QSTAR: SHORT PHYSICAL PREFIXES FOR TERMINAL-RISK PREDICTION AND CONDITIONAL LAYOUT SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

A layout that looks physically plausible can still fail a second later, when a stack settles, slides, and falls. A generator that must respect such constraints needs a world model, but only one answer from it: the probability that the constraint still holds at the horizon, calibrated well enough to serve as a sampling weight. We introduce QSTAR (**Q**uantum **S**ampling **T**hrough **A**nticipated **R**esponses), which obtains this answer by short physical imagination: an exact simulator runs for the first 0.25 s of a 1.5 s horizon, and a gradient-boosted head learns only the continuation of the terminal event. The predicted probabilities reweight a prior over candidate layouts into one conditional distribution. Sampling it is a rejection-sampling problem, solved exactly either by classical normalization or by fixed-point amplitude amplification, whose acceptance guarantee needs only the lower bound that weight softening supplies; we bound how score, sampler, and readout errors reach any downstream decoder. On 96 held-out contact-stack scenes, QSTAR reduces Brier error by 46% (three objects) and 36% (four objects) relative to a strong static predictor, uses 4.14 times less compute than full simulation, and keeps over 90% of the gain when only 10.4% of initial states are simulated. The error decreases in all 14 paired comparisons, covering new scenes, velocity shifts, sparse querying, and four further dynamical systems. All 48 compiled circuits accept with probability above 0.99 and reproduce the classical distribution to 10^{-14}, and a frozen diffusion editor preserves the selected layout in 95.6% of 768 images. As static selection is already near its ceiling here, the main benefit is calibrated risk at low simulation cost rather than higher top-choice survival.

1 INTRODUCTION

Generative models are increasingly expected to produce scenes that also behave correctly: stacked objects that stay up, bodies that do not collide, a pendulum that stays within range. Such requirements concern the future. A support arrangement can look fine when placed and still fail moments later, because a small initial velocity, an unequal mass, or a delayed rotation decides whether the contacts hold. To respect these constraints, a generator needs a world model that tells it what a candidate will do. It does not need the whole future, though, only the probability that the constraint still holds at the horizon, and because this number later serves as a sampling weight, it must be calibrated rather than merely correctly ranked.

Standard ways of obtaining it trade accuracy against cost. A static predictor maps geometry and mass to the outcome; it is cheap and often strong, but cannot see how contacts respond once the scene moves. Simulating every candidate and perturbation to the horizon gives direct evidence but repeats a long rollout many times. A learned world model that predicts the full future state avoids the simulator at test time, but its errors compound around contact transitions: in our probe, a learned transition model drifts to a state error of 0.57 within five recursive steps, no better than holding the state fixed (Figure 3a).

QSTAR (Quantum Sampling Through Anticipated Responses) takes a middle path: look briefly, then predict. For each candidate, the true simulator runs a short prefix (0.25 s of a 1.5 s horizon), and a learned head predicts only the continuation of one binary event, namely whether every required

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.

contact and pose constraint still holds at the horizon. The prefix is exact and cheap, and it already exposes the velocities and tilts that precede a collapse (Figure 6a); the long tail, which is costly to simulate and hard to learn step by step, is summarized by the head. Planners use the same split when they combine truncated rollouts with a learned value (Silver et al., 2016; Lowrey et al., 2019; Bhardwaj et al., 2021); here the learned quantity is a calibrated event probability. An exact prefix never hurts an oracle predictor (Proposition 1), so the empirical question is how short it can be while still paying for itself.

The probabilities turn layout selection into sampling from a prior reweighted by survival, which over $K = 2^n$ candidates is rejection sampling: propose a layout and accept it with its survival weight. Classical rejection needs on average $1/Z_\lambda$ proposals, where Z_λ is the mean acceptance, while amplitude amplification reaches the same accepted distribution with $O(Z_\lambda^{-1/2})$ oracle calls (Brassard et al., 2002; Ozols et al., 2013). Its usual obstacle, that the number of Grover steps depends on the unknown Z_λ, disappears with fixed-point amplification (Yoder et al., 2014), which needs only a lower bound; the softening that keeps classical normalization well conditioned provides exactly this bound. QSTAR therefore drives both samplers with one score table and bounds how score, sampler, and readout errors reach the output. Both samplers feed the same frozen depth-conditioned diffusion editor, so differences in the images trace back to the probabilities alone (Figure 1). Our contributions are:

- **Short-prefix terminal-risk prediction.** An exact prefix with a learned event continuation lowers Brier error by 36–46% relative to a strong static predictor on 96 held-out scenes, with 4.14 times less compute than full rollouts; simulating only 10.4% of the states keeps over 90% of the gain.
- **One distribution, two exact samplers.** We cast layout selection as survival-weighted rejection sampling, bound how score, sampler, and readout errors reach any bounded decoder utility, and verify a fixed-point quantum sampler on 48 compiled circuits whose acceptance matches theory to 10^{-13}.
- **A separated evaluation.** Probability accuracy, sampling fidelity, and visual preservation are measured separately on contact stacks, four further dynamical systems, replications, velocity shifts, and a 768-image diffusion pool. Brier error falls in all 14 paired comparisons (11 intervals exclude zero), and we report where decision-level gains stay small.

2 Related Work

Physical prediction and world models. People judge block-tower stability by approximate mental simulation (Battaglia et al., 2013), and learned models predict tower collapse from images (Lerer et al., 2016; Groth et al., 2018). Physion and CLEVRER benchmark physical prediction and causal reasoning from video (Bear et al., 2021; Yi et al., 2020), counterfactual world models query a pretrained predictor for physical structure (Bear et al., 2023; Venkatesh et al., 2024), and graph networks learn state updates (Sanchez-Gonzalez et al., 2020). Latent world models support planning by imagination (Ha & Schmidhuber, 2018; Hafner et al., 2020), while video generators still struggle to extrapolate physical laws (Kang et al., 2025). We assume known states and predict a terminal event after an exact prefix instead of rolling out a learned simulator.

Short rollouts with learned values. AlphaGo evaluates positions with truncated rollouts and a value network (Silver et al., 2016), POLO and MPC-with-value methods pair a short horizon with a terminal value (Lowrey et al., 2019; Bhardwaj et al., 2021), and MBPO keeps model rollouts short to limit compounding error (Janner et al., 2019). QSTAR applies this split to constrained generation: the exact part is a physics prefix, and the learned part is the probability of a terminal constraint event, scored with proper scoring rules (Brier, 1950; Guo et al., 2017) and used as a sampling weight.

Physics and constraints in generation. Diffusion models can be steered by physical residuals (Bastek et al., 2025), projections onto a feasible set (Christopher et al., 2024; Luan et al., 2026; Rochman-Sharabi & Louppe, 2026), or search during denoising (Ta et al., 2026). PhysGen couples rigid-body simulation with video diffusion (Liu et al., 2024), and PhyScene guides layout diffusion with collision and reachability terms (Yang et al., 2024). ControlNet (Zhang et al., 2023) on latent diffusion (Rombach et al., 2022) lets a fixed decoder follow a supplied geometry. Object-level evaluation matters because apparent physical gains can come from deleting requested objects (Ghosh

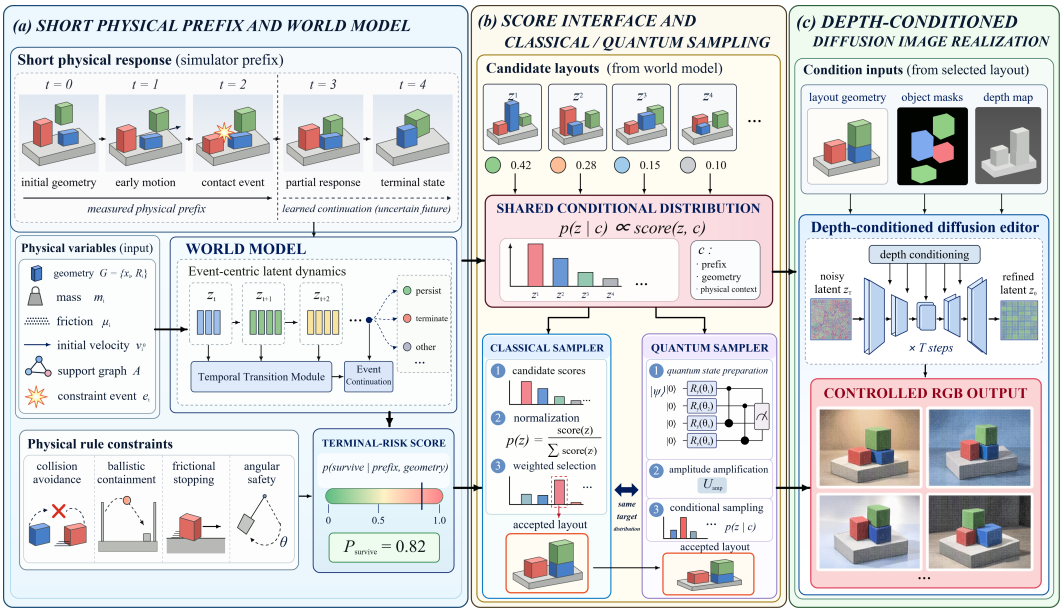

Figure 1: Overview of QSTAR. (a) The simulator runs a short prefix from the known initial state, and a learned continuation head turns the measured motion and contact events into a terminal survival probability; in our implementation the head is a gradient-boosted classifier on pooled prefix features. (b) The scores define one conditional distribution over candidate layouts, realized by exact classical sampling or by quantum state preparation with amplitude amplification. (c) A frozen depth-conditioned diffusion editor renders the selected geometry; the thumbnails are saved editor outputs.

et al., 2023); physical-commonsense benchmarks for text-to-image models target a broader endpoint (Xu et al., 2026). We decide before a shared decoder, so the decision's effect is measured apart from the decoder.

Quantum conditional sampling. Amplitude amplification boosts a marked subspace (Grover, 1996; Brassard et al., 2002), quantum rejection sampling reshapes amplitudes toward a target distribution (Ozols et al., 2013), and fixed-point variants avoid the overshoot of a fixed number of Grover iterations (Grover, 2005; Yoder et al., 2014); constraint-preserving ansatzes give distributions over feasible choices (Hadfield et al., 2019). We load scores with uniformly controlled rotations (Möttönen et al., 2005) and keep exact classical sampling as the reference, in line with calls for strong classical baselines (Bowles et al., 2024).

3 Method

3.1 Candidate layouts and terminal survival

A context c contains object geometry, masses, support structure, and a distribution of small initial perturbations η. A layout $z \in \{0,1\}^n$ chooses one of two relative horizontal offsets for each of n objects, giving $K = 2^n$ candidates with a uniform prior $p_0(z \mid c)$ and initial state $X_0(z, \eta)$. The cumulative event A_t equals one when all required parent contacts, tilt limits, and relative-displacement limits hold up to time t, and the target is the terminal survival probability $s(z, c) = \mathbb{E}_\eta[A_T \mid z, c]$.

3.2 Exact prefix, learned continuation

For a prefix length $h < T$, a short rollout returns the state X_h and the prefix event A_h. Because $A_T \le A_h$, survival factors as

$$\Pr(A_T = 1 \mid X_0, z, c) = \mathbb{E}[A_h \, v_h(X_0, X_h, c) \mid X_0, z, c], \tag{1}$$

where v_h is the probability of surviving the remaining horizon. We fit $\widehat{v}_\theta$ on terminal labels with log loss and average over m perturbations:

$$\widehat{s}_h(z, c) = \frac{1}{m} \sum_{j=1}^{m} A_h^{(j)} \, \widehat{v}_\theta\big(f(X_0^{(j)}, X_h^{(j)}, c)\big). \tag{2}$$

The prefix event is measured, so a candidate that already violates a constraint scores zero. The features f combine initial geometry, mass, and state with the prefix displacement, orientation, linear and angular velocity, and relative positions; minima, maxima, and means over objects make them independent of object count, so heads trained on three objects apply to larger stacks. The head is a histogram gradient-boosted classifier (Ke et al., 2017; Pedregosa et al., 2011) whose inference cost is negligible next to the prefix. Inference is charged for simulator setup and every prefix step.

Allocating physical queries. The static head gives a cheap estimate p_i for each candidate–perturbation pair. Within a scene we rank pairs by $p_i(1 - p_i)$, simulate the prefix only for the top fraction, and keep the static estimate elsewhere. The fraction is the smallest in $\{0.1, 0.25, 0.5\}$ whose development Brier is within a factor 1.1 of the dense prefix, which after rounding simulates 10.4% of the pairs; a learned gate lost to this simpler rule on development data.

3.3 One conditional distribution, two samplers

For $0 \le \lambda < 1$ we define softened weights and the conditional distribution

$$w_\lambda(z) = \lambda + (1 - \lambda)\widehat{s}_h(z), \qquad q_\lambda(z) = \frac{p_0(z) w_\lambda(z)}{Z_\lambda}, \qquad Z_\lambda = \mathbb{E}_{p_0}[w_\lambda] \ge \lambda. \tag{3}$$

This is the output of rejection sampling that proposes $z \sim p_0$ and accepts with probability $w_\lambda(z)$; Z_λ is its acceptance rate, which softening bounds from below. Selection results use $\lambda = 0$ (uniform fallback if all scores vanish), and a sharpened variant $\propto \widehat{s}_h^{\,b}$ with development-selected $b \in \{1, 2, 4, 8\}$ is reported for every selector.

The classical sampler normalizes the K weights exactly. The quantum sampler prepares

$$U|0\rangle = \sum_z \sqrt{p_0(z)}\,|z\rangle\big(\sqrt{1 - w_\lambda(z)}\,|0\rangle + \sqrt{w_\lambda(z)}\,|1\rangle\big) \tag{4}$$

with Hadamards and uniformly controlled rotations $R_y(\theta_z)$, $\theta_z = 2 \arcsin \sqrt{w_\lambda(z)}$ (Figure 5a). Measuring the flag in state one yields $z \sim q_\lambda$ with probability Z_λ, the coherent counterpart of one rejection trial. Amplitude amplification raises this probability without changing the accepted distribution. With the fixed-point phases of Yoder et al. (2014), the sequence length L depends only on λ and a target failure probability $\delta^2 = 0.01$, and the acceptance is at least $1 - \delta^2$ for every table with $Z_\lambda \ge \lambda$. In the oracle model, 99% success at a lower bound of 10^{-4} takes 301 calls against 46,050 classical proposals (Appendix B). Our score table is computed classically, so the circuit is an exact second realization of the same distribution; its query advantage matters when the event model itself can be evaluated coherently (Appendix B.4).

3.4 Image realization

The selected layout supplies an RGB reference and a depth map to a frozen SD1.5 depth-ControlNet editor with a shared prompt, settings, and paired seeds. We render every candidate once into a common pool and score each selector by its expectation over the pool, which removes sampling noise and unequal retry budgets. With $d(z)$ the fraction of images that preserve the intended layout, the joint proxy is $\sum_z q(z)s(z)d(z)$.

4 What the Interface Guarantees

Let $\mathrm{BS}(p) = \mathbb{E}[(A_T - p)^2]$ be the Brier score of a predictor p and $\mathcal{G}_h$ the information in the exact prefix of length h. Proofs are in Appendix A.

Proposition 1 (Longer exact prefixes never hurt the oracle)**.** *The oracle continuation* $p_h = \Pr(A_T = 1 \mid \mathcal{G}_h)$ *satisfies* $\mathrm{BS}(p_h) = \mathbb{E}[\mathrm{Var}(A_T \mid \mathcal{G}_h)]$*, and* $\mathrm{BS}(p_{h'}) \le \mathrm{BS}(p_h)$ *whenever* $h' \ge h$*.*

Table 1: Main results on 96 held-out scenes (48 per object count). Brier is the trajectory-level terminal-event error (lower is better); Survival is raw selector survival and Sharp. the development-selected sharpened survival, both in percent. Full simulation uses two separate draws (Table 3 gives a matched comparison).

	Three objects			Four objects		
Method	Brier	Survival	Sharp.	Brier	Survival	Sharp.
Uniform	–	74.40	74.40	–	63.35	63.35
Static COM rule	0.0068	98.63	98.63	0.0121	96.32	96.32
Static ExtraTrees	0.0201	91.81	99.77	0.0312	80.82	98.70
Static HGB	0.0061	98.87	99.50	0.0110	96.74	98.56
Prefix event only	0.2490	74.64	74.64	0.3346	64.02	64.02
Linear response rule	0.0065	99.09	99.09	0.0112	97.54	97.54
Prefix ExtraTrees	0.0046	97.10	99.80	0.0116	92.04	99.11
QSTAR (0.25 s + HGB)	0.0033	98.94	99.73	0.0071	97.00	98.78
Full simulator (2 draws)	–	99.07	99.07	–	97.06	97.06
Exact batch posterior	–	98.85	98.85	–	97.11	97.11

The static predictor is the case $h = 0$. The gain is free for an oracle; a learned head realizes it only if it can capture the information and the prefix is cheap, which Section 5 measures.

Proposition 2 (Improvement over the prior)**.** *Let $\mu = \mathbb{E}_{p_0}[s] > 0$ and let q^*_λ use the true survival s in place of $\widehat{s}_h$. Then $\mathbb{E}_{q^*_\lambda}[s] - \mu = (1-\lambda)\operatorname{Var}_{p_0}(s) \,/\, \big(\lambda + (1-\lambda)\mu\big) \geq 0$.*

Proposition 3 (Propagation of score, sampler, and readout error)**.** *Let $\lambda > 0$, $\epsilon_s = \mathbb{E}_{p_0}|\widehat{s}_h - s|$, and let a sampler realize q_λ with total-variation error ϵ_Q. If $\widetilde{q}_N$ is the empirical distribution of N accepted samples, then for any utility $u : \{0,1\}^n \to [0,1]$, with probability at least $1-\eta$,*

$$\big|\mathbb{E}_{\widetilde{q}_N} u - \mathbb{E}_{q^*_\lambda} u\big| \leq \min\Big\{1,\ \frac{(1-\lambda)\epsilon_s}{\lambda} + \epsilon_Q + \frac{1}{2}\sqrt{\frac{K-1}{N}} + \sqrt{\frac{\log(1/\eta)}{2N}}\Big\}. \tag{5}$$

The terms isolate score quality, sampler fidelity, and finite readout. Exact classical normalization and an exact amplified circuit both have $\epsilon_Q = 0$ and hence identical expected decoder quality. The bound is loose for small λ, but it separates three error sources that we measure one at a time; the exact posterior q^*_λ is a distributional reference, not a utility ceiling.

5 Experiments

Setup. We simulate rectangular stacks in PyBullet (Coumans & Bai, 2016) at 240 Hz (friction 0.6, 80 solver iterations), with widths in $[0.36, 0.62]$ m, heights in $[0.25, 0.75]$ m, density multipliers in $[0.5, 2]$, and easy, critical, and coupled strata of increasing offset. Initial positions and horizontal velocities are perturbed with standard deviations 0.005 m and 0.05 m/s. A candidate survives if its parent contacts hold, tilt stays below 0.20 rad, and relative displacement stays below 0.08 m up to $T = 1.5$ s. Both heads are trained on 48 three-object scenes, tuned on 12 development scenes, and frozen. The main evaluation uses 96 new scenes (48 with three and 48 with four objects) with 12 perturbations per candidate: ten form the evaluation ensemble and two feed a full-simulation control. We report trajectory Brier score and log loss, raw survival under the selector distribution, and TV to the exact posterior of the ensemble, with 95% intervals from 10,000 paired, stratified scene bootstraps (Appendix C).

