DECODING AND STEERING CONTACT STRUCTURE IN TRAJECTORY DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

When can contact structure read from a partially denoised trajectory support useful intervention? We study trajectory diffusion in three simulated multi-body systems and separate three quantities that are easily conflated: the error of a probe that reads intermediate states, how concentrated the final outcomes of a state are under repeated stochastic continuation, and the realized utility of acting on the readout. Repeated continuations from the same state show that outcomes concentrate during denoising and that, from the middle of the schedule onward, readout error rather than residual randomness dominates the probe's prediction error. The readout is nevertheless directionally useful: one-shot probe ascent transfers from a distance-based proxy target to an orientation-aware geometric diagnostic in all six fixed models tested, and selected multi-step guidance raises held-out joint target-and-quality success by 12.7–20.1 percentage points in the two three-dimensional systems and by 25–30 points in the two-dimensional system under a scene-consistent quality envelope. The same readout is not usable for selection: probability-gated early rejection establishes no gain in yield per computation, although a sixteen-branch empirical probability of the same event, if it were available at no cost, would. A short invariance argument shows what each use requires. Neither requires calibration, so the probe's overconfidence after optimization is beside the point; guidance needs the probe's gradient to align locally with the true success probability, whereas selection needs its scores to rank states like that probability, and the contact readout ranks poorly. The limitation is ranking, not calibration and not a structural limit of selection. We provide the branching, paired-intervention and held-out policy protocols that separate readout error, outcome concentration and intervention utility, with geometric correspondence, trajectory quality and computation reported explicitly.

1 INTRODUCTION

An intermediate state of a trajectory diffusion model can suggest which bodies will touch before the trajectory is finished. Such a signal invites two uses: change the trajectory so that a requested contact occurs, or reject an unpromising sample early to save computation. A rising contact-prediction accuracy establishes neither. It does not say whether the futures of that state already agree on the final structure, whether the probe's gradient points anywhere useful, or whether its probability is reliable enough to discard samples on. Our question is therefore operational: which uses of a contact readout does the readout actually support?

Three quantities have to be kept apart. *Readout error* is what a specified probe gets wrong about the conditional outcome of a state. *Outcome concentration* is how much the final structures obtained by continuing the same state with independent noise agree with each other. *Intervention utility* is how much a specified action changes target success, trajectory quality or computational yield. A probe can be wrong about a state whose futures are already concentrated; a gradient can point the right way while the score behind it overstates success; a decision rule that needs the probability can fail while a decision rule that needs only the direction succeeds.

We measure the three quantities on the same states in three simulated multi-body systems: two five-cube drops and a six-disc billiard. Repeated stochastic continuation, following the cloning principle used to study diffusion dynamics (Biroli et al., 2024), gives the conditional outcome distribution

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.

of a state and, with a finite-branch-corrected Brier decomposition, separates readout error from residual randomness. Paired perturbations with shared future noise give the actual change in target frequency caused by a probe direction. Independent selection and evaluation cohorts test multi-step guidance and probability-gated early rejection under explicit target, quality and computation criteria. A retrospective analysis replaces the probe's score by a recalibrated version and by an empirical probability from independent branches of the same state.

The answer is an asymmetry with a simple explanation. Neither use requires calibration: a strictly increasing recalibration of the score leaves the guidance direction and the family of gate decisions unchanged, and an isotonic one can only coarsen that family (Proposition 1). Guidance needs the probe's gradient to point where the true success probability rises; a gate needs the probe to rank states like that probability, by enough to repay the computation it discards. The contact readout has the first property but not the second. Its *direction* is usable: probe ascent raises the frequency of the requested event in every tested model, transfers to an orientation-aware geometric diagnostic in six fixed models, and selected multi-step guidance raises held-out joint target-and-quality success by 12.7–20.1 percentage points in the two three-dimensional scenes and by 25–30 points in the two-dimensional scene under a scene-consistent quality envelope. Its *ranking* is not: probability-gated early rejection establishes no gain in joint hits per computation, whereas a sixteen-branch empirical probability of the same event would if its own branch cost were not charged (yield ratios 1.08–1.26 in all eighteen model/time groups). The probe is also overconfident after optimization, by 24–46 points, but by the same argument this matters only for decisions that read absolute probabilities. Outcome concentration is real but explains neither finding: from roughly 500–600 remaining updates onward, readout error rather than residual randomness dominates the prediction error.

Contributions. (i) The direction-versus-ranking asymmetry and its explanation: an invariance argument with an exact break-even condition for gating shows that guidance consumes local alignment with the success probability, selection consumes ranking, and neither requires calibration. (ii) Evidence that contact readouts supply usable directions: local geometric transfer in six fixed models and held-out multi-step guidance gains in all three scenes, replicated on three denoisers per cube scene and extending to five-edge conjunctions. (iii) Evidence that the same readouts do not rank well enough to select: no prospectively selected gate establishes a benefit, and on saved states the probe clears the break-even condition by at most 0.09, whereas an uncharged empirical probability of the same event clears it by 0.12–0.33 in every model/time group. (iv) A measurement protocol that puts readout error, outcome concentration and intervention utility on the same states, with mothers as the inference unit and every attempt kept in the denominator. Branching and classifier guidance are established tools; the contribution is their joint, outcome-anchored use for contact structure, not a new guidance algorithm or a lock-in law. Methodologically, quality criteria must be set relative to each scene's real-data statistics; we report both envelopes and keep the frozen one primary.

Scope. Dense-time branching covers two scenes; local confirmation uses three denoiser seeds per cube scene; primary policies use one fixed model per scene, and the cube-scene policies are replicated on two more denoisers each (Appendix B.4). Targets are conjunctions of one or two edges of a distance-based ever-contact proxy, extended to five edges in the cube scenes on the primary models, not arbitrary complete graphs. Concentration is conditional on the stochastic sampler, and no quality check certifies physical validity.

2 Background and Related Work

Structure along denoising. Denoising diffusion models learn a reverse process from noisy observations (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021); trajectory and policy diffusion apply the construction to state or action sequences (Janner et al., 2022; Chi et al., 2023; Ajay et al., 2023). Studies of symmetry breaking and speciation examine when structural variables become organized along reverse time (Raya & Ambrogioni, 2023; Biroli et al., 2024; Sclocchi et al., 2025; Achilli et al., 2026). Biroli et al. (2024) already clone an intermediate trajectory, continue the copies with independent noise and measure the agreement of their final classes; we adopt that measurement rather than claim a new branching test, and we do not test a universal speciation criterion. Probing intermediate states with classifiers (Alain & Bengio, 2016; Belinkov, 2022) supplies the readout whose error we measure; its accuracy is conditional on the probe family and the label

object. A rise in probe accuracy along denoising is easily read as evidence that the probed structure has been committed; we treat such curves as readout measurements whose location depends on configuration (Section 4.7) and evaluate each use of the readout against actual outcomes.

Contact-aware trajectory diffusion. Classical contact-rich planning uses complementarity formulations or direct transcription over contact modes (Stewart, 2000; Anitescu, 2006; Posa et al., 2014; Mordatch et al., 2012). Contact-aware diffusion variants inject, predict or search contact modes as an explicit planning variable (Wang et al., 2025; Kumar et al., 2025; 2026; Huang et al., 2025; Liang et al., 2025) or compute the score from a dynamics model with contact (Pan et al., 2024). These methods supply contact from outside the model; we instead ask what a vanilla trajectory model, trained without contact supervision, already exposes to a readout and what that readout is good for. The distance-based proxy, the near-surface diagnostic and the simulator's force-threshold labels are kept as distinct objects.

Guidance and inference-time selection. Classifier guidance adds classifier gradients to the reverse update (Dhariwal & Nichol, 2021); training-free guidance generalizes it to differentiable objectives evaluated on intermediate or predicted-clean states (Chung et al., 2023; Yu et al., 2023; Bansal et al., 2024; Ye et al., 2024), and restricting guidance to an interval can improve sample quality (Kynkäänniemi et al., 2024). Our variance-scaled update is a standard member of this family; the sign-of-gradient variant (Goodfellow et al., 2015) is a control, not a proposal. Selection-based alternatives use an intermediate value or likelihood to resample or prune particles rather than to move them, from twisted sequential Monte Carlo (Wu et al., 2023) to value-based decoding and Feynman–Kac steering (Li et al., 2025; Singhal et al., 2025). The gate-and-respawn rule we evaluate is the simplest such selector; its failure, and the retrospective finding that an uncharged empirical probability would succeed where the probe does not, bear on what these methods require of their value functions, but we do not evaluate resampling schemes themselves.

3 Measurement and Experimental Design

3.1 Scenes and contact objects

The datasets contain bounded drops of five cubes, open cluster drops of five cubes, and a two-dimensional six-disc system. The two cube scenes have ten candidate pair edges and trajectories with 300 frames and 65 state coordinates; the disc scene has fifteen edges, 200 frames and 24 coordinates. We reuse trained trajectory diffusion models (one per scene in the primary tables; three denoiser seeds per cube scene where stated) and their matched pair probes.

Contact definitions. Let $a_{ij,f}(x)$ denote the center-distance contact proxy for bodies i, j at simulation frame f with a fixed scene-specific threshold. The ever-contact edge is $g_{ij}^{\mathrm{ever}}(x) = \max_f a_{ij,f}(x)$. A peak-contact graph selects instead the frame with the most contacts. These are different objects: the pair probe used throughout predicts ever-contact edges, whereas a separate class probe, used as a readout baseline and for exact-class guidance, predicts peak-contact classes. For cubes, an orientation-aware separating-axis near-surface label supplies a separate geometric diagnostic; it is not a force-contact oracle. The disc archive lacks the randomized radii, which limits retrospective physical contact reconstruction in 2D.

Diffusion time. We count remaining reverse updates r: at $r > 0$ the next denoiser index is $r - 1$; $r = 0$ is a clean endpoint. Probes receive r as their time input, following their training cache convention. All new branching uses stochastic ancestral DDPM continuation; deterministic DDIM controls do not estimate concentration.

3.2 Readout error and outcome concentration

From each fresh mother state we generate K independent stochastic branches. For a fixed edge, write s for its probe score and q for its probability under the continuation distribution. The conditional Brier error decomposes as

$$\mathbb{E}\big[(s - Y)^2 \mid x_r\big] = q(1 - q) + (s - q)^2, \tag{1}$$

continuation uncertainty plus readout error. With $\hat{q} = K^{-1}\sum_k Y_k$ the unbiased estimates are $K\hat{q}(1-\hat{q})/(K-1)$ and $(s-\hat{q})^2 - \hat{q}(1-\hat{q})/(K-1)$; noise-corrected estimates may be slightly negative and are not clipped. Edge components are averaged within a mother before mothers are resampled. We also report whole-graph agreement between distinct branches and different-mother baselines. Agreement is sensitive to graph size and label marginals and is not a fraction of committed states; all of these quantities are sampler-conditional.

3.3 Interventions and useful outcomes

Targets are assigned before their outcomes are observed. One-shot targets request presence ($y = 1$) or absence ($y = 0$) of an edge; the assigned-target score is $s_y = s$ or $1 - s$. One-shot experiments compare unmodified, probe log-probability ascent, descent and an isotropic random direction at matched normalized-coordinate RMS. Future noise is paired across conditions and independent across branches. The response is the actual change in target frequency, not a change in score.

Every attempted sample stays in the target and joint denominators; nonfinite endpoints earn no hit. Joint success requires the target event and a frozen empirical quality envelope calibrated on real TRAIN trajectories (99th percentiles with absolute floors, checked on VAL): a kinematic position–velocity residual, boundary and, for cubes, quaternion, penetration and extent limits. These checks do not certify physical validity. Because the kinematic floor is absolute (0.05 m/s) while scene speed scales differ, we also report a *scene-consistent secondary envelope* whose kinematic limit is the same multiple of the real TRAIN 99th percentile in every scene; it coincides with the frozen envelope in both cube scenes and relaxes only the disc scene's limit (Section 4.1). The frozen envelope remains primary; the secondary envelope is a post-hoc definition motivated by real-data statistics, not by generated outcomes.

3.4 What each use of a readout requires

Let $q(x)$ be the probability of a counted success under stochastic continuation from state x at remaining step r of T, and let $s(x) \in (0,1)$ be a differentiable single-edge target score. A *recalibration* is a differentiable map $f : (0,1) \to (0,1)$ with $f' > 0$, applied to s.

