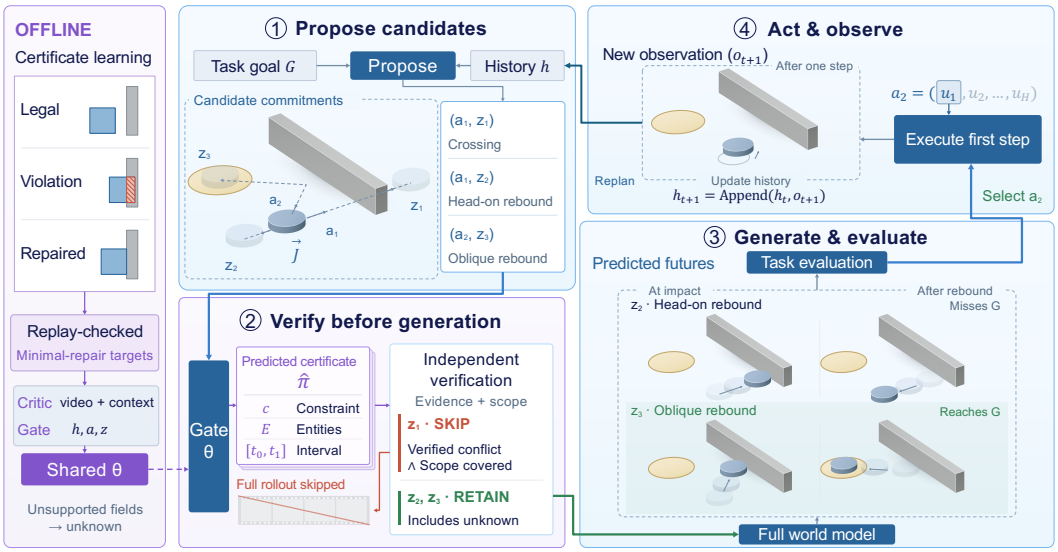

Figure 1: NegativeSpace-WM overview. Offline, replay-checked legal, violating, and repaired events supervise a shared Critic and Gate. Online, Gate guides independent checks before full rollout. Verification cancels the request for crossing commitment z_1 only when its entire acceptance set is excluded; the same-action rebound z_2 and unresolved candidates are retained. The world model generates retained futures, and task evaluation selects the goal-reaching oblique rebound z_3. The planner executes the first step of a_2, updates history, and replans.

3.1 Minimal Impossibility Certificates

Let $q = (h, a_{1:H}, K, H)$ specify observed history, candidate actions, physical premises, and horizon, with fixed task requirements. A *commitment* z specifies poses, contacts, or event orderings and denotes the futures $\mathcal{B}(z)$ satisfying them. Let $\mathcal{S}(h, K)$ contain the history-consistent states and $\mathcal{R}_K(s, a_{1:H})$ the task-valid futures reachable from state s. The reachable futures and the corresponding negative space of excluded commitments are

$$\mathcal{R}(q) = \bigcup_{s \in \mathcal{S}(h,K)} \mathcal{R}_K(s, a_{1:H}), \qquad \mathcal{N}(q) = \{z : \mathcal{R}(q) \cap \mathcal{B}(z) = \varnothing\}.$$

Our goal is to identify commitments in $\mathcal{N}(q)$ through checkable physical conflicts before generating their full futures.

The certificate $\pi = \langle K_\pi, c, E, [t_0, t_1], w, F, \sigma \rangle$ records the premises, constraint, entities, interval, witness, repair, and exclusion scope. Figure 2(a–b) illustrates a crossing commitment: the witness identifies penetration of a closed obstacle, the repair replaces the crossing segment with a contact-consistent continuation, and the scope identifies the futures excluded by that conflict.

Repairs edit commitment assertions while holding actions, initial state, dynamics, task, and random seed fixed. Let q_{rep} denote this context and $V_{K,H}$ return valid (1), invalid (0), or unknown. Validity requires an executable continuation satisfying the edited commitment and all physical and task constraints. An inclusion-minimal repair satisfies

$$\begin{aligned} V_{K,H}(\text{edit}(z, F); q_{\text{rep}}) &= 1, \\ V_{K,H}(\text{edit}(z, F'); q_{\text{rep}}) &= 0 \quad \forall F' \subsetneq F. \end{aligned} \tag{1}$$

Minimality is relative to the allowed edits and checking scope. Any repair meeting Equation 1 is accepted; an unknown subset result leaves minimality unestablished (Figure 2(c)).

3.2 Replay-Verified Counterfactual Supervision

Starting from a legal MuJoCo trajectory with commitment z^+, we edit a segment or event ordering and render a violating event with commitment z^-, keeping the scene, actions, and task fixed (Figure 1,

offline stage). The checker localizes the conflict and searches a fixed catalogue of commitment edits. A minimal repair F gives $z^{\mathrm{r}} = \mathrm{edit}(z^-, F)$, forming a verified counterfactual family:

$$V_{K,H}(z^+; q_{\mathrm{rep}}) = V_{K,H}(z^{\mathrm{r}}; q_{\mathrm{rep}}) = 1, \qquad V_{K,H}(z^-; q_{\mathrm{rep}}) = 0.$$

The catalogue contains reference values, injected alternatives, and interval relaxations containing legal values. Equation 1 verifies that F restores validity and that every proper subset fails. For k edited assertions, this requires at most 2^k replay checks; we use $k \leq 6$. The recorded constraint, entities, interval, witness, and repair provide certificate targets; unresolved fields receive unknown labels. These judgments concern the fixed replay context, with each learning mode receiving the targets supported by its observations.

Appearance variants preserve trajectories and evidence visibility, retaining the same physical target; successful repairs change the judgment at the checked scope. Legal events near constraint boundaries distinguish validity from violation, and multi-conflict events require coordinated repairs. These variants supply judgments and certificate fields for training, with every variant of one event kept in the same data split.

3.3 Shared Critic–Gate Learning

Each event supplies two observation modes (Figure 1): Critic receives the video v and context, $x_{\mathrm{Critic}} = (v, q)$, while Gate receives $x_{\mathrm{Gate}} = (h, a_{1:H}, z)$ before full rollout. They share a visual backbone, low-rank adapters, and an output decoder with parameters θ. A video violation can be known to Critic while remaining unknown to Gate. Gate exclusion targets require $z \in \mathcal{N}(q)$ to be established from its permitted inputs, covering all history-consistent states; repair fields must also be identifiable. Valid inputs have empty violation fields, and unsupported targets are marked unknown.

For mode $m \in \{\mathrm{Critic}, \mathrm{Gate}\}$, the target Y_m orders judgment, premises, constraint, entities, time, witness, scope, and repair. Given input x_m, training examples $\mathcal{D}_m$, and nonempty supervised token positions J_m, we minimize

$$\mathcal{L}_{\mathrm{cert}} = -\sum_m \omega_m \, \mathbb{E}_{\mathcal{D}_m} \left[\frac{1}{|J_m|} \sum_{j \in J_m} \log p_\theta(Y_{m,j} \mid x_m, Y_{m,<j}) \right], \qquad \omega_{\mathrm{Critic}} = \omega_{\mathrm{Gate}} = \tfrac{1}{2}. \tag{2}$$

Each event variant contributes its own supervised sequence. Training updates the shared adapters while keeping the base model and full world-model generator fixed.

3.4 Certificate-Guided Verification

Gate's constraint, entities, interval, and witness prioritize a finite inventory of physical checks prepared before prediction; missing fields invoke a fixed fallback order. Checks obtain independent measurements from observed masks, tracks, calibrated poses, and permitted geometry. State intervals and declared motion limits propagate observation uncertainty, with missing or ambiguous evidence widening the bounds or leaving a check unresolved (Figure 2(d)).

For a necessary constraint $g_c(\tau; q) \leq 0$, exclusion requires a conservative cover $\mathcal{A}(q, z)$ and a sound positive lower bound:

$$\mathcal{R}(q) \cap \mathcal{B}(z) \subseteq \mathcal{A}(q, z), \qquad 0 < \underline{g}_c(q, z) \leq \inf_{\tau \in \mathcal{A}(q,z)} g_c(\tau; q). \tag{3}$$

Proposition 1 (conditional exclusion). Suppose $g_c \leq 0$ is necessary under valid physical premises and Equation 3 provides a sound cover and bound. Then $\mathcal{R}(q) \cap \mathcal{B}(z) = \varnothing$, equivalently $z \in \mathcal{N}(q)$: any τ in this intersection would imply

$$0 \geq g_c(\tau; q) \geq \underline{g}_c(q, z) > 0, \qquad \text{a contradiction.}$$

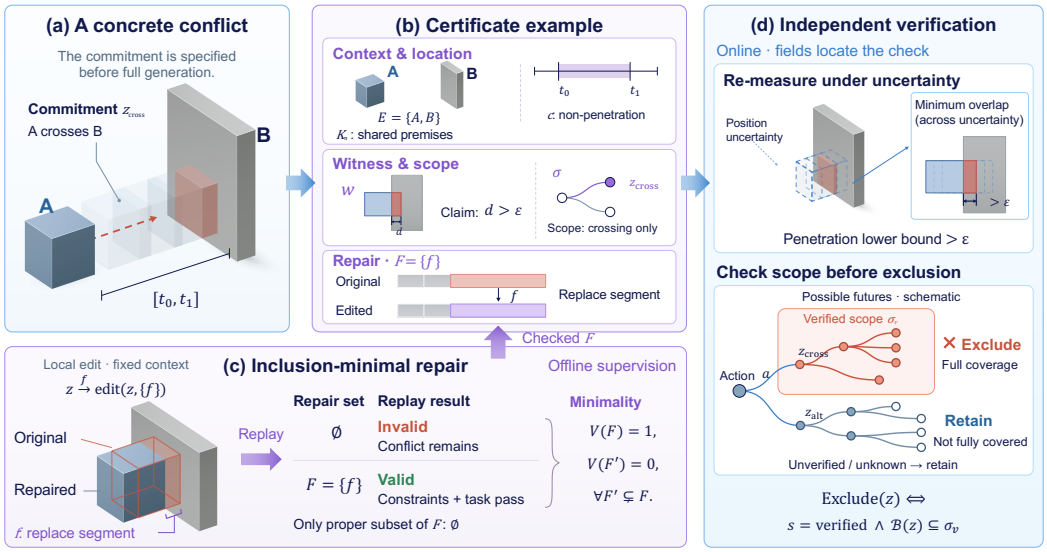

Figure 2: Minimal impossibility certificates and scoped verification. (a) A pre-generation commitment makes rigid body A cross a closed obstacle B. (b) The certificate specifies the shared premises K_π, constraint, entities, interval, witness, exclusion scope, and an executable repair. (c) Fixed-context replay establishes validity after repair and invalidity for every proper subset. Here $V(G) \equiv V_{K,H}(\text{edit}(z, G); q_{\text{rep}})$; unknown outcomes leave minimality unestablished. (d) Independent checks remeasure the penetration lower bound under uncertainty and establish the covered scope. Commitment exclusion requires a verified conflict covering that commitment; request cancellation requires coverage of its entire acceptance set. Feasible alternatives and unresolved requests are retained.

For non-penetration, $g_c(\tau; q) = \max_{t \in [t_0, t_1]} d_\tau(t) - \epsilon$, where $d_\tau(t)$ is penetration depth and ϵ is tolerance. The verified bound certifies a conflict over the full uncertainty set, as illustrated in Figure 2(d).

Verified exclusion establishes $\mathcal{R}(q) \cap \sigma_v = \varnothing$ for scope σ_v. A request accepts futures in $\mathcal{T}(\text{req})$; requiring $\mathcal{T}(\text{req}) \subseteq \sigma_v$ gives

$$\mathcal{R}(q) \cap \mathcal{T}(\text{req}) \subseteq \mathcal{R}(q) \cap \sigma_v = \varnothing. \tag{3a}$$

Thus the crossing conflict in Figure 1 cancels the request for z_1 while retaining the same-action rebound z_2. Broader requests accepting uncovered alternatives and requests with unresolved evidence remain eligible. Online cancellation uses verified evidence and scope; repair minimality supplies the offline learning target.

3.5 Selective Generation for Planning

A fixed proposer constructs candidate actions and inexpensive commitments from history, permitted geometry, and a nominal task path. Request i fixes an action sequence a_i, commitment z_i, and acceptance set $\mathcal{T}_i \subseteq \mathcal{B}(z_i)$ before screening; its query is $q_i = (h, a_i, K, H)$. The frozen world model generates from history and actions, and the planner checks its outputs against the commitment. Stopping, rebound, repaired, and residual requests together cover the original action-conditioned request's task-feasible alternatives, with the residual request covering futures outside the named commitments.

Gate guides budgeted verification of the proposed request pool $\mathcal{I}$; let $\mathcal{C}$ contain the requests cancelled under Equation 3a. Under the verification assumptions, each cancelled request contributes no feasible future, so

$$\bigcup_{i \in \mathcal{I} \setminus \mathcal{C}} (\mathcal{R}(q_i) \cap \mathcal{T}_i) = \bigcup_{i \in \mathcal{I}} (\mathcal{R}(q_i) \cap \mathcal{T}_i). \tag{3b}$$

Thus cancellation preserves the proposed pool's coverage of feasible futures. Retained requests proceed to generation and task evaluation within the computation budget. The planner executes the

first step of the highest-utility admissible action sequence and replans from the new observation. In Figure 1, both rebounds remain available, and task evaluation selects the goal-reaching z_3 and action a_2.