5.1 Terminal-risk prediction

QSTAR lowers Brier error from 0.00609 to 0.00328 for three objects (-46%; paired interval of the difference $[-0.00508, -0.00101]$) and from 0.01103 to 0.00710 for four objects (-36%; $[-0.00646, -0.00166]$), and log loss by 50% and 45% (Table 1). The four-object gain matters because both heads were trained on three objects only. Simple physical rules do not explain the improvement: the center-of-mass rule and a constant-velocity response rule are both worse than

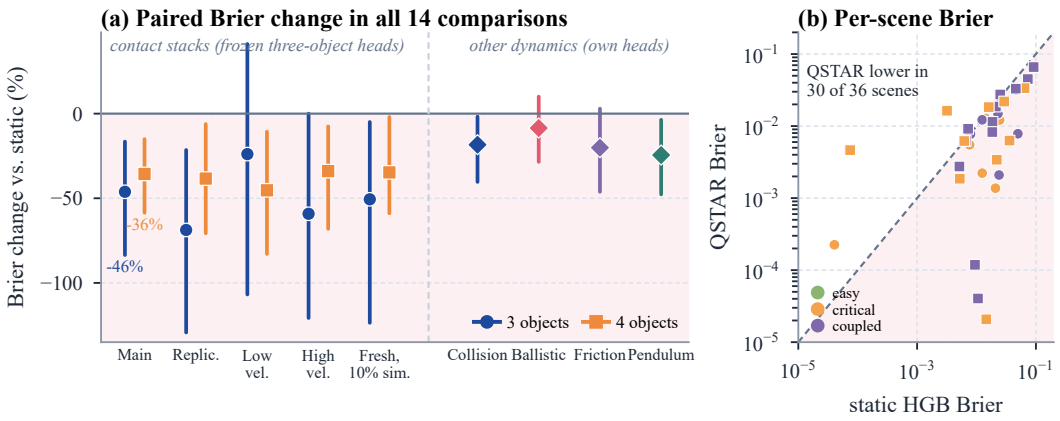

Figure 2: Terminal-risk results. (a) Relative Brier change of QSTAR against the static head, with paired 95% scene-bootstrap intervals divided by the static mean; the four procedural systems have separately fitted heads. (b) Per-scene Brier on the 36 main scenes where either head has non-negligible error; circles and squares are three and four objects.

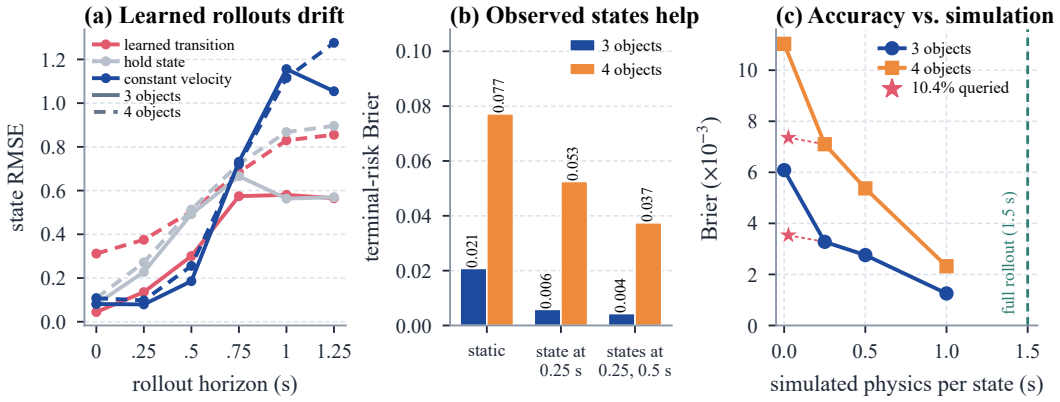

Figure 3: Why a short exact prefix. (a, b) Scene-held-out probe on 12 three-object and 12 four-object scenes: recursive rollouts of a learned transition model drift within a second, while heads that observe the true state at 0.25 s (and 0.5 s) are far more accurate than the static head. (c) Main scenes: Brier error against the physics simulated per state; the 0.25 s point is the frozen model, and stars simulate 10.4% of the states.

static HGB, and the prefix event alone is almost uninformative (Brier 0.249), because many failing candidates are still intact at 0.25 s. TV to the exact posterior falls from 0.00631 to 0.00499 and from 0.01850 to 0.01247.

Figure 2a adds every other paired comparison. On a 24-scene replication with new geometry seeds, Brier error falls by 69% and 38%; under lower and higher initial velocities by 24–59%, with the largest absolute gains at high velocity, where early motion is most revealing; and on 24 fresh scenes with only 10% of the states simulated by 51% and 35%. The error decreases in all 14 comparisons, and 11 intervals exclude zero; the exceptions are the low-velocity three-object setting, where static error is already small, and the ballistic and friction systems discussed below. The gain is spread across scenes: QSTAR is lower in 30 of the 36 scenes with non-negligible error, most strongly in the critical and coupled strata (Figure 2b; Appendix D.2).

5.2 Why a short exact prefix

Figure 3 compares two ways of adding dynamics to a static predictor. In a scene-held-out probe, a transition model trained on 0.25 s snapshots and rolled out recursively ends no better than holding the

Table 2: What the prefix contributes (main scenes). All heads share the terminal supervision; the 12- and 24-scene rows use nested training subsets.

	3 objects		4 objects	
Representation	Brier	Surv.	Brier	Surv.
Initial state only	0.0061	98.87	0.0110	96.74
Prefix pose	0.0044	99.00	0.0089	97.19
Prefix velocity	0.0026	98.89	0.0061	96.88
Prefix event only	0.2490	74.64	0.3346	64.02
No event gate	0.0034	98.96	0.0072	97.00
QSTAR (head + gate)	0.0034	98.96	0.0072	97.00
12 train scenes	0.0072	99.09	0.0101	97.30
24 train scenes	0.0038	98.95	0.0100	97.25

Table 3: Matched two-draw comparison: every method sees the same two initial conditions, and layouts are scored on ten independent perturbations. MSE compares candidate probabilities with independent-draw survival.

	3 objects		4 objects	
Method	Surv.	MSE	Surv.	MSE
Static HGB	98.86	0.0007	96.79	0.0049
QSTAR, 0.25 s	99.14	0.0045	97.00	0.0069
Full sim., 1.5 s	99.07	0.0041	97.06	0.0083

state fixed (panel a), whereas observing the true state at 0.25 s cuts terminal-risk Brier from 0.021 to 0.006 for three objects and from 0.077 to 0.053 for four (panel b). On the main scenes, accuracy keeps improving with longer prefixes, as Proposition 1 predicts (panel c): a 1.0 s prefix removes about 79% of the static error, but simulates four times longer than the frozen 0.25 s prefix, which removes 46% and 36% with one sixth of a full rollout.

Table 2 separates the response channels. The velocity at the end of the prefix carries most of the information (Brier 0.0026 and 0.0061 alone), pose helps less, and removing the event gate leaves the result unchanged, so the head already learns when an early violation is fatal. The ablations explain the frozen model rather than replace it. A head trained on 12 scenes is worse than static HGB for three objects, while 24 scenes recover most of the gain.

5.3 Compute and sparse physical queries

For 288 trajectories on a shared four-worker CPU pool, the 0.25 s prefix takes 0.712 s with setup, plus 0.003 s for inference, against 2.959 s for full rollouts: a 4.14-fold reduction. Simulating only the 10.4% of pairs with the largest static uncertainty keeps 90.8% and 93.4% of the dense-prefix improvement (Brier 0.00354 and 0.00736 against 0.00328 and 0.00710), and an end-to-end timing that re-executes the selected prefixes takes 0.104 s against 0.693 s dense and 2.963 s full, a 6.7-fold and 28.6-fold saving. With the rule frozen, 24 fresh scenes confirm the result: Brier falls from 0.01422 to 0.00702 and from 0.01011 to 0.00660, and posterior TV from 0.02462 to 0.01002 and from 0.02364 to 0.01452 (Appendix D.3). The rule works because it simulates the pairs whose outcome is uncertain, which is where the prefix is informative.

5.4 From probabilities to decisions

Better probabilities do not automatically raise top-choice survival, because static HGB already selects near-perfect layouts (MAP survival 100% and 99.4%). Raw selector survival rises from 98.87% to 98.94% (interval $[-0.04, 0.18]$ points) and from 96.74% to 97.00% ($[0.03, 0.55]$), and the constant-velocity rule reaches higher raw survival despite worse Brier error. The distribution improves more clearly: with $\lambda = 0.25$, four-object TV to the exact posterior falls from 0.00949 to 0.00653 ($[-0.00574, -0.00051]$). When selection and evaluation use independent draws (Table 3), QSTAR matches full simulation (99.14% and 97.00% against 99.07% and 97.06%) with 4.14 times less compute and beats static HGB by 0.28 points for three objects ($[0.06, 0.54]$). Its candidate MSE is higher than static HGB's, because a response prediction tracks the simulated perturbations and two draws summarize unseen ones poorly. On the fresh scenes, more accurate probabilities move mass toward genuinely uncertain layouts that the biased static model suppresses, and raw survival falls below static selection (96.37% against 97.54%, 97.22% against 97.84%). The robust outcome is a closer conditional distribution; a rule that wants the safest layout can act on the calibrated scores, e.g., by sharpening, which gives 99.73% three-object survival (Appendix D.4).

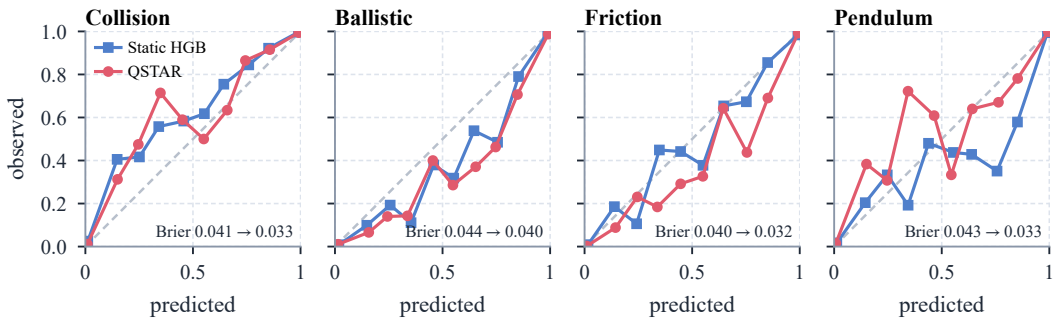

Figure 4: Four procedural systems with their own dynamics and terminal events (32 test scenes, 3,072 trajectories each): binned calibration of terminal survival, with the Brier error of each head.

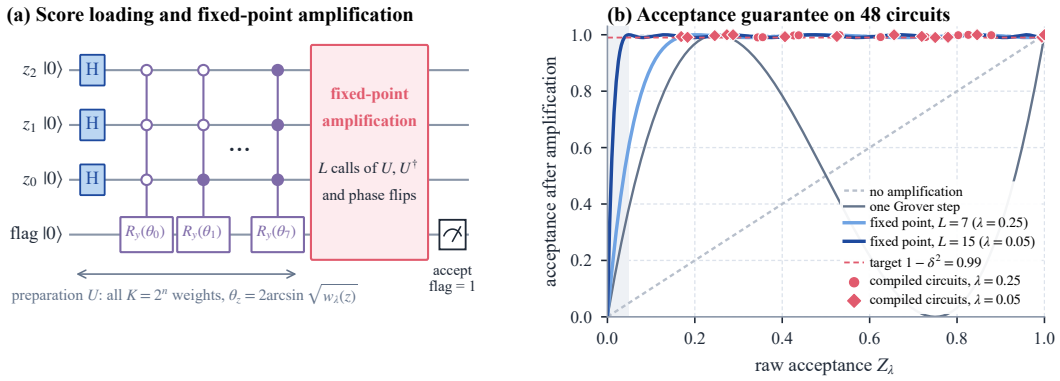

Figure 5: Quantum realization of the score interface. (a) Hadamards prepare the uniform prior, uniformly controlled rotations load all K weights into a flag, and a fixed-point sequence amplifies the accepted subspace. (b) Final against raw acceptance Z_λ for fixed-point sequences ($\delta = 0.1$), one ordinary Grover step, and the 48 compiled circuits, which lie on the theoretical curves to within 1.6×10^{-13}; the shaded band lies below the smallest softening used.

5.5 Beyond contact stacks

Four procedural systems reuse the event interface with different dynamics: collision avoidance, ballistic containment, frictional stopping, and angular safety of a damped pendulum, each with 64 training, 16 development, and 32 test scenes and separately fitted heads. QSTAR lowers Brier error in all four, by 18%, 9%, 20%, and 25%, and TV to the oracle posterior in every system (Figure 4; Appendix E). The intervals exclude zero for collision and pendulum, where the prefix reveals an imminent boundary crossing or a growing swing; for ballistic and friction, the initial state already fixes much of the outcome.

5.6 Quantum sampler

We compile the score-table circuit for the first 12 main scenes of each object count at $\lambda \in \{0.05, 0.25\}$: 48 circuits with four or five qubits and 1,180 to 6,068 CX gates. Raw acceptance Z_λ is as low as 0.169, and below 0.5 in 13 circuits. After fixed-point amplification every circuit accepts with probability at least 0.99029, matching the Chebyshev polynomial of Yoder et al. (2014) to within 1.6×10^{-13}, whereas one ordinary Grover step would overshoot for any $Z_\lambda > 0.25$ (Figure 5b). The accepted distribution matches exact classical normalization to a maximum TV of 1.14×10^{-14}, and finite-shot error follows the $\frac{1}{2}\sqrt{(K-1)/N}$ term of Proposition 3, with mean TV 0.0201 at 4,096 samples. All circuits run in statevector simulation on classical score tables, so these results establish correctness and resource counts, not a hardware speed-up.

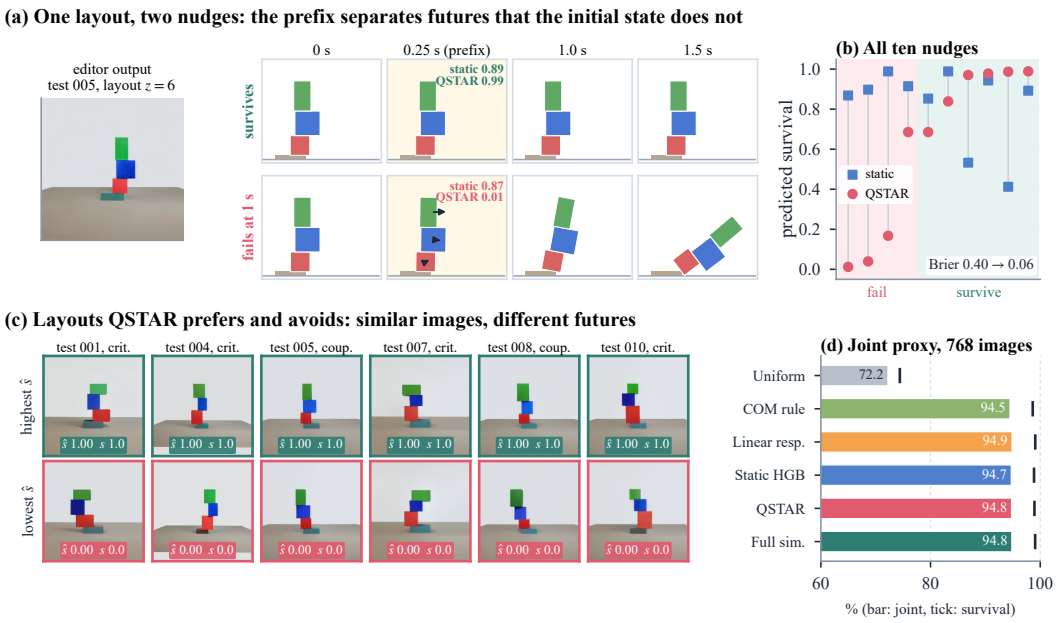

Figure 6: Seeing versus simulating. (a) The held-out layout with the largest trajectory-Brier reduction: editor output and simulated side views of a surviving and a failing nudge, with velocity arrows at the end of the 0.25 s prefix. (b) Static and QSTAR predictions for all ten nudges of this layout. (c) Editor outputs of the highest- and lowest-scored layouts in six critical or coupled scenes, with QSTAR score $\hat{s}$ and ensemble survival s. (d) Joint proxy over the shared 768-image pool; ticks mark layout survival.

5.7 Seeing versus simulating

We render every candidate of the 48 three-object main scenes with two paired seeds, 768 images in total. Figure 6a shows the layout with the largest per-layout Brier reduction. The static head, which sees each nudge's initial state, gives a surviving and a failing nudge nearly the same score (0.89 and 0.87). At 0.25 s the failing stack already slides at 0.08 m/s while the survivor is almost at rest, and QSTAR assigns 0.01 and 0.99, although the violation itself occurs only at 1.0 s; over all ten nudges, Brier error drops from 0.40 to 0.06 (panel b). The editor renders preferred and avoided layouts as equally plausible images (panel c), so appearance alone cannot separate their futures. Of the 768 images, 734 (95.57%) pass the prespecified preservation test, with mean object box IoU 0.930 (Appendix F). The joint proxy rises from 72.2% for a uniform selector to 94.5–94.9% for every physics-aware selector (panel d), while QSTAR and static HGB differ by only 0.046 points ($[-0.066, 0.146]$), in line with their nearly identical raw survival.

6 Conclusion

QSTAR treats the world model of a constrained generator as a question about one terminal event and answers it with a short exact prefix and a learned continuation. On contact stacks this reduces terminal-risk error by 36–46% at a fraction of the cost of full simulation, sparse querying keeps most of the gain, and the improvement holds on replications, velocity shifts, and four further dynamical systems. The probabilities define one distribution that an exact classical sampler and a verified quantum sampler realize identically, and a shared decoder carries the choice into images. The evidence also marks the scope. Our scenes have known states and controlled geometry, and the image proxy measures layout preservation rather than re-simulating pixels. Where static prediction is near its ceiling, better risk estimates change top-choice survival little. The score table is computed classically, so the quantum sampler's query advantage matters only once the event model can be evaluated coherently. Within this scope, spending physical computation on the first moments of motion and learning the rest is a simple, measurable recipe for generators whose constraints are defined by what happens next.

AI USE STATEMENT

In this study, generative-AI tools were only employed for grammatical error correction and manuscript language polishing. No generative-AI tools were utilized to generate research concepts, propose hypotheses, design experimental schemes, perform data processing, generate figures and tables, analyse experimental findings, or derive conclusions. All AI-produced revisions were fully examined and revised by the authors to guarantee technical correctness and integrity of academic arguments. The authors bear full responsibility for all contents presented in this manuscript.

REPRODUCIBILITY STATEMENT

Section 5 and Appendix C specify the simulator settings, scene seeds, splits, head hyperparameters, and evaluation contracts, and Appendices A and B contain all proofs. The anonymous code package includes the configurations, data schemas, experiment scripts, the script that writes the table fragments, and a single script that regenerates every data figure from the saved result files. Per-scene, per-circuit, and per-permutation records behind the aggregate numbers are listed in Appendix G.

REFERENCES

Jan-Hendrik Bastek, WaiChing Sun, and Dennis M. Kochmann. Physics-informed diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Peter W. Battaglia, Jessica B. Hamrick, and Joshua B. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

Daniel M. Bear, Elias Wang, Damian Mrowca, Felix J. Binder, Hsiao-Yu Fish Tung, R. T. Pramod, Cameron Holdaway, Sirui Tao, Kevin Smith, Fan-Yun Sun, Li Fei-Fei, Nancy Kanwisher, Joshua B. Tenenbaum, Daniel L. K. Yamins, and Judith E. Fan. Physion: Evaluating physical prediction from vision in humans and machines. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021.