Proposition 1 (Calibration invariance and gate break-even)**.** *(a) $\nabla_x \log f(s(x)) = c(x)\,\nabla_x \log s(x)$ with $c(x) = f'(s(x))\,s(x)/f(s(x)) > 0$. Recalibration therefore leaves normalized one-shot ascent unchanged and changes variance-scaled guidance only by a positive, state-dependent step length. (b) For every threshold τ in the range of f, $\{x : f(s(x)) \geq \tau\} = \{x : s(x) \geq f^{-1}(\tau)\}$, and other thresholds keep every or no state, so a threshold gate has the same family of keep sets, and the same best attainable yield, before and after recalibration. A continuous non-decreasing map with flat segments, such as isotonic regression, can only coarsen this family. (c) Let a gate at step r keep a fraction $p > 0$ of independent proposals, charge each rejected proposal the $T - r$ updates already spent and nothing for scoring, and let $L = \mathbb{E}[q \mid \text{kept}]/\mathbb{E}[q]$, with $\mathbb{E}[q] > 0$, be the lift among kept states. The gate raises long-run successes per denoiser update over no gate if and only if*

$$L > 1 + \Big(1 - \frac{r}{T}\Big)\frac{1-p}{p}. \tag{2}$$

The proof is in Appendix A. Parts (a) and (b) mean that neither use requires calibration. A step ϵu changes the success probability by $\epsilon\langle \nabla q, u\rangle$ to first order, so guidance needs *local alignment* $\langle \nabla q, \nabla s\rangle > 0$ where it is applied, while a random direction has zero expected first-order effect; the paired ascent-minus-random contrasts estimate its mother-averaged effect at a finite step. A gate needs *ranking*: by (c) its yield depends on s only through which states it keeps, and the required lift grows as the gate moves later or rejects more. Calibration matters only for rules that read absolute values, such as a fixed stopping level or a coarse threshold grid set in advance.

3.5 Confirmation and held-out policies

The local confirmation uses 128 mothers and 32 branches at $r \in \{600, 400, 200\}$ with amplitude 0.05 for each of three denoiser seeds in each cube scene; its primary contrast is ascent minus random on the geometric target at $r = 600$, a three-comparison family per scene. Mothers are the resampling units.

	Ever-edge micro F1		Exact ever-graph match		Kinematic residual (m/s)			
Scene	Distance	SAT	Distance	SAT	TRAIN p99	Frozen limit	Ratio	Secondary limit
BOUNDED DROP	0.842	0.922	9.38%	34.96%	0.0109	0.050	4.59	0.050
CLUSTER DROP	0.829	0.915	17.97%	34.38%	0.0108	0.050	4.62	0.050
2D DISCS	–	–	–	–	0.0290	0.050	1.72	0.133

Table 1: Measurement object and quality criterion. Left: correspondence to stored simulator labels on real validation data, for the cube scenes only; SAT is the orientation-aware near-surface band, not a recovered force oracle, and the 2D archive lacks the radii it needs. Right: Ratio is the frozen limit over the TRAIN p99; the secondary envelope uses one common multiple of each scene's real TRAIN 99th percentile and changes nothing in the cube scenes; neither envelope certifies physical validity.

Policies are selected and evaluated on independent streams of one- and two-edge targets drawn from a supported catalog (256 endpoints, event rates 0.05–0.95). Selection uses 128 proposals per complexity over no action and variance-scaled classifier guidance $\lambda\sigma_r^2\nabla_{x_r}\log s_y(x_r)$ with $\lambda \in \{1, 10, 100\}$ from $r = 600$ or 200, choosing the largest joint target/quality success (no action wins ties); a conjunction's $\log s_y$ is the mean of marginal target log scores, not a joint probability. Evaluation uses 512 fresh paired proposals per complexity with matched random, matched sign and one-shot controls. Gate selection compares no gate with $r \in \{600, 400, 200\}$ and thresholds $\{0.25, 0.5, 0.75, 0.9\}$ by joint hits per expected sample NFE; prospective evaluation runs 32 paired budget blocks of 32,000 sample NFE with one fixed target per block, rejected prefixes charged and unused budget retained in the denominator. Table 3 in Appendix B lists the cohorts.

Retrospective gate-score sources. To test the ranking requirement directly, we reuse the confirmation cohorts: branch slots 0–15 estimate each mother's target probability and slots 16–31 supply outcomes. On identical states we compare the frozen probe score s_y, its cross-fitted isotonic recalibration and the first-half empirical probability, a sixteen-branch oracle whose own cost is not charged. Thresholds are chosen on first-half yields and evaluated on second-half joint hits per executed sample NFE; the bootstrap refits the map and reselects the threshold in every replicate.

Statistical scope. Intervals are percentile bootstraps over mothers with the presence/absence mixture fixed, over paired proposals within catalog-rate strata, or over paired budget blocks, with Bonferroni-adjusted quantiles within declared families: three per cube scene for confirmation, eight for the three-dimensional policies, four for the two-dimensional policies, and eighteen per score source for the exploratory retrospective contrasts. Secondary comparisons use pointwise intervals; Appendices B and C specify the post-hoc analyses.

4 RESULTS

4.1 THE MEASUREMENT OBJECT AND THE QUALITY CRITERION MATTER

On separate 512-trajectory validation cohorts, the orientation-aware cube diagnostic corresponds to stored force-threshold labels better than the center-distance proxy (Table 1), but exact ever-graph agreement stays near 34–35%, so the definitions are not interchangeable. Framewise, ever-contact and peak-contact objects also differ; the separate peak-class probe's clean-endpoint accuracies of 14.3%, 10.6% and 5.3% characterize that readout, not the ever-edge probe used here.

The quality envelope also has to be read relative to each scene. Its kinematic floor of 0.05 m/s sits at 4.6 times the real TRAIN 99th percentile of the residual in both cube scenes but at only 1.7 times in the disc scene, whose speed scale is 3.3–4.4 times larger (Table 1). Real and generated residuals differ by similar factors where measured (about 2.5 times in bounded drop and 2.8 times in 2D), so the frozen envelope rejected 95% of generated 2D trajectories for a scene-specific reason. The secondary envelope applies the smaller cube multiple to every scene: both cube limits stay frozen and the 2D limit becomes 0.133 m/s, under which 94–96% of unmodified 2D trajectories pass. The two envelopes are not equally strict: the 2D envelope has no penetration or overlap component because the disc radii are unavailable.

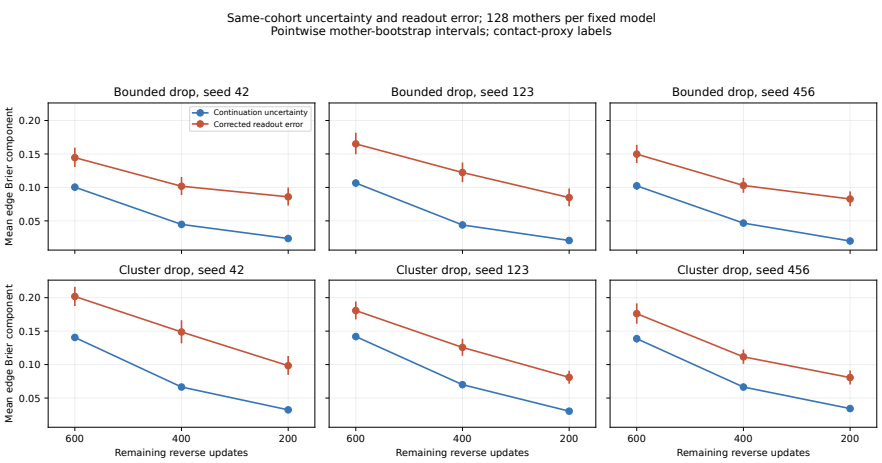

Figure 1: Readout error remains substantial as continuation uncertainty declines in six fixed models (128 mothers, 32 branches per state). Bars are pointwise mother-bootstrap intervals for finite-branch-corrected edge Brier components.

4.2 Concentrated outcomes coexist with readout error

The pair and peak-class probe readouts (Figure 14, Appendix F) show pair-F1 rising steeply between about 600 and 400 remaining updates in all scenes while peak-class accuracy stays low throughout. The branching pilot (64 mothers, 16 branches, bounded drop and 2D) measures what such curves cannot: at $r = 600$ the probability that two distinct branches of the same state share their entire proxy graph is 13.7% and 1.5%; at $r = 200$ it is 67.4% and 53.4%; at $r = 50$ it is 86.3% and 83.6%. Futures concentrate during denoising; residual randomness is the larger error term early ($r = 800$ in both scenes, $r = 600$ in 2D), but from the middle of the schedule onward the finite-branch-corrected readout term dominates, accounting for about 79% of the mean edge Brier error at $r = 200$ in both scenes. The six confirmation cohorts confirm the separation on fixed models (Figure 1): at $r = 200$ the readout component is 0.081–0.098 against an intrinsic component of 0.020–0.034. A large readout error and high branch agreement together do not establish an irreversible commitment; they say that the specified readout, not residual randomness, is the larger source of prediction error.

4.3 Probe directions improve outcomes and transfer to geometry

In the one-shot pilot, RMS-0.05 ascent at $r = 600$ raises single-edge proxy success from 53.4% to 66.0% in bounded drop and from 51.5% to 59.3% in 2D, with positive ascent-minus-random contrasts at all three times. The prespecified confirmation, 128 new mothers and 32 branches for each of three denoiser seeds per cube scene, finds the geometric ascent-minus-random contrast at $r = 600$ positive in all six models (Figure 2): gains of 12.1, 11.8 and 9.5 points in bounded drop and 14.5, 14.3 and 13.9 points in cluster drop, every within-scene family interval excluding zero. These are fixed-model effects with mother-level inference on a geometric diagnostic without a quality condition.

A usable direction is not a usable forecast. After ascent the mean probe probability is 99.2% in the bounded-drop pilot while observed success is 66.0%, and the confirmation cohorts show mean score-minus-success of 0.25–0.34 at $r = 600$ and up to 0.46 at $r = 200$: optimization moves states to where the probe is confident, not to where its confidence is earned.

4.4 Selected guidance improves held-out joint success

All four three-dimensional selection tasks choose variance-scaled guidance with $\lambda = 100$ from $r = 600$, the largest scale and earliest start in the grid. On fresh proposals every primary joint effect is positive after the eight-comparison adjustment (Table 2); gains also appear on the geometric diagnostic, and quality-pass rates change by at most 3.5 points with intervals that include zero except

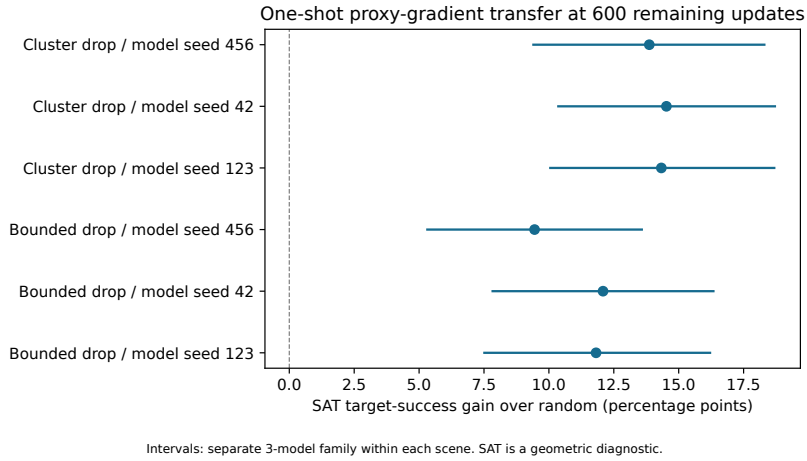

Figure 2: Geometric transfer confirmation: ascent versus matched random at $r = 600$, RMS 0.05, 128 mothers and 32 paired branches per model, separate three-comparison families per scene. SAT is a diagnostic, not force contact.