For the same pool $\mathcal{I}$, let C_{prop} be the common proposal cost and $c_{G,i}$ the generation cost of request i. Screening cost C_{screen} includes prediction and verification; C_{extra} covers incremental commitment checks, discarded outputs, and retries. Subtracting screened generation cost from full generation cost gives

$$\begin{aligned} C_{\text{full}} &= C_{\text{prop}} + \sum_{i \in \mathcal{I}} c_{G,i}, \\ C_{\text{NS}} &= C_{\text{prop}} + C_{\text{screen}} + \sum_{i \in \mathcal{I} \setminus \mathcal{C}} c_{G,i} + C_{\text{extra}}, \\ \Delta C &= C_{\text{full}} - C_{\text{NS}} = \sum_{i \in \mathcal{C}} c_{G,i} - C_{\text{screen}} - C_{\text{extra}}. \end{aligned} \tag{4}$$

For a nonempty pool with uniform generation cost $c_G > 0$, the required cancellation rate is

$$\Delta C > 0 \quad \Longleftrightarrow \quad r > \frac{C_{\text{screen}} + C_{\text{extra}}}{|\mathcal{I}| c_G}, \qquad r = \frac{|\mathcal{C}|}{|\mathcal{I}|}.$$

Complete-episode cost sums all proposal, screening, generation, and task-evaluation work across replanning rounds, including failures and timeouts.

4 Experiments

We evaluate whether minimal certificates improve physical representations, whether their content guides independent verification, and whether these benefits translate into more effective world-model planning.

4.1 Experimental Setup

We evaluate replay-verified MuJoCo events spanning conservation, contact and solidity, permanence, causal ordering, and resource or reachability constraints, with 4800 paired planning tasks across four families. Cosmos3-Nano provides the LoRA-adapted Critic–Gate backbone and frozen generator; SAM 2, CoTracker3, and FoundationPose provide perception, while FCL and IBEX support geometric and interval checks. All policies share the proposer, generator, and task evaluator. Event families are split before constructing variants, and learning comparisons use eight paired seeds. We report in-domain (ID) and out-of-distribution (OOD) balanced accuracy (BA), certificate quality, task success, and complete-episode GPU time, including failures and timeouts. Online metrics measure feasible rejection (FR), infeasible retention (MR), and verification resolution. Settings are selected on development data; intervals are paired 95% intervals. Appendix B details configurations, transfer conditions, metric definitions, and statistical procedures.

4.2 Main Comparisons

Table 1 compares supervision formats on common events and permitted inputs, and planning policies under common request-acceptance rules. Free-form and Multi-head targets contain the same physical facts as certificates. Verdict Gate uses the judgment alone; Discard content decodes the complete certificate but applies the same decision rule. Verifier Search checks the same physical inventory in a fixed order, while Certificate Gate uses predicted fields to prioritize it. Full Rollout retains every proposed request for generation.

Representation quality and transfer. Minimal certificates improve ID BA over information-matched free-form targets by 4.0 points [1.8, 6.3] and joint certificate quality by 21.4 points [17.1, 25.7]. They also improve mean OOD BA over redundant certificates, from 73.1% to 75.5%, with a paired difference 2.45 points [0.3, 4.6]. Minimal targets also exceed Multi-head supervision by 3.3 points in mean OOD BA and 12.4 points in joint certificate quality. These gains extend the benefit beyond

predicting the same physical facts through separate output heads. The length-controlled ablation below tests whether this transfer benefit comes from repair semantics rather than shorter descriptions.

Complete planning performance. Certificate Gate uses 33.20% less GPU time than Full Rollout, with 81.6% versus 82.5% task success. The paired success difference is -0.92 points $[-1.98, 0.17]$, leaving the prespecified one-point noninferiority criterion unmet. Certificate also achieves higher success at lower total cost than Score Gate and Cheap preview in the reported means. Relative to Cheap preview, it saves 16.87 GPU seconds per episode while increasing success by 0.7 points. Random pruning saves computation but drops success to 63.5%, showing the cost of removing useful alternatives without physical selection. Because total GPU time includes unsuccessful episodes and repeated planning rounds, this comparison reflects both screening overhead and the downstream work caused by retained or removed requests. Verifier Search reaches similar success and complete cost: its cost difference from Certificate is 0.49 seconds $[-2.55, 3.59]$. Discard content matches Verdict's 79.9% success while adding decoding cost; using the fields for verification raises success to 81.6%. These comparisons motivate the following tests of what the certificates learn and how that information is used.

Table 1: Main comparisons. A: ID and mean OOD BA, and ID joint certificate quality J_{cert}. B: task success and GPU seconds per complete episode, including failures and timeouts. Rates are percentages. Bold marks our method.

A. Representation learning

Supervision	ID BA	OOD BA	J_{cert}
Binary	77.2	66.8	–
Scalar	78.0	67.8	–
Free-form	80.6	70.7	49.0
Multi-head	82.1	72.2	58.0
Redundant	84.1	73.1	50.0
Minimal (ours)	**84.6**	**75.5**	**70.4**

B. Complete planning

Policy	Success ↑	GPU s ↓
Full Rollout	82.5	75.93
Random pruning	63.5	55.28
Score Gate	78.6	58.88
Cheap preview	80.9	67.59
Verdict Gate	79.9	52.60
Discard content	79.9	52.94
Verifier Search	81.6	51.21
Certificate Gate	**81.6**	**50.72**

4.3 Ablation Studies

Table 2 isolates five parts of the learning design: repair semantics, counterfactual data, shared training, field supervision, and Critic observations. Each comparison holds the event population and permitted inputs fixed except for the component being tested.

Repair semantics and counterfactual data (A–B). On the common population admitting all four target formats, minimal repairs improve OOD BA by 2.2 points at both lengths, with a cross-bin interval [0.3, 4.1]; changing length alone changes BA by only 0.1 points. Validity and task success also improve, linking the benefit to repair semantics. The data ablations identify complementary roles: removing repaired inputs reduces repair-pair success by 14.6 points, removing near-boundary legal inputs increases boundary FR by 4.9 points, and removing multi-factor violations reduces composition BA by 6.2 points. Replacements preserve label, source, scene, and exposure.

Shared learning and future observations (C, E). Sharing increases infeasible rejection more with certificate targets than with binary targets. The interaction is +5.0 points at equal Gate exposure and +3.7 points [2.2, 5.3] at equal total training compute. Restricting Critic to Gate-visible inputs retains part of this benefit; full video adds a further 5.5 points for certificates at equal exposure and 4.75 points at equal compute. Thus both shared learning and access to future physical evidence contribute to the Gate's screening ability.

Field supervision (D). Removing any field target reduces transfer accuracy and task success. Witness removal causes the largest rejection loss (8.3 points) and task loss (3.28 points before rounding; paired-seed interval [2.74, 3.82]). Removing repair targets retains repaired inputs and witness verification, yet lowers OOD BA by 3.4 points. Together with B, this separates learning from repaired events from learning to specify a repair. Complete metrics and paired contrasts are reported in Appendix C.2 and C.8–C.9. Together, these ablations connect the training design to its two intended uses: minimal repair targets improve generalization in physical recognition, while witness supervision and shared learning strengthen the screening of proposed futures before generation.

Table 2: Ablations of the learning design. A: repair semantics at matched mean token lengths. B: counterfactual input types. C: shared training at equal Gate exposure or total training compute. D: field supervision. E: rejection gain from full-video versus restricted Critic inputs. A–D report percentages; E reports percentage-point gains. J_{valid} is joint validity without requiring minimality; Reject is infeasible rejection, Task is task success, and Comp. denotes composition. Gray highlights results that improve on the matched ablations.

A. Repair semantics and length

Target (tokens)	OOD BA ↑	OOD J_{valid} ↑	Task ↑
Minimal (120)	**75.2**	**66.4**	**81.6**
Redundant (120)	73.0	64.0	80.2
Minimal (220)	**75.3**	**66.5**	**81.7**
Redundant (220)	73.1	64.1	80.3

B. Counterfactual supervision

Training data	Repair pair ↑	Boundary FR ↓	Comp. BA ↑
Full data	**64.2**	**4.8**	**72.1**
w/o repaired	49.6	5.6	69.0
w/o boundary legal	61.0	9.7	70.2
w/o multi-factor	58.7	5.2	65.9

C. Shared Critic–Gate training

Training	Reject ↑ Exposure	Reject ↑ Compute	Task ↑ Exposure
Gate / binary	55.0	55.0	77.9
Shared / binary	61.0	60.8	80.1
Gate / certificate	58.0	58.0	79.2
Shared / certificate	**69.0**	**67.5**	**81.8**

D. Field supervision

Targets	OOD BA ↑	Reject ↑	Task ↑
Complete	**75.5**	**69.0**	**81.8**
w/o entity	73.4	63.2	79.5
w/o time	72.8	61.9	79.0
w/o repair	72.1	65.3	79.7
w/o witness	72.5	60.7	78.5

E. Future-video contribution (rejection gain, percentage points)

Supervision	Exposure	Compute
Binary	+3.0	+2.90
Certificate	**+5.5**	**+4.75**

4.4 Certificate Use and Scoped Verification

We hold the trained model fixed and vary the fields supplied to the verifier. Table 3 tests certificate use at a fixed verification budget and matched pruning rates.

Content directs useful checks. Permuting entities or witnesses lowers resolution by 9.25 points [8.81, 9.68] or 15.35 points [14.86, 15.84], respectively, while FR stays below 1%. The check inventory is unchanged; certificate fields help locate scope-covering conflicts within the budget, while independent measurements establish exclusion.

Branch selection and exclusion scope. At 48.6% pruning, all policies remove 9796 requests from 1269 attainable pools. Certificate has 0.742% FR versus 2.689% for Verdict, and 31.538% MR versus 32.367%. In exact-geometry tests, all 400 crossing requests are cancelled while same-action stopping alternatives remain available; all 400 broader requests accepting either continuation are retained. Appendices C.4, C.11, and C.13 provide details.

Table 3: Online certificate use. A: resolution (Res.), feasible rejection (FR), and infeasible retention (MR) at 16 checks on a common candidate pool. B: FR at matched pruning rates. All rates are percentages. Gray highlights improvements from certificate use.

A. Fixed-budget verification

Verifier input	Res. ↑	FR ↓	MR ↓
Full certificate	**84.64**	0.69	**30.34**
Permute entities	75.38	0.60	39.60
Permute time	72.60	0.58	42.38
Permute witness	69.29	0.53	45.69
Search (16 checks)	70.48	0.53	44.50

B. Matched-pruning FR

Policy	30.0%	40.0%	48.6%
Score	2.943	5.740	9.404
Verdict	0.659	1.448	2.689
Certificate	**0.395**	**0.554**	**0.742**
Random	29.735	39.904	48.292

4.5 Computational Efficiency

Table 4 separates the computation needed for verification from task performance under a common episode budget. This connects certificate-guided checking to the planning objective without conflating a fixed request workload with a complete episode. Figure 3 shows the verification, pruning, and complete-task comparisons.

Table 4: Computation and budgeted task performance. A: resolution and screening GPU seconds per 16 requests; b is the check budget. Certificate uses the prespecified budget of 16; Search at 64 is selected on development data. B: test success under a common 70-GPU-second episode cap; candidate count M is selected on development data. Complete episodes include failures and timeouts.

A. Verification work

Policy	b	Res. ↑	Screen s ↓
Certificate	16	84.64	0.898
Search	16	70.48	0.680
Search	64	85.12	1.266

B. Success at 70 GPU seconds

Policy	M	Success (%) ↑
Certificate	16	79.60
Verdict	16	77.15
Cheap preview	8	76.58

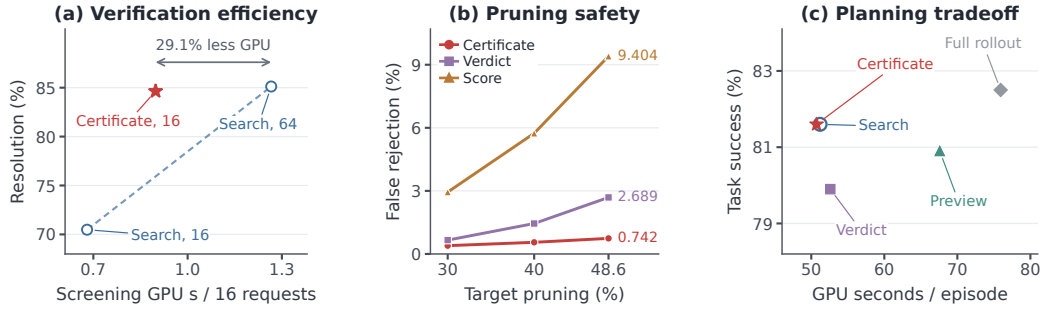

Figure 3: Verification, pruning, and complete planning. (a) Resolution versus screening GPU time per 16 requests (Table 4A); labels give check budgets. (b) Feasible-branch rejection at matched pruning targets (Table 3B). (c) Complete-episode success versus total GPU time (Table 1B). Points use the displayed table values; segments connect reported settings.

Verification work. At 16 checks, Certificate reaches 84.64% resolution versus 70.48% for Search. Search at its development-selected 64-check budget reaches 85.12%, but costs 1.266 GPU seconds per 16 requests compared with 0.898 for Certificate. Across complete episodes, the screening costs are 6.44 and 4.73 seconds, respectively. Different episode workloads offset much of this stage-level saving, yielding the similar total costs in Table 1.

Generation and budgeted success. Executed cancellation reduces the fixed workload from 32000 to 16331 full generations and from 640000 to 326620 network evaluations; the 15669 cancellations include 66 feasible requests, counted in FR. Veto-only logging dispatches every job. With a common 70-GPU-second episode cap, Certificate improves task success over Verdict by 2.46 points [1.31, 3.60] across 4800 paired tasks. This budgeted comparison measures how screening translates into completed tasks. Appendix C.13 gives the dispatch ledger, and C.10 reports the complete verification-budget scan, including unsuccessful development selections.