Daniel M. Bear, Kevin Feigelis, Honglin Chen, Wanhee Lee, Rahul Venkatesh, Klemen Kotar, Alex Durango, and Daniel L. K. Yamins. Unifying (machine) vision via counterfactual world modeling. arXiv:2306.01828, 2023.

Mohak Bhardwaj, Sanjiban Choudhury, and Byron Boots. Blending MPC & value function approximation for efficient reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2021.

Joseph Bowles, Shahnawaz Ahmed, and Maria Schuld. Better than classical? The subtle art of benchmarking quantum machine learning models. arXiv:2403.07059, 2024.

Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. *Contemporary Mathematics*, 305:53–74, 2002.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Jacob K. Christopher, Stephen Baek, and Ferdinando Fioretto. Constrained synthesis with projected diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Erwin Coumans and Yunfei Bai. PyBullet, a Python module for physics simulation for games, robotics and machine learning. `http://pybullet.org`, 2016.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Oliver Groth, Fabian B. Fuchs, Ingmar Posner, and Andrea Vedaldi. ShapeStacks: Learning vision-based physical intuition for generalised object stacking. In *European Conference on Computer Vision (ECCV)*, 2018.

Lov K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 212–219, 1996.

Lov K. Grover. Fixed-point quantum search. *Physical Review Letters*, 95(15):150501, 2005.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Stuart Hadfield, Zhihui Wang, Bryan O'Gorman, Eleanor G. Rieffel, Davide Venturelli, and Rupak Biswas. From the quantum approximate optimization algorithm to a quantum alternating operator ansatz. *Algorithms*, 12(2):34, 2019.

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

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ali Javadi-Abhari, Matthew Treinish, Kevin Krsulich, Christopher J. Wood, Jake Lishman, Julien Gacon, Simon Martiel, Paul D. Nation, Lev S. Bishop, Andrew W. Cross, et al. Quantum computing with Qiskit. arXiv:2405.08810, 2024.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *International Conference on Machine Learning (ICML)*, 2025.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Adam Lerer, Sam Gross, and Rob Fergus. Learning physical intuition of block towers by example. In *International Conference on Machine Learning (ICML)*, pp. 430–438, 2016.

Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. PhysGen: Rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, pp. 360–378, 2024.

Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, and Igor Mordatch. Plan online, learn offline: Efficient learning and exploration via model-based control. In *International Conference on Learning Representations (ICLR)*, 2019.

Hao Luan, Yi Xian Goh, See-Kiong Ng, and Chun Kai Ling. Projected coupled diffusion for test-time constrained joint generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Mikko Möttönen, Juha J. Vartiainen, Ville Bergholm, and Martti M. Salomaa. Transformation of quantum states using uniformly controlled rotations. *Quantum Information and Computation*, 5(6): 467–473, 2005.

Maris Ozols, Martin Roetteler, and Jérémie Roland. Quantum rejection sampling. *ACM Transactions on Computation Theory*, 5(3):11:1–11:33, 2013.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Omer Rochman-Sharabi and Gilles Louppe. Predict-project-renoise: Sampling diffusion models under hard constraints. arXiv:2601.21033, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *International Conference on Machine Learning (ICML)*, pp. 8459–8468, 2020.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Huu Binh Ta, Michael Cardei, Alvaro Velasquez, and Ferdinando Fioretto. Search-augmented masked diffusion models for constrained generation. arXiv:2602.02727, 2026.

Rahul Venkatesh, Honglin Chen, Kevin Feigelis, Daniel M. Bear, Khaled Jedoui, Klemen Kotar, Felix Binder, Wanhee Lee, Sherry Liu, Kevin A. Smith, Judith E. Fan, and Daniel L. K. Yamins. Understanding physical dynamics with counterfactual world modeling. In *European Conference on Computer Vision (ECCV)*, 2024.

Yajing Xu, Yarong Lan, Jiaoyan Chen, Yichi Zhang, Jeff Z. Pan, Mingchen Tu, Zhizhen Liu, Wen Zhang, and Huajun Chen. OmniPhys: Knowledge-graph-driven benchmarking and collective optimization for physical commonsense in text-to-image generation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026. doi: 10.1145/3770855.3817552.

Yandan Yang, Baoxiong Jia, Peiyuan Zhi, and Siyuan Huang. PhyScene: Physically interactable 3D scene synthesis for embodied AI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16262–16272, 2024.

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 *International Conference on Learning Representations (ICLR)*, 2020.

Theodore J. Yoder, Guang Hao Low, and Isaac L. Chuang. Fixed-point quantum search with an optimal number of queries. *Physical Review Letters*, 113(21):210501, 2014. doi: 10.1103/PhysRevLett.113.210501.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

A PROOFS

Throughout, p_0 is the uniform prior over the $K = 2^n$ layouts, $s(z)$ the true terminal survival probability, $\widehat{s}(z)$ its estimate, and w_λ, $\widehat{w}_\lambda$, Z_λ, $\widehat{Z}_\lambda$ the corresponding softened weights and normalizers of Equation (3).

Proposition 1. For a binary event, the conditional mean minimizes squared error among all $\mathcal{G}_h$-measurable predictors: for any such f, $\mathbb{E}[(A_T - f)^2] = \mathbb{E}[(A_T - p_h)^2] + \mathbb{E}[(p_h - f)^2]$, because the cross term has zero conditional expectation. Hence $\mathrm{BS}(p_h) = \mathbb{E}[\mathrm{Var}(A_T \mid \mathcal{G}_h)]$. If $h' \geq h$, the longer prefix contains the shorter one, $\mathcal{G}_h \subseteq \mathcal{G}_{h'}$, and the law of total variance gives

$$\mathbb{E}[\mathrm{Var}(A_T \mid \mathcal{G}_h)] = \mathbb{E}[\mathrm{Var}(A_T \mid \mathcal{G}_{h'})] + \mathbb{E}\big[\,\mathrm{Var}(p_{h'} \mid \mathcal{G}_h)\big] \geq \mathrm{BS}(p_{h'}).$$

The statement concerns the oracle only. A learned head has finite capacity and data, and a longer prefix also costs more simulation, so the empirical curve in Figure 3c need not be monotone in general.

Proposition 2. With oracle weights, $Z_\lambda = \lambda + (1-\lambda)\mu$ and

$$\mathbb{E}_{q^*_\lambda}[s] = \frac{\lambda\mu + (1-\lambda)\mathbb{E}_{p_0}[s^2]}{\lambda + (1-\lambda)\mu} = \mu + \frac{(1-\lambda)\,\mathrm{Var}_{p_0}(s)}{\lambda + (1-\lambda)\mu},$$

using $\mathbb{E}[s^2] = \mu^2 + \mathrm{Var}(s)$. The identity holds whenever the normalizer is positive, including $\lambda = 0$ with $\mu > 0$. The oracle posterior is not a maximizer: with two equally likely candidates of survival 1 and 0.5, oracle conditioning gives $5/6$, while choosing the first candidate gives 1. A biased predictor can therefore exceed oracle-posterior survival while being farther from its distribution, which is why we report probability accuracy and survival separately.

Proposition 3. Write $w = w_\lambda$, $\widehat{w} = \widehat{w}_\lambda$, $Z = Z_\lambda$, $\widehat{Z} = \widehat{Z}_\lambda$. Since p_0 is normalized, $|Z - \widehat{Z}| \leq \mathbb{E}_{p_0}|w - \widehat{w}| = (1-\lambda)\epsilon_s$, and $\widehat{Z} \geq \lambda$. Adding and subtracting $p_0 w/\widehat{Z}$,

$$2\,\mathrm{TV}(q_\lambda, q^*_\lambda) \leq \frac{\mathbb{E}_{p_0}|\widehat{w} - w| + |Z - \widehat{Z}|}{\widehat{Z}} \leq \frac{2(1-\lambda)\epsilon_s}{\lambda}.$$

For $u \in [0,1]$, the variational form of total variation gives $|\mathbb{E}_q u - \mathbb{E}_{q'} u| \leq \mathrm{TV}(q, q')$, and the triangle inequality adds the sampler error ϵ_Q. For the readout term, let $\widetilde{q}_N$ be the empirical distribution of N independent accepted samples from q. Cauchy–Schwarz gives $\|\widetilde{q}_N - q\|_1^2 \leq K\|\widetilde{q}_N - q\|_2^2$, and the multinomial variance sum is $(1 - \sum_z q_z^2)/N \leq (K-1)/(KN)$, so $\mathbb{E}\ \mathrm{TV}(\widetilde{q}_N, q) \leq \frac{1}{2}\sqrt{(K-1)/N}$. Changing one sample changes the TV by at most $1/N$, so McDiarmid's inequality adds $\sqrt{\log(1/\eta)/(2N)}$ with probability at least $1-\eta$. Here N counts accepted samples; the rejection overhead depends separately on the acceptance probability.

B QUANTUM CONSTRUCTION AND ERROR ACCOUNTING

The circuit represents the same softened distribution q_λ as the classical sampler. The physical predictor computes the weights, and the circuit encodes them coherently behind a measurable acceptance flag. Fixed-point amplification follows Yoder et al. (2014); our contribution is the event-score interface and its verification. Circuits are built and compiled with Qiskit (Javadi-Abhari et al., 2024).

B.1 FROM AN EVENT SCORE TO AN AMPLITUDE

Define the block-diagonal score rotation

$$V_w = \sum_z |z\rangle\langle z| \otimes R_y(\theta_z), \qquad \theta_z = 2\arcsin\sqrt{w_\lambda(z)}.$$

Each block is unitary, so V_w is unitary. With $P_0|0\rangle = \sum_z \sqrt{p_0(z)}|z\rangle$ (Hadamards for the uniform prior), the preparation $U = V_w(P_0 \otimes I)$ gives

$$|\psi\rangle = U|0\rangle = \sum_z \sqrt{p_0(z)}\,|z\rangle\big(\sqrt{1 - w_\lambda(z)}\,|0\rangle + \sqrt{w_\lambda(z)}\,|1\rangle\big). \tag{6}$$

All K weights are loaded, so the multiplexed rotation in Figure 5a represents an explicit table rather than a constant-cost oracle.

Proposition 4 (Accepted state)**.** *Measuring the flag of Equation* (6) *accepts with probability Z_λ, and conditioned on acceptance the layout register is in the pure state $|g\rangle = \sum_z \sqrt{q_\lambda(z)}\,|z\rangle$.*

Proof. Projecting onto flag one leaves $\sum_z \sqrt{p_0(z)w_\lambda(z)}\,|z\rangle|1\rangle$, whose squared norm is Z_λ. Dividing by $\sqrt{Z_\lambda}$ and applying the Born rule gives the claim. □

B.2 Amplification preserves the conditional law

Let Π project onto flag one and assume $0 < Z_\lambda < 1$. The states $|G\rangle = \Pi|\psi\rangle/\sqrt{Z_\lambda}$ and $|B\rangle = (I-\Pi)|\psi\rangle/\sqrt{1-Z_\lambda}$ are orthonormal, and $|\psi\rangle = \sqrt{1-Z_\lambda}\,|B\rangle + \sqrt{Z_\lambda}\,|G\rangle$. The generalized reflections

$$S_G(\beta) = I + (e^{i\beta}-1)\Pi, \qquad S_\psi(\alpha) = I + (e^{-i\alpha}-1)|\psi\rangle\langle\psi| = U\,S_0(\alpha)\,U^\dagger$$

both preserve $\mathrm{span}\{|B\rangle, |G\rangle\}$. Within an iteration we apply the flag phase, the inverse preparation, the source phase, and the preparation. Every iterate therefore has the form $a|B\rangle + b|G\rangle$, and conditioning on flag one returns $|G\rangle$ up to a global phase. The raw and amplified accepted distributions thus agree even though their acceptance probabilities differ; the case $Z_\lambda = 1$ is trivial.

B.3 Fixed-point length, phases, and resources

For target failure amplitude $\delta = 0.1$ we choose the smallest odd L with

$$1 - \mathrm{sech}^2\big(\,\mathrm{arccosh}(1/\delta)/L\big) \le \lambda.$$

With $\ell = (L-1)/2$ and $\gamma = \mathrm{sech}(\mathrm{arccosh}(1/\delta)/L)$, the phases are $\alpha_j = 2\,\mathrm{atan2}\left(1, \tan(2\pi j/L)\sqrt{1-\gamma^2}\right)$ and $\beta_j = -\alpha_{\ell-j+1}$ for $j = 1, \ldots, \ell$. The final acceptance is

$$P_L(Z_\lambda) = 1 - \delta^2\,T_L\big(\gamma^{-1}\sqrt{1-Z_\lambda}\big)^2,$$

with T_L the Chebyshev polynomial of degree L. Since $Z_\lambda \ge \lambda \ge 1-\gamma^2$, the argument lies in $[0,1]$ and $P_L \ge 1-\delta^2$. This gives $L = 7$ at $\lambda = 0.25$ and $L = 15$ at $\lambda = 0.05$, with $\ell + 1$ calls to U and ℓ calls to $U^\dagger$. The abstract query count is $O(\lambda^{-1/2}\log(1/\delta))$; circuit size also includes the full score loader and the reflections, and a query bound is not a wall-clock comparison with classical categorical sampling. We compile to the $\{U, \mathrm{CX}\}$ basis with optimization level one and seed 921, and simulate the complete compiled circuit as a statevector. Across the 48 circuits, the measured acceptance agrees with $P_L(Z_\lambda)$ to within 1.6×10^{-13} (Figure 5b). Table 4 lists the compiled sizes and Figure 8 shows the per-circuit audit.

Figure 7a compares the oracle cost of the two samplers when the acceptance rate is only known to exceed λ. To guarantee 99% success, classical rejection needs $\lceil \log(0.01)/\log(1-\lambda) \rceil$ proposals, which grows as $1/\lambda$, while the fixed-point sequence needs L calls, which grows as $\lambda^{-1/2}$: 17 against 7 calls at $\lambda = 0.25$, and 46,050 against 301 at $\lambda = 10^{-4}$. The values are computed from the two-dimensional invariant subspace and checked against the closed form to 10^{-13}. They count oracle calls and assume that the weights can be evaluated coherently, which our score table does not provide. Figure 7b shows the finite-shot readout of the 48 main circuits, which follows the expected-TV bound of Proposition 3.

Table 4: Complete score-table circuits in the $\{U, \mathrm{CX}\}$ basis, including inverse preparations and phase reflections. Classical construction of the score table is additional work.

Objects	Qubits	λ	L (calls to $U, U^\dagger$)	CX gates	Depth
3	4	0.25	7	1180	1656
3	4	0.05	15	2540	3564
4	5	0.25	7	2820	3788
4	5	0.05	15	6068	8152

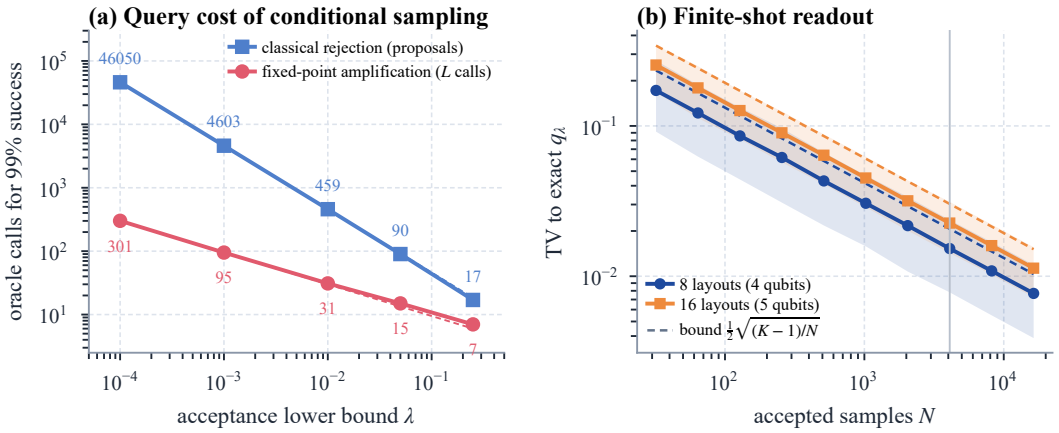

Figure 7: (a) Oracle calls needed for 99% success when the acceptance rate is only known to exceed λ: classical rejection sampling against fixed-point amplification; dashed lines show $1/\lambda$ and $\lambda^{-1/2}$ scaling. (b) Finite-shot TV of the 48 main circuits over 256 multinomial repetitions per circuit and shot budget (mean and 5–95% band) with the expected-TV bound.

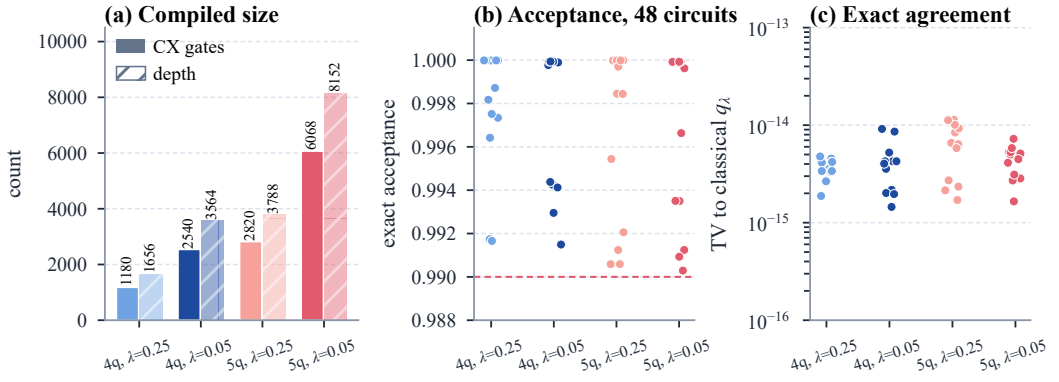

Figure 8: Audit of the 48 main circuits, grouped by qubit count and softening. (a) Compiled CX count and depth. (b) Exact acceptance after fixed-point amplification; the dashed line is the $1 - \delta^2 = 0.99$ target. (c) Total variation between the accepted statevector distribution and exact classical normalization.

B.4 A STRUCTURED-EVENT CONSTRUCTION

The main circuits load a precomputed score table. We also compiled a construction in which a classical factorized model of a future $y = (y_1, \ldots, y_r)$ given a layout is loaded explicitly. Controlled rotations load the local factors in topological order,

$$|\Psi\rangle = \sum_{z,y} \sqrt{p_0(z) \prod_j p(y_j \mid \mathrm{pa}_j, z)}\, |z, y\rangle,$$

a coin $b \sim \mathrm{Bernoulli}(\lambda)$ and a reversible flag computation mark $e = b \vee A_T(y)$, and marginalizing b and y gives $\Pr(e = 1 \mid z) = \lambda + (1 - \lambda) \Pr(A_T = 1 \mid z)$. The accepted layout marginal therefore has the same normalized form as the direct loader. Keeping the future register can entangle the accepted layout with its future; the marginal is still correct, but the reduced layout state need not be pure. Table 5 compares the two constructions and Figure 10 shows the registers. The structured preparation compiles to 5,883 CX gates and one ordinary amplification iterate to 18,229. An optimized compilation of the same model, with a 243-CX preparation, gives a complete fixed-point circuit at $\lambda = 0.25$ with 3,441 CX gates on six development contexts, all accepting above 0.99. This model admits efficient classical contraction, so the construction is a compilation study, not evidence that the physics engine or the HGB head runs coherently.