Scene	Edges	Base	Sel.	Joint gain [family], pp	SAT∧quality, pp	Quality, pp
BOUNDED DROP	1	150	216	12.9 [8.4, 17.6]	8.2 [4.1, 12.5]	1.6 [−1.0, 4.3]
BOUNDED DROP	2	127	192	12.7 [8.1, 17.6]	6.8 [3.3, 10.4]	−1.2 [−3.9, 1.6]
CLUSTER DROP	1	133	229	18.8 [13.5, 24.0]	11.1 [6.8, 15.4]	0.0 [−2.3, 2.3]
CLUSTER DROP	2	109	212	20.1 [15.2, 25.4]	13.7 [9.6, 17.8]	3.5 [1.2, 6.1]
2D DISCS(frozen)	1	10	19	1.8 [0.6, 3.3]	n/a	0.0 [−1.2, 1.2]
2D DISCS(frozen)	2	10	10	no action selected	n/a	identity
2D DISCS(secondary)	1	222	377	30.3 [25.8, 34.8]	n/a	0.0 [−1.2, 1.2]
2D DISCS(secondary)	2	182	310	25.0 [20.5, 29.5]	n/a	−1.0 [−2.1, 0.0]

Table 2: Held-out joint target-and-quality successes out of 512 proposals (baseline, selected policy) with family-adjusted gains; SAT∧quality and quality-pass changes are secondary with pointwise intervals. The 3D family has eight members and the frozen 2D family four; the secondary 2D rows evaluate the configuration selected under the secondary envelope ($\lambda = 100$ from $r = 600$) on the replayed held-out proposals, a family of two. The 2D two-edge frozen row replays no action.

for a small positive change in cluster drop. In 2D the frozen-envelope selection chose $\lambda = 10$ for single edges and no action for two edges because joint success was 2–8% for every candidate; under the secondary envelope the same selection cohorts choose $\lambda = 100$ from $r = 600$ for both, and that configuration gains 30.3 and 25.0 points on the replayed held-out proposals (family of two; 31.5 and 25.6 over matched random, 3.1 over matched sign, no quality change), versus 2.2 and 2.3 points under the frozen envelope, which only 22–31 of 512 trajectories pass.

Replication across denoiser seeds. Rerunning the complete protocol on the two other denoisers per cube scene (Appendix B.4) selects the same configuration in all twelve tasks and gives positive held-out joint gains in all twelve (10.7–23.6 points, per-seed family intervals excluding zero); no selected gate has a positive family interval.

Timing and scale. A fixed-arm sweep on 256 replayed proposals per task (Appendix B.5) shows the selection grid stopped at an edge: starts at 1000 or 800 give 15–20 points in all four tasks, $\lambda = 300$ beats $\lambda = 100$ everywhere, and $\lambda = 1000$ reaches 27–28 points in bounded drop with unchanged quality but falls to 8 points in single-edge cluster drop. Guidance confined to $(1000, 800]$ does nothing while every later window contributes 3–11 points, so there is no narrow window; equal-magnitude random steps give no gain and at $\lambda = 1000$ cut quality passes by 32–56 points, so the direction carries the effect.

Larger conjunctions. With the same fixed policy, three-, four- and five-edge targets (Appendix B.6) give declining but positive joint gains of 12.7, 9.8 and 7.0 points in bounded drop and 18.4, 16.2 and 10.5 in cluster drop (family intervals excluding zero, matched random near zero): control reaches half of the ten-edge graph with diminishing effect.

The matched sign update also has positive secondary joint effects, so the contrast with exact-class guidance, which gives no population-level benefit (Appendix H), is not attributable to variance scaling alone (Section 5).

4.5 Probability gating fails because the readout ranks poorly

The prospective budget evaluation gives bounded-drop gate effects of -14.6 [-34.2, 3.9] and 1.0 [-14.6, 16.6] joint hits per million allocated sample NFE; both cluster selectors choose no gate (identity zeros); the 2D effects are -2.0 [-6.8, 2.0] and -7.8 [-13.7, -2.0] and stay negative under the secondary envelope. No selected gate establishes a benefit.

The retrospective analysis says why (Table 4 in Appendix B.2). On identical saved states, a gate scored by the sixteen-branch empirical probability raises joint hits per executed NFE over no gate in all eighteen model/time groups (family intervals excluding zero, ratios 1.08–1.26): its kept states exceed the break-even lift of equation 2 by 0.12–0.33 in every group. The frozen probe, whose rank correlation with the independent estimate is only 0.27–0.59 at $r = 600$, gives 0.97–1.07 with two family intervals excluding zero; its kept states exceed break-even by at most 0.09 in thirteen groups and fall short in one, and four select no gate. Its isotonic recalibration does no better, 1.00–1.06 with none, consistent with Proposition 1(b), although it lowers the corrected squared error by 21–50% at $r = 600$. Oracle-minus-probe contrasts exclude zero in fifteen of eighteen groups and are largest early, where, at a given keep fraction, the required lift is smallest. The oracle's branch cost is not charged, so it is an information-availability reference, not a gate. Guidance gains are not runtime gains: each selected 3D guided cohort adds 306,688 probe evaluations to 512,000 denoiser evaluations.

4.6 Post-hoc branch analyses

Appendix C reports three exploratory analyses of the saved cohorts: within-cell associations between first-half uncertainty and second-half gain (positive in all eighteen groups, but descriptive and ceiling-limited, not an adaptive policy); direct mother-paired time comparisons of the geometric effect, whose eighteen family intervals all include zero, so no control window is claimed; and direct paired policy contrasts placing selected guidance above matched random on all four 3D tasks and above sign and one-shot ascent on three.

4.7 Existing controls delimit the readout finding

Forward-renoising the *same generated endpoints* with the same labels leaves a reverse-state readout advantage at intermediate noise (pair-F1 differences of 0.040 ± 0.002, 0.061 ± 0.005 and 0.166 ± 0.008 at $r = 500$ in the probes' cache convention, mean and SD over three probe seeds); comparing against forward-noised real data instead would mix endpoint populations. The readout's location is configuration dependent: a cosine schedule moves the sharpest interval to 950–900, deterministic DDIM moves the 2D interval to 400–350, and probe capacity and model width change amplitude and location.

5 Discussion

Direction versus probability. The results separate along the line drawn by Proposition 1. Uses that consume local alignment work: one-shot ascent raises event frequency and transfers to geometry, and multi-step guidance improves joint success on held-out targets in every scene once quality is measured on a scene-consistent scale. The use that consumes ranking does not: early rejection on the probe establishes no benefit prospectively and clears the break-even lift by at most 0.09 retrospectively, whereas a sixteen-branch empirical probability, whose rank correlation with an independent estimate of the same quantity exceeds 0.9, would clear it by 0.12–0.33 at every time if its branches were free. Selection on these states is therefore possible in principle and limited by discrimination,

which motivates readouts trained or estimated for ranking rather than a change of decision rule. The probe's overconfidence after optimization is real but matters only for decisions that read absolute probabilities.

Target granularity. Classifier guidance toward exact peak-contact classes, among 449–559 realized graphs per scene, gives no population-level benefit (Appendices H and I). The target object is the most visible difference from the policies above: single edges and small conjunctions are events with supported base rates on the proxy the probe was trained on, and the update is variance-scaled from a selected step. The matched sign update also helps, so the difference is not variance scaling alone; gains decline with conjunction size but stay positive up to five edges, and the exact-class result stands for its own object.

Concentration is measured, not explanatory. Repeated continuation gives the distribution of futures under the sampler and shows that it narrows with r. Nothing in the data links concentration to either finding: readout error dominates prediction error from the middle of the schedule onward, and direct time contrasts of the intervention effect include zero. A concentrated unmodified distribution also need not resist a deliberate intervention, so concentration should not be read as lock-in.

Quality has to be scene-relative. An absolute kinematic floor made the disc scene look unsteerable: raw gains were comparable to those in bounded drop, but 95% of every arm failed a limit that sat at 1.7 times the real data's 99th percentile instead of 4.6 times. Reporting the envelope relative to each scene's real statistics, and both envelopes when they disagree, prevents a measurement choice from being read as a property of the generator.

6 LIMITATIONS

The three controlled systems permit exact labels and paired interventions but do not establish generalization to larger scenes, other architectures or real manipulation. Local confirmation and the retrospective gate analysis use three fixed denoiser seeds per cube scene that share one probe seed; the primary policy tables use one fixed model per scene, and the seed replications are supplementary. Inference over mothers or proposals is not inference over a population of trained models. Conjunctions of up to five edges do not demonstrate arbitrary graph control, and repeated catalog targets are not unseen-target generalization.

Distance, near-surface, peak-frame and force-threshold contacts are different objects, and missing disc radii prevent overlap checks in 2D. Both quality envelopes are empirical; the frozen cube envelope permits substantial penetration, and the secondary envelope is a post-hoc definition that changes only the 2D kinematic limit. No pass rate certifies dynamical validity.

The retrospective oracle is noisy and uncharged, and its target mix differs from the prospective blocks, so it indicates what a readout could provide rather than proposing a gate. Concentration is sampler-relative, post-hoc analyses remain exploratory, and the study establishes no lock-in, control window or optimal intervention time.

7 CONCLUSION

A contact readout of a partially denoised trajectory can be useful for steering and useless for selection at the same time. Its gradient improves actual outcomes in every scene once quality is measured on a scene-consistent scale; its scores do not rank states well enough for early rejection to establish a benefit, although a free empirical probability of the same event would; and calibration, which neither use requires, is beside the point. None of this is visible in an accuracy curve.

AI USE STATEMENT

In this work, we used generative AI tools to assist with drafting sections of this paper and to improve the clarity, organization, and readability of the writing. We did not use generative AI tools to run the experiments. We reviewed and revised all AI-assisted text and checked technical claims, citations, and descriptions of experiments and results against the underlying work and original sources.

We take responsibility for the final content of this work, including text produced with the aid of generative AI.

Reproducibility statement

Section 3 defines the scenes, contact labels, estimators, interventions, policies and statistical procedures, and Appendix B.1 lists every cohort with its inference unit. Proposition 1 is proved in Appendix A. Scene construction, model training, probes and caches, and the noise schedule are described in Appendices D, E, F and G; the supplementary and post-hoc analyses, with their estimators, in Appendices B and C; and the exact-class experiments in Appendices H and I. Appendix J describes the code in the supplementary material and lists all trial counts.

References

Beatrice Achilli, Marco Benedetti, Giulio Biroli, and Marc Mézard. Theory of speciation transitions in diffusion models with general class structure. *Journal of Statistical Mechanics: Theory and Experiment*, 2026(4):043304, 2026. doi: 10.1088/1742-5468/ae5c94. arXiv:2602.04404.

Anurag Ajay, Yilun Du, Abhi Gupta, Joshua Tenenbaum, Tommi Jaakkola, and Pulkit Agrawal. Is conditional generative modeling all you need for decision-making? In *International Conference on Learning Representations (ICLR)*, 2023.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Mihai Anitescu. Optimization-based simulation of nonsmooth rigid multibody dynamics. *Mathematical Programming*, 105(1):113–143, 2006. doi: 10.1007/s10107-005-0590-7.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2302.07121.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15:9957, 2024. doi: 10.1038/s41467-024-54281-3.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023.

Hyungjin Chung, Jeongsol Kim, Michael T. McCann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.14687.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Zixuan Huang, Yinong He, Yating Lin, and Dmitry Berenson. Implicit contact diffuser: Sequential contact reasoning with latent point cloud diffusion. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10846–10853, 2025. doi: 10.1109/ICRA55743.2025.11128192. arXiv:2410.16571.

Michael Janner, Yilun Du, Joshua Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning (ICML)*, 2022.

Abhinav Kumar, Thomas Power, Fan Yang, Sergio Aguilera Marinovic, Soshi Iba, Rana Soltani Zarrin, and Dmitry Berenson. Diffusion-informed probabilistic contact search for multi-finger manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14277–14283, 2025. doi: 10.1109/ICRA55743.2025.11127844. arXiv:2410.00841.

Abhinav Kumar, Fan Yang, Sergio Aguilera Marinovic, Soshi Iba, Rana Soltani Zarrin, and Dmitry Berenson. Diffusing trajectory optimization problems for recovery during multi-finger manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2026. arXiv:2510.07030.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. doi: 10.52202/085713-3194. arXiv:2408.08252.

Zhixuan Liang, Yao Mu, Yixiao Wang, Tianxing Chen, Wenqi Shao, Wei Zhan, Masayoshi Tomizuka, Ping Luo, and Mingyu Ding. DexHandDiff: Interaction-aware diffusion planning for adaptive dexterous manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.18562.

Igor Mordatch, Emanuel Todorov, and Zoran Popović. Discovery of complex behaviors through contact-invariant optimization. *ACM Transactions on Graphics (SIGGRAPH)*, 31(4):43, 2012. doi: 10.1145/2185520.2185539.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, 2021.

Chaoyi Pan, Zeji Yi, Guanya Shi, and Guannan Qu. Model-based diffusion for trajectory optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Michael Posa, Cecilia Cantu, and Russ Tedrake. A direct method for trajectory optimization of rigid bodies through contact. *International Journal of Robotics Research (IJRR)*, 33(1):69–81, 2014. doi: 10.1177/0278364913506757.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences (PNAS)*, 122(1):e2408799121, 2025. doi: 10.1073/pnas.2408799121.

Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55810–55827, 2025. arXiv:2501.06848.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

David E. Stewart. Rigid-body dynamics with friction and impact. *SIAM Review*, 42(1):3–39, 2000. doi: 10.1137/S0036144599360110.

Dexin Wang, Chunsheng Liu, Faliang Chang, and Yichen Xu. Hierarchical diffusion policy: Manipulation trajectory generation via contact guidance. *IEEE Transactions on Robotics*, 41:2086–2104, 2025. doi: 10.1109/TRO.2025.3547272. arXiv:2411.12982.

Luhuan Wu, Brian L. Trippe, Christian A. Naesseth, David M. Blei, and John P. Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. arXiv:2306.17775.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2409.15761.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2303.09833.

A Proof of Proposition 1

Part (a). By the chain rule, $\nabla_x \log f(s(x)) = f'(s(x))\,\nabla_x s(x)/f(s(x)) = c(x)\,\nabla_x \log s(x)$ with $c(x) = f'(s(x))\,s(x)/f(s(x))$, which is positive because $f' > 0$ and $s, f(s) \in (0,1)$. One-shot ascent uses the gradient of the log score divided by its root-mean-square, which cancels $c(x)$. Variance-scaled guidance adds $\lambda\sigma_r^2$ times the gradient, which is multiplied by $c(x) > 0$ and keeps its direction. For conjunctions scored by the mean of marginal log scores, the same argument holds for a map applied to the combined score but not for maps applied edge by edge, even a common one, which reweight the edges by $c_e(x)$.

Part (b). Because $f' > 0$, f is continuous and strictly increasing, so $f(s) \geq \tau$ holds exactly when $s \geq f^{-1}(\tau)$ for τ in the range of f; thresholds below or above the range keep every or no state, which score thresholds also produce. Conversely $\{s \geq t\} = \{f(s) \geq f(t)\}$ for every t. The two families of keep sets therefore coincide, and because a gate's yield depends on the score only through its keep set, the best yield over thresholds is the same. If f is continuous and non-decreasing with flat segments, as an interpolated isotonic fit is, every set $\{f(s) \geq \tau\}$ is still of the form $\{s \geq t\}$, but score thresholds inside a flat segment cannot be produced, so the family is a coarsening and its best yield cannot increase. On a finite threshold grid the two scores reach different members of the same family, so the equality of best yields, and for non-decreasing maps the inequality, hold only up to that discretization.

Part (c). Each proposal is kept with probability p and then succeeds with probability $\mathbb{E}[q \mid \text{kept}]$; a rejected proposal yields no success. The expected number of successes per proposal is therefore $p\,\mathbb{E}[q \mid \text{kept}]$, and the expected number of denoiser updates per proposal is $pT+(1-p)(T-r) = T-(1-p)r$. Without a gate these are $\mathbb{E}[q]$ and T. By the renewal-reward theorem, the long-run number of successes per update equals the ratio of these expectations, so the gate is better exactly when $p\,\mathbb{E}[q \mid \text{kept}]/\{T-(1-p)r\} > \mathbb{E}[q]/T$, that is, $pL > 1-(1-p)r/T$, which rearranges to equation 2. Fixed budgets add only the truncation of the final proposal. The retrospective yields of Appendix B.2 satisfy this accounting exactly: for every source, model and time, $pL/\{1-(1-p)r/T\}$ reproduces the reported yield ratio.

B Supplementary analyses: gate-score sources and the 2D quality envelope

The retrospective gate analysis and the envelope reanalysis reuse saved trajectories and records without new sampling; the held-out 2D evaluation, the seed replications and the two sweeps report fresh GPU cohorts with their own reviewed protocols. No training, threshold or label change occurred, and the frozen primary families are unchanged.

B.1 Cohorts and inference units

Study	Scenes	Models per scene	Evaluation units
Branching pilot	Bounded drop, 2D	1	64 mothers × 16 branches
Local confirmation	Both cube scenes	3	128 mothers × 32 branches
Guidance policies	All three	1 (+2, suppl.)	512 paired proposals per task
Prospective gates	All three	1 (+2, suppl.)	32 paired budget blocks per task
Retrospective gate scores	Both cube scenes	3	128 mothers, 16+16 branches

Table 3: Cohorts and inference units. Branches are repeated futures of a mother state, not independent models. The two additional policy seeds (supplementary) cover the cube scenes only.

B.2 Retrospective gate-score sources

For each of the six confirmation cohorts (three denoiser seeds per cube scene) and each $r \in \{600, 400, 200\}$, branch slots 0–15 of the unmodified condition estimate the first-half target probability $\hat{q}_A$ and slots 16–31 supply the second-half outcomes. The three score sources are the frozen assigned-target probe score s_y, its cross-fitted isotonic recalibration (pool-adjacent-violators

fit on the other mother fold against $\hat{q}_A$, folds alternating within desired-event strata) and $\hat{q}_A$ itself. For a threshold τ the retrospective gate keeps mothers with score at least τ; its yield is $10^6 \sum_i \mathrm{keep}_i\, o_i / \sum_i [1000 - r(1 - \mathrm{keep}_i)]$ with o_i the second-half joint target-and-quality frequency, and no gate is $10^6\, \bar{o}/1000$. The threshold is chosen on first-half yields over a twenty-value grid starting at zero (no gate wins ties) and evaluated on the second half. Mothers are resampled within desired strata in 10,000 replicates, refitting the map and reselecting the threshold each time. Table 4 summarizes the selected-threshold results by source and time; Figures 3 and 4 show the fixed-threshold curves and the selected-threshold effects.

Score source	r	Yield ratio range	Median	Families > 0	Spearman
Frozen probe	600	0.97–1.04	1.00	0/6	0.27–0.59
Frozen probe	400	1.00–1.07	1.04	0/6	0.63–0.72
Frozen probe	200	1.02–1.06	1.04	2/6	0.79–0.84
Recalibrated probe	all	1.00–1.06	1.03	0/18	0.13–0.83
16-branch empirical probability	600	1.09–1.26	1.19	6/6	0.91–0.94
16-branch empirical probability	400	1.12–1.21	1.16	6/6	0.96–0.98
16-branch empirical probability	200	1.08–1.12	1.09	6/6	0.96–0.99

Table 4: Retrospective gate-and-respawn on the six confirmation cohorts: second-half joint hits per executed NFE relative to no gate, threshold selected on the first half; families of eighteen per source. Spearman is against the independent sixteen-branch second-half estimate. Ratios of 1.00 include no-gate identities; the oracle's branch cost is not charged.

The oracle's own branch cost is not charged, so it is an information-availability reference. The retrospective population mixes balanced presence and absence targets, unlike the fixed-target prospective blocks, and no counterfactual outcome is known for prospectively rejected proposals; the analysis therefore explains what a gate needs, not the specific prospective numbers. Recalibration reduces the noise-corrected squared error by 21–50% at $r = 600$ but lowers the rank correlation in all eighteen groups because plateaus merge scores. Within each mother fold the recalibrated score is a continuous non-decreasing map of s_y, so by Proposition 1(b) it can only coarsen that fold's family of keep sets over all thresholds. On the twenty-value grid, and because the two folds use different maps, the attainable pooled keep sets differ from the probe's, but each map is fitted without the evaluated fold's data, so these differences carry no information about the evaluated mothers' outcomes. The recalibrated rows are therefore a consistency check rather than independent evidence. Wherever the probe gates, its selected threshold is 0.05 or 0.1, the two lowest nonzero grid values, so its retrospective ratios are conditional on the grid's resolution near zero. Rows whose selected threshold is zero are no-gate identities.

B.3 Scene-consistent secondary quality envelope

The secondary kinematic limit is $\max(\text{frozen}, m \cdot p_{99})$ with m the smaller of the two cube scenes' frozen-to-p_{99} ratios (4.59), so that both cube limits remain exactly frozen and the 2D limit becomes 0.1332 m/s. Re-measuring every saved 2D held-out, selection and prospective trajectory under both envelopes reproduces all frozen masks, selections and block hits exactly. Figure 6 shows real and generated residual distributions with both limits; Figure 7 shows the held-out arms under both envelopes. Under the secondary envelope the selection cohorts would have chosen $\lambda = 100$ from $r = 600$ for both complexities (joint success 0.656 and 0.680 versus 0.320 and 0.375 unmodified), whereas the frozen-envelope selections were made on joint successes of 2–8%. The frozen-selected gates re-scored under the secondary envelope remain negative; the gates the secondary envelope would select are untested.

Held-out evaluation of the secondary-envelope selection. The configuration selected under the secondary envelope ($\lambda = 100$ from $r = 600$) was evaluated on the replayed held-out proposals of the completed 2D run (the unmodified arm reproduces the completed cohort bit-exactly), with matched random and matched pre-step sign controls, measured under both envelopes (Table 5, Figure 5). Selection and evaluation cohorts are disjoint; the frozen envelope remains primary; the 2D envelope has no overlap component, so its joint success is close to raw success.

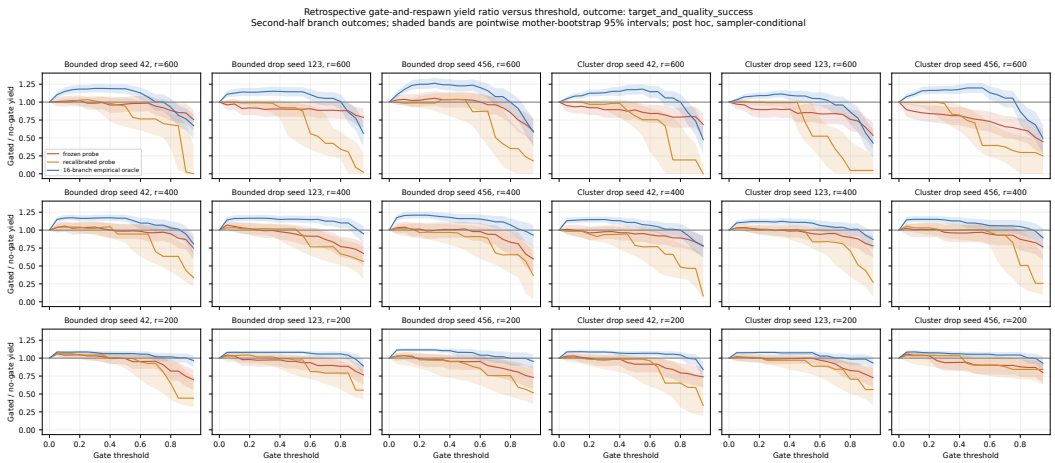

Figure 3: Retrospective gate yield relative to no gate versus threshold for the three score sources, joint outcome, second-half branches; shaded bands are pointwise mother-bootstrap intervals.

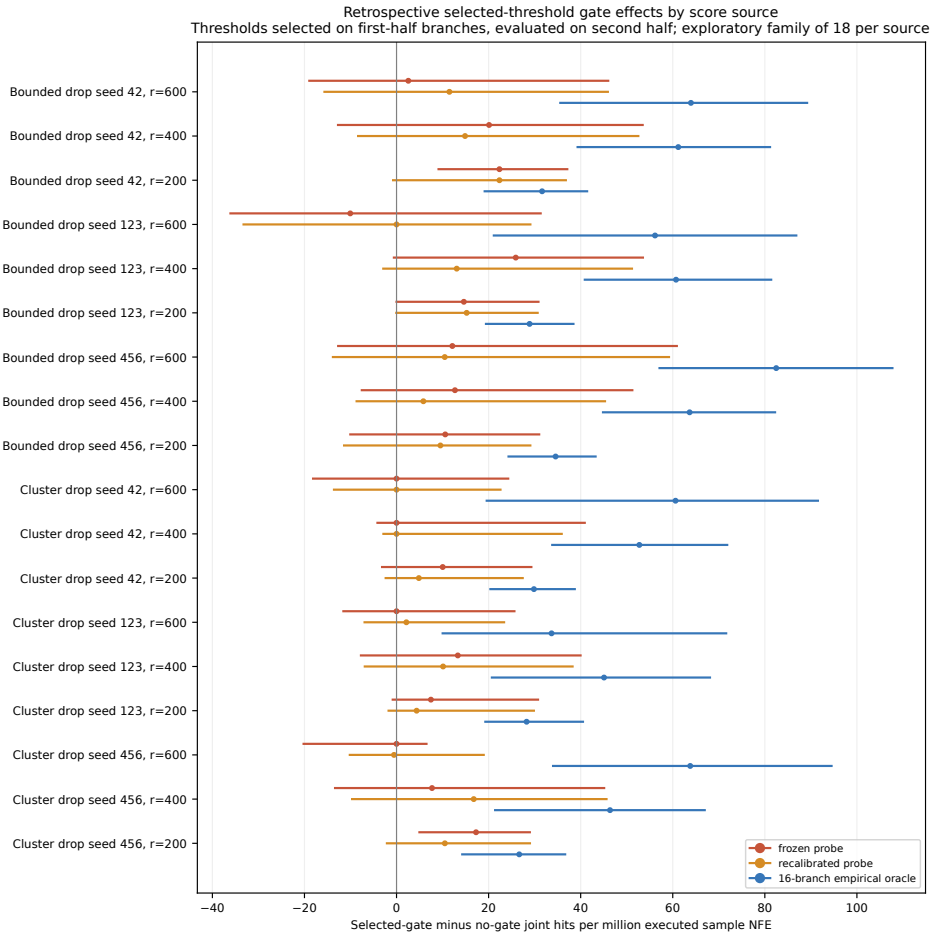

Figure 4: Selected-threshold gate effects by score source, joint hits per million executed sample NFE; exploratory families of eighteen per source. Rows at exactly zero (four probe and three recalibrated rows at $r = 600$ or 400) selected no gate and are identities, not measured null effects.