5 Limitations

Evaluation is limited to controlled physical environments; robustness to incomplete observations, unfamiliar dynamics, and longer planning horizons in open-ended settings remains to be established. Verified exclusion depends on valid physical premises and conservative uncertainty bounds; incomplete evidence leaves more candidates unresolved. Net savings also depend on whether avoided generation outweighs certificate prediction and verification costs.

6 Conclusion

NegativeSpace-WM learns minimal impossibility certificates from executable repairs and shares them across Critic and Gate to guide independent verification before generation. At matched description lengths, minimal repairs improve transfer accuracy, linking the semantics of supervision to generalization. Certificate guidance improves transfer and verification resolution and reduces screening work relative to Search, with similar complete-task cost. Compared with Full Rollout, it uses 33.20% less GPU time at 81.6% versus 82.5% task success. Its effectiveness depends on observation coverage and sound physical bounds. The central principle is to learn which futures can be excluded, verify the evidence and scope, and allocate generation to the remaining alternatives.

AI USE STATEMENT

AI tools assisted with manuscript drafting, language polishing, and identifying and summarizing related work. The authors are responsible for the accuracy of the references, the interpretation of prior work, and all scientific claims in the final manuscript.

ETHICS STATEMENT

The study design uses MuJoCo physical events and independently checks interventions against fixed task conditions. Human review concerns model-output correctness under a fixed rubric; it does not collect sensitive personal attributes. Any review participants receive the study purpose and annotation instructions. Distribution of third-party videos, model weights, and derived annotations remains subject to their respective access and license terms. The scoped verifier is a research planning component whose guarantees are conditional on its observation and physical premises.

REPRODUCIBILITY STATEMENT

Appendix A gives definitions and derivations. Appendix B records the complete experimental configurations, algorithms, and evaluation protocols. Appendix C provides supplementary representation, planning, selection, recognition, and reinspection results. Code will be available at `https://anonymous.4open.science/r/NSWM-7C6F/`.

REFERENCES

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *NeurIPS*, 2021. URL https://arxiv.org/abs/2108.13264.

Mido Assran, Adrien Bardes, David Fan, et al. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning, 2025. URL https://arxiv.org/abs/2506.09985. arXiv:2506.09985.

Zechen Bai, Hai Ci, and Mike Zheng Shou. Impossible Videos, 2025. URL https://arxiv.org/abs/2503.14378. arXiv:2503.14378.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A Challenging Action-Centric Physical Commonsense Evaluation in Video Generation. In *ICLR*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/c02f6a1d5c55e16db50d339dad905b4d-Abstract-Conference.html.

Florian Bordes, Quentin Garrido, Justine T. Kao, Adina Williams, Michael Rabbat, and Emmanuel Dupoux. IntPhys 2: Benchmarking Intuitive Physics Understanding In Complex Synthetic Environments, 2025. URL https://arxiv.org/abs/2506.09849. arXiv:2506.09849.

Gilles Chabert and Luc Jaulin. Contractor Programming. *Artificial Intelligence*, 173(11):1079–1100, 2009. URL https://doi.org/10.1016/j.artint.2009.03.002.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin CM Burchfiel, and Shuran Song. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. In *Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026. URL https://roboticsproceedings.org/rss19/p026.html.

Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Campagnolo Guizilini, and Yue Wang. PhysBench: Benchmarking and Enhancing Vision-Language Models for Physical World Understanding. In *ICLR*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f38cb4cf9a5eaa92b3cfa481832719c6-Abstract-Conference.html.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep Reinforcement Learning in a Handful of Trials using Probabilistic Dynamics Models. In *NeurIPS*, 2018. URL https://arxiv.org/abs/1805.12114.

C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934. URL https://doi.org/10.1093/biomet/26.4.404.

Pieter-Tjerk de Boer, Dirk P. Kroese, Shie Mannor, and Reuven Y. Rubinstein. A Tutorial on the Cross-Entropy Method. *Annals of Operations Research*, 134:19–67, 2005. URL https://doi.org/10.1007/s10479-005-5724-z.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In *ICML*, 2019. URL https://arxiv.org/abs/1811.04551.

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

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training Agents Inside of Scalable World Models, 2025b. URL https://arxiv.org/abs/2509.24527. arXiv:2509.24527.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2): 65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022. URL https://arxiv.org/abs/2106.09685.

Serwan Jassim, Mario Holubar, Annika Richter, Cornelius Wolff, Xenia Ohmer, and Elia Bruni. GRASP: A Novel Benchmark for Evaluating Language GRounding and Situated Physics Understanding in Multimodal Language Models. In *IJCAI*, 2024. URL https://www.ijcai.org/proceedings/2024/696.

Ulrich Junker. QUICKXPLAIN: Preferred Explanations and Relaxations for Over-Constrained Problems. In *AAAI*, 2004. URL https://cdn.aaai.org/AAAI/2004/AAAI04-027.pdf.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *ICCV*, pp. 6013–6022, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Karaev_CoTracker3_Simpler_and_Better_Point_Tracking_by_Pseudo-Labelling_Real_Videos_ICCV_2025_paper.html.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018. URL https://arxiv.org/abs/1604.04173.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging Video Generation Models As World Models. In *NeurIPS, Datasets and Benchmarks Track*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/4ec03ed08a3fcb59e1c815b5598beff1-Abstract-Datasets_and_Benchmarks_Track.html.

Yuejiang Liu, Fan Feng, Lingjing Kong, Weifeng Lu, Jinzhou Tang, Kun Zhang, Kevin Murphy, Chelsea Finn, and Yilun Du. World Action Verifier: Self-Improving World Models via Forward-Inverse Asymmetry, 2026. URL https://arxiv.org/abs/2604.01985. arXiv:2604.01985.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. URL https://arxiv.org/abs/1711.05101.

Saman Motamed, Minghao Chen, Luc Van Gool, and Iro Laina. TRAVL: A Recipe for Making Video-Language Models Better Judges of Physics Implausibility, 2025. URL https://arxiv.org/abs/2510.07550. arXiv:2510.07550.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do Generative Video Models Understand Physical Principles? In *WACV*, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Motamed_Do_Generative_Video_Models_Understand_Physical_Principles_WACV_2026_paper.html.

NVIDIA. Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning, 2025. URL https://arxiv.org/abs/2503.15558. arXiv:2503.15558.

NVIDIA. Cosmos3-Nano: Model Card and Public Release, 2026. URL https://huggingface.co/nvidia/Cosmos3-Nano. Technical report and model release.

Art B. Owen and Dean Eckles. Bootstrapping Data Arrays of Arbitrary Order. *Annals of Applied Statistics*, 6: 895–927, 2012. URL https://arxiv.org/abs/1106.2125.

Jia Pan, Sachin Chitta, and Dinesh Manocha. FCL: A General Purpose Library for Collision and Proximity Queries. In *IEEE International Conference on Robotics and Automation*, 2012. URL https://gamma.cs.unc.edu/FCL/fcl_docs/webpage/pdfs/fcl_icra2012.pdf.

PAN Team. PAN: A World Model for General, Actionable, and Long-Horizon World Simulation, 2025. URL https://arxiv.org/abs/2511.09057v4. arXiv:2511.09057v4; Revised September 2026.

Adam Paszke, Sam Gross, Francisco Massa, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019. URL https://arxiv.org/abs/1912.01703.

Luis S. Piloto, Ari Weinstein, Peter Battaglia, and Matthew Botvinick. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, 6:1257–1267, 2022. URL https://doi.org/10.1038/s41562-022-01394-8.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. SAM 2: Segment Anything in Images and Videos. In *ICLR*, 2025. URL https://arxiv.org/abs/2408.00714.

Raymond Reiter. A Theory of Diagnosis from First Principles. *Artificial Intelligence*, 32(1):57–95, 1987. URL https://doi.org/10.1016/0004-3702(87)90062-2.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A Framework and Benchmark for Visual Intuitive Physics Reasoning, 2018. URL https://arxiv.org/abs/1803.07616. arXiv:1803.07616.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012. URL https://mujoco.org/.

Antonios Tragoudaras, Chenyu Zhang, Daniil Cherniavskii, Antonios Vozikis, Thijmen Nijdam, Derck W. E. Prinzhorn, Mark Bodracska, Nicu Sebe, Andrii Zadaianchuk, and Efstratios Gavves. Evaluating Newtonian Mechanics in Video Generative Models with Real Physical Systems. In *ICML*, 2026. URL https://arxiv.org/abs/2504.02918v3.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, Yuhui Yin, and Xiaodan Liang. WISA: World Simulator Assistant for Physics-Aware Text-to-Video Generation, 2025. URL https://arxiv.org/abs/2503.08153. arXiv:2503.08153.

Luca Weihs, Amanda Rose Yuile, Renée Baillargeon, Cynthia Fisher, Gary Marcus, Roozbeh Mottaghi, and Aniruddha Kembhavi. Benchmarking Progress to Infant-Level Physical Reasoning in AI. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=9NjqD9i48M.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024. URL https://arxiv.org/abs/2312.08344.

Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-Art Natural Language Processing. In *EMNLP System Demonstrations*, 2020. URL https://aclanthology.org/2020.emnlp-demos.6/.

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, 2026. URL https://arxiv.org/abs/2605.10434. arXiv:2605.10434.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi "Jim" Fan, and Joel Jang. World Action Models are Zero-shot Policies, 2026. URL https://arxiv.org/abs/2602.15922. arXiv:2602.15922.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: CoLlision Events for Video REpresentation and Reasoning. In *ICLR*, 2020. URL https://arxiv.org/abs/1910.01442.

Jianhao Yuan, Fabio Pizzati, Francesco Pinto, Lars Kunze, Ivan Laptev, Paul Newman, Philip Torr, and Daniele De Martini. LikePhys: Evaluating Intuitive Physics Understanding in Video Diffusion Models via Likelihood Preference. In *ICLR*, 2026. URL https://arxiv.org/abs/2510.11512.

Chuhan Zhang, Seiji Ito, Kenta Hoshino, Satoshi Ikehata, and Ikuro Sato. World-Coherent Decoding: Self-Verifying Test-Time Planning for World Action Models, 2026. URL https://arxiv.org/abs/2609.02159. arXiv:2609.02159.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World Models on Pre-trained Visual Features enable Zero-shot Planning. In *ICML*, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Christos Ziakas, Amir Bar, and Alessandra Russo. Grounding Generated Videos in Feasible Plans via World Models, 2026. URL https://arxiv.org/abs/2602.01960. arXiv:2602.01960.

Appendix

Contents of Appendix

A Definitions and Derivations **14**
- A.1 Reachable Futures and Exclusion Scope . . . 14
- A.2 Edit Language and Minimality . . . 14
- A.3 Proof and Observation Coverage . . . 15
- A.4 Complete Cost . . . 15

B Experimental Details **16**
- B.1 Components, Observations, and Generation . . . 16
- B.2 Verification Primitives and Budgets . . . 17
- B.3 Planner, Baselines, and Complete Accounting . . . 17
- B.4 Data, Splits, and Annotation Populations . . . 19
- B.5 Transfer and Counterfactual Controls . . . 19
- B.6 Training and Supervision . . . 20
- B.7 Calibration and Intervention Protocols . . . 20
- B.8 Metrics and Statistical Analysis . . . 21
- B.9 Selection and Pilot Configurations . . . 22

C Supplementary Results **22**
- C.1 Recognition and Transfer . . . 22
- C.2 Physical Response and Learning Ablations . . . 23
- C.3 Planning Cost and Decoding . . . 24
- C.4 Planning Diagnostics and Working Points . . . 25
- C.5 Selection Budgets . . . 26
- C.6 Recognition Pilots . . . 26
- C.7 Local Reinspection . . . 27
- C.8 Restricted-Input and Field-Supervision Results . . . 27
- C.9 Paired Seed Contrasts . . . 28
- C.10 Verification-Budget Scan . . . 29
- C.11 Matched-Pruning Populations and Counts . . . 30
- C.12 Multiplicity Analysis . . . 30
- C.13 Online Diagnostics and Request Dispatch . . . 31

A Definitions and Derivations

A.1 Reachable Futures and Exclusion Scope

For the query $q = (h, a_{1:H}, K, H)$ and fixed task in Section 3.1, history-consistent states $\mathcal{S}(h, K)$ give

$$\mathcal{R}_K(h, a_{1:H}) = \bigcup_{s \in \mathcal{S}(h,K)} \mathcal{R}_K(s, a_{1:H}), \qquad \mathcal{R}_K(h, a_{1:H}) \cap \mathcal{B}(z) = \varnothing. \tag{A1}$$

A request fixes $\mathcal{T}(\mathrm{req})$ before screening; conditioning or acceptance enforces $\mathcal{T}(\mathrm{req}) \subseteq \mathcal{B}(z)$ (Equation 3a). Stopping/repaired/residual branches preserve native task-feasible alternatives, with every branch charged. Missing observations enlarge $\mathcal{S}(h, K)$; unresolved requests remain eligible.

A.2 Edit Language and Minimality

Commitments conjoin typed poses, contacts, identities, and event precedences over complete trajectories, using immutable slot IDs, coordinates, time units, identities, and actions. Empty/malformed commitments authorize no action-wide veto.

Each edited slot takes one value from a fixed typed catalogue: legal-reference values, injected alternatives, or intervals containing legal values. Canonical ordering removes duplicates/no-ops. Replay fixes actions, initial state, material laws, goal, observations, and exogenous seed in q_{rep}; success satisfies every edited assertion, constraint, and task requirement.