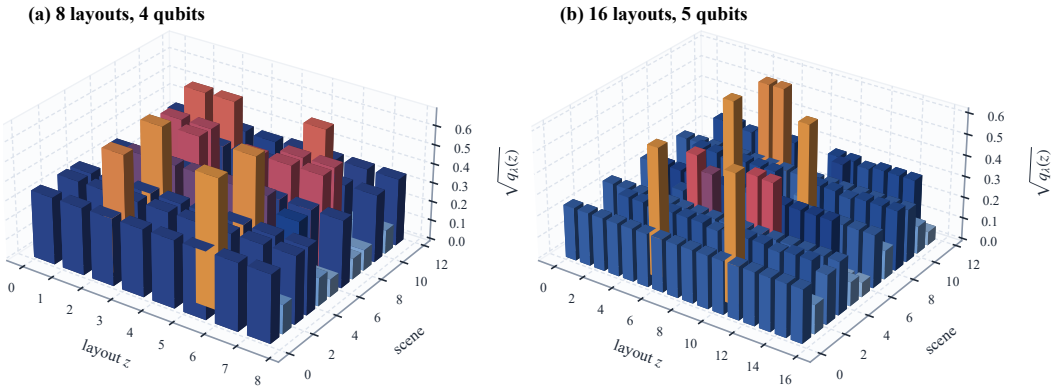

Figure 9: Accepted amplitudes $\sqrt{q_\lambda(z)}$ at $\lambda = 0.05$ for the 12 compiled three-object scenes (a) and four-object scenes (b). Uniform bars mark scenes where every layout survives; tall, isolated bars mark scenes where the score concentrates mass on a few safe layouts.

Table 5: The two implemented constructions. They use different score models, so their qubit counts and acceptance values should not be compared directly.

Property	Main score table	Structured-event construction
Layout register	3 or 4 qubits	3 qubits
Future register	none	6 qubits
Auxiliary registers	1 acceptance flag	1 softening coin, 1 event flag
Total	4 or 5 qubits	11 qubits
Input	prefix HGB score table	local conditional probability tables
Verification set	48 circuits	saved development context
Acceptance	≥ 0.99029 after fixed-point amplification	raw 0.47343; one ordinary iterate 0.57942
Accepted TV	max. 1.14×10^{-14}	raw 2.91×10^{-15}; amplified 4.54×10^{-15}

B.5 Finite precision

Suppose the loader uses angles $\widetilde{\theta}_z$ with $\max_z |\widetilde{\theta}_z - \theta_z| \leq \rho$. The direct-sum structure gives $\|V_{\widetilde{w}} - V_w\|_{\mathrm{op}} = \max_z \|R_y(\widetilde{\theta}_z) - R_y(\theta_z)\|_{\mathrm{op}} \leq \rho/2$. For a sequence of L approximate preparations or inverses with per-call operator error at most ε_U, a telescoping argument bounds the state error by $L\varepsilon_U$, and trace-distance contraction bounds the TV distance τ between the full measured distributions by the same quantity; approximate reflections add their own terms. If both implementations accept with probability at least $a > 0$, then $\mathrm{TV}(\widetilde{q}, q) \leq \min\{1, 2\tau/a\}$: the accepted joint ℓ_1 error and the acceptance-mass error are each at most 2τ, and conditioning divides by twice the acceptance mass. Amplification can therefore stabilize readout, but it cannot repair a biased physical score.

C Experimental Details

C.1 Contact-stack scenes

Geometry is sampled before perturbations. Each object receives two offsets of opposite sign, with magnitudes relative to half the parent width: $[0.02, 0.25]$ for easy scenes, $[0.60, 1.02]$ for critical scenes, and $[0.30, 0.78]$ for coupled scenes. Candidate bits select offsets, and cumulative offsets give object centers. The pedestal is 0.65 m wide, objects are 0.5 m deep, and masses are proportional to volume times the density multiplier. Perturbation draws are paired across the candidates of a scene. Figure 11 shows examples. The replication uses geometry seeds 81000/82000 for three/four objects and the main expansion 91000/92000; the low- and high-velocity runs reuse the first 12 geometries of each main group and change only the velocity scale, so they are paired robustness checks rather than

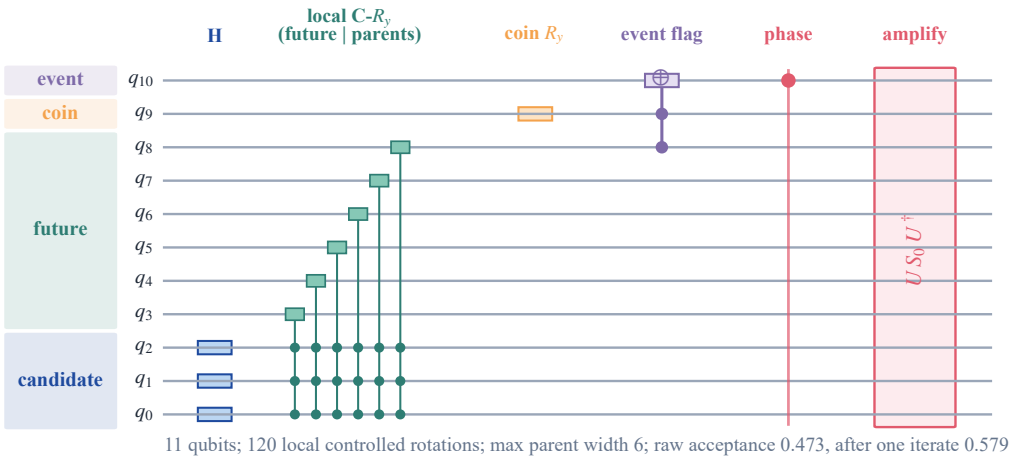

Figure 10: Registers of the 11-qubit structured-event construction. Gate blocks summarize the 120 local controlled rotations (maximum parent width six) rather than showing every decomposed gate.

new geometries. The replication, expansion, and velocity conditions required 24,192 new trajectories; the original training and development data were reused with frozen heads.

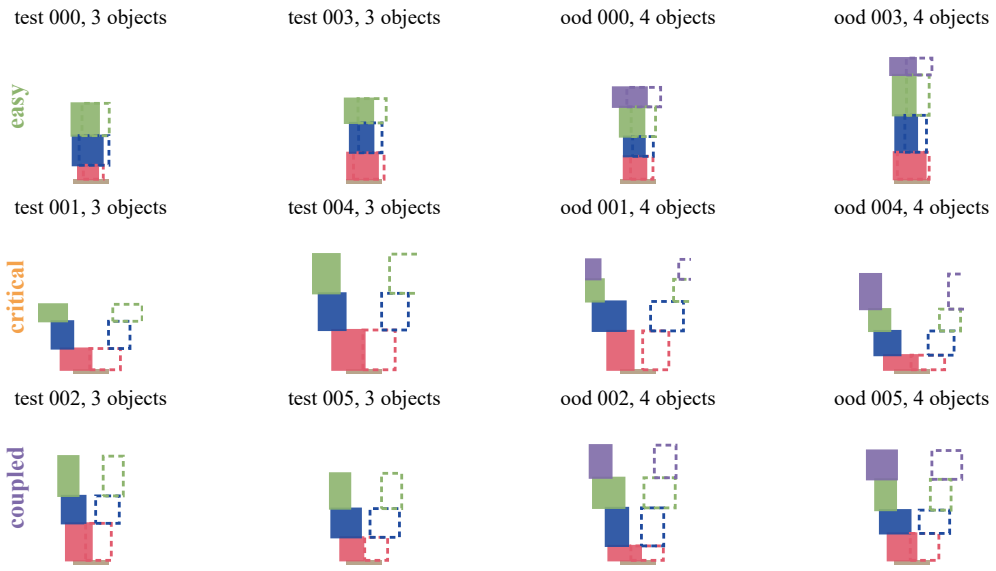

Figure 11: Contact-stack scenes from the three strata. Filled boxes show layout $z = 0$ and dashed outlines layout $z = 2^n - 1$, which flips every relative offset; object colors give the order from the bottom.

C.2 Features, heads, and baselines

The state has 12 channels: position, Euler orientation, linear velocity, and angular velocity. Initial features include dimensions, scaled mass, parent width, object and supported-substack center-of-mass margins, relative positions, velocities, and normalized margins; prefix features add displacement and response channels. Pooling ignores padded objects, which lets heads trained on three objects score larger stacks. No terminal outcome is used as an input. The static and prefix heads use 200 boosting iterations with development-selected leaves (7 and 15). ExtraTrees controls use 300 trees, minimum leaf size two, and seed zero. Response ablations use 200 iterations, 15 leaves, L2 regularization one, no early stopping, and seed zero; the 0.5 s and 1.0 s heads use the same labels and training scenes, and the training-size ablations use the first 12 and 24 scenes. The linear response rule extrapolates the

prefix position and orientation over the remaining 1.25 s with the instantaneous velocities, recomputes overlap and supported-substack margins, and checks them together with prefix survival; the static COM rule checks the initial substack margins.

C.3 Information contracts and statistics

Two contracts are kept separate. The main comparison conditions on a fixed ensemble of ten known initial states and scores each prediction against its own future. The independent-draw comparison uses two initial states to choose a layout and measures its survival on ten other perturbations. The two answer different questions and are never pooled. In both, test outcomes are used only for evaluation, and the prefix is executed as part of inference. For every paired effect the bootstrap unit is the scene (Efron, 1979): with d_i the QSTAR-minus-static effect of scene i, we resample the N scene rows 10,000 times, stratified by scene type, and report percentile intervals. Perturbations are never resampled independently. TV is computed against the exact posterior of the finite evaluation ensemble, not an infinite-sample oracle. Table 6 lists the evaluation blocks.

Table 6: Evaluation blocks. Scene counts are distinct geometries or simulator contexts; image counts are generated outputs.

Block	Count	Modality	Contract
Contact stacks, main	96 scenes	simulator	3/4 objects, frozen static and prefix heads
Replication	24 scenes	simulator	new geometry seeds, frozen heads
Velocity shifts	48 scenes	simulator	low/high velocity, paired geometries
Fresh sparse-query set	24 scenes	simulator	frozen 10% allocation rule
Object-count transfer	12 scenes	simulator	5 and 6 objects, frozen heads
Procedural systems	128 scenes	simulator	four equations of motion, separate heads
Image pool	768 images	frozen diffusion editor	48 scenes, 8 layouts, 2 paired seeds

C.4 Query allocation and timing

The uncertainty policy ranks all candidate–perturbation states of a scene by the frozen static prediction and rounds the requested count upward. Because the ranking covers all 12 available draws, including the two reserved for the simulator control, some queried states lie outside the ten-state evaluation ensemble; all queries are charged. The policy never reads a future response before deciding to query it. The learned-value alternative uses three scene-level folds of consecutive three-scene groups to keep strata balanced; out-of-fold static and prefix models provide the difference of their squared errors as a regression target, and a 150-iteration, seven-leaf HGB regressor predicts it from initial features. Only training scenes fit this gate, and the original development set chose uncertainty allocation over it. The end-to-end timing executes the selected queries in the simulator: static features and prediction, ranking, simulator initialization, prefix execution, response features, continuation prediction, and score replacement. A shared four-worker pool processes three alternating repeats; selected responses are checked outside the timed region against independently generated full trajectories and agree exactly. Median adaptive, dense, and full times are 0.10358, 0.69271, and 2.96344 s for 288 states, with 30 physical queries in the adaptive run. This benchmark uses six training geometries for timing only; quality is always measured on held-out scenes.

C.5 Image generation and the preservation test

References use a common frontal camera and red, blue, and green identities from bottom to top at 512×512. The editor is SD1.5 with a depth ControlNet (Rombach et al., 2022; Zhang et al., 2023) in image-to-image mode, with strength 0.65, 25 steps, guidance 7.5, conditioning scale 1.0, and paired seeds. The detector takes the largest HSV component of each color. An image passes if every component has at least 128 pixels, the minimum object box IoU is at least 0.35, the mean IoU at least 0.5, the maximum center error at most 0.08 of the image width, and the vertical order is correct; references must pass the same test before generation. All candidates and both seeds are kept, including failures, so selector-weighted preservation can differ even on a shared pool. Visual inspection was not

used to choose images or tune thresholds. The 24-image pilot preceded a 192-image first expansion, and the remaining 576 images completed all 48 scenes without regenerating the initial pool.

D Additional Contact-Stack Results

D.1 Replication and velocity shifts

Table 7 and Figure 12 report the frozen-model replication and the velocity shifts. Risk improvements are consistent across settings; survival changes are small and their intervals are wide, except for the four-object main setting and the three-object high-velocity setting.

Table 7: Frozen-model replication and velocity shifts. Static and Prefix give Brier error; ΔS is the change in raw survival in percentage points. The low- and high-velocity sets share geometries with the main set.

	Three objects			Four objects		
Setting	Static	Prefix	ΔS	Static	Prefix	ΔS
Replication	0.0193	0.0060	+1.22	0.0103	0.0063	+0.32
Low velocity	0.0023	0.0018	+0.07	0.0066	0.0036	+0.17
Main experiment	0.0061	0.0033	+0.07	0.0110	0.0071	+0.26
High velocity	0.0109	0.0045	+0.33	0.0183	0.0121	+0.04

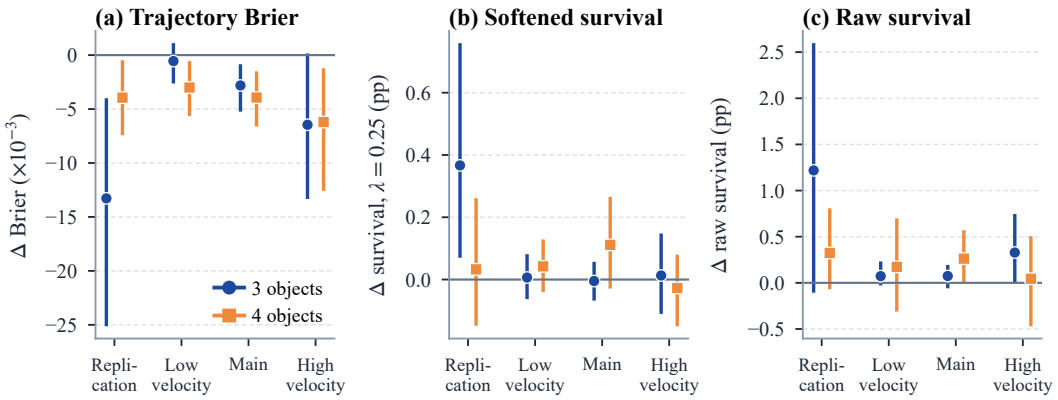

Figure 12: Prefix-minus-static effects with paired 95% scene-bootstrap intervals. The replication and velocity settings have 12 scenes per object count and the main experiment 48.

D.2 Calibration and per-scene structure

Figure 13 shows reliability curves and the tail of candidate-level errors. Both heads are close to the diagonal, and QSTAR mainly removes the large errors: the fraction of candidates with absolute error above 0.1 drops for both object counts. Figure 14 breaks the gain down by scene. Easy scenes have essentially no error for either head, and the improvement is concentrated in critical and coupled scenes. Figure 15 shows the candidate score tables of the 12 three-object scenes with the largest static error: QSTAR moves most uncertain candidates toward their ensemble survival, with a few cells in the opposite direction. Tables 15 and 16 give the stratified means and scene ranges.

D.3 Sparse physical queries

Table 8 lists the allocation rules on the main scenes, Table 9 the fresh-scene validation, and Figure 16 all four conditions. At 25% of the states the rule matches dense querying in every condition; at 10% it keeps at least 83% of the gain everywhere except the three-object replication scenes (68%). On

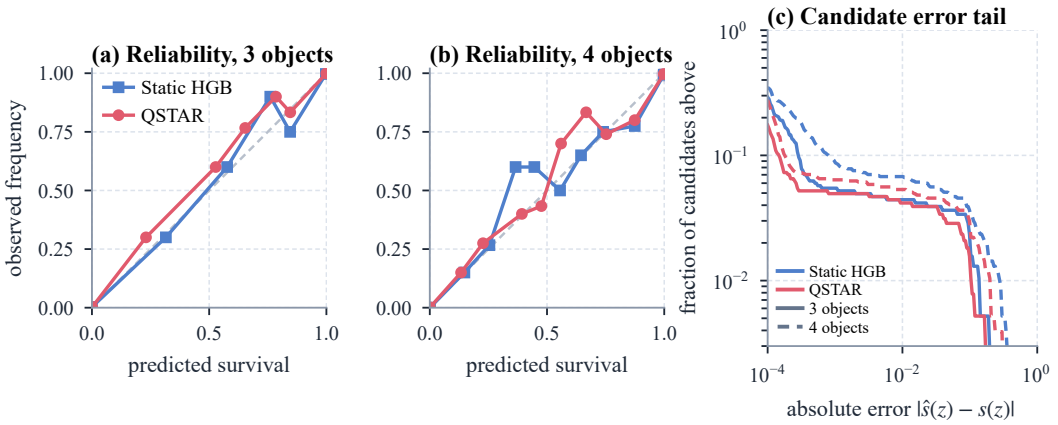

Figure 13: (a, b) Reliability of candidate survival probabilities on the main scenes; marker area grows with the number of candidates in a bin. (c) Fraction of candidates whose absolute error exceeds a threshold, for both heads and object counts.

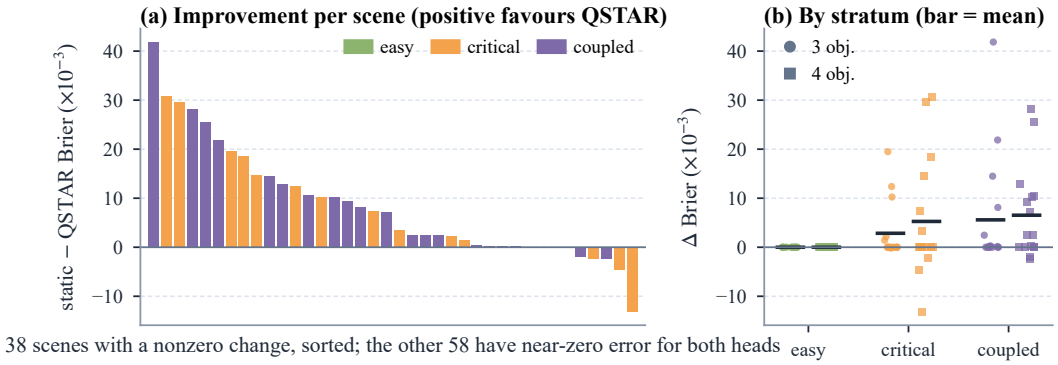

Figure 14: Per-scene Brier improvement on the 96 main scenes. (a) Scenes with a nonzero change, sorted and colored by stratum. (b) All scenes by stratum and object count; bars mark the mean.

the low-velocity three-object set, the 10% rule is even slightly better than dense querying, probably because it keeps the accurate static estimate on confident states.

Table 8: Allocation rules on the main scenes. Survival is a fraction; budgets are nominal fractions before rounding.