Edges	Arm	Raw	Joint (fr.)	Joint (sec.)	Sec. gain vs unmodified	vs random
1	unmodified	232	10	222		
1	$\lambda = 100$ from 600	393	21	377	30.3 [25.8, 34.8]	31.5 [27.5, 35.5]
1	matched random	229	9	216	-1.2 [-3.3, 1.0]	
1	matched sign	375	20	361	27.1 [23.2, 30.9]	
2	unmodified	193	10	182		
2	$\lambda = 100$ from 600	332	22	310	25.0 [20.5, 29.5]	25.6 [21.7, 29.5]
2	matched random	190	14	179	-0.6 [-2.1, 0.8]	
2	matched sign	315	21	294	21.9 [18.2, 25.6]	

Table 5: 2D held-out evaluation, 512 replayed proposals per arm. Counts are raw proxy-target hits and joint target-and-quality hits under each envelope; gains are percentage points with the family-of-two interval for the selected arm versus unmodified and pointwise intervals otherwise. Under the frozen envelope the selected arm gains 2.2 and 2.3 points.

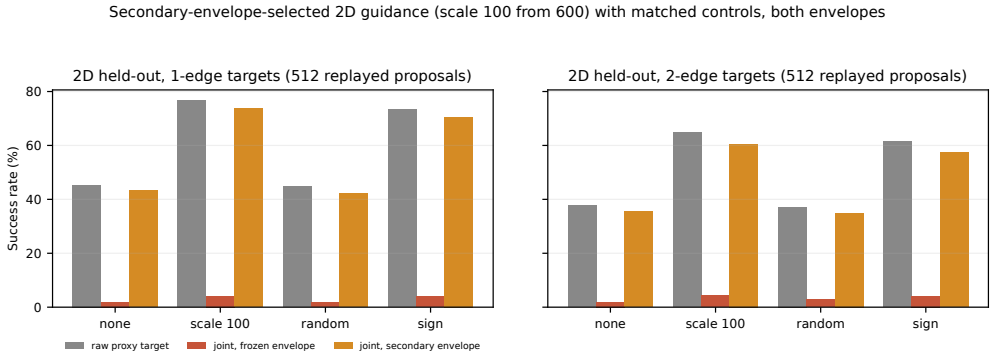

Figure 5: 2D held-out evaluation of the secondary-envelope selection with matched controls, raw and joint success under both envelopes.

B.4 Replication of the held-out policy protocol across denoiser seeds

The complete policy protocol (catalog, selection, held-out evaluation and prospective gating) was rerun on the two other independently trained denoisers per cube scene with their own matched pair probes and new random seeds. Each seed forms its own eight-comparison family; cross-seed statements are descriptive counts. All twelve selection tasks chose variance guidance at scale 100 from remaining step 600. The selected gate configurations differ across seeds, so no single gate policy transfers.

Scene	Edges	Seed	Joint gain [family], pp	Quality, pp	Gate (r/τ)	Gate effect [family]
Bounded drop	1	42	12.9 [8.4, 17.6]	+1.6	200/0.25	-14.6 [-34.2, 3.9]
Bounded drop	1	123	16.0 [11.3, 20.7]	+2.3	400/0.5	-23.4 [-50.5, 3.7]
Bounded drop	1	456	11.7 [7.2, 16.4]	+0.6	400/0.5	-22.5 [-50.8, 4.9]
Bounded drop	2	42	12.7 [8.1, 17.6]	-1.2	600/0.25	1.0 [-14.6, 16.6]
Bounded drop	2	123	10.9 [6.6, 15.4]	-0.6	200/0.25	-4.9 [-19.5, 10.7]
Bounded drop	2	456	10.7 [6.6, 15.2]	+0.6	400/0.25	6.8 [-9.5, 21.5]
Cluster drop	1	42	18.8 [13.5, 24.0]	+0.0	none	identity
Cluster drop	1	123	21.3 [15.8, 26.6]	-0.4	600/0.25	-34.2 [-64.5, -5.9]
Cluster drop	1	456	19.1 [14.1, 24.2]	+2.0	200/0.25	-2.9 [-19.5, 12.7]
Cluster drop	2	42	20.1 [15.2, 25.4]	+3.5	none	identity
Cluster drop	2	123	23.2 [17.8, 28.9]	+1.4	400/0.25	-6.8 [-22.5, 7.8]
Cluster drop	2	456	23.6 [18.4, 29.1]	+3.5	200/0.25	-1.0 [-15.6, 13.7]

Table 6: Held-out selected-guidance joint gains and selected-gate effects (joint hits per million allocated sample NFE) for three denoiser seeds per cube scene; seed 42 rows are the primary results of Table 2. "Identity" marks tasks whose selector chose no gate.

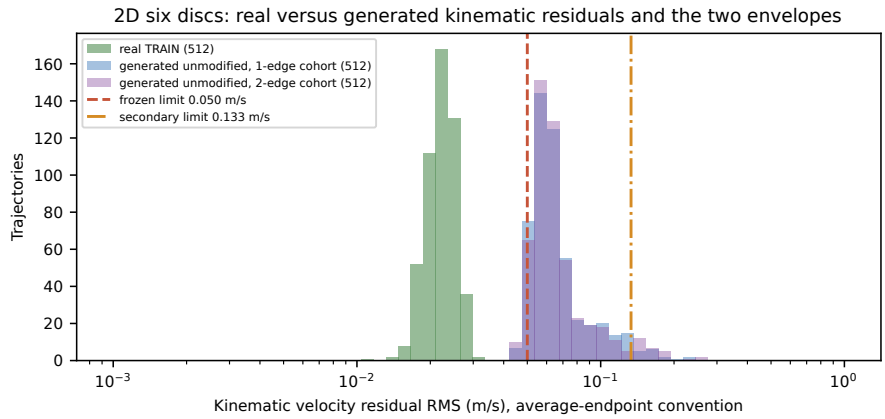

Figure 6: 2D kinematic residuals of real TRAIN and generated unmodified trajectories with the frozen and secondary limits.

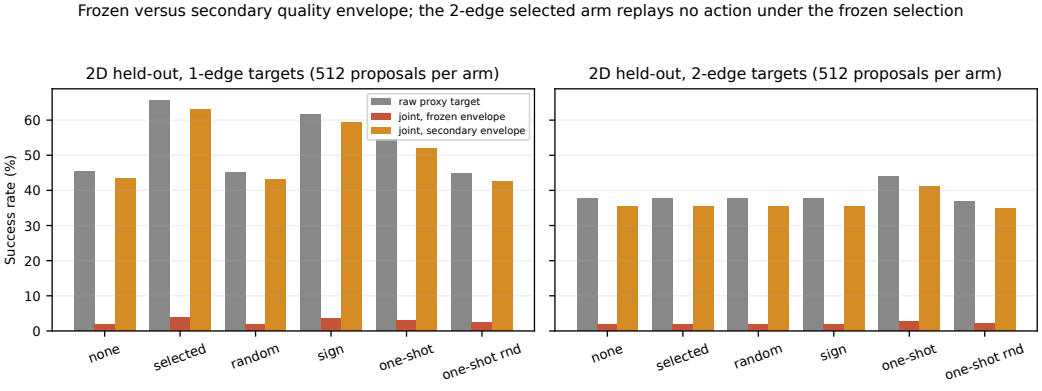

Figure 7: 2D held-out arms under both envelopes (512 proposals each). The two-edge selected, random and sign arms replay no action under the frozen selection.

B.5 Guidance timing and scale sweep

On the first 256 of the completed held-out evaluation proposals per task (same targets and future-noise streams; the unmodified arm reproduces the completed batches bit-exactly and the $\lambda = 100$-from-600 arm reproduces its completed cohort exactly), fifteen additional arms vary the guidance start, restrict guidance to a window $(\text{start}, \text{end}]$ of remaining updates, or vary the scale from remaining update 600, with matched random controls for the two most aggressive settings. All contrasts are paired against the same unmodified proposals within catalog-rate strata; the sixty joint contrasts form one exploratory Bonferroni family (Table 7). Quality is the frozen envelope.

Earlier starts help or do not hurt: with $\lambda = 100$, gains from 1000 or 800 are 15–20 points in all four tasks versus 10–20 from 600 and 3–11 from 200. Guidance confined to $(1000, 800]$ produces no gain, while every later window contributes, so the effect accrues over most of the schedule rather than in a narrow band. Larger scales help up to a scene-dependent limit: $\lambda = 300$ exceeds $\lambda = 100$ in all four tasks, and $\lambda = 1000$ reaches 27–28 points in bounded drop with unchanged quality but falls to 8 points with a 3.5-point quality loss in single-edge cluster drop. Matched random steps of the same magnitude give no gain from 1000 at $\lambda = 100$ and, at $\lambda = 1000$, reduce quality passes by 32–56 points with visibly diverging states (median normalized-state RMS 1.0–1.4 versus 0.67–0.80), whereas the gradient arms keep it: the direction, not the perturbation size, carries the effect. The selection grid of the primary experiments therefore stopped at a grid edge, and the primary gains are conservative.