For $k \leq 6$ edits, test the full set and all $2^k - 1$ proper subsets (at most 64 replay queries), allowing nonmonotone effects. Unknown subsets leave minimality unresolved; any verified minimal repair counts. Search orders cardinality then slot/value, caches up to 4096 distinct queries per family, and stores every verified minimal set encountered. Exhaustion without a repair yields `repair_unavailable`; unfinished search yields unknown.

Redundant targets use replay-valid $F^+ \supsetneq F$, e.g., a crossing correction plus an already-satisfied interval relaxation. Short/long versions keep their edit sets.

Offline checks fix replay state; Gate labels cover all history-consistent states and require input-identifiable repairs. One task-valid continuation establishes feasibility; exclusion covers all continuations. Oracle-unresolved cases leave every binary FR/MR denominator, are reported separately, and remain in task/abstention accounting. Critic labels concern its video.

A.3 Proof and Observation Coverage

For Proposition 1, independent measurement constructs a cover satisfying

$$\mathcal{R}_K(h, a_{1:H}) \cap \mathcal{B}(z) \subseteq \mathcal{A}(q, z). \tag{A2}$$

Every reachable future satisfies the necessary constraint $g_c(\tau; q) \leq 0$. A positive sound lower bound in Equation 3 would therefore give

$$0 \geq g_c(\tau; q) \geq \underline{g}_c(q, z) > 0, \tag{A3}$$

a contradiction. An independently certified empty cover also establishes exclusion. □

For non-penetration, g_c is maximum interval penetration minus tolerance, bounded over object and trajectory uncertainty. With inventory-wide observation-coverage event $\mathcal{E}$ fixed before adaptive ordering,

$$\begin{aligned} P(\text{reject} \mid \text{feasible}) \leq\ & P(\mathcal{E}^c \mid \text{feasible}) \\ & + P(\text{invalid premise or unsound bound} \mid \text{feasible}). \end{aligned} \tag{A4}$$

Empirical FR evaluates perception and verification jointly.

Split conformal calibration (Lei et al., 2018) scores each independent history by its maximum normalized error over accessible objects/times. At $\alpha = 0.01$, rank $\lceil (n+1)(1-\alpha) \rceil$ among n exchangeable histories sets the bound (infinite if the rank exceeds n). Fixed quantity scales yield joint history boxes, propagated without future measurements. Episodes require joint trajectory coverage or valid error allocation. Physical tolerance and roundoff remain separate; feasible-unit FR intervals follow Clopper & Pearson (1934).

A.4 Complete Cost

For the same predeclared request pool $\mathcal{I}$, cancelled-job set $\mathcal{C}$, and retained set $\mathcal{S} = \mathcal{I} \setminus \mathcal{C}$,

$$\begin{aligned} C_{\text{full}} &= C_{\text{prop}} + \sum_{i \in \mathcal{I}} c_{G,i}, \\ C_{\text{NS}} &= C_{\text{prop}} + C_{\text{screen}} + \sum_{i \in \mathcal{S}} c_{G,i} + C_{\text{extra}}. \end{aligned} \tag{A5}$$

Subtraction gives Equation 4. Weights 90/7 for generation/screening yield break-even pruning $7/90 = 7.78\%$, or $\beta > 15.88\%$ at $r = 0.48965625$. Full cancellation saves 44.07 generation and 37.07 net index units. Native savings require $\gamma[97 - 90r\beta] + C_{\text{extra}} < 90$; Table A1 gives indices, C6 gives task GPU seconds.

Cancellation rate and native-request cost. Let β be the fraction of vetoes realized as cancelled jobs, γ the branch expansion per native reference request, and $r = 0.48965625$ the observed veto fraction. The fixed-workload cost index is

$$C_{\text{index}} = 10 + \gamma[7 + 90(1 - r\beta)] + C_{\text{extra}}. \tag{A6}$$

Here C_{extra} records additional work in the same index units. At full cancellation, $\gamma = 1$ gives 62.931, $\gamma = 1.5$ gives 89.396, and $\gamma = 2$ gives 115.862. Positive saving against Native Full requires $\gamma < 1.70033$ with zero extra work. Table A1 gives integer cancellation counts and the full sweep. These indices quantify request workload; Table C6 reports the separate complete-episode timing.

Table A1. Fixed-workload index, Full = 100. First five rows vary target cancellation realization β; integer cancellations are rounded once and costs use the realized counts. Last four rows use full cancellation and vary branch expansion γ; the fourth column then reports saving versus Native Full in percent. Extra work is zero.

β / γ	Cancelled	Remaining	Pruning / saving (%)	Cost index
0.00	0	32000	0.000	107.000
0.25	3917	28083	12.241	95.983
0.50	7834	24166	24.481	84.967
0.75	11752	20248	36.725	73.947
1.00	15669	16331	48.966	62.931
$\gamma = 1.00$	—	—	37.069	62.931
$\gamma = 1.25$	—	—	23.836	76.164
$\gamma = 1.50$	—	—	10.604	89.396
$\gamma = 2.00$	—	—	-15.862	115.862

B Experimental Details

B.1 Components, Observations, and Generation

Critic sees candidate video; Gate sees history/actions/commitments. Future frames and hidden state supply offline labels only. Unsupported fields are typed unknowns; verified scope determines cancellation independently of decoded judgment.

Table B1. Public components and permitted inputs.

Module	Component and inputs
Physics replay	MuJoCo, with MuJoCo Menagerie robot assets (Todorov et al., 2012). Legal trajectories, controlled trajectory edits, and fixed-context replay.
Certificate backbone	Cosmos3-Nano autoregressive reasoner tower (NVIDIA, 2026). Critic video or Gate history, context, and structured text output.
Full generation	Frozen Cosmos3-Nano forward-dynamics interface. Current image, native embodiment identifier, and action trajectory.
Segmentation	SAM 2 (Ravi et al., 2025). Image-space entity masks propagated using observed frames.
Point tracking	CoTracker3, online variant (Karaev et al., 2025). Tracks and visibility flags within observed history.
Metric pose measurement	FoundationPose (Wen et al., 2024). Calibrated RGB-D and object geometry/reference observations.
Geometric queries	FCL (Pan et al., 2012). Candidate collision and distance queries.
Conservative bounds	IBEX (Chabert & Jaulin, 2009). Interval evaluation with explicit state, geometry, and roundoff bounds.

Replay/proposal. Event construction uses MuJoCo, UR5e, Robotiq, 240 Hz physics, and 15 Hz control. Render legal trajectories, edit trajectories/events, and re-render; replay repairs in fixed context. Appearance variants preserve trajectories/evidence visibility. Appendix A.2 defines slots and the six-atom limit.

The frozen proposer draws 16 smooth end-effector sequences around an observation-based task path, pairing draws, path, perturbation scales, inverse-kinematics limits, and admission order. Clip translation/rotation to 2 cm/5 degrees per step and gripper commands to range. The cross-entropy variant (de Boer et al., 2005) charges optimization.

Generation. Frozen Cosmos3-Nano takes current image, embodiment ID, and actions: 3-D translation, 6-D rotation, gripper state, with embodiment-specific transforms/normalization. Sixteen transitions give 17 observations. Full/preview generation uses 20/4 denoising steps, guidance 1.0; NFE counts all network calls, including extra conditional evaluations. Policies share weights, processor, sampler, encoding, and draws.

Dispatch. Native Full accepts task-valid action futures. Each branch wraps one native job with a fixed z-consuming predicate; `z_hash` identifies it. Branch Full and veto arms share contracts. Veto-only dispatches all jobs; cancellation removes scope-covered jobs before denoising. Charge acceptance/scoring, unusable outputs, and retries.

Interface checks use exact geometry: five sets of 400 independent histories. Workload checks use 2000 histories/32000 requests; disjoint 500-group selection/calibration partitions freeze settings. Commitments use past observations, permitted geometry, actions, and task path. Stopping/repaired/residual branches preserve native alternatives. Log request/parent-native IDs, history, action, z, acceptance, scope, cancellation reason, dispatch/completion, backend jobs, NFE, compliance, and cost. Jobs are full sampling sequences;

batched calls/retries have separate IDs. Added branches enter γ; incremental binding/checking/retries enter C_{extra}.

Measurements. Observed frames, calibration, and permitted geometry give masks, tracks, metric poses, and uncertainty. Pose intervals bound velocity; declared acceleration/control limits bound future positions. Missing depth/scale, occlusion, or ambiguous identity widens bounds or leaves unknowns. Uncalibrated RGB supports image diagnostics only; oracle verification and Search both use exact state.

FCL proposes queries; interval checks cover pose/geometry uncertainty. Inscribed spheres certify overlap when the lower summed-radii bound minus upper center-distance bound exceeds tolerance. Other shapes use sound relaxations or remain unknown. Training uses PyTorch (Paszke et al., 2019), shared Transformers tokenizers/processors (Wolf et al., 2020), LoRA, and AdamW; record model/processor revisions, geometry, MuJoCo settings, and calibration data.

B.2 Verification Primitives and Budgets

Before prediction, fix identities, intervals, predicates, tolerances, measurements, and refinement limits. Fields rank this inventory; Search supplies deterministic, duplicate-free fallback.

Table B2. Independent checks; inputs precede the rejection condition.

Check family	Inputs and conservative rejection test
Contact/solidity	Object shapes, pose tubes, contact assertions, and permitted openings. Lower bound on penetration minus the fixed geometric tolerance is positive for every realization covered by the asserted crossing.
Conservation	Observed state intervals, mass/inertia intervals, and bounds on external impulse or work. The required momentum/energy change lies outside the interval permitted by the declared external exchanges; unbounded exchanges make the check unknown.
Permanence	Persistent identity, observed occupancy, occlusion regions, and reachable position tubes. An asserted appearance/disappearance or identity transition conflicts with every continuation permitted by the occlusion and motion bounds; lack of visible pixels alone is insufficient.
Causal ordering	Event-time intervals and minimum predecessor delays. Precedence constraints have no feasible time assignment, or a positive-delay cycle makes their sum contradictory.
Resource/reachability	Capacity intervals, demand lower bounds, action limits, travel-time or reachable-set bounds. Minimum required demand exceeds maximum capacity, or the goal commitment is disjoint from an outer bound on the action-conditioned reachable set.

Checks return outward-rounded intervals, units, premises, and scope. A continuous-time maximum is bounded above by the largest sampled upper bound plus $L_g \Delta t/2$; any sampled lower bound bounds it below. Without residual derivative bound L_g, use continuous collision checks or return unknown. Scope containment includes every unobserved coordinate in its boxes/finite unions.

Training-fixed scales normalize penetration by object length, conservation by momentum/energy, event time by control-step duration, and resources by capacity/reachable distance; replace zero scales with positive unit scales. Tolerance is 10^{-3} plus conservative discretization margins. Rigid-body, force, occlusion, and capacity premises are explicit.

Budgets $b \in \{4, 8, 16, 32, 64\}$ (default 16) read nested request-order prefixes. Search-16 equals judgment-only fallback; development selects Search-target (B.7). Fixed-pool tests have no episode cutoff. Terminal states are:

- V: independently verified conflict covers the cancellation target; reject immediately.
- S: inventory completed, with a local conflict of insufficient scope; retain.
- N: inventory completed without an established conflict; retain.
- X: any required entry remains unavailable or unchecked when search ends; retain, including cases with a partial local conflict.

Unavailable entries leave other checks active; legal/unknown judgments receive the same budget. Checks do not update one another: reordering preserves completed S/N, while V depends on prefix position. Log episode-time truncation separately; it can add X outcomes.

B.3 Planner, Baselines, and Complete Accounting

Core planning uses $H = M = 16$, one proposal batch per round, no cache/re-proposals. Sweeps share $P_\psi(h)$, generator, admission order, commitment test, utility, tie-breaker, action alphabet, deadline, and exogenous randomness.

The immutable goal allows at most 64 control steps, plus GPU caps on budget curves: contact reaches targets without penetration; persistence selects the designated occluded object; ordering completes a precedence graph; resources satisfy capacity/reachability.

Table B3. Planning policies; charged operations follow each decision rule.

Policy	Decision and charged work
Full Rollout (Branch Full)	Retain all branch requests under the shared commitment-acceptance rules and episode budget. Proposal and common setup.
Score Gate	A scalar physical-risk predictor and fixed residual checks; prune by its development-calibrated scalar threshold. Scalar prediction and residual evaluation.
Verdict Gate	Threshold the shared model's judgment probability; stop decoding after the judgment. Judgment prediction.
Discard content	Decode the complete certificate, then apply exactly the Verdict rule. Complete certificate prediction.
Verifier Search	Search the common inventory in its fixed fallback order; veto only for V. All independent checks and search.
Cheap preview	Generate and score the short preview, then threshold its risk before complete generation. Preview, scoring, and residual evaluation.
Certificate Gate	Rank the common inventory by predicted fields; veto only for V. Complete certificate prediction and independent checks.
Random pruning	Remove a label-independent subset at the declared matched rate. Random selection and common setup.

Score/preview residual checks enter Verify. Previews add four evaluations per candidate plus scoring and lie outside information-matched representation comparisons. Training targets 1% development FR; Verdict targets 2%, reaching 1.93% test FR.

Serialization. Greedy decoding orders judgment, premises, constraint, entities, time, witness, scope, repair, with typed unknowns. Common tokenization and deterministic checks make witness/scope a decision-preserving prefix. Repair-supervision removal is a separate intervention.