	Three objects		Four objects	
Allocation	Brier	Survival	Brier	Survival
Static	0.00609	0.98869	0.01103	0.96739
Uncertainty 10%	0.00354	0.98929	0.00736	0.96994
Uncertainty 25%	0.00328	0.98943	0.00712	0.96997
Uncertainty 50%	0.00328	0.98943	0.00711	0.97002
Learned value 10%	0.00380	0.98918	0.00839	0.96953
Dense prefix	0.00328	0.98943	0.00710	0.97002

D.4 Decision-utility stress test

To check whether the probability gain reaches a less saturated decision endpoint, we compute, from the saved candidate-level survival vectors, the expected survival under the raw selector distribution, the probability mass on candidates whose ensemble survival is at least 0.90 (safe mass), the regret

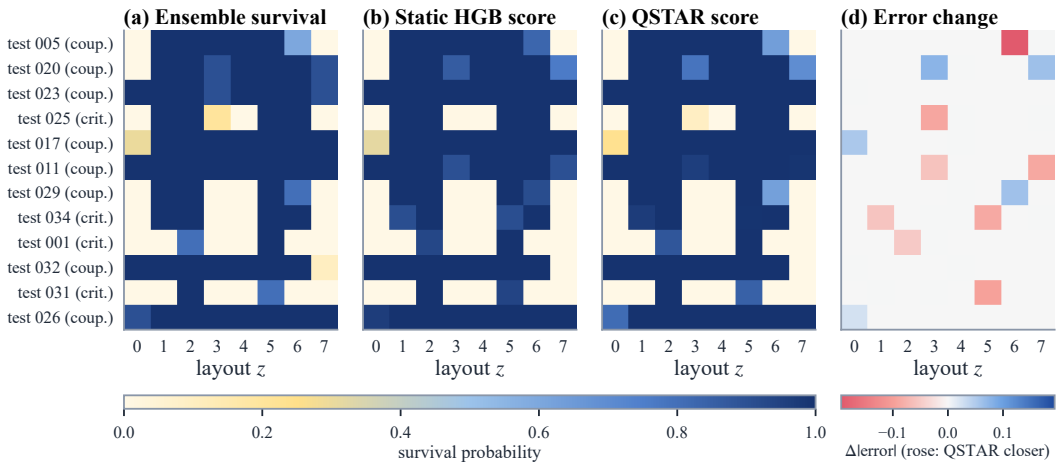

Figure 15: Candidate-level tables for the 12 three-object scenes with the largest static error: (a) survival frequency in the evaluation ensemble, (b) static score, (c) QSTAR score, and (d) change in absolute error, where rose cells mark candidates on which QSTAR is closer to the ensemble.

Table 9: Fresh 24-scene validation (seeds 101000/102000) after freezing the allocation rule. Probability errors decrease while raw survival is lower than static selection. TV is measured against the exact finite-ensemble posterior.

	Three objects			Four objects		
Method	Brier	TV	Survival	Brier	TV	Survival
Static	0.01422	0.02462	0.97539	0.01011	0.02364	0.97839
Dense prefix	0.00680	0.00675	0.96226	0.00655	0.01441	0.97225
Uncertainty 10%	0.00702	0.01002	0.96372	0.00660	0.01452	0.97215

against the best candidate in the same finite set, and the mean survival of the four highest-scored candidates. Thresholds and the top-4 cutoff were fixed before this analysis. On the main four-object scenes, QSTAR changes expected survival by $+0.262$ points and regret by -0.262 points, with intervals that include zero. On the replication scenes the changes are $+1.218$ and -1.218 points, and the safe-mass interval excludes zero. Top-choice statistics barely move, because both heads usually rank the same safe candidate first. The utility gain is therefore smaller and less stable than the probability gain, as the main text states.

Table 10: Decision-utility stress test, in percent. Safe mass is the selector probability on candidates with ensemble survival at least 0.90; regret is the best-candidate survival minus the expected selector survival.

Setting / selector	Expected survival	Safe mass	Regret	Top-4 survival
Main, 4 obj. / Static	96.739	92.159	2.844	89.167
Main, 4 obj. / QSTAR	97.002	92.554	2.582	89.167
Replication, 3 obj. / Static	87.914	84.741	4.586	85.833
Replication, 3 obj. / QSTAR	89.132	86.256	3.368	85.833

D.5 Transfer to five and six objects

The frozen three-object heads are evaluated on six five-object and six six-object scenes (seeds 111000/112000, 12 perturbations per candidate, 6,912 additional trajectories); only the simulator's storage padding changes. For five objects, static/prefix Brier is 0.01200/0.00560, raw survival 93.17%/93.75%, and sharpened survival 97.58%/99.01%. For six objects, Brier is 0.02106/0.01836,

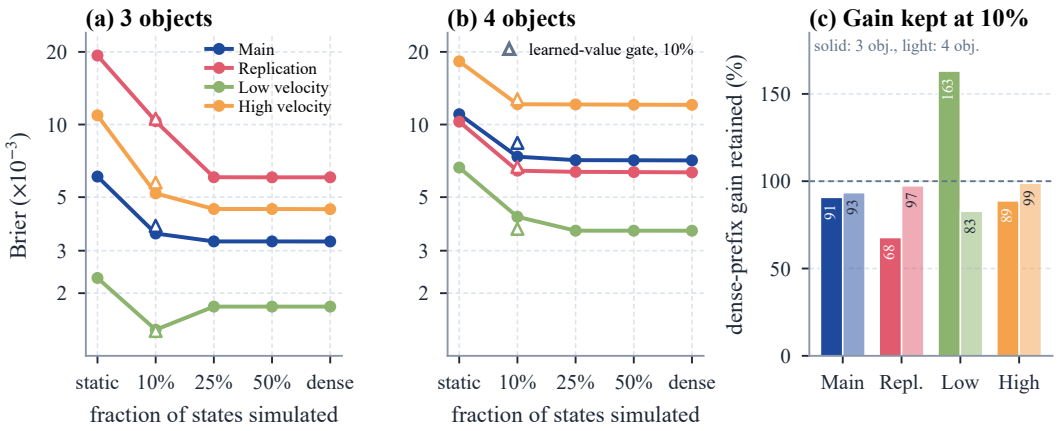

Figure 16: Brier error against the fraction of states simulated, for the main, replication, and velocity-shift conditions (a, b); triangles mark the learned-value gate at 10%. (c) Share of the dense-prefix improvement kept by the 10% uncertainty rule.

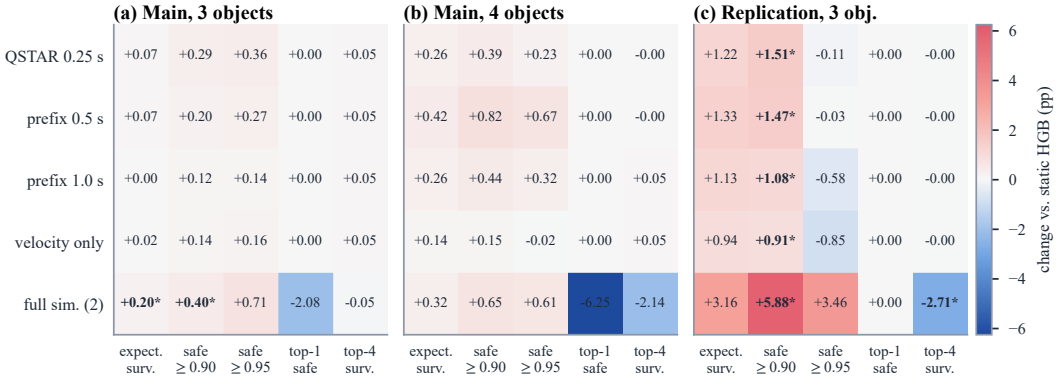

Figure 17: Paired changes against static HGB for several selectors and decision statistics (percentage points; rose favours the selector). Starred bold cells have 95% scene-bootstrap intervals that exclude zero.

raw survival 94.01%/93.04%, and sharpened survival 98.82%/98.80%. Both heads reach 100% MAP survival on these small sets. The risk improvement persists but shrinks at six objects, where raw selection reverses (Figure 18); with six scenes per count these results are exploratory. Tables 17 and 18 list the strata and every scene.

D.6 Learned rollouts versus short prefixes

To test the alternative of learning the future state instead of observing it, we trained an ExtraTrees transition model on 0.25 s snapshots of 48 training scenes (180 trees, minimum leaf size two) and evaluated it on 12 held-out three-object and 12 four-object scenes. Recursive rollouts drift quickly: the three-object state RMSE grows from 0.04 to 0.57 in five steps, ending no better than holding the state fixed, and the four-object error starts at 0.31 (Figure 3a). Terminal-risk heads that use one or two observed future states instead reduce Brier error from 0.021 to 0.006 and 0.004 for three objects (Figure 3b). Figure 19 repeats these panels with the one-step error of each state channel: velocity and orientation errors grow fastest, while one-step position error stays small. Observed short responses are informative; learned long rollouts are not reliable enough to replace them in this setting.

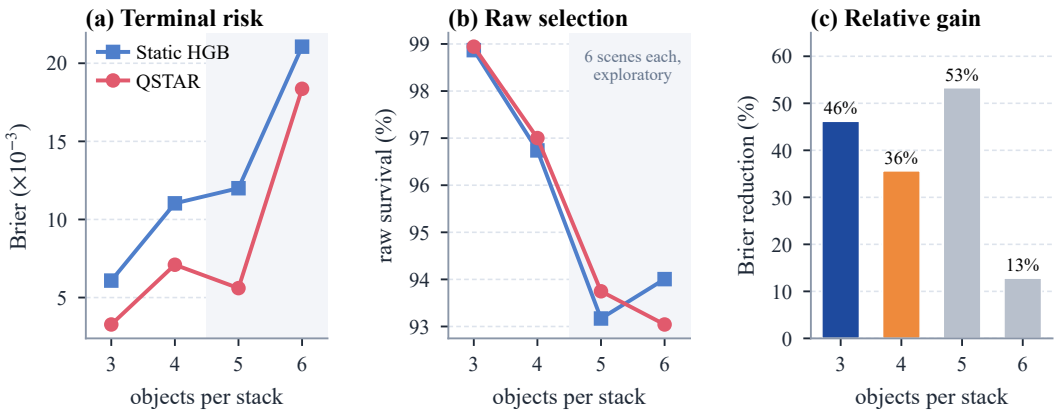

Figure 18: Transfer of the frozen heads across object counts. Three and four objects use the 48-scene main sets; five and six objects (shaded) use six scenes each. (a) Brier error. (b) Raw survival. (c) Relative Brier reduction.

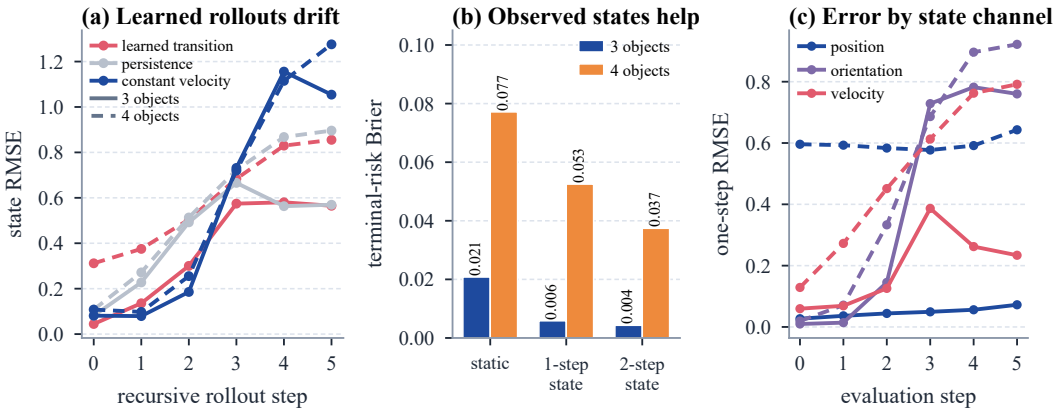

Figure 19: Scene-held-out world-model probe. (a) State RMSE of recursive rollouts from a learned transition model, persistence, and constant velocity; solid lines are three-object and dashed lines four-object scenes. (b) Terminal-risk Brier with static features or with one or two observed future states. (c) One-step error by state channel.

D.7 Label-permutation control

We shuffled the terminal training labels globally across all training rows, 100 times, keeping the features, test labels, fixed hyperparameters (7 and 15 leaves, no development selection), and evaluation scenes unchanged, with the prefix event gate disabled. We report prefix-minus-static Brier, so positive values favour the static head. Under permutation the three-object difference has mean $+0.0086$ (standard deviation 0.0086, 2.5–97.5% range $[-0.0069, 0.0244]$) and the four-object difference mean $+0.0150$ (0.0134, $[-0.0096, 0.0427]$). The positive mean shows that the richer prefix representation overfits permuted labels more than the static one, so the null is not centered at zero. The real-label differences (-0.0028 and -0.0039) lie in the lower 10% and 6% of the permutation distributions (Figure 20). Three smaller controls with 20 seeds each (global permutations with leaves $(7, 7)$ or $(15, 15)$, and within-scene permutations with $(7, 15)$) give mean differences between -0.0025 and $+0.0007$, and all their ranges cross zero. Because the permuted heads are much worse in absolute terms, this control does not isolate the prefix effect on its own; we read the main result together with the frozen splits, the strong static baseline, and the independent replications.

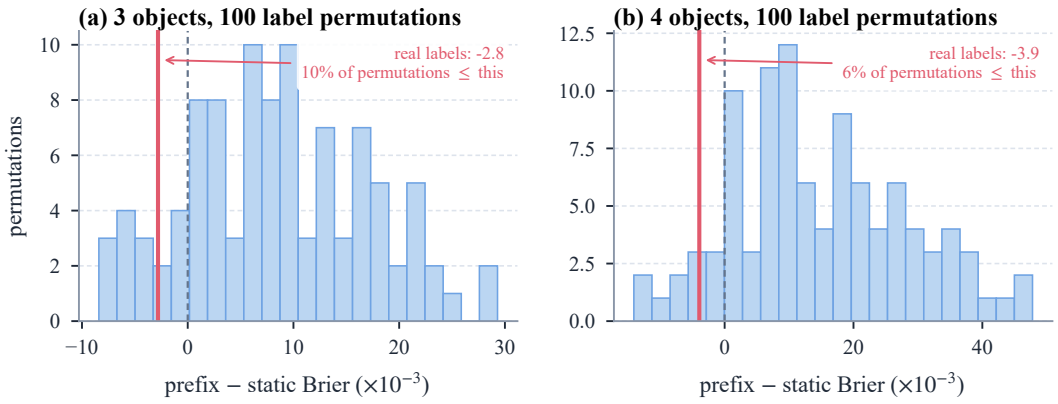

Figure 20: Global label-permutation control: prefix-minus-static Brier over 100 permutations, with the real-label difference marked.

D.8 Earlier alternatives

Several earlier designs did not beat strong classical models consistently. In a continuous-control pilot with its own information contract (12 test and 12 OOD scenes, not pooled with the main experiment), a neural rollout scorer had clearly higher probability error than invariant boosting (Tables 11 and 12), and its survival depended on the scoring convention. Trajectory supervision, contact-aware representations, mild moving-table interventions, and convex fusion of static and rollout scores gave no robust gain either. An earlier learned dynamic Bayesian network selector reached 98.53% raw survival but had candidate-distribution Brier 0.026543, against 0.010035 for a direct risk model and 0.005697 for a four-query simulator control; its high survival came from a concentrated selector, not a better probability model. Visual counterfactual-response experiments on new videos did not show a reliable dynamics-specific benefit either. These results led us to use actual short physical responses rather than larger learned rollout models, and to rank candidate methods by Brier error and log loss rather than raw survival.

Table 11: Continuous-control pilot, point estimates. Brier and BCE are lower-is-better; survival columns are percentages, where Power is the development-selected sharpened survival.

Method	Split	Brier	BCE	Raw (%)	Power (%)	MAP (%)
Invariant boost	Test	0.00481	0.01353	98.01	99.76	100.00
Rollout	Test	0.02977	0.14606	97.50	98.42	99.44
Invariant boost	OOD	0.00993	0.03159	93.92	99.37	100.00
Rollout	OOD	0.05069	0.21846	93.98	97.64	100.00

Table 12: Continuous-control pilot, paired scene-bootstrap intervals (rollout minus invariant boosting; 10,000 stratified resamples after averaging three neural seeds within scene). Positive Brier or BCE favours invariant boosting; survival differences are in percentage points.

Split	Metric	Difference	95% interval
Test	Brier	+0.02496	[0.01027, 0.04060]
Test	BCE	+0.13253	[0.05188, 0.22407]
Test	Raw survival (pp)	-0.508	[-1.884, 0.912]
Test	Power survival (pp)	-1.345	[-3.447, 0.340]
Test	MAP survival (pp)	-0.556	[-1.667, 0.000]
OOD	Brier	+0.04075	[0.02632, 0.05550]
OOD	BCE	+0.18686	[0.09999, 0.27487]
OOD	Raw survival (pp)	+0.061	[-3.292, 3.327]
OOD	Power survival (pp)	-1.734	[-3.387, -0.291]
OOD	MAP survival (pp)	0.000	[0.000, 0.000]

E PROCEDURAL DYNAMICAL SYSTEMS

The procedural experiment asks whether the benefit comes from the measured short response rather than from contact-specific features. Each family has its own simulator and terminal event, and the protocol is fixed: 64 training, 16 development, and 32 test scenes; eight candidate settings per scene; 12 perturbations per candidate; and a 1.5 s horizon with a 0.25 s prefix. Static and QSTAR heads are fitted separately within each family on identical splits, with the leaf count chosen from $\{7, 15, 31\}$ by development log loss, and the test set is used once.

Events. All systems use $\Delta t = 0.01$ and $T = 1.5\,\mathrm{s}$. *Collision avoidance*: a disk moves as $p_{t+\Delta t} = p_t + \Delta t\, v$ past a static disk of radius r at the origin, and survives if $\|p_t\|_2 > r$ for all $t \le T$. *Ballistic containment*: $x_t = x_0 + v_x t$ and $y_t = y_0 + v_y t - \frac{1}{2} g t^2$, with survival if $y_t > 0$ and $|x_t| < b$ throughout. *Frictional stopping*: $v_{t+\Delta t} = \max\{0, v_t - \mu g \Delta t\}$ and $x_{t+\Delta t} = x_t + \frac{1}{2}(v_t + v_{t+\Delta t})\Delta t$, with survival if $x_t < b$ for all t. *Damped pendulum*: $\omega_{t+\Delta t} = \omega_t + \Delta t(-\frac{g}{\ell}\sin\phi_t - c\,\omega_t)$ and $\phi_{t+\Delta t} = \phi_t + \Delta t\,\omega_{t+\Delta t}$, with survival if $|\phi_t| < \phi_{\max}$ throughout. The prefix gate uses the first 25 steps, and a prefix that already leaves the admissible region gets score zero, exactly as in the contact experiment.

Protocol and results. Each family and method has 3,072 test trajectories with complete states at 151 time points. A fixed seed draws one candidate per test scene from the QSTAR posterior for the galleries, once, with no retries or success filtering. Intervals resample the 32 scene-level Brier differences 10,000 times. Brier error decreases in all four families (Table 13, Figure 21); the intervals are wholly below zero for collision and pendulum, where the prefix exposes an imminent boundary crossing or a growing angular excursion, and include zero for ballistic and friction, whose outcome is largely fixed by the initial state. Calibration curves and per-scene differences are shown in Figures 4 and 22, and all 128 conditional samples in Figures 23–26. These are simulator renderings, not diffusion outputs.

Table 13: Procedural systems. Brier and posterior TV are lower-is-better; survival is the expected terminal survival under the posterior. The interval is for the paired scene-level Brier difference, QSTAR minus static.