Arm	BOUNDED DROP, 1 edge	BOUNDED DROP, 2 edges	CLUSTER DROP, 1 edge	CLUSTER DROP, 2 edges
$\lambda = 100$ from 600 (replay)	10.2 [5.9, 14.8] (+0.4)	12.9 [8.2, 17.6] (-0.8)	19.5 [14.5, 25.0] (+0.4)	19.5 [14.5, 24.6] (+4.3)
$\lambda = 100$ from 1000	17.2 [11.7, 22.7] (+0.4)	15.2 [10.5, 19.9] (+0.8)	19.9 [14.1, 25.8] (+1.2)	18.4 [13.3, 23.8] (+1.2)
$\lambda = 100$ from 800	16.4 [10.9, 21.9] (+1.2)	17.2 [12.1, 22.3] (+2.0)	18.0 [12.5, 23.4] (+1.6)	19.9 [14.8, 25.4] (+1.6)
$\lambda = 100$ from 400	4.7 [2.0, 7.8] (+2.0)	8.2 [4.7, 12.1] (-0.4)	14.8 [10.5, 19.1] (-0.4)	14.5 [10.2, 19.1] (+0.0)
$\lambda = 100$ from 200	2.7 [0.8, 5.1] (+1.2)	5.5 [2.7, 8.6] (+0.4)	7.4 [4.3, 10.9] (-0.8)	10.9 [7.0, 14.8] (+0.4)
window (1000,800]	1.2 [-2.0, 4.7] (-2.7)	2.7 [0.0, 5.9] (-1.2)	-2.0 [-4.7, 0.8] (-1.6)	1.2 [-1.2, 3.5] (-2.3)
window (800,600]	10.9 [6.6, 15.6] (+1.6)	8.6 [4.3, 12.9] (-1.6)	4.7 [1.2, 8.2] (+1.6)	4.7 [1.6, 7.8] (+0.0)
window (600,400]	4.7 [1.2, 8.2] (+0.4)	7.8 [3.9, 12.1] (-0.4)	9.4 [5.1, 13.7] (-1.6)	8.6 [5.1, 12.5] (+0.4)
window (400,200]	2.7 [0.8, 5.1] (+0.4)	5.9 [3.1, 9.0] (+0.0)	6.6 [3.9, 9.8] (+0.4)	8.2 [4.7, 11.7] (+1.2)
window (200,1]	2.7 [0.8, 5.1] (+1.2)	5.5 [2.7, 8.6] (+0.4)	7.4 [4.3, 10.9] (-0.8)	10.9 [7.4, 14.8] (+0.4)
$\lambda = 10$ from 600	5.5 [2.3, 9.0] (+2.3)	8.6 [5.1, 12.5] (-0.4)	12.1 [7.8, 16.4] (+0.8)	11.7 [7.8, 16.0] (+0.4)
$\lambda = 30$ from 600	7.8 [4.3, 11.3] (+2.7)	10.9 [6.6, 15.2] (-0.8)	16.0 [11.3, 21.1] (+0.4)	15.2 [10.5, 19.9] (+2.3)
$\lambda = 300$ from 600	16.0 [10.9, 21.5] (-0.4)	16.8 [11.7, 22.3] (-1.6)	21.5 [15.2, 27.7] (-2.3)	22.3 [17.2, 27.7] (+1.6)
$\lambda = 1000$ from 600	27.7 [21.5, 34.0] (-0.8)	27.0 [20.3, 33.2] (-0.4)	8.2 [1.2, 15.2] (-3.5)	18.8 [12.5, 25.0] (-2.7)
random, $\lambda = 100$ from 1000	-2.7 [-6.2, 0.8] (-7.4)	0.4 [-3.1, 3.9] (+0.4)	-7.4 [-12.9, -2.0] (-18.8)	-4.7 [-9.4, -0.4] (-8.6)
random, $\lambda = 1000$ from 600	-14.5 [-19.9, -9.4] (-52.7)	-3.9 [-8.2, 0.4] (-32.4)	-21.1 [-26.6, -15.6] (-56.2)	-13.7 [-18.8, -9.0] (-51.6)

Table 7: Joint target-and-quality gain versus unmodified in percentage points with pointwise 95% intervals (family-of-60 intervals in the run tables) and, in parentheses, the change in quality-pass rate. Unmodified joint successes are 72, 62, 68 and 53 of 256. Random arms use the same step RMS as the corresponding gradient arm at each step.

B.6 CONJUNCTION SIZE SWEEP

Supported three-, four- and five-edge conjunction targets were built per complexity from the saved 256-endpoint catalog cohorts with the completed rate bounds and strata (cluster drop has only the lowest-rate stratum at four and five edges), and 512 fresh proposals per task were evaluated under a new seed with the fixed selected policy ($\lambda = 100$ from 600), unmodified and matched random arms. Conjunction scores remain mean marginal log scores, not joint probabilities; catalog target identities can recur. Table 8 and Figure 8 combine these with the completed one- and two-edge results.

Scene	Edges	Unmod.	Selected	Random	Gain [family of 6], pp	vs random, pp	Quality, pp
BOUNDED DROP	3	61	126	65	12.7 [8.8, 16.8]	11.9 [8.8, 15.0]	+0.2
BOUNDED DROP	4	55	105	55	9.8 [6.1, 13.7]	9.8 [7.0, 12.7]	+1.6
BOUNDED DROP	5	56	92	54	7.0 [3.5, 10.7]	7.4 [4.9, 10.2]	+0.2
CLUSTER DROP	3	60	154	58	18.4 [13.7, 23.0]	18.8 [15.4, 22.3]	+1.6
CLUSTER DROP	4	22	105	24	16.2 [11.9, 20.7]	15.8 [12.7, 19.1]	+2.1
CLUSTER DROP	5	29	83	26	10.5 [7.0, 14.3]	11.1 [8.2, 14.1]	-0.8

Table 8: Joint target-and-quality successes out of 512 for three- to five-edge conjunctions, selected-minus-unmodified gains with the exploratory family-of-six interval, selected-minus-matched-random gains with pointwise intervals, and the change in quality-pass rate.

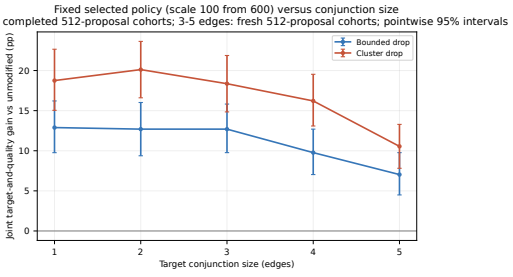

Figure 8: Joint gain of the fixed selected policy versus conjunction size; one- and two-edge points are the completed 512-proposal cohorts, three- to five-edge points the fresh cohorts; pointwise 95% intervals.

Gains decline with conjunction size but remain positive at five edges, half of the ten-edge graph, with family intervals excluding zero in all six tasks; matched random gives none. Raw target gains decline similarly and quality is unchanged. Control therefore extends beyond the single edges and pairs of the primary tables, with diminishing but not vanishing effect.

B.7 ASSEMBLED FIGURES

Figure 9 shows the three-seed replication and Figure 10 the guidance timing and scale sweep of Table 7.

C EXPLORATORY ANALYSIS OF SAVED BRANCHES AND POLICY RECORDS

This analysis was specified after the aggregate confirmation and policy outcomes were inspected. It reuses saved data, without new sampling, model updates, target changes or quality-threshold tuning. Its comparison families are distinct from the primary families.

Consistency and units. The analysis checks 180 consumed confirmation-record hashes against producer indices and reconstructs 2,304 mother/time records, including graph agreement, finite-branch Brier components and the saved independent-half measurements. Policy records receive current input hashes. All 384 prospective budget blocks are reconstructed from saved proposal/batch records, including target identity through respawns, rejected prefixes and unused budget. The six primary mean gate contrasts are reproduced exactly. In a block with gate time r, executed sample NFE is $1000N_{\text{completed}} + (1000 - r)N_{\text{rejected}}$. The denominator for budget yield is the allocated budget, and quality rates among completed trajectories are reported separately.

Independent branch halves. For each mother, let A and B denote disjoint sets of 16 branch slots. Conditions share future noise within a slot; slots are independent conditional on the mother, target and intervention. Let $\hat{q}_A, \hat{q}_B$ be baseline target frequencies, and let Δ_B be the ascent-minus-random difference estimated using slots B. Uncertainty estimates are $\hat{u}_A = (16/15)\hat{q}_A(1-\hat{q}_A)$ and analogously for B. For fixed model/time and target-pair/desired-event cells c, define

$$\widehat{C}(X, Z) = \frac{\sum_c \sum_{i \in c} (X_i - \bar{X}_c)(Z_i - \bar{Z}_c)}{\sum_c (n_c - 1)}. \tag{3}$$

The three separate predictor analyses use $(X_A, X_B) = (s_y, s_y)$, $(\hat{q}_A, \hat{q}_B)$, or $(\hat{u}_A, \hat{u}_B)$, where s_y is the score for the assigned presence or absence target. The linear projection estimate is $\widehat{C}(X_A, \Delta_B)/\widehat{C}(X_A, X_B)$. Cross-replicate covariance estimates the predictor's signal variance in expectation; neither half is a noise-free oracle. The estimand is weighted by $n_c - 1$, with approximately six mothers per target cell. There are no singleton cells in these cohorts.

We resample entire paired mother rows within target cells, recomputing cell means and both covariances in each of 3,000 bootstrap replicates. Slopes are reported only when the point denominator and its lower pointwise 95% bound exceed 10^{-8} and every bootstrap ratio is finite. Finite negative-denominator draws are not discarded. All 162 predictor/outcome/model/time combinations meet this numerical criterion in the saved cohorts. This is not evidence that all associations are nonzero. Covariances, denominator diagnostics and all outcome definitions remain available, including effects on SAT and joint target/quality outcomes.

The probability-bin display in Figure 11 uses fixed edges $0, 0.25, 0.5, 0.75, 1$: bins are left-closed and right-open except that the final bin includes 1. Bins depend on the noisy first-half frequency and are unadjusted descriptive displays. Their effects are measured on the independent second half. These results neither test incremental prediction beyond probe score nor validate a policy that chooses when or where to intervene. Estimating branch probabilities itself has a computational cost.

Direct contrasts. Time comparisons resample mothers within the assigned presence/absence strata and preserve each mother's targets and times. Their 18 SAT contrasts form one exploratory Bonferroni family. Policy comparisons resample paired proposals within catalog-rate bins; their 18 joint-quality contrasts form another family. Both use 20,000 percentile-bootstrap replicates. The small effective number of extreme-tail draws and sparse discordances limit the precision of adjusted intervals. Figures 12 and 13 retain every contrast, including no-action identities. Secondary outcomes use pointwise intervals, without promotion to the primary families.

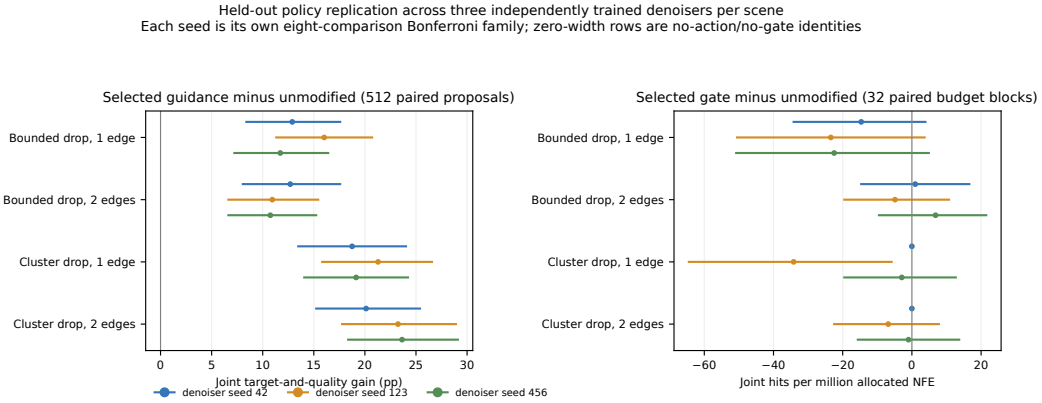

Figure 9: Held-out policy replication across three denoiser seeds per cube scene: selected guidance joint gains (left) and selected gate effects (right) with per-seed eight-comparison family intervals; zero-width rows are no-gate identities.

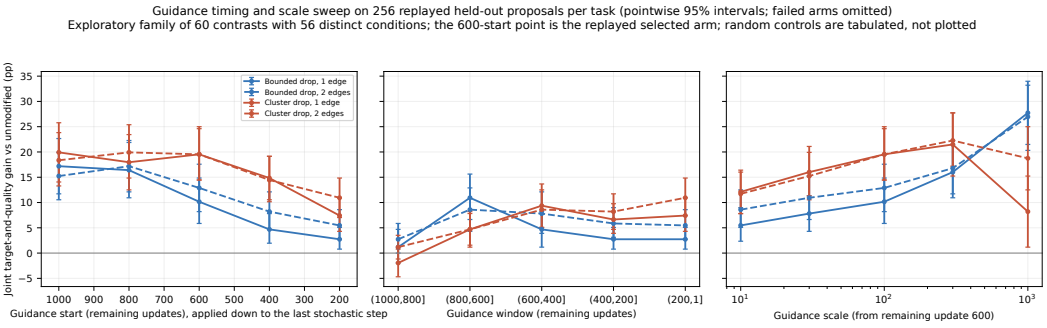

Figure 10: Guidance timing and scale sweep on 256 replayed proposals per task: joint gain versus guidance start, restricted window and scale, pointwise 95% intervals. The 600-start and scale-100 points are the replayed selected arm; the (200,1] window equals the 200 start; random controls are in Table 7.