Algorithm 1 Certificate-guided planning

```
Require: h, K, task, H, episode GPU budget B, verifier budget b,
         shared proposal/admission orders and task terminal rules
 1: Set episode ledger L = 0 and status = running
 2: while running and proposal plus screening fits B − L:
 3:   Reserve proposal and screening; propose requests and charge work
        Include stopping/residual requests covering native alternatives
 4:   S = empty; completed futures = empty
 5:   for request i in the common proposal order:
 6:     If admissible, decode the declared certificate prefix and charge
 ↪  it
 7:       Otherwise use an empty certificate
 8:     Verify within the remaining local and episode budgets
 9:       Charge checks; record status and established exclusion scope
10:     If verified exclusion covers all of T(req_i), cancel i:
          remove its queued backend job and log the cancellation
        Otherwise add i to S
11:   for retained i in the common generator-admission order:
12:     If admissible, dispatch; log backend work and NFE; charge all
 ↪  work
        Check fixed acceptance and task rules; score usable outputs
13:     Stop admitting work at the budget boundary
14:   If a complete actionable future is available:
15:     Execute one step of the highest-utility admissible action
16:     Update history; check success and unrecoverable task failure
17:   Otherwise fail if all requests have verified exclusion
18:     or all completed outputs are unusable; time out if none completed
19: end while
20: Assign one terminal outcome; return it with the complete ledger
```

Admission. Reserve proposal/screening jointly, then admit individual generation jobs whose development-calibrated cost fits the remaining budget. Charge actual completion; log overrun indicators/amounts and

stop admission. Strict caps use interruptible kernels or deterministic costs. Partial batches act on completed admissible outputs. Skipped prediction gives an empty certificate; skipped verification retains the request.

Termination. Assign exactly one outcome, prioritizing success. Failure means all requests excluded, all completed outputs unusable, or an unrecoverable violation. GPU exhaustion without an actionable completion, or the control deadline, means timeout.

Timing. One H100 80-GB GPU, BF16, generation batch one, serialized GPU stages, eight CPU-verifier threads. Assign each device-active interval once to proposal, prediction, preview, verification, or generation/acceptance/scoring. Wall time includes CPU checks/scope/acceptance, dispatch, synchronization, and environment execution. Log every proposal/job, retry, failure, timeout, and partial-batch completion; exclude loading/warm-up. Index weights follow A.4.

Cache/repair. Scope keys include history/observations, geometry/identity intervals, premises, actions, time origin, horizon, constraints. Reuse requires identical premises or sound containment; changed actions/geometry/time or wider uncertainty invalidates it. Repairs receive new IDs and observation/scope/budget checks; charge decoding, replay, generation, and proposal.

B.4 Data, Splits, and Annotation Populations

Families share scene, action template, identities, task, and exogenous seed across legal/violating/repaired/appearance/renderer variants. Split before rendering/annotation; near-duplicates and cross-condition trajectories share IDs. Categories: conservation, contact/solidity, permanence, causal ordering, resource/reachability.

Quotas are 10000/1000/2000 train/development/ID-test families and 1000 per transfer condition, equally divided by category. Development has 500 selection and 500 calibration families; selection, severity bins, and tolerances use training/development only.

Pair violations with same-task legal references; appearance/repair variants form pair/quartet endpoints. Average same-label variants within family, then families within category, then categories equally.

Before prediction, certificate-eligible $\mathcal{C} \subseteq \mathcal{V}$ requires witness, entity/time truth, admissible repair, and complete subset replay: $|\mathcal{C}|/|\mathcal{V}| = 80.0\%$. Remaining violations contribute judgment/emission/resolution. Four-target $\mathcal{Q} \subseteq \mathcal{C}$ has minimal/redundant repairs within six edits and both lengths: $|\mathcal{Q}|/|\mathcal{C}| = 74.0\%$. BA adds class-balanced legal references; joint quality/repair use eligible violations.

Stratify by a canonical minimum-cardinality minimal repair (3+ means three–six edits); accept any minimal repair. Single-solution labels require exhaustive search; two minimal sets establish multiplicity. Unknown uniqueness leaves only this diagnostic.

B.5 Transfer and Counterfactual Controls

Holdouts change a principle's realization (implementation), withhold a conjunction but retain constituents (composition), or omit a principle entirely. Scene varies geometry/assets/layout; renderer pairs renderings of fixed trajectories; composition-plus-scene combines both. Worst-group BA is the minimum fixed category/shift cell, including joint shifts, requiring both labels.

Severity $u = (r - \epsilon_c)/s_c$ uses checked residual r, tolerance ϵ_c, and positive training-fixed scale s_c; nonscalar cases use ordered intervention magnitude. Training/development fixes Boundary $(0, 0.25]$, Small $(0.25, 0.5]$, Medium $(0.5, 1]$, Large $(1, \infty)$. Each violation has a checker-valid legal reference in its bin. Transfer training omits Boundary violations, retaining hard legal cases. Disjoint bins share renderings, category weights, and seed draws.

Appearance changes lighting/texture/rendering while preserving trajectory/evidence visibility. Necessary-factor edits belong to a minimal repair; matched-scale unrelated edits preserve the relation. Pair success requires both judgments and valid mapped constraint/entity/time/witness/repair fields at violation endpoints; quartet success requires all four original/appearance/injected/repaired endpoints. Non-emission fails.

Data ablations replace repaired, near-boundary legal, or multi-factor inputs at matched label/source/scene/family exposure. Repaired-input removal keeps repair targets; repair-supervision removal keeps repaired inputs. There is no added group-order or paired-consistency loss.

B.6 Training and Supervision

Table B4. Training settings; initialization, family order, and augmentations are paired.

Component	Setting
Backbone and adaptation	Frozen Cosmos3-Nano base and visual encoder; LoRA (Hu et al., 2022) on language-attention query/key/value/output projections: rank 64, scaling 128, dropout 0.05.
Independent repetitions	Eight seeds (0–7), paired initialization, family order, and augmentation draws.
Optimizer	AdamW (Loshchilov & Hutter, 2019) adapters: learning rate 10^{-4}, betas (0.9, 0.999), epsilon 10^{-8}, weight decay 0.01, gradient-norm cap 1.0.
Schedule	Two family-list passes; linear warmup for 5% of updates, then cosine decay to zero.
Effective batch	16 family examples per mode update; accumulation changes memory use, preserving effective batch and loss normalization.
Arithmetic	BF16 model operations, FP32 optimizer states; common arithmetic/accumulation rules.
Critic view	17 uniform frames, including endpoints, with timestamps and short-clip padding masks.
Gate view	At most five uniform past frames ending at the current observation, plus actions/commitments; no candidate-future pixels.
Image processing	Preserve aspect ratio, resize longest side to 448 pixels, pad spatially; shared processor/normalization.
Text and output	4096 input text tokens and 512 output tokens maximum; greedy decoding; retain all commitment assertions.
Shared training	One Gate and one Critic update per cycle, weights $1/2$ each; Gate-only has one Gate update.
Checkpoints	Match exposure fractions within seed pairs; primary checkpoint chosen on selection development data, with earlier-checkpoint tie-break.

Exposure fixes event order, missing-variant masks, and class/category mix. Teacher forcing applies Equation 2 to judgment/supported fields, excluding input/padding; unavailable decisions target unknown.

Exposure matching fixes Gate lists/exposure/augmentation and adds paired Critic updates. Compute matching fixes total accelerator time, selects the last complete in-budget checkpoint, and schedules learning rate by consumed budget fraction.

Binary predicts judgment/unknown; Scalar predicts category-normalized signed residual with zero validity boundary; Free-form states the same facts in prose. Multi-head uses parallel field queries, shared backbone/tied vocabulary-output weights, summed loss per common supervised-token count, and charges extra queries. Minimal/redundant targets share schema/inputs, differing only in repairs.

The factorial crosses Gate-only/shared with Binary/Certificate. Restricted Critic sees Gate history/actions/commitments; fixed masks and information-free placeholders preserve full-video update counts/shapes. Exposure-matched Binary/Certificate cost $1.72T_0/2.18T_0$ (T_0: Gate/Binary); compute-matched arms share one budget. Table 2E isolates video information; 2C reports sharing.

Field removal masks its loss and later teacher-forced occurrences, retaining shapes with information-free padding. Missing entity/time/witness invokes bounded search; missing repair supervision keeps witness checks and disables repair re-proposals.

Length bands are 100–140/200–240 tokens (means 120/220), with common families/exposure/class mix; relative training costs 1.00/1.24 and decoding 38/66 ms. Timing retains actual candidate counts, output distributions, batching, and synchronization.

B.7 Calibration and Intervention Protocols

On disjoint calibration data, maximize rejection with empirical FR $\leq 1\%$, breaking ties conservatively; inventory, tolerance, and scope stay common.

GPU caps: 50/60/70/85/100 seconds. Freeze selection/calibration before testing. The 82%-success working point minimizes development mean cost, ties by FR; unmet targets stay unselected.

Fields. Post-decoding deletion/permutation preserves prediction cost, judgment, candidates, inventory, fallback, and budget. Derange donors within mechanism/difficulty/eligibility strata; remove singletons from every arm. Preserve syntax/local entity maps; charge truncated-prefix decoding. Among $b \in \{4, 8, 16, 32, 64\}$, development selects the cheapest Res $\geq 83\%$, FR $\leq 1\%$ point for test. Restoration requires paired resolution-difference lower bound above -1 pp; report failures. Task restoration uses complete episodes.

Matched pruning. Targets $r \in \{30, 40, 48.6\}\%$ remove $k = \lfloor rm \rfloor$ or $k + 1$, using a common label-independent draw with upper-count probability $rm - k$. Score/Verdict rank risk, ties by ID; Random samples uniformly; Certificate takes verified exclusions in admission order. Exclude quota-unattainable

pools from every arm and report their fraction. Labels serve scoring only; closed-loop thresholds stay development-fixed. Budget-aware Random fills the cap without quotas.

Scope. Local-conflict/missing-observation/wide-commitment/changed-premise cases retain a witnessed same-action/task feasible continuation. Score false vetoes against it, compare verdict-only/stale-cache controls, and check each changed premise's invalidation.

B.8 Metrics and Statistical Analysis

With fixed nonnegative weights summing to one over complete-truth violations, define W_i as valid constraint/entity/time/witness fields, R_i as successful repair, and M_i as inclusion-minimality:

$$J_{\text{valid}} = \sum_i w_i \mathbf{1}[\hat{y}_i = 1 \wedge W_i \wedge R_i], \qquad J_{\text{cert}} = \sum_i w_i \mathbf{1}[\hat{y}_i = 1 \wedge W_i \wedge R_i \wedge M_i]. \tag{B1}$$

Non-emission scores zero for joint quality, Repair, and unconditional components. Minimality conditions on valid repair; emission includes incomplete truth. Gate-emitted witness validity is separate. Entity F1 matches one-to-one; temporal IoU threshold 0.5 and entities must identify the same event. Use fixed constraint types, independent witnesses, task-preserving replay, and all repair subsets.

Violations have $y = 1$; missing/malformed/unknown judgments fail. AUROC uses judgment probability or sigmoid residual, half-credit ties, and missing score 0.5. For seed s, shift a, and $C = 5$ categories,

$$\widehat{\mathrm{BA}}_{s,a} = \frac{1}{2C} \sum_{c=1}^{C} \left[\frac{1}{n_{a,c,1}} \sum_{f \in \mathcal{F}_{a,c,1}} \bar{u}_{s,f,1} + \frac{1}{n_{a,c,0}} \sum_{f \in \mathcal{F}_{a,c,0}} \bar{u}_{s,f,0} \right]. \tag{B2}$$

$\bar{u}$ averages same-label family variants. Weight four axes/eight seeds equally, using common class/category weights for severity and one family weight across axes. Joint quality/Repair balances categories and family multiplicity within eligible populations. Empty/single-label cells leave every arm.

Task success equally weights four families of 1200 paired tasks (4800 total), or 600 in the separate 2400-task design. Task/environment pairs are units; rounds/candidates repeat within them. For common candidate weights and infeasible fraction ρ,

$$\mathrm{FR} = P(\text{reject} \mid \text{feasible}), \quad \mathrm{MR} = P(\text{retain} \mid \text{infeasible}), \qquad r = \rho(1 - \mathrm{MR}) + (1 - \rho)\mathrm{FR}, \qquad \mathrm{Res} = V + S + N = 1 - X, \quad V = 1 - \mathrm{MR}. \tag{B3}$$

B3 uses infeasible-state fractions; table counts give $\mathrm{Res} = (V + S + N)/n_-$, $\mathrm{MR} = 1 - V/n_-$, $\mathrm{FR} = \mathrm{FP}/n_+$. Only pre-dispatch removals count as cancellations; budget omissions are separate. Call pruning is βr (Table C23B: $\beta = 1$).

Contrasts. Use A minus B with common weights. Sharing is $(R_{\text{shared,cert}} - R_{\text{gate,cert}}) - (R_{\text{shared,binary}} - R_{\text{gate,binary}})$ within seed/protocol. Length averages short/long Minimal–Redundant differences; interaction is long minus short. Severity interaction is Boundary minus Large; weak/strong means Boundary/Small versus Medium/Large.

Intervals. Use 10000 paired crossed-bootstrap replicates: independently resample training seeds and base families within category/split, sharing multiplicities across methods/fields/axes and recomputing weights. Deployment resamples task/environment clusters within family, carrying all rounds/candidates and timing/input shapes. Retraining adds paired policy seeds; selection crosses list IDs; field interventions resample histories. Fixed-policy intervals condition on the models (Agarwal et al., 2021; Owen & Eckles, 2012).

Cost saving is $100(1 - \bar{C}_A/\bar{C}_B)$; recompute means per replicate. Log-ratio interval $[\ell, u]$ maps to $[100(1 - e^u), 100(1 - e^\ell)]$. Screening ratios include prediction/verification on common workloads.

C16 reports mean, sample SD, and $\bar{d} \pm 2.364624 s_d/\sqrt{8}$ (t_7) across eight paired seeds on fixed data. Crossed intervals also resample families; C6 uses fixed-policy task pairs.