Family	Method	Brier	Posterior TV	Survival (%)	Paired Brier interval
Collision	Static	0.0407	0.0554	85.30	–
Collision	QSTAR	0.0332	0.0463	85.21	[-0.0164, -0.0007]
Ballistic	Static	0.0440	0.0951	75.48	–
Ballistic	QSTAR	0.0402	0.0851	75.60	[-0.0125, 0.0044]
Friction	Static	0.0401	0.1114	68.33	–
Friction	QSTAR	0.0321	0.0943	69.05	[-0.0185, 0.0012]
Pendulum	Static	0.0432	0.0753	67.53	–
Pendulum	QSTAR	0.0326	0.0644	68.80	[-0.0206, -0.0016]

F IMAGE ENDPOINT

Table 14 separates intended-layout survival, preservation, and the joint proxy for every selector on the shared pool. Figure 27 breaks preservation down by stratum and shows how pass and fail cases separate on the two geometric criteria. Critical scenes, whose large offsets push the upper objects far from the pedestal, account for most failures (88.7% pass against 98.8% and 99.2% for easy and coupled scenes). Five of the 34 failures have an undetected object, including blank outputs; the other 29 fail the overlap or center-error criteria because an object is displaced or largely occluded. Appearance changes that keep the measured geometry do not count as failures. A visually preserved image can still correspond to a physically failing candidate, which is why the joint proxy multiplies preservation by simulated survival.

Image atlas. Figures 29–32 show 128 images from the existing pool: for 32 evenly spaced scene IDs, the four preservation-passing outputs with the highest mean object IoU. The atlas is therefore a

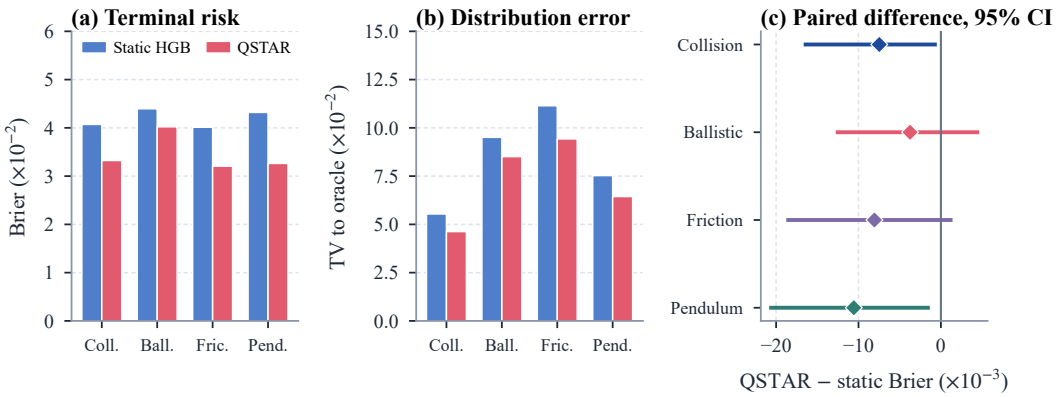

Figure 21: Procedural systems. (a) Terminal-risk Brier. (b) TV to the oracle posterior. (c) Paired Brier difference with 95% scene-bootstrap intervals.

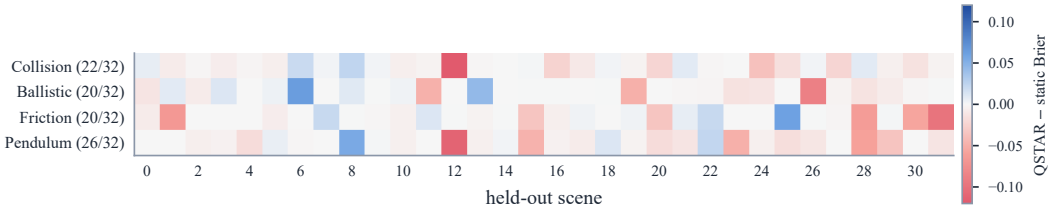

Figure 22: Scene-level Brier differences for all 128 held-out procedural scenes; rose cells favour QSTAR. Row labels give the number of scenes where QSTAR is better.

curated display of successful realizations, not an unbiased sample; the full-pool rate of 734/768 and the failures in Figure 28a give the unfiltered picture. No new images were generated for the atlas.

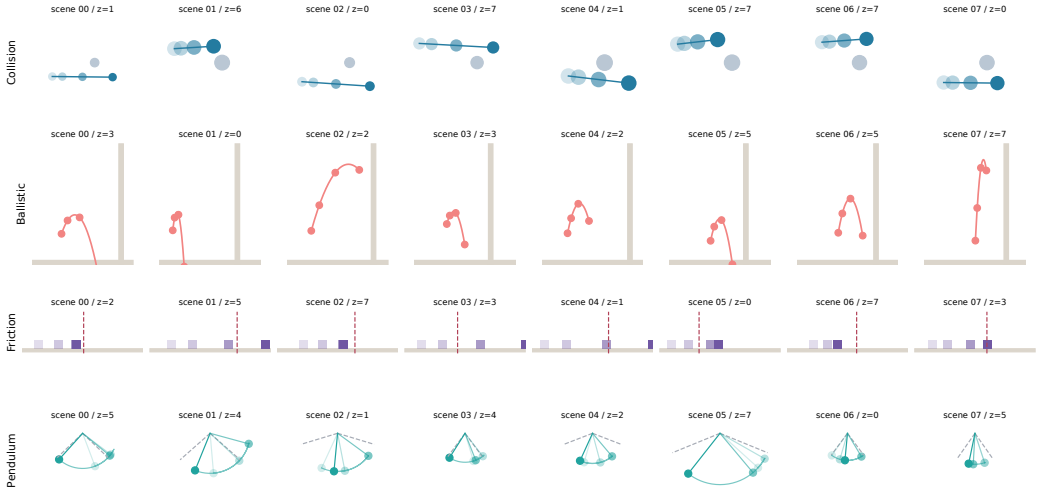

Figure 23: Conditional samples from the procedural systems, page 1 of 4. Rows are collision, ballistic, friction, and pendulum; columns are fixed-seed test scenes.

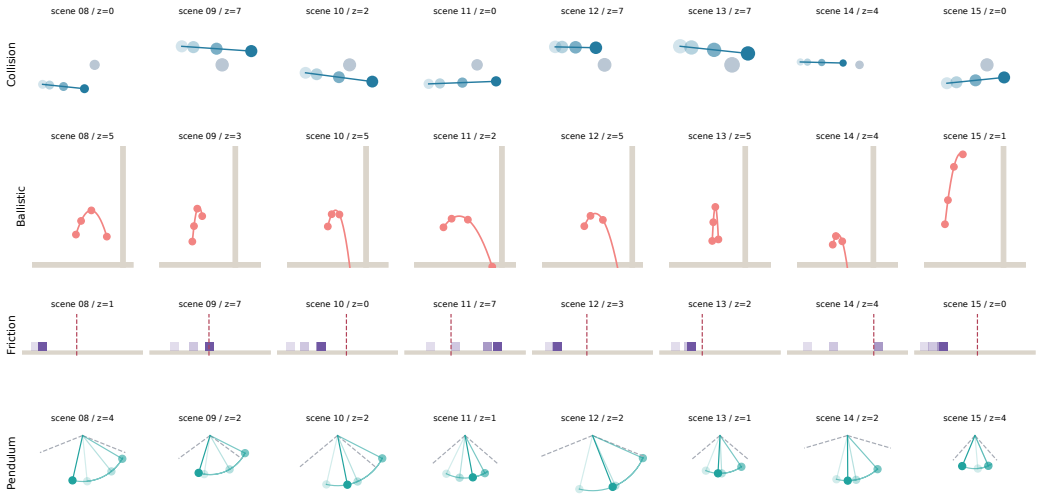

Figure 24: Conditional samples from the procedural systems, page 2 of 4.

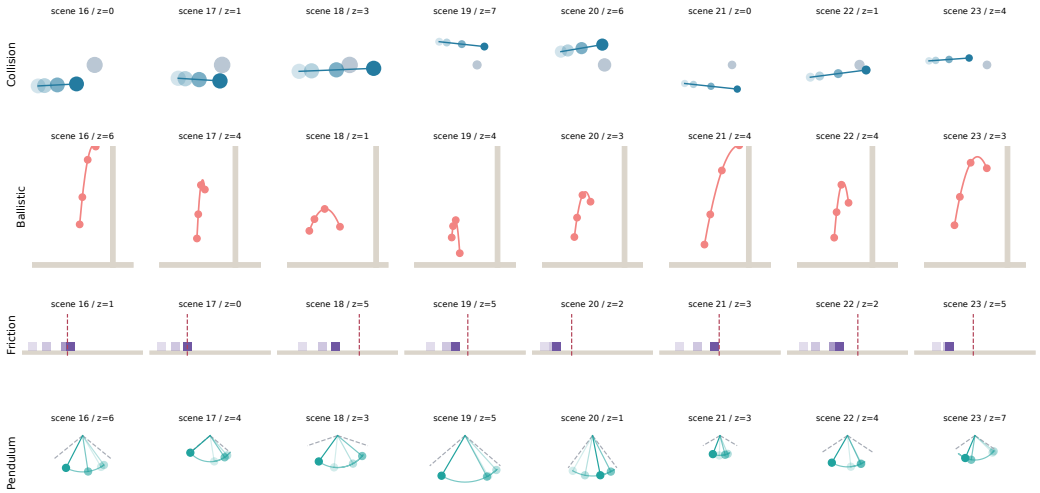

Figure 25: Conditional samples from the procedural systems, page 3 of 4.

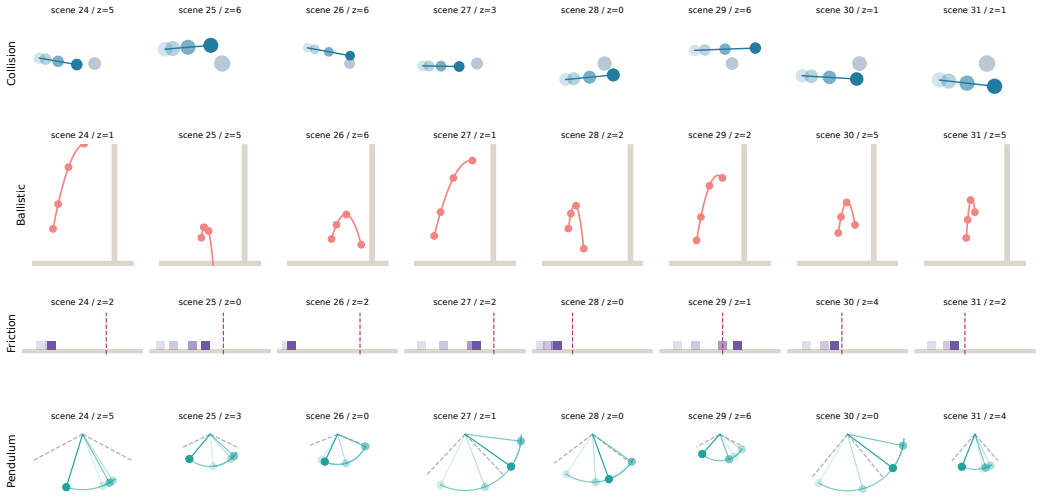

Figure 26: Conditional samples from the procedural systems, page 4 of 4.

Table 14: Shared image pool, in percent. Joint is intended-layout survival times visible preservation; Sharp. uses the development-selected sharpened distribution.

Selector	Layout	Preservation	Joint	Sharp.
Uniform	74.40	95.57	72.24	72.24
Static COM rule	98.63	95.69	94.49	94.49
Static HGB	98.87	95.70	94.72	95.28
Linear response rule	99.09	95.62	94.87	94.87
QSTAR	98.94	95.68	94.77	95.38
Full simulator (2 draws)	99.07	95.61	94.82	94.82

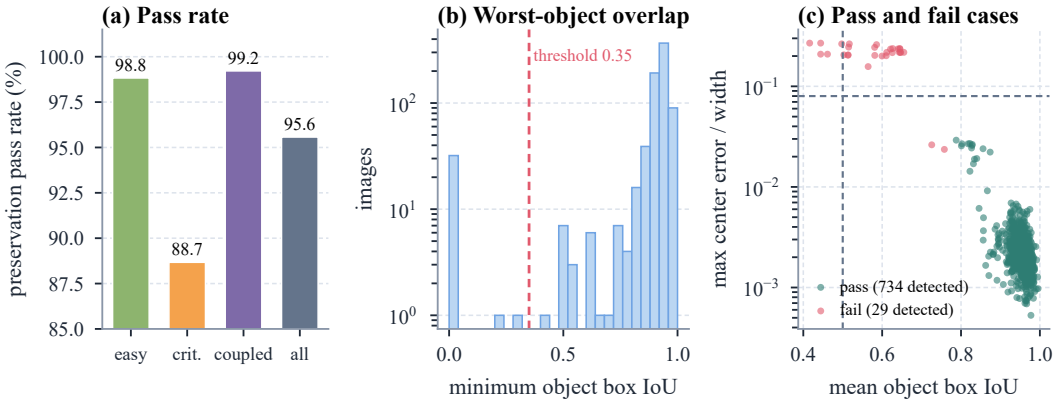

Figure 27: Preservation on the 768-image pool. (a) Pass rate by stratum. (b) Minimum object box IoU per image (log scale); the leftmost bin holds images with an undetected object. (c) Mean IoU against maximum center error for images where every object is detected; dashed lines are the thresholds.

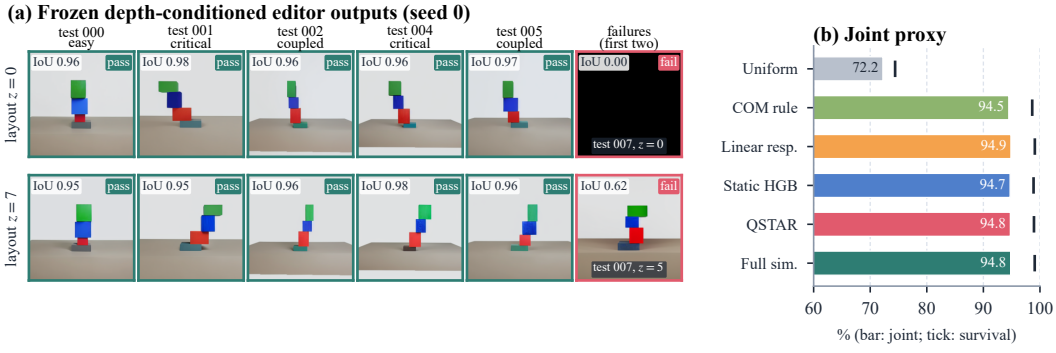

Figure 28: Frozen-editor outputs for layouts $z = 0$ and $z = 7$ of five held-out scenes and the first two preservation failures in sorted order; badges give mean object box IoU and the preservation outcome. (b) Joint proxy for each selector, as in Figure 6d.

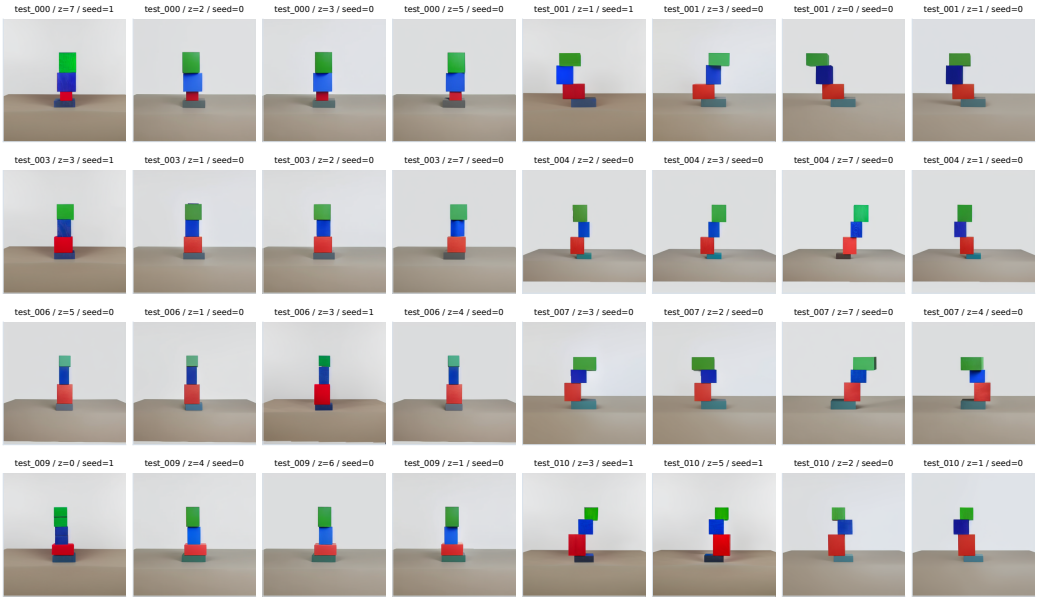

Figure 29: Diffusion-editor outputs, atlas page 1 of 4. Labels give the held-out scene, candidate layout, and editor seed.

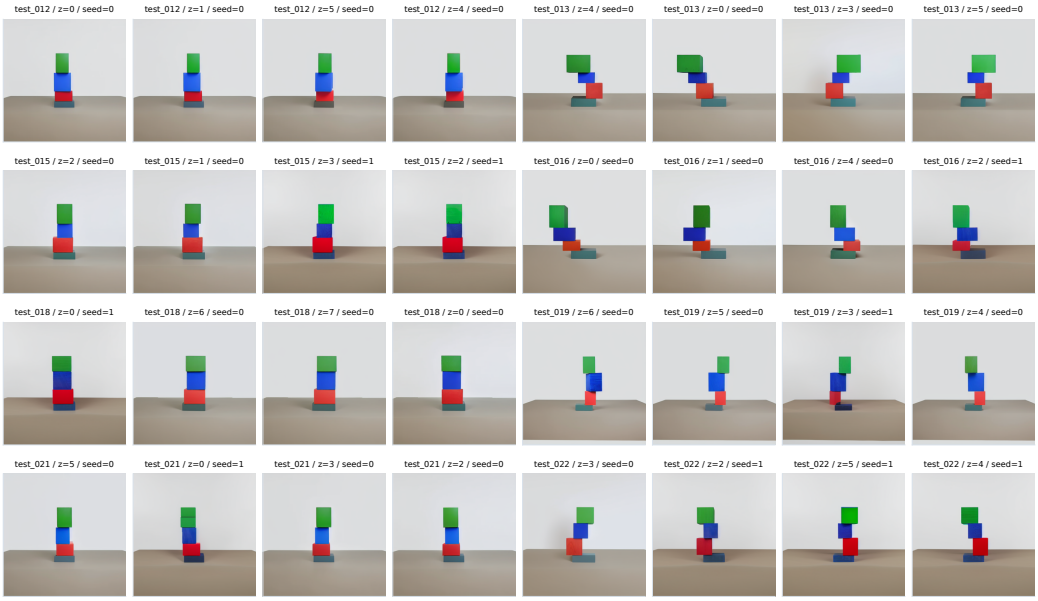

Figure 30: Diffusion-editor outputs, atlas page 2 of 4.

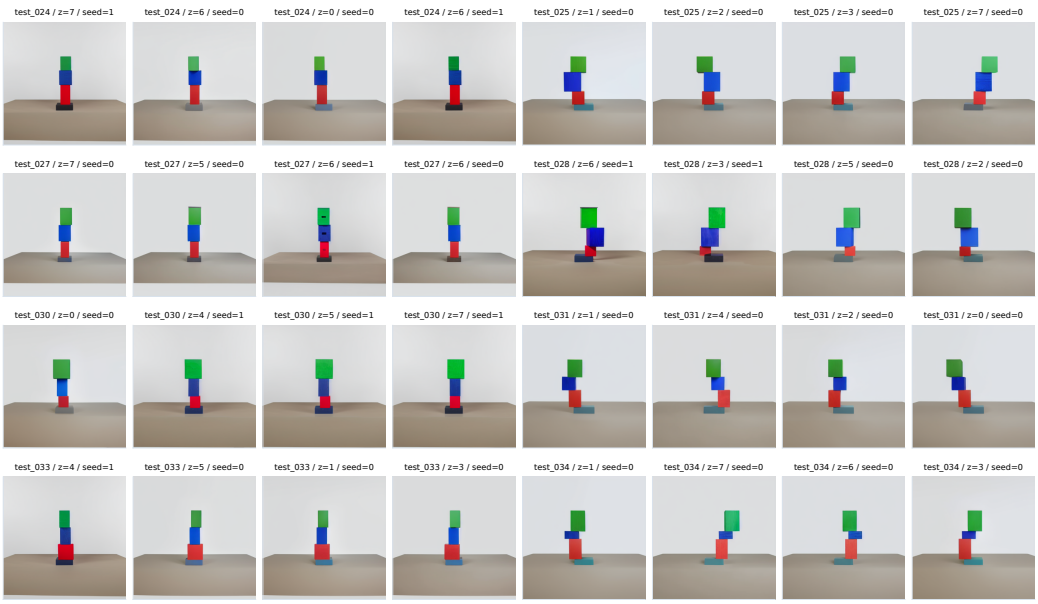

Figure 31: Diffusion-editor outputs, atlas page 3 of 4.