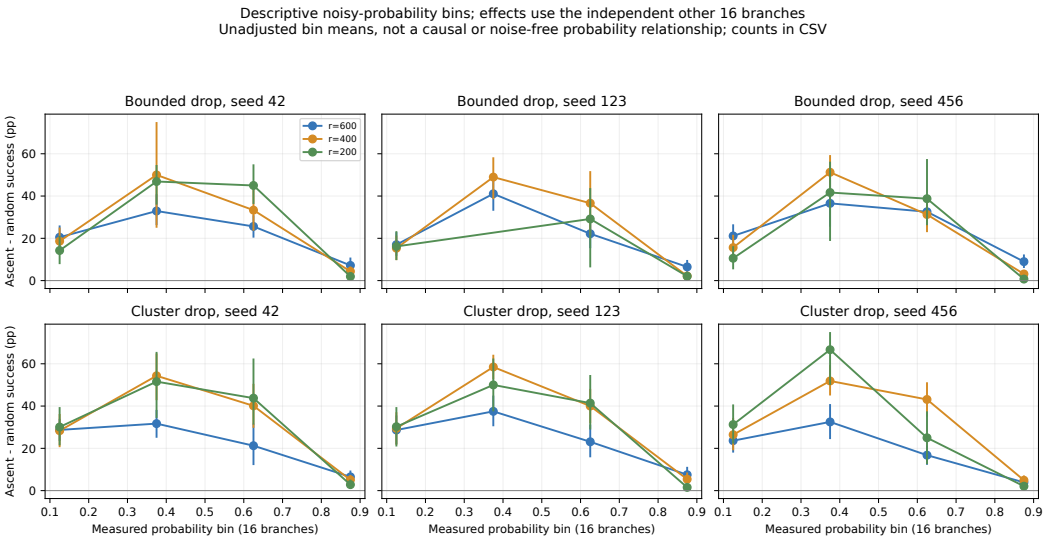

Figure 11: Descriptive association between measured baseline target probability and intervention gain. Bin membership uses 16 branches and is noisy; effects use the other 16 branches. Curves do not control for target composition within bins and should not be interpreted as a causal response to changing q. Table 9 gives the mother counts, including sparse and empty bins.

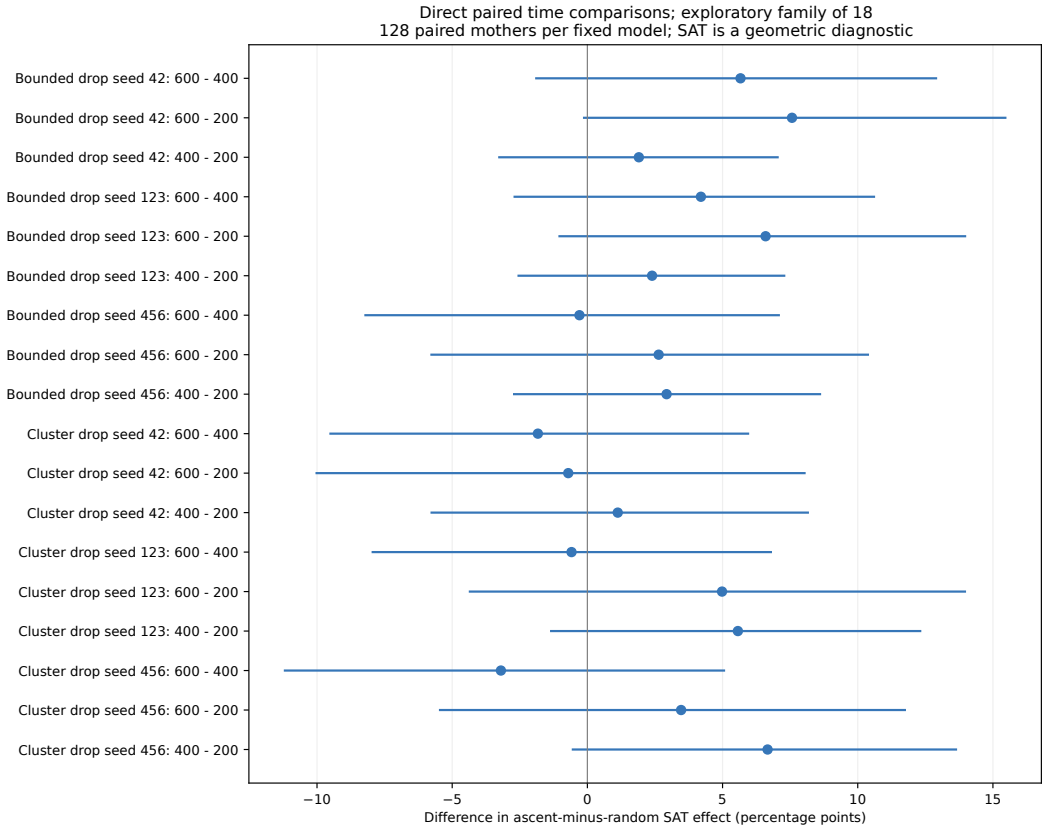

Figure 12: Direct mother-paired time differences in SAT ascent-minus-random effects. All 18 exploratory family intervals include zero. This is neither evidence of equal effects nor support for a universal intervention window.

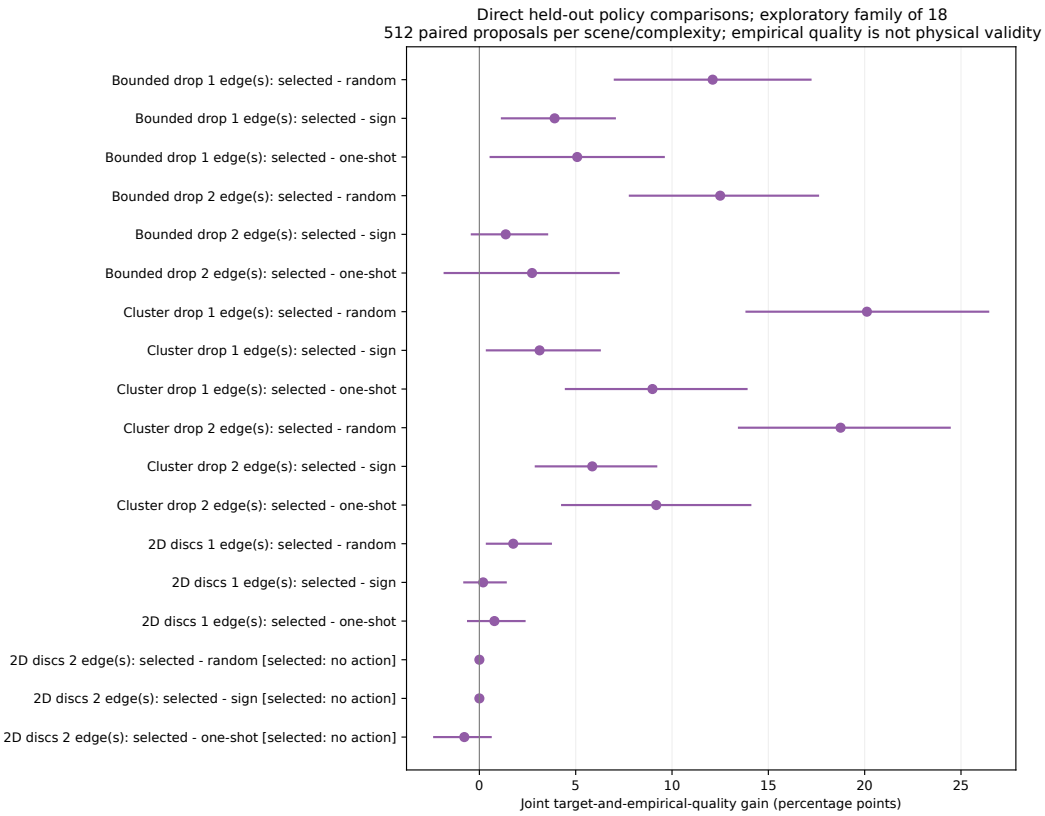

Figure 13: Direct paired comparisons of the selected policy against matched random, matched sign and one-shot ascent controls. These joint target/quality contrasts use a separate exploratory family of 18. In 2D two-edge tasks, the selected policy and its matched random/sign arms all replay no action; their zero contrasts are identities. The one-shot arm remains an active intervention.

Scene	Seed	r	$[0, .25)$	$[.25, .5)$	$[.5, .75)$	$[.75, 1]$
BOUNDED DROP	42	600	46	15	20	47
	42	400	59	4	9	56
	42	200	61	4	5	58
	123	600	39	21	11	57
	123	400	51	6	7	64
	123	200	58	0	3	67
	456	600	50	13	19	46
	456	400	62	10	6	50
	456	200	68	3	5	52
CLUSTER DROP	42	600	37	28	15	48
	42	400	45	13	12	58
	42	200	59	4	3	62
	123	600	29	31	23	45
	123	400	45	14	10	59
	123	200	56	3	8	61
	456	600	41	20	22	45
	456	400	52	10	10	56
	456	200	63	3	2	60

Table 9: Mother counts in the noisy probability bins of Figure 11. Each row has 128 mothers. Membership depends on 16 baseline branches; counts apply to all three outcome definitions in the saved tables. Sparse middle bins, including one empty bin, limit interpretation.

D FURTHER DETAILS ON SCENES

D.1 BOUNDED DROP: BOUNDED DROP OF FIVE CUBES

Five 2-cm cubes, each initialized at a random position in a $6 \times 6 \times 6$ cm walled box with a random initial velocity of magnitude 0.2–0.45 m/s in a random direction; gravity 9.81 m/s^2; restitution inelastic; friction 0.6. $T_{\text{sim}} = 300$ steps at $dt = 0.003$ s $= 0.9$ s. The walls keep cubes in a collision-dense regime, producing rich per-trajectory contact dynamics.

D.2 CLUSTER DROP: OPEN-CLUSTER DROP OF FIVE CUBES

Five 2-cm cubes initialized in a tight xy ± 1.5 cm, $z \in [2.5, 5]$ cm cluster on an open floor; gravity 9.81 m/s^2; restitution inelastic; friction 0.9; initial velocity zero. $T_{\text{sim}} = 300$ steps at $dt = 0.003$ s. Most trajectories end with cubes separated on the floor, yielding a sparser contact signal than BOUNDED DROP.

D.3 2D DISCS: ZERO-GRAVITY SIX-DISC BILLIARD

Six discs of radius 2–3.5 cm in a 30×30 cm 2D box; no gravity; restitution 0.8 (near-elastic); friction 0.3; initial velocities random. $T_{\text{sim}} = 200$ steps at $dt = 0.002$ s.

E TRAINING AND SANITY DETAILS

Linear schedule DDPM, $T = 1000$, $\beta_1 = 10^{-4}$, $\beta_T = 2 \cdot 10^{-2}$. TemporalUNet with base dim 256, channel multipliers $(1, 2, 2)$, time embedding dim 128. AdamW, learning rate $2 \cdot 10^{-4}$, batch size 128, gradient clip 1.0, EMA decay 0.9999. Trained 500K steps per scene on a single RTX 4090, approximately 10 hours for 3D and 8 hours for 2D. Validation loss final: 0.01291 (BOUNDED DROP), 0.00615 (CLUSTER DROP), 0.01118 (2D DISCS). Sanity (100 unconditional samples): normalized_std 0.778/0.700/0.934, spread p95 6.8/8.8/27.1 cm, contact_rate 100%/100%/100%. All scenes pass the sanity gates.

F PROBE AND CACHE DETAILS

Scope note. These probes are trained once per scene and reused throughout. The pair probe predicts ever-contact edges of the center-distance proxy; the class probe predicts peak-contact classes and serves as a readout baseline and for exact-class guidance.

Intermediates cached at $t \in \{0, 50, 100, \ldots, 1000\}$ (21 timesteps) per scene, 10K intermediates per timestep (approximately 16 GB per 3D scene, 10 GB for 2D DISCS). For each scene, probes are trained with 3 seeds (42, 123, 456) to separate probe-training noise from denoiser-state noise. Probe architecture (main): 3-hidden-layer MLP with widths $(512, 256, 128)$, ReLU activations, a learnable timestep embedding concatenated to the state embedding. Training: AdamW, learning rate 10^{-3}, batch size 256, 100 epochs, 80/10/10 train/val/test split. Contact threshold calibration: the proxy threshold τ was chosen on training data to maximize pair-F1 of the center-distance proxy against the simulator's contact labels; it is frozen per scene (0.028 m, 0.022 m and 0.06 m).

Table 10: Per-seed pair-F1 at $t=600$ and $t=400$ and the end-point ($t=0$). The main text quotes the mean across these seeds; individual seeds differ by ≤ 0.02 throughout.