Multiplicity. For C21's eight primary contrasts, Holm (1979) tests ordered $p_{(k)} \leq 0.05/(9 - k)$ until failure. Adjust normal and mixed tails (t_7 for five training, normal for three fixed-policy contrasts) separately. Bootstrap supplies fixed-policy SEs/percentile intervals; 95% intervals are marginal. Mechanism, severity, selection, and individual deletions are secondary.

Noninferiority uses a 1 pp margin at one-sided 0.025: the 95% interval must start above -1 pp; 2 pp is sensitivity. Joint cost/noninferiority requires both endpoints.

Semantic audit. Two model/score-blind annotators review 200 outputs (50/format), sharing entity maps, time rules, edit language, and replay evidence. Score judgment, constraint, entities, time, witness, repair, minimality; adjudicated joint labels define evaluator agreement. Field disagreements diagnose errors. Equal-format paired resampling gives agreement/Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$ intervals. Compute effects unrounded; round SDs/endpoints separately.

B.9 Selection and Pilot Configurations

Component studies use disjoint billiards/dominoes/wrecking-ball runs: train/validation positives 7240/891, negatives 9905/1216. Training negatives: solidity 3484, permanence 2999, conservation 1499 (including scale change), continuity 1459, causality 464. Cross-dataset categories follow physical definitions.

Table B5. Component-study populations and evaluation roles.

Population	Variation and data role	Size
Fixed-background validation	New base runs within the three training scene families	2,107
IntPhys2 Main	Target scenes; labels used in selector fitting	1,012
InfLevel-Lab	Real videos; native three-principle controlled accuracy	5,772
IntPhys2 HeldOut	New target scenes; official scoring deferred	344

Main: 253 four-video scenes, 292 solidity and 240 each permanence/immutability/continuity. Lab: 2268 continuity, 2592 gravity, 912 solidity videos, equal principle weights. External pool: 4096 GRASP, 920 LikePhys, 5000 Physical Concepts; byte checks exclude overlap with Main/validation/Lab/HeldOut.

Selection. A training pool exceeding 10000 families supports source/label/category quotas. Independent non-nested lists contain 2500/5000/10000 families at matched exposure. Random samples uniformly without replacement; Representation ranks cosine similarity to a disjoint training-side legal/violating boundary centroid; Coverage selects farthest-first in the same normalized frozen space. Exclude centroid families; break ties deterministically. Three list seeds cross eight training seeds; deduplication removes repeats. Record mechanism-cell coverage and cached/uncached extraction/normalization/ranking/construction costs.

Pilots. Cosmos3-Nano: primary iteration 500, fixed prompt, native frames/processor, greedy decoding, 512-token cap. Background-only trains on background; Mixed-source adds external data (seeds 0–2); GRASP-source replaces that source. Negative masking retains negative-media exposure/background supervision. Representation/Random/Low-score add 360 videos (180/180 labels), at 19-fold external exposure with differing source mixes. Main is selector-fitting development, scored by scene-bootstrap AUROC; Lab uses native controlled accuracy and 10000 paired $(\mathrm{camera_loc}, \mathrm{cover}, \mathrm{obj})$ replicates within principle. Iteration 300 is secondary; masking uses seed 0.

Source-matched follow-up fixes 180 positives and source/label-matches 180 negatives: 151 GRASP/29 LikePhys/180 Physical Concepts per arm, 6840 external/23985 total exposure slots. Three frozen lists cross four seeds; record coverage and within-source redundancy separately.

Stock/reinspection. LikePhys AUROC uses 920 videos (800 invalid/120 valid), averaging flow residual over noise 0.2–0.8. Reinspect 2107 validation videos with iteration-500 Mixed-source. R@90 retains at least 90% legal cases at a discrete threshold. Token increments include timestamps/control symbols/common cached first pass; annotation guidance centers on labeled time. Clip onset index to $\min(16, \max(0, \lfloor 8t \rfloor))$; score continuous-time errors separately.

C Supplementary Results

Figure C1 summarizes the semantic, matched-pruning, and complete-task comparisons.

C.1 Recognition and Transfer

Equal-compute ID J_{cert} is 49.8/58.7/49.6/68.8% for Free-form/Multi-head/Redundant/Minimal. OOD J_{valid} is 66.8/64.2% for Minimal/Redundant. The prompted control shares backbone, inputs, schema, verifier, development-selected prompts/thresholds, and charged inference; BA/Res is 76.6/70.0%, versus 84.6/84.0% after training.

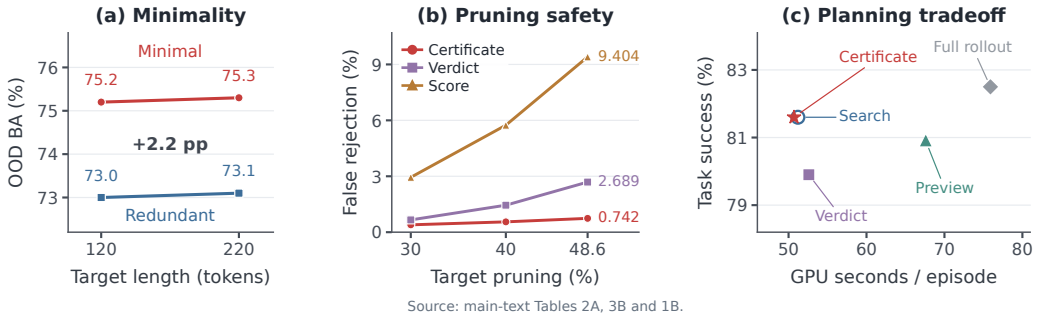

Figure C1: Representation, pruning, and planning. (a) OOD BA for minimal and redundant repairs at matched lengths (Table 2A); the gap is 2.2 percentage points at either length. (b) Feasible-branch false rejection for learned pruning policies at matched target pruning rates (Table 3B); policies share the same attainable pools at each rate. (c) Complete-episode success versus GPU time at the common configuration (Table 1B). Points summarize the corresponding main-text tables; GPU time is rounded to one decimal in the plot. Lines connect the reported settings.

Table C1. Recognition and certificate components (percent; Tokens is a count). Rep. is repair success; Con./Wit./Min./Emit. are constraint, witness, minimality, and emission scores. Minimality conditions on valid repair; dashes mark absent outputs.

Supervision	AUROC	Rep.	E-F1	tIoU	Con.	Wit.	Min.	Emit.	Tokens
Binary	84.0	–	–	–	–	–	–	–	–
Scalar	84.9	–	–	–	–	–	–	–	–
Free-form	87.3	74.0	83.1	61.0	84.0	80.0	69.0	94.0	186
Multi-head	88.8	79.0	87.2	66.0	88.2	84.5	77.0	95.1	–
Redundant	90.5	85.0	90.1	71.0	91.4	88.2	61.0	96.5	198
Minimal	91.1	84.0	90.5	72.0	91.8	88.6	90.2	96.8	148

Table C1 (B). Additional recognition and transfer endpoints (percent). Joint validity uses ID complete-truth violations; worst-group BA uses the fixed category/shift groups.

Supervision	ID J_{valid}	OOD worst-group BA
Binary	–	53.1
Scalar	–	54.0
Free-form	65.0	56.8
Multi-head	71.0	58.4
Redundant	77.5	59.2
Minimal	77.0	61.8

Table C2. Transfer BA / J_{cert} (percent). Mechanism averages Composition and Implementation; worst-group BA is in Table C1B. All columns use one frozen checkpoint.

Supervision	Comp.	Impl.	Comp. + scene	Principle	Scene	Severity	Renderer
Binary	63.0 / –	65.4 / –	60.0 / –	57.8 / –	70.1 / –	60.4 / –	72.3 / –
Scalar	64.0 / –	66.2 / –	61.0 / –	58.5 / –	71.0 / –	61.6 / –	73.4 / –
Free-form	67.2 / 38.0	69.0 / 41.0	64.5 / 33.6	60.5 / 27.2	74.2 / 47.2	63.7 / 33.0	76.8 / 50.1
Multi-head	69.0 / 45.6	71.4 / 49.0	66.1 / 40.3	62.2 / 33.8	75.6 / 53.6	65.1 / 39.8	78.0 / 57.2
Redundant	69.5 / 37.8	72.1 / 40.6	67.2 / 34.0	63.0 / 28.9	76.4 / 45.7	65.9 / 32.8	79.1 / 48.1
Minimal	72.1 / 55.0	74.3 / 58.6	70.1 / 49.2	65.8 / 41.5	78.8 / 62.5	68.7 / 49.1	81.3 / 65.6

C.2 Physical Response and Learning Ablations

The Boundary–Large interaction in Minimal–Redundant BA is 2.4 pp $[-1.0, 5.8]$; its dependence on evidence strength remains unresolved (Table C3).

Table C3. Physical response (percent). A: appearance, repair, necessary-factor, unrelated-factor pair success, and whole-quartet success (Ap./Rep./Nec./Unr./Q). B: BA at Boundary/Small/Medium/Large violation magnitude, with paired legal controls.

Supervision	A: Ap.	Rep.	Nec.	Unr.	Q	B: Bnd.	Small	Med.	Large
Binary	–	–	–	–	–	60.4	68.5	77.2	83.6
Scalar	–	–	–	–	–	61.6	70.1	78.6	85.1
Multi-head	–	–	–	–	–	65.1	74.3	83.2	89.6
Free-form	45.3	43.8	52.4	61.1	35.9	63.7	73.0	81.0	87.7
Redundant	59.3	62.2	66.9	74.5	46.8	65.9	75.0	84.5	90.7
Minimal	60.8	64.2	71.6	75.3	52.6	68.7	77.4	85.2	91.1

ID J_{cert} is 78.2/67.5/55.1% for one/two/three-or-more factors; Table C5 reports OOD strata.

Table C4. Metrics supplementing Table 2: A, length controls; B, data ablations; C, relative training cost and emitted-witness validity; D, test FR and resolution at equal Gate exposure. Performance is percent; Edits is mean repair size.

A. Target	Repair	Edits
Minimal (short)	82.8	1.6
Redundant (short)	83.6	2.8
Minimal (long)	83.0	1.6
Redundant (long)	83.8	2.8

Table C4 (B). Additional data-ablation metrics (percent).

Training data	ID BA	OOD J_{valid}	OOD BA
Full counterfactual set	84.6	66.8	75.5
Without repaired examples	82.5	60.9	72.7
Without near-boundary legal examples	82.9	63.8	73.4
Without multi-factor violations	83.3	61.8	72.9

Table C4 (C–D). Training cost, emitted-witness validity, FR, resolution, and task success at equal Gate exposure. Cost is relative to Gate/Binary; rates are percentages.

Training	Cost	Witness	FR	Res.	Task
Gate / Binary	1.00	–	1.1	–	77.9
Critic+Gate / Binary	1.72	–	1.0	–	80.1
Gate / Certificate	1.28	90.0	1.0	73.0	79.2
Critic+Gate / Certificate	2.18	95.0	1.0	84.0	81.8

Table C5. Within-stratum OOD results on the common short-target population (percent). M/R denote Minimal/Redundant supervision.

Stratum	M BA	R BA	M J_{valid}	R J_{valid}
1 factor	81.0	79.2	74.2	72.0
2 factors	74.6	72.0	65.0	62.4
3+ factors	66.8	63.7	54.1	50.8
Single solution	76.1	73.8	67.9	65.5
Multiple solutions	73.8	71.4	64.2	61.7

C.3 Planning Cost and Decoding

Table C6. All 4800 paired tasks. A, mean GPU seconds and total SE (Generate includes acceptance/scoring). B, rounds, proposed requests, jobs, NFE, CPU seconds, and terminal counts; 16 requests per round. C, success differences/SE/95% intervals (pp). Totals use unrounded records and include failures/timeouts.

A. Policy	Propose	Predict	Verify	Preview	Generate	Total $\pm$ SE
Full Rollout	7.24	0.00	0.00	0.00	68.68	75.93 $\pm$ 1.66
Verdict Gate	7.96	1.71	0.00	0.00	42.92	52.60 $\pm$ 1.24
Verifier Search	7.59	0.00	6.44	0.00	37.18	51.21 $\pm$ 1.14
Certificate Gate	7.87	2.03	2.71	0.00	38.11	50.72 $\pm$ 1.19
Score Gate	8.44	2.17	0.79	0.00	47.47	58.88 $\pm$ 1.43
Cheap preview	7.87	2.03	0.74	16.06	40.89	67.59 $\pm$ 1.56

Table C6 (continued)

A. Policy	Propose	Predict	Verify	Preview	Generate	Total ± SE
Random pruning	9.46	0.00	0.00	0.00	45.82	55.28 ± 1.40
Discard content	7.96	2.05	0.00	0.00	42.92	52.94 ± 1.25

B. Policy	Rounds	Proposed	Gen. jobs	NFE	CPU s	Success / failure / timeout
Full Rollout	4.86	77.82	77.82	1556.47	1.12	3962 / 768 / 70
Verdict Gate	5.33	85.24	48.48	969.54	0.86	3835 / 870 / 95
Verifier Search	5.08	81.22	41.93	838.66	2.91	3917 / 807 / 76
Certificate Gate	5.27	84.26	43.03	860.51	1.78	3918 / 790 / 92
Score Gate	5.67	90.75	53.87	1077.42	0.93	3772 / 906 / 122
Cheap preview	5.29	84.64	46.39	1266.40	0.83	3882 / 824 / 94
Random pruning	6.33	101.28	51.74	1034.74	0.96	3047 / 1589 / 164
Discard content	5.33	85.24	48.48	969.54	0.86	3835 / 870 / 95

C. Task-success contrast	Paired tasks	Difference	SE	95% interval
Certificate − Full, common configuration	4800	-0.92	0.55	[-1.98, 0.17]
Certificate − Verdict, 70 GPU seconds	4800	2.46	0.59	[1.31, 3.60]

Table C6 (D). Success by task family (percent) and complete-episode wall seconds at the common configuration. Each family has 1200 paired tasks.