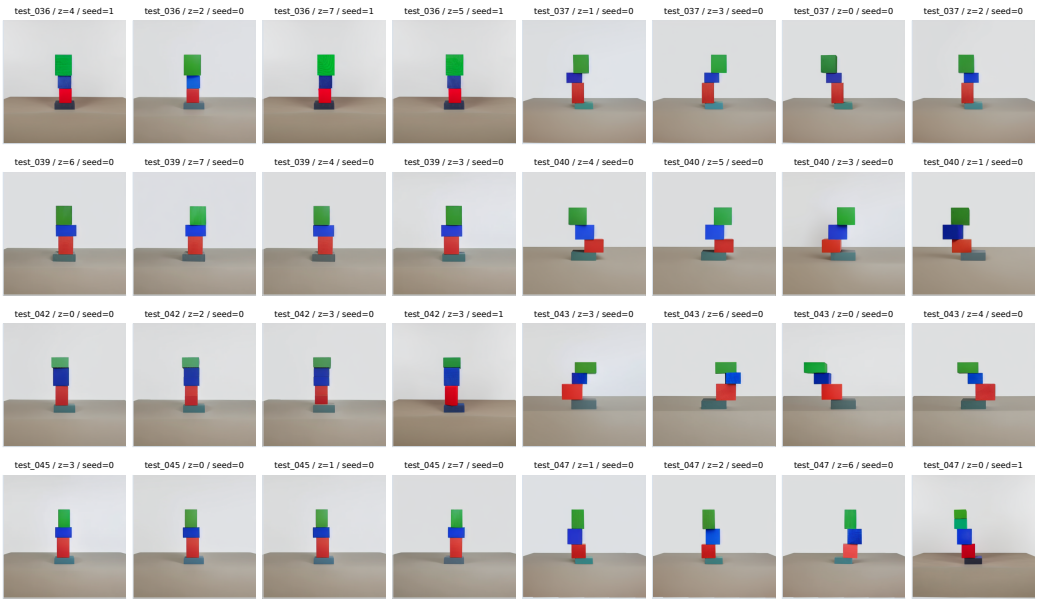

Figure 32: Diffusion-editor outputs, atlas page 4 of 4.

G PER-SCENE, PER-CIRCUIT, AND PER-PERMUTATION RECORDS

The tables below are written directly from the saved result files. In them, TEST denotes the three-object split and OOD the four-object split of the main experiment; in the transfer tables they denote the five- and six-object scenes. A negative Brier delta means that QSTAR has lower error, and a positive survival delta means higher raw survival. Each scene row aggregates the twelve perturbations of every candidate, so the rows keep the dependence structure used by the scene bootstrap.

Data and evaluation contract

Artifact	Role	Scope
qstar_prefix_main/results.json	main risk experiment	48 test + 48 OOD scenes; 3/4 objects
qstar_prefix_more_objects/results.json	frozen transfer	5/6-object exploratory test and OOD
qstar_prefix_budget/results.json	query allocation	main, confirmation, low/high shifts
qstar_prefix_negative_control_global_fixed_7_15_100/results.json	null control	100 global label permutations
qstar_prefix_quantum.json	quantum interface	48 circuits; two softening values
qstar_prefix_images*/	image endpoint	768 PNG outputs and preservation evaluation

Main TEST scene audit (1–16)

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
test-000	easy	0.000000	0.000000	-0.000000	+0.000
test-001	critical	0.014199	0.012780	-0.001419	+0.291
test-002	coupled	0.000000	0.000000	-0.000000	+0.000
test-003	easy	0.000000	0.000000	-0.000000	+0.000
test-004	critical	0.000000	0.000000	-0.000000	+0.008
test-005	coupled	0.049677	0.007829	-0.041848	+1.194
test-006	easy	0.000000	0.000000	+0.000000	+0.000
test-007	critical	0.000041	0.000225	+0.000184	+0.005
test-008	coupled	0.000000	0.000000	-0.000000	+0.001
test-009	easy	0.000000	0.000000	+0.000000	+0.000
test-010	critical	0.000000	0.000000	-0.000000	+0.001
test-011	coupled	0.023960	0.002091	-0.021869	+0.000
test-012	easy	0.000000	0.000000	-0.000000	+0.000
test-013	critical	0.000000	0.000000	+0.000000	+0.003
test-014	coupled	0.000000	0.000000	-0.000000	+0.000
test-015	easy	0.000000	0.000000	-0.000000	+0.000

Main TEST scene audit (17–32)

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
test-016	critical	0.000000	0.000000	-0.000000	+0.027
test-017	coupled	0.024108	0.021653	-0.002456	+0.786
test-018	easy	0.000000	0.000000	-0.000000	+0.000
test-019	critical	0.000002	0.000006	+0.000005	+0.012
test-020	coupled	0.046288	0.031799	-0.014488	+0.165
test-021	easy	0.000000	0.000000	-0.000000	+0.000
test-022	critical	0.000005	0.000000	-0.000005	+0.073
test-023	coupled	0.024997	0.024850	-0.000146	+0.001
test-024	easy	0.000000	0.000000	-0.000000	+0.000
test-025	critical	0.024497	0.012114	-0.012383	-1.766
test-026	coupled	0.007927	0.007734	-0.000193	+0.170
test-027	easy	0.000000	0.000000	-0.000000	+0.000
test-028	critical	0.000000	0.000000	-0.000000	-0.002
test-029	coupled	0.022956	0.014841	-0.008115	+1.193
test-030	easy	0.000000	0.000000	+0.000000	+0.000
test-031	critical	0.012469	0.002224	-0.010245	+0.526

Main TEST scene audit (33–48)

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
test-032	coupled	0.012472	0.012215	-0.000257	-0.010
test-033	easy	0.000000	0.000000	-0.000000	+0.000
test-034	critical	0.020858	0.001373	-0.019486	+0.004
test-035	coupled	0.000000	0.000000	-0.000000	+0.000
test-036	easy	0.000000	0.000000	-0.000000	+0.000
test-037	critical	0.000000	0.000000	-0.000000	-0.001
test-038	coupled	0.000000	0.000000	-0.000000	+0.000
test-039	easy	0.000000	0.000000	+0.000000	+0.000
test-040	critical	0.000000	0.000000	-0.000000	+0.016
test-041	coupled	0.000000	0.000000	+0.000000	-0.001
test-042	easy	0.000000	0.000000	-0.000000	+0.000
test-043	critical	0.007662	0.005490	-0.002172	+0.837
test-044	coupled	0.000000	0.000000	-0.000000	-0.000
test-045	easy	0.000000	0.000000	-0.000000	+0.000
test-046	critical	0.000000	0.000000	-0.000000	+0.011
test-047	coupled	0.000000	0.000000	-0.000000	+0.003

Main OOD scene audit (1–16)

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
ood-000	easy	0.000000	0.000000	-0.000000	+0.000
ood-001	critical	0.000000	0.000000	-0.000000	+0.095
ood-002	coupled	0.073031	0.044939	-0.028093	+0.644
ood-003	easy	0.000000	0.000000	+0.000000	+0.000
ood-004	critical	0.064350	0.033619	-0.030730	+2.183
ood-005	coupled	0.005156	0.002747	-0.002409	+0.553
ood-006	easy	0.000000	0.000000	-0.000000	+0.000
ood-007	critical	0.000000	0.000000	-0.000000	+0.059
ood-008	coupled	0.009411	0.000119	-0.009293	-0.087
ood-009	easy	0.000000	0.000000	-0.000000	+0.000
ood-010	critical	0.005231	0.001859	-0.003372	-0.808
ood-011	coupled	0.006574	0.006250	-0.000324	-0.003
ood-012	easy	0.000000	0.000000	-0.000000	+0.000
ood-013	critical	0.000000	0.000000	-0.000000	-0.013
ood-014	coupled	0.000000	0.000000	-0.000000	-0.000
ood-015	easy	0.000000	0.000000	-0.000000	+0.000

Main OOD scene audit (17–32)

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
ood-016	critical	0.000000	0.000000	-0.000000	+0.230
ood-017	coupled	0.007217	0.009170	+0.001953	-0.448
ood-018	easy	0.000000	0.000000	-0.000000	+0.000
ood-019	critical	0.014622	0.000021	-0.014601	-0.637
ood-020	coupled	0.025087	0.027449	+0.002361	+0.003
ood-021	easy	0.000000	0.000000	+0.000000	+0.000
ood-022	critical	0.000000	0.000000	-0.000000	+0.181
ood-023	coupled	0.000000	0.000000	+0.000000	+0.002
ood-024	easy	0.000000	0.000000	+0.000000	+0.000
ood-025	critical	0.035940	0.006296	-0.029644	-1.254
ood-026	coupled	0.045872	0.033072	-0.012801	+0.344
ood-027	easy	0.000000	0.000000	-0.000000	+0.000
ood-028	critical	0.021902	0.003412	-0.018490	+0.892
ood-029	coupled	0.018579	0.011440	-0.007139	+3.656
ood-030	easy	0.000000	0.000000	+0.000000	+0.000
ood-031	critical	0.000000	0.000000	-0.000000	+0.023

Main OOD scene audit (33–48)

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
ood-032	coupled	0.021181	0.018677	-0.002504	+0.118
ood-033	easy	0.000000	0.000000	+0.000000	+0.000
ood-034	critical	0.000075	0.004637	+0.004562	+0.097
ood-035	coupled	0.000000	0.000000	-0.000000	+0.013
ood-036	easy	0.000000	0.000000	+0.000000	+0.000
ood-037	critical	0.006227	0.006204	-0.000023	+0.006
ood-038	coupled	0.000000	0.000000	+0.000000	+0.001
ood-039	easy	0.000000	0.000000	-0.000000	+0.000
ood-040	critical	0.003188	0.016343	+0.013154	+0.021
ood-041	coupled	0.018455	0.008268	-0.010187	+0.640
ood-042	easy	0.000000	0.000000	-0.000000	+0.000
ood-043	critical	0.029239	0.021927	-0.007312	+4.817
ood-044	coupled	0.010590	0.000040	-0.010550	+1.251
ood-045	easy	0.000000	0.000000	-0.000000	+0.000
ood-046	critical	0.016087	0.018346	+0.002259	+0.142
ood-047	coupled	0.091503	0.065987	-0.025516	-0.134

Frozen transfer TEST scene audit

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
test-000	easy	0.000000	0.000000	+0.000000	+0.000
test-001	critical	0.011599	0.002919	-0.008680	+0.916
test-002	coupled	0.021303	0.007142	-0.014161	+0.609
test-003	easy	0.000000	0.000000	-0.000000	+0.000
test-004	critical	0.017464	0.002391	-0.015074	+1.881
test-005	coupled	0.021631	0.021163	-0.000468	+0.062

Frozen transfer OOD scene audit

Scene	Kind	Static Brier	QSTAR Brier	Δ Brier	Raw Δ (pp)
ood-000	easy	0.000000	0.000000	+0.000000	-0.000
ood-001	critical	0.021512	0.014664	-0.006848	-3.944
ood-002	coupled	0.034708	0.035952	+0.001243	-1.126
ood-003	easy	0.000000	0.000000	+0.000000	-0.000
ood-004	critical	0.030912	0.025356	-0.005556	-2.148
ood-005	coupled	0.039242	0.034216	-0.005025	+1.438

Permutation TEST seed audit (1–50)

Seed	Δ Brier	Δ BCE	Raw Δ (pp)
0	+0.006877	+0.038783	+0.143
1	+0.011213	+0.038179	-0.004
2	+0.021815	+0.066292	-0.348
3	+0.007627	+0.042297	+0.195
4	+0.023714	+0.085913	-0.179
5	+0.010171	+0.055221	+0.418
6	+0.002577	+0.023193	+0.188
7	+0.014736	+0.042570	-0.432
8	+0.005495	+0.020930	-0.071
9	+0.004378	+0.031730	+0.135
10	-0.002171	+0.001366	+0.395
11	+0.016357	+0.071909	+0.074
12	+0.012881	+0.054802	-0.211
13	+0.006096	+0.027682	+0.060
14	+0.005475	+0.058700	+0.432
15	-0.005849	-0.004876	+0.626
16	+0.001689	+0.010458	+0.252
17	+0.010676	+0.038120	+0.187
18	+0.006249	+0.023922	+0.576
19	+0.023554	+0.100052	-0.649
20	+0.021443	+0.080029	-0.051
21	+0.002103	+0.012432	+0.146
22	+0.016825	+0.057509	-0.324
23	+0.014039	+0.056893	+0.118
24	+0.009932	+0.044052	-0.100
25	+0.012603	+0.057838	+0.051
26	-0.005476	-0.007613	+0.105
27	+0.001604	+0.010540	+0.648
28	-0.006928	-0.006908	+0.498
29	+0.002649	+0.039176	+0.070
30	+0.014390	+0.052859	+0.191
31	+0.009388	+0.041767	+0.096
32	+0.005913	+0.031950	+0.120
33	+0.006512	+0.031733	+0.129
34	+0.013143	+0.048094	-0.372
35	+0.009523	+0.033461	-0.013
36	-0.005187	+0.012155	+0.656
37	+0.028376	+0.104558	-0.532
38	+0.004993	+0.024624	-0.157
39	-0.001462	+0.002224	+0.265
40	+0.017567	+0.064949	-0.002
41	+0.016756	+0.053363	-0.293
42	+0.001728	+0.022993	+0.364
43	-0.001424	+0.009181	+0.222
44	+0.003405	+0.017915	-0.012
45	+0.012855	+0.081042	+0.152
46	+0.007125	+0.033541	-0.326
47	+0.007862	+0.046131	-0.125
48	-0.004470	-0.000582	+0.532
49	+0.007090	+0.032946	+0.174

Permutation TEST seed audit (51–100)

Seed	Δ Brier	Δ BCE	Raw Δ (pp)
50	+0.005926	+0.032064	+0.278
51	+0.021839	+0.082804	+0.097
52	+0.017901	+0.071318	-0.295
53	+0.010411	+0.036009	-0.121
54	+0.001063	+0.028039	+0.159
55	-0.007523	-0.002823	+0.474
56	-0.003682	-0.017366	+0.344
57	+0.009628	+0.047305	-0.046
58	+0.006375	+0.037933	+0.140
59	-0.000524	+0.013932	+0.296
60	+0.007851	+0.034987	+0.003
61	+0.018774	+0.076866	-0.215
62	+0.009854	+0.059651	+0.497
63	-0.001625	+0.007041	+0.428
64	+0.007636	+0.042592	+0.117
65	+0.019390	+0.081626	-0.208
66	+0.009777	+0.036404	-0.125
67	+0.001282	+0.015461	+0.238
68	+0.003417	+0.029010	+0.174
69	-0.003739	+0.009132	-0.145
70	+0.002700	+0.024840	+0.135
71	-0.000504	+0.011415	+0.417
72	+0.011547	+0.040344	+0.343
73	+0.007472	+0.034238	+0.032
74	+0.000793	+0.025654	+0.363
75	+0.003535	+0.020999	+0.282
76	+0.016113	+0.063184	-0.188
77	+0.021401	+0.088358	-0.029
78	+0.010443	+0.048953	+0.180
79	+0.008917	+0.054024	-0.223
80	+0.017243	+0.064752	-0.126
81	-0.005728	-0.001143	+0.289
82	+0.000451	+0.022241	+0.170
83	+0.003174	+0.029760	+0.108
84	+0.013287	+0.047162	-0.061
85	+0.008433	+0.032873	-0.067
86	+0.017690	+0.082436	-0.333
87	+0.016792	+0.084783	-0.325
88	+0.006080	+0.026849	+0.062
89	+0.022077	+0.079341	-0.476
90	+0.029301	+0.102320	-0.518
91	+0.001496	+0.017836	+0.040
92	+0.019338	+0.093862	-0.283
93	+0.024368	+0.073491	-0.440
94	+0.004868	+0.033761	-0.215
95	+0.013864	+0.044248	+0.221
96	-0.008403	-0.007135	+0.518
97	+0.018868	+0.090442	+0.030
98	+0.012503	+0.054289	+0.147
99	+0.017074	+0.081747	-0.185

Permutation OOD seed audit (1–50)

Seed	Δ Brier	Δ BCE	Raw Δ (pp)
0	+0.026289	+0.098350	-0.117
1	+0.030060	+0.105713	-0.137
2	+0.018385	+0.070837	-0.324
3	+0.008809	+0.046740	+0.053
4	+0.025329	+0.114927	-0.138
5	+0.019147	+0.089657	+0.055
6	-0.001002	+0.024033	+0.292
7	-0.003446	-0.003103	-0.045
8	+0.012412	+0.041852	-0.132
9	+0.010633	+0.052463	-0.046
10	+0.007995	+0.037402	+0.006
11	+0.005887	+0.050467	+0.342
12	+0.008391	+0.054283	+0.594
13	+0.005485	+0.046937	+0.355
14	+0.000712	+0.046198	+0.400
15	+0.010987	+0.046676	+0.123
16	+0.008325	+0.041139	+0.354
17	+0.016246	+0.067630	+0.191
18	+0.004734	+0.027605	+0.391
19	+0.042686	+0.186440	-0.511
20	+0.041456	+0.164633	-0.077
21	+0.008055	+0.049124	+0.408
22	+0.016993	+0.056142	+0.038
23	+0.005645	+0.035520	+0.250
24	+0.011626	+0.056054	-0.027
25	+0.021370	+0.095947	+0.410
26	-0.006150	+0.005291	+0.587
27	+0.031677	+0.089361	-0.610
28	-0.007817	+0.000314	+0.400
29	+0.006454	+0.057422	+0.158
30	-0.009620	+0.012624	+0.822
31	+0.009883	+0.046198	-0.223
32	+0.034177	+0.127593	-0.567
33	+0.022690	+0.098089	-0.068
34	+0.027188	+0.091167	-0.760
35	+0.000551	+0.021532	+0.563
36	+0.010730	+0.056394	+0.093
37	+0.018528	+0.095610	-0.112
38	+0.020692	+0.100373	-0.298
39	+0.021319	+0.074733	-0.411
40	+0.033193	+0.118799	+0.040
41	+0.019432	+0.083478	+0.205
42	+0.016069	+0.073330	+0.149
43	+0.004277	+0.026845	+0.148
44	+0.012834	+0.054998	+0.327
45	+0.035616	+0.147614	-0.428
46	+0.009084	+0.046560	-0.456
47	+0.025651	+0.113156	-0.119
48	+0.013030	+0.062242	-0.216
49	+0.038119	+0.122867	-0.981

Permutation OOD seed audit (51–100)