Policy	Contact	Persistence	Ordering	Resource	Wall s
Full Rollout	84.2	83.0	81.2	81.8	80.3
Random pruning	66.5	65.7	62.2	59.5	60.4
Score Gate	79.8	80.6	77.0	77.0	63.6
Cheap preview	83.0	82.7	79.8	78.0	71.9
Verdict Gate	82.8	82.2	77.0	77.7	57.0
Verifier Search	84.2	84.4	77.8	80.0	57.5
Certificate Gate	84.9	82.7	79.6	79.3	56.0

Witness/scope truncation saves 18% [17%,19%] of post-judgment decoding with identical decisions (Table C7); Tables 1 and 3 charge full decoding.

Table C7. Decoded prefixes under the same 16-check budget. No-field fallback and Search-16 have identical decisions. Witness/scope and full decoding share verifier inputs. Rates and post-judgment decoding savings are percentages.

Prefix / verifier input	Res.	MR	Decoding saving
Judgment + fallback	70.48	44.50	100
+ constraint + fallback	75.53	39.46	86
+ entity/time + fallback	80.71	34.28	60
+ witness and scope; repair omitted	84.64	30.34	18
Full certificate	84.64	30.34	0

C.4 Planning Diagnostics and Working Points

Table C8 uses the fixed request pool (matched-pruning populations: C19).

Table C8. Paired fixed-pool statistics, per 16 requests: 2000 histories, 22400 infeasible and 9600 feasible requests. Complete generation uses 20 NFE; previews add 64 NFE. Verifier Search uses the development-selected 64-check budget; Search-16 is the no-field prefix reference. Non-success comes from the separate complete-episode evaluation. Rates are percentages.

Method	FR	MR	Full jobs	Preview	NFE	Non-success
Full Rollout	0.00	100.00	16.000	0	320.000	17.46
Score Gate	2.82	42.55	9.431	0	188.610	21.42
Verdict Gate	1.93	38.55	9.025	0	180.500	20.10
Verifier Search	0.67	30.87	8.225	0	164.500	18.40
Search-16	0.53	44.50	9.759	0	195.180	—
Cheap preview	1.44	35.73	8.733	16	238.660	19.12
Certificate Gate	0.69	30.34	8.165	0	163.310	18.38

Table C8 (continued)

Method	FR	MR	Full jobs	Preview	NFE	Non-success
Random pruning	49.07	50.71	8.124	0	162.480	36.52
Discard content	1.93	38.55	9.025	0	180.500	20.10

Table C9. Complete-task working points and one-axis sweeps. A, development selects the lowest-cost candidate count meeting 82% success; a dash means no configuration qualifies. B, GPU ratios use Full Rollout at the identical candidate count and horizon. All test episodes include failures and timeouts.

A. Method	M	Success	GPU s	Wall s
Full Rollout	16	82.54	75.93	80.26
Verdict Gate	32	82.06	99.67	107.48
Cheap preview	32	82.21	130.94	138.74
Certificate Gate	32	83.31	92.69	101.90

B. Sweep	Setting	GPU ratio	Success
Candidates	8	0.645	77.06
Candidates	16	0.668	81.62
Candidates	32	0.620	83.31
Horizon	8	0.719	79.06
Horizon	16	0.668	81.62
Horizon	32	0.591	79.75

Cache and repair re-proposals are disabled in Tables C6–C9. On the missing-observation challenge, unknown emission is 92.4% and feasible retention 99.0% (Table C22B).

C.5 Selection Budgets

Table C10 crosses list and training-seed variation; supervision comparisons fix lists.

Table C10. Event-budget learning (percent). Budgets 25/50/100 correspond to 2500/5000/10000 families; selection preprocessing is charged separately.

Selection	Budget	OOD BA	OOD J_{valid}	OOD J_{cert}	ID J_{cert}
Random	25	68.4	56.0	40.5	56.1
Random	50	71.5	60.9	48.9	62.7
Representation	25	70.2	58.7	43.1	58.8
Representation	50	73.5	63.1	51.7	65.1
Representation	100	75.5	66.8	58.5	70.4
Coverage/diversity	25	69.8	58.0	42.4	58.2
Coverage/diversity	50	72.9	62.5	50.8	64.5

C.6 Recognition Pilots

Main's development AUROC difference, Representation minus Random, is 0.0558 [0.0316,0.0807]. The secondary iteration-300 Lab difference is -1.62 pp $[-2.97, -0.36]$. Mixed-source Lab averages 79.3% over three seeds (sample SD 1.66 pp); masking contrasts use seed 0. Primary paired Lab contrasts are in Table C11B.

Table C11. Pilot recognition: A, iteration-500 Main AUROC and Lab controlled accuracy; B, paired Lab contrasts and 95% intervals in pp. Main is a development population. AUROC and accuracy retain their native scales.

A. Model	Main	Lab macro	Continuity	Gravity	Solidity
Background-only	0.5599	75.6	100.0	63.3	63.5
Mixed-source (seed 0)	0.7186	80.6	97.6	85.5	58.6
Mixed-source (seed 1)	0.7346	77.4	99.1	67.0	66.1
Mixed-source (seed 2)	0.7221	79.8	98.3	78.0	63.3
Mixed-source, negative targets masked	0.5804	83.4	100.0	80.4	69.9
GRASP-source	0.5873	74.6	99.7	64.5	59.8
GRASP-source, negative targets masked	0.5603	75.2	98.0	69.9	57.6
Representation selection	0.6231	86.7	99.3	88.9	71.9

Table C11 (continued)

A. Model	Main	Lab macro	Continuity	Gravity	Solidity
Random selection	0.5674	86.7	99.5	88.7	71.8
Low-score selection	0.4365	72.6	99.5	61.3	56.8

B. Test	Comparison	Difference	95% CI
C1	Mixed-source − Mixed-source, negative targets masked	−2.86	[−4.49, −1.28]
C2	Representation selection − Random selection	+0.02	[−1.10, 1.15]
C3	Representation selection − Low-score selection	+14.14	[12.64, 15.61]
C4	GRASP-source − GRASP-source, negative targets masked	−0.52	[−1.81, 0.81]
C5	C1 − C4	−2.35	[−4.41, −0.32]

Table C12. Stock pilots, using their native decoded-verdict subsets. Videos/scored gives input and valid-verdict counts; beneath each percentage is its exact numerator/denominator. Missing/malformed verdicts leave these pilot rates. Core evaluation uses B.8 instead; dashes are unreported endpoints.

Population and condition	Videos / scored	Accuracy (%)	Violation recall (%)	Legal FR (%)	AUROC
		54.4	25.8	17.1	
IntPhys2 Main, stock reasoner	1012 / 1006	547/1006	130/503	86/503	–
			28.7	21.2	
Motion diagnostic, unguarded	247 / 205	–	25/87	25/118	–
			15.1	5.9	
Motion diagnostic, guarded	247 / 242	–	16/106	8/136	–
LikePhys, stock diffusion residual	920 / 920	–	–	–	0.4750

C.7 Local Reinspection

On 1024 issued-negative records, onset is within one/two frames for 653/868 cases, and within 0.125/0.250 seconds for 473/747. Table C13 reports reinspection and paired contrasts.

Table C13. Local reinspection: A, R@90, AUROC, and cumulative input-token increment; B, paired differences and marginal 95% intervals in pp. All policies include first-pass cost; annotation guidance centers on the labeled event time.

A. Input policy	R@90 (%)	AUROC	Input increment
Cached first pass	26.1	0.6655	240
Uniform five frames	40.6	0.7388	1020
Certificate-guided five frames	30.5	0.6803	1020
Annotation-guided five frames	35.2	0.7093	1020
Full seventeen frames	44.1	0.7458	2580

B. Comparison	Difference	95% CI
Certificate-guided five − uniform five	−10.02	[−14.27, −6.04]
Certificate-guided five − within-mechanism shuffled guidance	+3.80	[1.64, 6.23]
Annotation-guided five − uniform five	−5.39	[−9.77, −1.30]
Certificate-guided seven − uniform seven	−5.44	[−9.86, −1.09]
Seven-frame guidance with global anchors − uniform seven	−5.12	[−8.00, −0.49]

C.8 Restricted-Input and Field-Supervision Results

Tables C14–C15 extend Table 2; T_0 is Gate/Binary training cost.

Table C14. Restricted Critic at equal Gate exposure. Rates are percentages; training budgets are relative to Gate/Binary. Binary emits no verification-resolution output.

Restricted-input supervision	Reject	FR	Res.	Budget / T_0
Binary / restricted Critic	58.0	1.05	–	1.72
Certificate / restricted Critic	63.5	1.00	78.5	2.18

At equal total training compute, restricted-Critic rejection is 57.90% with binary targets and 62.75% with certificates. Full video raises these to 60.8% and 67.5%, respectively (Table 2C–E).

Table C15. Field-training task success, feasible rejection, fixed-workload cost index (Full Rollout = 100), and resolution, averaged over eight seeds. Table 2D gives OOD BA and rejection.

Supervision	Task	FR	Cost index	Resolution
Complete certificate	81.80	1.02	63.255	84.0
Without entity targets	79.51	1.04	66.902	78.2
Without time targets	78.95	1.06	67.716	76.9
Without repair targets	79.70	0.96	65.600	80.3
Without witness targets	78.52	1.02	68.483	75.7

Table C15 averages eight seeds with 9600 feasible candidates each. Using Table 2D rejection and $\rho = 0.70$, its cost index is $17 + 90[0.7(1 - \mathrm{Reject}/100) + 0.3(1 - \mathrm{FR}/100)]$, with proposal/prediction/verification/generation weights 10/5.97/1.03/90. Rates are averaged before rounding.

C.9 Paired Seed Contrasts

Table C16 covers 37 contrasts in 33 rows: four Resolution contrasts equal Reject. Positive performance and negative Boundary FR differences favor the first condition.

Table C16. Paired seed contrasts: mean, sample SD, and 95% t_7 interval (pp), eight seeds. Minimal–Redundant averages lengths; omitted-input/field rows compare Full with the named removal. Reject/Res. rows apply identically to both endpoints.

Contrast	**Endpoint**	**Mean**	**SD**	**95% interval**
Minimal − Redundant	OOD BA	2.20	0.74	[1.58, 2.82]
Semantics × length	OOD BA	0.00	2.36	[-1.97, 1.97]
Minimal − Redundant	J_{valid}	2.40	0.98	[1.58, 3.22]
Semantics × length	J_{valid}	0.00	1.93	[-1.61, 1.61]
Minimal − Redundant	Task	1.40	0.72	[0.80, 2.00]
Semantics × length	Task	0.00	1.29	[-1.08, 1.08]
Sharing × supervision	Reject	5.00	1.43	[3.81, 6.19]
Sharing × supervision	Equal-compute Reject	3.70	1.43	[2.50, 4.90]
Sharing × supervision	Task	0.40	1.22	[-0.62, 1.42]
Full − repaired inputs	OOD BA	2.80	1.02	[1.94, 3.66]
Full − repaired inputs	Repair-pair	14.60	1.77	[13.12, 16.08]
Full − repaired inputs	Boundary FR	-0.80	0.35	[-1.10, -0.50]
Full − repaired inputs	Composition BA	3.10	0.92	[2.33, 3.87]
Full − boundary inputs	OOD BA	2.10	1.03	[1.24, 2.96]
Full − boundary inputs	Repair-pair	3.20	1.42	[2.01, 4.39]
Full − boundary inputs	Boundary FR	-4.90	0.71	[-5.50, -4.30]
Full − boundary inputs	Composition BA	1.90	0.72	[1.30, 2.50]
Full − multi-factor inputs	OOD BA	2.60	0.89	[1.85, 3.35]
Full − multi-factor inputs	Repair-pair	5.50	2.16	[3.69, 7.31]
Full − multi-factor inputs	Boundary FR	-0.40	0.50	[-0.82, 0.02]
Full − multi-factor inputs	Composition BA	6.20	1.12	[5.26, 7.14]
Full − entity	OOD BA	2.10	1.43	[0.91, 3.29]
Full − entity	Reject / Res.	5.80	1.78	[4.31, 7.29]
Full − entity	Task	2.29	0.71	[1.70, 2.89]
Full − time	OOD BA	2.70	0.83	[2.00, 3.40]
Full − time	Reject / Res.	7.10	1.52	[5.83, 8.37]
Full − time	Task	2.85	0.66	[2.29, 3.40]
Full − repair	OOD BA	3.40	1.53	[2.12, 4.68]
Full − repair	Reject / Res.	3.70	1.47	[2.47, 4.93]
Full − repair	Task	2.10	1.10	[1.18, 3.02]
Full − witness	OOD BA	3.00	1.32	[1.90, 4.10]
Full − witness	Reject / Res.	8.30	1.19	[7.31, 9.29]
Full − witness	Task	3.28	0.65	[2.74, 3.82]

C.10 VERIFICATION-BUDGET SCAN

C17–C18 use 500/2000 development/test histories, 16 requests each: $n_- = 5600/22400$, $n_+ = 2400/9600$. A separate 500-family calibration set fixes settings. Within split, one fixed policy shares candidates/truth/inventory/randomness across conditions.

Table C18B reports B.7's development-selected test points, including failed selections. Table 4A contrasts the prespecified 16-check certificate condition with Search at 16 and at its development-selected budget.

Table C17. Verifier counts by budget, development / test. Each condition cell gives V on the first line and feasible FP on the second. Completed S/N counts are shared across conditions; $X = n_- - V - S - N$. Denominators are $n_- = 5600/22400$, $n_+ = 2400/9600$; Equation B3 recovers every rate.

Content	$b = 4$	$b = 8$	$b = 16$	$b = 32$	$b = 64$
	3042 / 12557	3462 / 14353	3787 / 15603	3979 / 16310	4051 / 16589
Full	19 / 51	22 / 60	24 / 66	25 / 67	25 / 68
	2638 / 10957	3154 / 13051	3556 / 14635	3839 / 15748	3994 / 16296
Delete entity	17 / 49	19 / 57	22 / 62	24 / 66	25 / 68
	2496 / 10620	2999 / 12633	3467 / 14299	3782 / 15538	3943 / 16148
Delete time	17 / 44	18 / 53	23 / 55	24 / 62	24 / 65
	2499 / 10291	2960 / 12321	3386 / 14071	3699 / 15288	3874 / 15985
Delete witness	13 / 45	17 / 51	22 / 59	23 / 62	24 / 64
	2285 / 9613	2773 / 11699	3243 / 13530	3617 / 14981	3854 / 15870
Permute entity	15 / 46	19 / 51	21 / 58	23 / 60	23 / 66
	2182 / 8986	2654 / 11015	3135 / 12906	3517 / 14540	3793 / 15624
Permute time	15 / 41	19 / 51	21 / 56	21 / 61	24 / 62
	1952 / 8268	2447 / 10214	2931 / 12165	3360 / 13863	3661 / 15088
Permute witness	14 / 40	18 / 45	19 / 51	21 / 59	24 / 63
	1570 / 6544	2286 / 9605	2980 / 12431	3471 / 14405	3760 / 15486
Search	10 / 35	15 / 44	20 / 51	23 / 61	24 / 64
Shared S	392 / 1526	468 / 1797	514 / 1980	540 / 2083	553 / 2135
Shared N	266 / 1099	303 / 1258	334 / 1376	351 / 1417	359 / 1447

Table C18. Verifier work by budget, development / test. Each cell gives executed checks per request (first line) and prediction-plus-verification GPU seconds per 16 requests (second line). Search has zero prediction cost. Budgets are nested prefixes; exclusion and inventory completion stop early.

Content	$b = 4$	$b = 8$	$b = 16$	$b = 32$	$b = 64$
	2.98 / 2.96	4.12 / 4.02	5.61 / 5.36	7.55 / 7.08	10.41 / 9.58
Full	0.695 / 0.698	0.790 / 0.787	0.914 / 0.898	1.078 / 1.041	1.319 / 1.249
	3.06 / 3.03	4.38 / 4.28	6.15 / 5.92	8.46 / 8.02	11.74 / 10.96
Delete entity	0.701 / 0.705	0.811 / 0.809	0.959 / 0.946	1.153 / 1.121	1.429 / 1.365
	3.08 / 3.05	4.48 / 4.35	6.37 / 6.07	8.83 / 8.32	12.32 / 11.45
Delete time	0.704 / 0.706	0.820 / 0.815	0.977 / 0.959	1.183 / 1.146	1.476 / 1.406
	3.09 / 3.06	4.49 / 4.40	6.43 / 6.20	9.05 / 8.56	12.86 / 11.89
Delete witness	0.704 / 0.708	0.821 / 0.819	0.982 / 0.970	1.202 / 1.167	1.525 / 1.446
	3.13 / 3.10	4.63 / 4.52	6.74 / 6.46	9.59 / 9.04	13.62 / 12.58
Permute entity	0.708 / 0.710	0.833 / 0.829	1.009 / 0.991	1.247 / 1.207	1.584 / 1.502
	3.15 / 3.13	4.71 / 4.63	6.92 / 6.74	9.99 / 9.58	14.31 / 13.47
Permute time	0.709 / 0.713	0.839 / 0.838	1.023 / 1.015	1.280 / 1.253	1.642 / 1.578
	3.19 / 3.16	4.85 / 4.75	7.29 / 7.06	10.72 / 10.27	15.64 / 14.75
Permute witness	0.712 / 0.715	0.851 / 0.849	1.054 / 1.042	1.341 / 1.311	1.752 / 1.687
	3.27 / 3.24	5.08 / 5.00	7.58 / 7.36	10.85 / 10.38	15.33 / 14.40
Search	0.335 / 0.335	0.486 / 0.482	0.694 / 0.680	0.967 / 0.932	1.341 / 1.266

Table C18 (B). Development-selected verification budgets and frozen test outcomes. Select the cheapest point with development Res. $\geq 83\%$ and FR $\leq 1\%$. If none qualifies, display the first point reaching the resolution target, or 64 if none does. Screen is GPU seconds per 16 requests; rates are percentages.

Content	Budget	Res.	Reject	FR	Screen s	Dev. rule
Full	32	88.44	72.81	0.70	1.041	No joint point
Delete entity	32	85.93	70.30	0.69	1.121	Selected
Delete time	32	84.99	69.37	0.65	1.146	Selected
Delete witness	64	87.35	71.36	0.67	1.446	Selected
Permute entity	64	86.84	70.85	0.69	1.502	Selected
Permute time	64	85.74	69.75	0.65	1.578	Selected
Permute witness	64	83.35	67.36	0.66	1.687	No joint point
Search	64	85.12	69.13	0.67	1.266	Selected

Full content and witness permutation have no qualifying development point. For Full, the prespecified 16-check condition has development resolution below 83%, while larger budgets exceed the 1% development-FR threshold. The selected conditions are entity/time deletion at 32, witness deletion and entity/time permutation at 64, and Search at 64. Test outcomes are evaluated separately from development selection.

C.11 Matched-Pruning Populations and Counts

Targets 30/40/48.6% use 4–5/6–7/7–8 removals with upper-count probabilities 0.800/0.400/0.776. Attainable-pool conditioning changes mixture/rate: C19 gives denominators, C20 integer errors.

Table C19. Attainable populations for matched pruning. A pool is included for every arm only when Certificate can meet the shared integer removal quota. Mixture counts and realized rates refer to those included pools. Development and test use independent histories.

Split	Target	Pools	Excluded (%)	n_-	n_+	Removed	Retained	Actual (%)
Dev	30.0	473/500	5.40	5301	2267	2260	5308	29.863
Dev	40.0	388/500	22.40	4355	1853	2474	3734	39.852
Dev	48.6	295/500	41.00	3320	1400	2291	2429	48.538
Test	30.0	1899/2000	5.05	21277	9107	9086	21298	29.904
Test	40.0	1660/2000	17.00	18616	7944	10578	15982	39.827
Test	48.6	1269/2000	36.55	14243	6061	9796	10508	48.247

Table C20. Matched-pruning counts, development / test. All policies remove the same within-pool quota. With Table C19 denominators, FR $= 100\mathrm{FP}/n_+$ and MR $= 100(n_- - \mathrm{TP})/n_-$. FP + TP equals Removed; Table 3B reports test FR, and Table C20B gives both rates.

Target	Policy	False vetoes	True vetoes
30.0	Score	57 / 268	2203 / 8818
30.0	Verdict	7 / 60	2253 / 9026
30.0	Certificate	13 / 36	2247 / 9050
30.0	Random	696 / 2708	1564 / 6378
40.0	Score	91 / 456	2383 / 10122
40.0	Verdict	15 / 115	2459 / 10463
40.0	Certificate	15 / 44	2459 / 10534
40.0	Random	748 / 3170	1726 / 7408
48.6	Score	112 / 570	2179 / 9226
48.6	Verdict	28 / 163	2263 / 9633
48.6	Certificate	17 / 45	2274 / 9751
48.6	Random	702 / 2927	1589 / 6869

Table C20 (B). Test FR / MR (percent) at matched target pruning rates. Aggregate realized pruning is 29.904%, 39.827%, and 48.247%, respectively; policies share the included pools and within-pool removal counts.

Policy	30.0% target	40.0% target	48.6% target
Score	2.943 / 58.556	5.740 / 45.627	9.404 / 35.224
Verdict	0.659 / 57.579	1.448 / 43.796	2.689 / 32.367
Certificate	0.395 / 57.466	0.554 / 43.414	0.742 / 31.538
Random	29.735 / 70.024	39.904 / 60.206	48.292 / 51.773

C.12 Multiplicity Analysis

C21 applies B.8's separate eight-test Holm adjustments, with $p_z = 2\Phi(-|d/\mathrm{SE}|)$.

Table C21. Eight-comparison Holm analysis. Differences and SEs are percentage points. Fixed-model effects use paired histories; the budgeted task effect uses paired task/environment units. Normal and mixed-tail probabilities are adjusted jointly across all eight contrasts.

Primary contrast	Difference / SE	p_z	Holm-z	p_{mix}	Holm-mix
ID BA: Minimal − Free-form	4.00 / 1.15	0.00050	0.00151	0.01029	0.03087
J_{cert}: Minimal − Free-form	21.40 / 2.20	$< 10^{-12}$	$< 10^{-12}$	2.6×10^{-5}	0.00015
Four-axis BA: Minimal − Redundant	2.45 / 0.90	0.00648	0.01297	0.02967	0.05934
Length-controlled BA: Minimal − Redundant	2.20 / 0.96	0.02192	0.02192	0.05567	0.05934

Table C21 (continued)

Primary contrast	**Difference / SE**	p_z	**Holm-z**	p_{mix}	**Holm-mix**
Compute-matched sharing interaction	3.70 / 0.79	2.8×10^{-6}	1.4×10^{-5}	0.00225	0.00901
Res.: full − entity permutation	9.25 / 0.22	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$
Res.: full − witness permutation	15.35 / 0.25	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$	$< 10^{-12}$
70-second Task: Certificate − Verdict	2.46 / 0.59	3×10^{-5}	0.00012	3×10^{-5}	0.00015

All eight contrasts pass normal-tail Holm adjustment; six pass mixed-tail adjustment. The two remaining adjusted probabilities are 0.05934. The Certificate–Full task interval in Table C6 remains above the secondary -2 pp boundary but crosses the primary -1 pp boundary.

C.13 Online Diagnostics and Request Dispatch

Table C22. Extended online diagnostics. A: fixed-model field interventions at 16 checks. B: false vetoes in feasible-continuation challenges; the cache challenge is separate from the uncached core. All rates are percentages.

A. Field intervention	Resolution	FR	MR
Full certificate	84.64	0.69	30.34
Delete entities → fallback	80.32	0.65	34.67
Permute entities	75.38	0.60	39.60
Delete time → fallback	78.82	0.57	36.17
Permute time	72.60	0.58	42.38
Delete witness → fallback	77.80	0.61	37.18
Permute witness	69.29	0.53	45.69

Table C22 (B). Feasible-continuation challenges.

Challenge	Scoped false veto	Control false veto
Local conflict; broader feasible branch	1.4	8.6
Missing critical observation	1.0	10.8
Wide commitment with feasible continuation	1.6	11.2
Changed premise; invalidate cache	1.2	9.4

Cancellation follows request scope. In all 400 crossing/stopping cases, the verifier cancels the crossing-bound request and retains the native request and stopping alternative (Table C23A). Requests accepting either continuation are retained in all 400 cases. Action-wide cancellation occurs only in the 400 cases with no task-valid future, which are repair-unavailable. Feasible and unresolved requests are retained. For a body radius $r \in [0.049, 0.051]$ m, a solid half-space boundary $w \in [0.299, 0.301]$ m, and a crossing commitment $x \in [0.305, 0.315]$ m, the penetration lower bound is $0.305 + 0.049 - 0.301 = 0.053$ m, above the 0.0001 m tolerance. Contact at $x = w - r$ remains a same-action continuation; its nominal center is 0.25 m. The crossing certificate therefore covers the branch request alone.

From vetoes to avoided generation. The 32000-request pool contains 22400 infeasible and 9600 feasible branches. Certificate Gate cancels 15603 infeasible and 66 feasible requests, giving 69.66% rejection and 0.69% FR. Veto-only logging dispatches every job. Executed cancellation removes 15669 jobs and leaves 16331 complete generations and 326620 NFE (Table C23B). Call pruning is 48.966%. At fixed reference weights of 10 for proposal, 7 for screening, and 90 for full generation, the total index is 62.931 versus 100 for Branch Full and 107 for veto-only logging. The 66 erroneous cancellations remain part of FR. Output compliance and task success are separate endpoints.

Table C23. Request scope and backend dispatch. A, exact-geometry interface tests, 400 histories per case; entries count correctly cancelled requests. B, the same 32000 branch requests and acceptance rules in all three arms; each dispatched job completes 20 NFE. Costs are a fixed-workload index with Branch Full = 100; complete-task GPU seconds are reported separately in Table C6.

A. Interface case	**Native**	**Branch**	**Retained requests**
Crossing impossible; same-action stopping feasible	0	400	Stopping and Native
Request accepts crossing or stopping; only crossing excluded	0	0	All 400 broad requests
No task-valid continuation	400	400	None; repair unavailable
Feasible commitment	0	0	All 400 feasible requests
Missing observations; scope unresolved	0	0	All 400 unresolved requests

B. Dispatch and cost	**Branch Full**	**Veto only**	**Cancellation**
Branch vetoes	—	15669	15669
Cancelled jobs	0	0	15669
Complete generations	32000	32000	16331
Correct / erroneous cancellations	0 / 0	0 / 0	15603 / 66
Retained infeasible / feasible	22400 / 9600	22400 / 9600	6797 / 9534
Full-generation NFE	640000	640000	326620
Actual call pruning (%)	0	0	48.966
Generation index	90.000	90.000	45.931
Screening index	0.000	7.000	7.000
Proposal index	10.000	10.000	10.000
Total index	100.000	107.000	62.931