Seed	Δ Brier	Δ BCE	Raw Δ (pp)
50	+0.008751	+0.060642	+0.128
51	+0.030295	+0.107202	-0.069
52	+0.037276	+0.121328	-0.428
53	+0.002423	+0.015254	+0.510
54	+0.023989	+0.133013	-0.214
55	+0.008322	+0.055084	+0.481
56	-0.011527	-0.031224	+0.873
57	+0.021695	+0.092961	+0.325
58	+0.000763	+0.028148	+0.330
59	+0.002653	+0.033661	+0.656
60	+0.000638	+0.018817	+0.737
61	+0.038859	+0.178527	-0.206
62	+0.029821	+0.108906	-0.190
63	+0.008216	+0.036342	+0.161
64	-0.000067	+0.016896	+0.524
65	+0.027122	+0.135787	-0.289
66	+0.002066	+0.020850	+0.395
67	+0.001001	+0.018771	+0.283
68	+0.008379	+0.036298	+0.317
69	+0.010867	+0.053170	-0.423
70	+0.020298	+0.087839	+0.267
71	+0.011426	+0.065849	+0.480
72	-0.004416	-0.000691	+0.459
73	+0.001462	+0.021352	+0.383
74	-0.000131	+0.027868	+0.577
75	+0.021981	+0.080104	-0.018
76	+0.019318	+0.083924	-0.200
77	+0.047786	+0.179742	-0.562
78	+0.009034	+0.056785	+0.206
79	+0.018304	+0.074863	-0.417
80	+0.008327	+0.056296	-0.288
81	-0.014091	-0.010942	+0.097
82	+0.002370	+0.035743	+0.269
83	+0.027515	+0.107642	-0.298
84	+0.028724	+0.097081	-0.351
85	+0.034028	+0.119444	-0.391
86	+0.016040	+0.084993	-0.223
87	+0.009851	+0.078990	+0.240
88	+0.032674	+0.100715	-0.922
89	+0.009866	+0.045389	+0.281
90	+0.023163	+0.095924	-0.408
91	+0.008648	+0.046165	+0.216
92	+0.018764	+0.087329	-0.208
93	+0.019583	+0.073456	+0.661
94	+0.011959	+0.066814	+0.215
95	+0.023712	+0.082434	+0.004
96	-0.003923	+0.005100	+0.759
97	+0.034037	+0.165441	-0.320
98	+0.015949	+0.084257	+0.073
99	+0.045581	+0.190326	-0.627

Quantum circuit audit (1–24)

Scene	λ	Qubits	FP length	Acceptance	Exact TV	4096-shot TV	CX
test-000	0.05	4	15	0.999936	3.57e-15	0.0205	2540
test-000	0.25	4	7	0.999987	1.88e-15	0.0205	1180
test-001	0.05	4	15	0.999894	9.15e-15	0.0127	2540
test-001	0.25	4	7	0.997342	4.54e-15	0.0146	1180
test-002	0.05	4	15	0.999907	2.17e-15	0.0203	2540
test-002	0.25	4	7	0.999982	3.58e-15	0.0203	1180
test-003	0.05	4	15	0.999936	4.26e-15	0.0205	2540
test-003	0.25	4	7	0.999987	3.39e-15	0.0205	1180
test-004	0.05	4	15	0.999771	8.59e-15	0.0127	2540
test-004	0.25	4	7	0.998174	4.15e-15	0.0194	1180
test-005	0.05	4	15	0.994257	2.01e-15	0.0180	2540
test-005	0.25	4	7	0.996419	3.31e-15	0.0241	1180
test-006	0.05	4	15	0.999936	4.26e-15	0.0205	2540
test-006	0.25	4	7	0.999987	3.39e-15	0.0205	1180
test-007	0.05	4	15	0.994380	1.45e-15	0.0197	2540
test-007	0.25	4	7	0.991741	4.77e-15	0.0213	1180
test-008	0.05	4	15	0.994126	1.96e-15	0.0151	2540
test-008	0.25	4	7	0.991659	4.21e-15	0.0174	1180
test-009	0.05	4	15	0.999936	4.26e-15	0.0205	2540
test-009	0.25	4	7	0.999987	3.39e-15	0.0205	1180
test-010	0.05	4	15	0.992952	4.03e-15	0.0137	2540
test-010	0.25	4	7	0.998719	4.19e-15	0.0131	1180
test-011	0.05	4	15	0.991492	5.24e-15	0.0204	2540
test-011	0.25	4	7	0.997517	2.66e-15	0.0204	1180

Quantum circuit audit (25–48)

Scene	λ	Qubits	FP length	Acceptance	Exact TV	4096-shot TV	CX
ood-000	0.05	5	15	0.999928	5.12e-15	0.0205	6068
ood-000	0.25	5	7	0.999986	2.35e-15	0.0205	2820
ood-001	0.05	5	15	0.993499	5.28e-15	0.0120	6068
ood-001	0.25	5	7	0.990588	6.59e-15	0.0276	2820
ood-002	0.05	5	15	0.999890	2.84e-15	0.0257	6068
ood-002	0.25	5	7	0.998451	1.14e-14	0.0256	2820
ood-003	0.05	5	15	0.999920	4.98e-15	0.0205	6068
ood-003	0.25	5	7	0.999984	2.15e-15	0.0205	2820
ood-004	0.05	5	15	0.996633	7.25e-15	0.0175	6068
ood-004	0.25	5	7	0.995438	8.41e-15	0.0239	2820
ood-005	0.05	5	15	0.999622	2.71e-15	0.0171	6068
ood-005	0.25	5	7	0.992056	9.28e-15	0.0214	2820
ood-006	0.05	5	15	0.999927	4.48e-15	0.0205	6068
ood-006	0.25	5	7	0.999986	1.71e-15	0.0205	2820
ood-007	0.05	5	15	0.993507	5.27e-15	0.0133	6068
ood-007	0.25	5	7	0.990587	6.36e-15	0.0279	2820
ood-008	0.05	5	15	0.991240	1.65e-15	0.0242	6068
ood-008	0.25	5	7	0.999702	1.01e-14	0.0254	2820
ood-009	0.05	5	15	0.999927	4.12e-15	0.0205	6068
ood-009	0.25	5	7	0.999986	2.72e-15	0.0205	2820
ood-010	0.05	5	15	0.990930	5.82e-15	0.0198	6068
ood-010	0.25	5	7	0.991239	5.81e-15	0.0279	2820
ood-011	0.05	5	15	0.990291	3.11e-15	0.0194	6068
ood-011	0.25	5	7	0.998441	1.12e-14	0.0249	2820

Table 15: Stratified scene-level means for the frozen heads on the main scenes. The largest reductions occur in the coupled and critical strata; easy scenes are near zero error for both heads.

Split	Regime	Scenes	Static Brier	QSTAR Brier	Δ Brier	Static BCE	QSTAR BCE	Static raw (%)	QSTAR raw (%)
TEST	easy	16	0.0000	0.0000	-0.0000	0.0000	0.0000	100.00	100.00
TEST	critical	16	0.0050	0.0021	-0.0028	0.0225	0.0096	97.83	97.83
TEST	coupled	16	0.0133	0.0077	-0.0056	0.0657	0.0344	98.78	98.99
OOD	easy	16	0.0000	0.0000	-0.0000	0.0000	0.0000	100.00	100.00
OOD	critical	16	0.0123	0.0070	-0.0053	0.0590	0.0332	93.44	93.81
OOD	coupled	16	0.0208	0.0143	-0.0065	0.1222	0.0667	96.78	97.19

Table 16: Mean and 10–90% scene ranges of Brier and log loss. The gains vary widely across scenes within a stratum.

Head	Split	Regime	Mean Brier	Brier 10–90%	Mean BCE	BCE 10–90%
static	TEST	easy	0.0000	0.0000–0.0000	0.0000	0.0000–0.0001
QSTAR	TEST	easy	0.0000	0.0000–0.0000	0.0000	0.0000–0.0000
static	TEST	critical	0.0050	0.0000–0.0175	0.0225	0.0001–0.0898
QSTAR	TEST	critical	0.0021	0.0000–0.0088	0.0096	0.0001–0.0328
static	TEST	coupled	0.0133	0.0000–0.0356	0.0657	0.0000–0.2020
QSTAR	TEST	coupled	0.0077	0.0000–0.0233	0.0344	0.0000–0.1158
static	OOD	easy	0.0000	0.0000–0.0000	0.0000	0.0000–0.0000
QSTAR	OOD	easy	0.0000	0.0000–0.0000	0.0000	0.0000–0.0000
static	OOD	critical	0.0123	0.0000–0.0326	0.0590	0.0001–0.1418
QSTAR	OOD	critical	0.0070	0.0000–0.0201	0.0332	0.0001–0.1036
static	OOD	coupled	0.0208	0.0000–0.0595	0.1222	0.0001–0.3988
QSTAR	OOD	coupled	0.0143	0.0000–0.0390	0.0667	0.0001–0.2068

Query allocation audit

Condition	Rule	Query fraction	Test Brier	OOD Brier
main	static	0.0000	0.006086	0.011032
main	uncertainty_0.1	0.1042	0.003535	0.007358
main	uncertainty_0.25	0.2500	0.003278	0.007118
main	uncertainty_0.5	0.5000	0.003278	0.007110
main	full_prefix	1.0000	0.003276	0.007100
confirmation	static	0.0000	0.019319	0.010277
confirmation	uncertainty_0.1	0.1042	0.010323	0.006438
confirmation	uncertainty_0.25	0.2500	0.006048	0.006365
confirmation	uncertainty_0.5	0.5000	0.006042	0.006352
confirmation	full_prefix	1.0000	0.006042	0.006335
low	static	0.0000	0.002313	0.006634
low	uncertainty_0.1	0.1042	0.001411	0.004146
low	uncertainty_0.25	0.2500	0.001760	0.003631
low	uncertainty_0.5	0.5000	0.001760	0.003631
low	full_prefix	1.0000	0.001760	0.003630
high	static	0.0000	0.010916	0.018269
high	uncertainty_0.1	0.1042	0.005184	0.012134
high	uncertainty_0.25	0.2500	0.004466	0.012111
high	uncertainty_0.5	0.5000	0.004466	0.012082
high	full_prefix	1.0000	0.004458	0.012065

Image endpoint audit

Subset/metric	Count	Pass	Rate/value	IoU/interval	Interpretation
All saved outputs	768	734	95.57	0.930	preservation evaluator
Test/OOD contact sheet	representative subset	display only	not an estimator	visual audit	same editor/camera
QSTAR joint proxy	all selectors	shared image pool	94.771	–	scene bootstrap
Static HGB joint proxy	all selectors	shared image pool	94.725	–	scene bootstrap
QSTAR minus static	paired scenes	same PNG pool	+0.046 pp	[-0.066,+0.146]	not resolved

H STRATIFIED SUMMARIES

These summaries are computed from the stored scene-level records and group them by stratum, object count, run, or rule. They are descriptive aggregates, not additional replications.

Table 17: Transfer of the frozen heads to five and six objects, by stratum. Only the simulator padding changes; the six-object raw-survival reversal comes from the critical scenes.

Split	Objects	Regime	Static Brier	QSTAR Brier	Δ Brier	Static raw (%)	QSTAR raw (%)
TEST	5	easy	0.0000	0.0000	-0.0000	100.00	100.00
TEST	5	critical	0.0145	0.0027	-0.0119	83.48	84.88
TEST	5	coupled	0.0215	0.0142	-0.0073	96.03	96.36
OOD	6	easy	0.0000	0.0000	0.0000	100.00	100.00
OOD	6	critical	0.0262	0.0200	-0.0062	86.73	83.68
OOD	6	coupled	0.0370	0.0351	-0.0019	95.29	95.44

Table 18: Every five- and six-object scene. With six scenes per count these sets are exploratory.

Scene	Split	Objects	Regime	Static Brier	QSTAR Brier	Δ Brier	Static raw (%)	QSTAR raw (%)
ood_000	OOD	6	easy	0.0000	0.0000	0.0000	100.00	100.00
ood_001	OOD	6	critical	0.0215	0.0147	-0.0068	83.14	79.20
ood_002	OOD	6	coupled	0.0347	0.0360	0.0012	98.50	97.38
ood_003	OOD	6	easy	0.0000	0.0000	0.0000	100.00	100.00
ood_004	OOD	6	critical	0.0309	0.0254	-0.0056	90.32	88.17
ood_005	OOD	6	coupled	0.0392	0.0342	-0.0050	92.07	93.51
test_000	TEST	5	easy	0.0000	0.0000	0.0000	100.00	100.00
test_001	TEST	5	critical	0.0116	0.0029	-0.0087	76.94	77.86
test_002	TEST	5	coupled	0.0213	0.0071	-0.0142	95.31	95.92
test_003	TEST	5	easy	0.0000	0.0000	-0.0000	100.00	100.00
test_004	TEST	5	critical	0.0175	0.0024	-0.0151	90.02	91.90
test_005	TEST	5	coupled	0.0216	0.0212	-0.0005	96.75	96.81

Table 19: Allocation rules on the three-object split of each condition. Query fractions are measured from the saved allocation records; the uncertainty rule was chosen on development data.

Condition	Rule	Query frac.	Brier	BCE	Raw (%)	MAP (%)
main	static	0.0000	0.0061	0.0294	98.87	–
main	uncertainty_0.1	0.1042	0.0035	0.0202	98.93	–
main	uncertainty_0.25	0.2500	0.0033	0.0156	98.94	–
main	uncertainty_0.5	0.5000	0.0033	0.0156	98.94	–
main	full_prefix	1.0000	0.0033	0.0147	98.94	–
confirmation	static	0.0000	0.0193	0.0636	87.91	–
confirmation	uncertainty_0.1	0.1042	0.0103	0.0388	89.01	–
confirmation	uncertainty_0.25	0.2500	0.0060	0.0258	89.13	–
confirmation	uncertainty_0.5	0.5000	0.0060	0.0213	89.13	–
confirmation	full_prefix	1.0000	0.0060	0.0213	89.13	–
low	static	0.0000	0.0023	0.0070	99.02	–
low	uncertainty_0.1	0.1042	0.0014	0.0046	99.07	–
low	uncertainty_0.25	0.2500	0.0018	0.0055	99.10	–
low	uncertainty_0.5	0.5000	0.0018	0.0055	99.10	–
low	full_prefix	1.0000	0.0018	0.0055	99.10	–
high	static	0.0000	0.0109	0.0569	97.53	–
high	uncertainty_0.1	0.1042	0.0052	0.0334	97.78	–
high	uncertainty_0.25	0.2500	0.0045	0.0260	97.86	–
high	uncertainty_0.5	0.5000	0.0045	0.0260	97.86	–
high	full_prefix	1.0000	0.0045	0.0215	97.86	–

Table 20: Allocation rules on the four-object split of each condition.

Condition	Rule	Query frac.	Brier	BCE	Raw (%)	MAP (%)
main	static	0.0000	0.0110	0.0604	96.74	–
main	uncertainty_0.1	0.1042	0.0074	0.0473	96.99	–
main	uncertainty_0.25	0.2500	0.0071	0.0446	97.00	–
main	uncertainty_0.5	0.5000	0.0071	0.0399	97.00	–
main	full_prefix	1.0000	0.0071	0.0333	97.00	–
confirmation	static	0.0000	0.0103	0.0719	96.16	–
confirmation	uncertainty_0.1	0.1042	0.0064	0.0530	96.47	–
confirmation	uncertainty_0.25	0.2500	0.0064	0.0488	96.47	–
confirmation	uncertainty_0.5	0.5000	0.0064	0.0431	96.48	–
confirmation	full_prefix	1.0000	0.0063	0.0321	96.48	–
low	static	0.0000	0.0066	0.0259	96.93	–
low	uncertainty_0.1	0.1042	0.0041	0.0186	96.98	–
low	uncertainty_0.25	0.2500	0.0036	0.0156	97.09	–
low	uncertainty_0.5	0.5000	0.0036	0.0156	97.10	–
low	full_prefix	1.0000	0.0036	0.0136	97.10	–
high	static	0.0000	0.0183	0.1222	93.69	–
high	uncertainty_0.1	0.1042	0.0121	0.1002	93.72	–
high	uncertainty_0.25	0.2500	0.0121	0.0921	93.73	–
high	uncertainty_0.5	0.5000	0.0121	0.0770	93.73	–
high	full_prefix	1.0000	0.0121	0.0666	93.74	–

Table 21: Circuit audit means by split and softening value, from statevector simulation of the compiled circuits.

Split	λ	Circuits	Acceptance	Exact TV	4096-shot TV	Length	CX	Depth
OOD	0.05	12	0.99628	4.4e-15	0.0193	15.0	6068	8152
OOD	0.25	12	0.99637	6.5e-15	0.0239	7.0	2820	3788
TEST	0.05	12	0.99721	4.2e-15	0.0179	15.0	2540	3564
TEST	0.25	12	0.99763	3.6e-15	0.0194	7.0	1180	1656

Table 22: Preservation by generation run and stratum. Critical scenes account for most failures.

Subset	Run	Regime	Images	Detected	Preserved	Preserved rate (%)	Mean IoU	Mean center error
base-96	0	coupled	32	32	32	100.00	0.962	0.0016
base-96	0	critical	32	31	31	96.88	0.917	0.0333
base-96	0	easy	32	32	32	100.00	0.950	0.0017
base-96	1	coupled	32	32	32	100.00	0.952	0.0012
base-96	1	critical	32	32	29	90.62	0.896	0.0097
base-96	1	easy	32	32	32	100.00	0.943	0.0016
extra-144	0	coupled	48	48	48	100.00	0.957	0.0014
extra-144	0	critical	48	48	46	95.83	0.953	0.0030
extra-144	0	easy	48	48	48	100.00	0.950	0.0020
extra-144	1	coupled	48	48	48	100.00	0.944	0.0016
extra-144	1	critical	48	48	38	79.17	0.860	0.0176
extra-144	1	easy	48	47	47	97.92	0.925	0.0226
extra-144	2	coupled	48	47	47	97.92	0.943	0.0222
extra-144	2	critical	48	48	46	95.83	0.955	0.0029
extra-144	2	easy	48	46	46	95.83	0.908	0.0438
extra-144	3	coupled	48	48	47	97.92	0.942	0.0026
extra-144	3	critical	48	48	37	77.08	0.845	0.0197
extra-144	3	easy	48	48	48	100.00	0.946	0.0018

Table 23: Global label-permutation control, prefix minus static. Quantiles use linear interpolation; "Prefix better" counts permutations with a negative difference.

Split	Metric	Permutations	Mean	Std.	2.5%	97.5%	Prefix better
TEST	Brier	100	0.0086	0.0086	-0.0064	0.0241	16
TEST	BCE	100	0.0410	0.0282	-0.0070	0.0971	8
TEST	raw survival	100	0.0006	0.0028	-0.0050	0.0060	60
OOD	Brier	100	0.0150	0.0134	-0.0088	0.0421	11
OOD	BCE	100	0.0696	0.0455	-0.0020	0.1792	4
OOD	raw survival	100	0.0005	0.0039	-0.0070	0.0075	56

Table 24: Summary of the main effects. Brier changes are relative reductions; survival changes are in percentage points.

Slice	Metric	Static	QSTAR	Delta	Interpretation
3-object TEST	Brier	0.0061	0.0033	46.18%	risk prediction
4-object OOD	Brier	0.0110	0.0071	35.64%	risk prediction
3-object TEST	raw survival	98.87	98.94	0.07 pp	decision utility
4-object OOD	raw survival	96.74	97.00	0.26 pp	decision utility
5-object TEST	Brier	0.0120	0.0056	53.31%	frozen transfer
6-object OOD	raw survival	94.01	93.04	-0.96 pp	reversal
Quantum interface	acceptance	–	99.687	–	statevector
Quantum interface	shot TV	–	0.0201	–	finite-shot diagnostic