Scene	Seed	$t=600$	$t=400$	$t=0$
BOUNDED DROP	42	0.773	0.859	0.908
	123	0.770	0.859	0.911
	456	0.769	0.862	0.910
CLUSTER DROP	42	0.589	0.752	0.873
	123	0.588	0.749	0.865
	456	0.587	0.748	0.857
2D DISCS	42	0.360	0.644	0.765
	123	0.357	0.644	0.765
	456	0.362	0.655	0.768

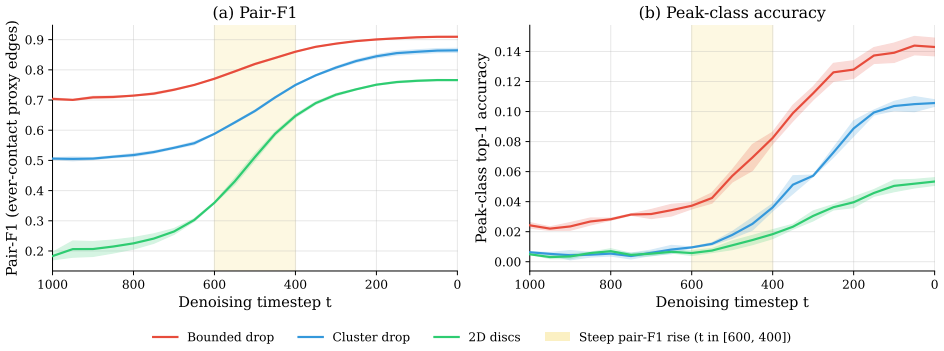

Figure 14: Probe readouts versus timestep t (equal to remaining updates under the linear schedule): pair-F1 of the ever-edge probe and top-1 accuracy of the peak-class probe, three probe seeds. The figure is shown as context. The shaded band marks where pair-F1 rises steeply; its location is configuration dependent (Section 4.7) and is not a lock-in claim.

G MOTIVATION FOR LINEAR OVER COSINE SCHEDULE

During early piloting we observed that the cosine schedule of Nichol & Dhariwal (2021), with β capped at 0.999 as in their implementation, produces $1/\sqrt{\alpha_T} = 31.6$ at $t = T$, which amplifies prediction errors by $31.6\times$ per step. Over 1000 ancestral sampling steps this compounded into catastrophic sample divergence (e.g. a 5-cube box giving *normalized_std* of 217 and spread p95 of 3437 cm on a scene with a 6 cm physical scale). With the linear schedule ($\beta_T = 2 \cdot 10^{-2}$, $1/\sqrt{\alpha_T} \approx 1.01$), standard ancestral sampling produces stable outputs with no workaround, so we can attribute all downstream effects to the network's intermediate-state representation rather than to any inference-time remedy (DDIM, x_0-clipping, dynamic thresholding, etc.).

H EXACT-CLASS PAIRED INTERVENTIONS AND CLASS-LEVEL GUIDANCE

Scope note. This appendix reports paired single-step interventions measured by mode-flip rates of the exact peak-contact class, and a class-level sign-guidance protocol. They concern a different target object (exact peak-contact classes) and a different update from the single-edge and conjunction policies of the main text, and their flip-rate peaks are not used to claim a control window. Timesteps t equal remaining updates under the linear schedule.

H.1 FULL CAUSAL-INTERVENTION TABLES

Table 11: Causal intervention — sparse 9-point grid on all three scenes ($n = 1000$ paired trials per cell). The sparse grid suggests a peak near $t = 500$; dense sampling (below) shows the absolute peak is earlier.

	BOUNDED DROP		CLUSTER DROP		2D DISCS	
t	plausible	random	plausible	random	plausible	random
999	0.243	0.072	0.255	0.075	0.120	0.057
950	0.310	0.101	0.325	0.096	0.171	0.110
900	0.389	0.126	0.424	0.135	0.243	0.134
500	0.474	0.086	0.585	0.120	0.247	0.134
200	0.268	0.048	0.445	0.054	0.127	0.060
100	0.262	0.041	0.369	0.067	0.101	0.059
Directed/random @ $t = 500$	5.5×		4.9×		1.84×	

Table 12: Causal intervention — dense 13-point grid on all three scenes ($n = 500$ paired trials per cell). Bold entries mark the absolute-flip-rate peak per scene. The dense scan reveals that this peak is earlier than the sparse-sample $t = 500$ on every scene (BOUNDED DROP: $t = 650$, CLUSTER DROP: $t = 600$, 2D DISCS: $t = 650$). The directed-vs-random ratio peaks where pair-F1 rises steeply.

	BOUNDED DROP			CLUSTER DROP			2D DISCS		
t	plaus.	rand.	d/r	plaus.	rand.	d/r	plaus.	rand.	d/r
650	**0.548**	0.166	3.30×	0.595	0.173	3.44×	**0.392**	0.211	1.86×
625	0.538	0.149	3.62×	0.621	0.151	4.11×	0.343	0.225	1.52×
600	0.508	0.140	3.63×	**0.628**	0.163	3.86×	0.354	0.178	1.99×
575	0.487	0.125	3.89×	0.597	0.144	4.14×	0.316	0.174	1.82×
550	0.479	0.092	5.23×	0.604	0.144	4.19×	0.269	0.151	1.78×
525	0.455	0.111	4.11×	0.580	0.105	5.52×	0.229	0.126	1.82×
500	0.486	0.078	**6.23**×	0.575	0.115	5.00×	0.225	0.103	2.19×
475	0.453	0.097	4.68×	0.601	0.103	**5.85**×	0.203	0.115	1.76×
450	0.381	0.088	4.35×	0.530	0.097	5.46×	0.232	0.097	**2.38**×
425	0.366	0.078	4.70×	0.555	0.075	7.40×	0.194	0.096	2.03×
400	0.394	0.042	9.29×	0.533	0.084	6.38×	0.155	0.078	1.98×
375	0.354	0.061	5.80×	0.500	0.109	4.57×	0.132	0.083	1.59×
350	0.358	0.056	6.36×	0.514	0.094	5.47×	0.143	0.080	1.79×

H.2 BOUNDARY SWEEP ON BOUNDED DROP

ReproducibleNoise. We implement a seeded noise generator that, given a master seed and a trajectory index i, deterministically produces the noise sequence $\{\epsilon_T^{(i)}, \epsilon_{T-1}^{(i)}, \ldots\}$ used by ancestral sampling. Two trajectories sharing the same i share the same noise at every step; the only difference between a control and a perturbed trajectory is the presence of the perturbation at the intervention step.

Perturbation scales. For each directed condition we scale the perturbation so that its norm matches the expected noise norm at that denoising step, $|\Delta|_2 = \lambda\sigma_t$ with $\lambda = 0.5$. The random condition is a Gaussian of the same norm but random direction.

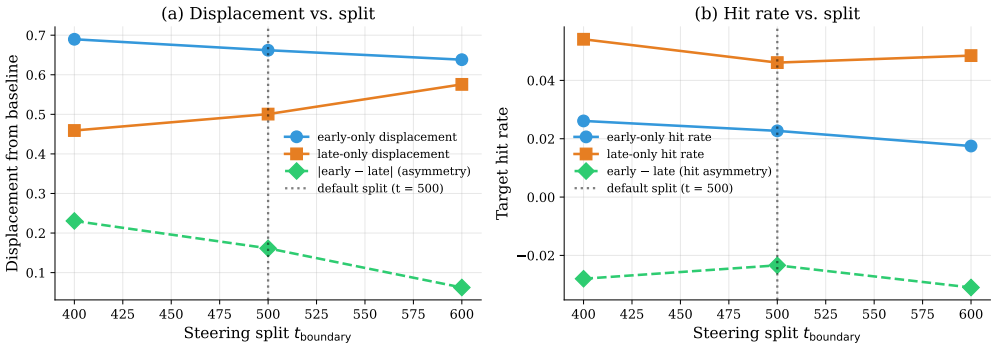

Figure 15: Boundary sweep on BOUNDED DROP. Displacement from baseline, one minus the rate at which a guided sample's peak-contact class agrees with the unguided baseline, for early-only guidance (blue), late-only guidance (orange), and their absolute asymmetry (green dashed) as the split point t_{boundary} shifts among $\{400, 500, 600\}$; the dotted line marks the default split $t_{\text{boundary}} = 500$. Asymmetry monotonically decreases as t_{boundary} moves to larger t, consistent with an extended window rather than a single split point; we do not interpret it as lock-in.

Steering objective. Targeted steering uses the peak-class probe trained on the same scene. The guidance gradient is FGSM-style: the sign of $\nabla_{x_t} \log P_\theta(g_{\text{target}} \mid x_t, t)$, scaled by η_t, is added to the denoising update at each step in the window (*early_only*, *late_only*, *full*). The wrong_target condition targets a graph that is uniformly sampled from the space of *non-realized* contact graphs, i.e. graphs that do not appear in the training distribution.

I ADDITIONAL CROSS-SCENE READOUT AND GUIDANCE TABLES

Scope note. Tables for the readout controls and the exact-class guidance experiments. The forward-noising row compares against forward-noised *real* trajectories, which mixes endpoint populations; the same-endpoint renoise control of Section 4.7 is the appropriate comparison. The controllability-gap tables concern exact peak-contact class targets under sign guidance, discussed in Section 5.

I.1 FULL CAUSAL TABLE

Table 13: Full flip-rate table, all four conditions, all scenes. $n = 1000$ paired trials per cell.

	$t=999$	995	990	980	950	900	**500**	200	100
BOUNDED DROP									
plausible	0.243	0.266	0.274	0.264	0.310	0.389	**0.474**	0.268	0.262
distant	0.274	0.297	0.294	0.266	0.341	0.439	0.457	0.288	0.262
random	0.072	0.079	0.079	0.073	0.101	0.126	0.086	0.048	0.041
wrong_mode	0.253	0.265	0.283	0.281	0.318	0.441	0.432	0.269	0.244
CLUSTER DROP									
plausible	0.255	0.254	0.284	0.318	0.325	0.424	**0.585**	0.445	0.369
distant	0.257	0.301	0.295	0.301	0.352	0.447	0.608	0.429	0.431
random	0.075	0.070	0.068	0.070	0.096	0.135	0.120	0.054	0.067
wrong_mode	0.273	0.280	0.285	0.280	0.356	0.458	0.596	0.442	0.416
2D DISCS									
plausible	0.120	0.133	0.138	0.134	0.171	0.243	**0.247**	0.127	0.101
distant	0.133	0.135	0.117	0.127	0.151	0.217	0.250	0.117	0.100
random	0.057	0.061	0.057	0.083	0.110	0.134	0.134	0.060	0.059
wrong_mode	0.105	0.134	0.126	0.152	0.154	0.223	0.246	0.124	0.110

I.2 FULL ARTIFACT-EXCLUSION MATRIX

Table 14: Summary of all artifact-exclusion tests across scenes.

Test	BOUNDED DROP	CLUSTER DROP	2D DISCS
Geometric baseline (window jump)	$+0.004$ vs probe $+0.087$	$+0.012$ vs $+0.163$	$+0.012$ vs $+0.283$
Forward-noising of real data (window jump; mixes populations)	$+0.045$ vs $+0.087$	$+0.133$ vs $+0.163$	$+0.075$ vs $+0.283$
Random labels (total jump)	$+0.011$ vs $+0.119$	$+0.002$ vs $+0.099$	$+0.0002$ vs $+0.051$
Balanced accuracy jump	$+0.035$	$+0.039$	$+0.031$
Macro-F1 jump	$+0.025$	$+0.031$	$+0.025$
NMI jump	$+0.076$	$+0.025$	$+0.011$
Linear probe jump	-0.019 (flat)	-0.005 (flat)	—
MLP1 probe jump	$+0.068$	$+0.076$	—
MLP3 probe jump (main)	$+0.119$	$+0.099$	$+0.051$
Coarse label jump	$+0.109$	$+0.092$	—
Medium label jump (main)	$+0.119$	$+0.099$	$+0.051$
Fine label jump	$+0.092$	$+0.015$	—
Sharpest pair-F1 interval	$t=550$-500	$t=500$-450	$t=550$-500

"—" indicates not run on that scene: the probe architecture and label granularity sweeps were executed only on the 3D scenes because each sweep requires re-training ~ 10 additional probe checkpoints; the main MLP3 on medium labels is shown for 2D to allow direct comparison with the 3D entries.

I.3 BOUNDED DROP CONTROLLABILITY-GAP DETAIL (CONSISTENT MASK)

Table 15: Controllability gap on BOUNDED DROP ($n = 1000$ samples; consistent baseline $\cap$ condition valid mask). All probe-grad conditions have $\Delta \mid \text{hit} \approx +1.0$ to $+1.5$ Hamming bits (target-directed on the hit subset), but the population mean is dragged into the negative range by the much larger non-hit subset. Both narrow late-window probe-grad rows (scale 0.02 and scale 0.05) have CIs that include zero — statistically indistinguishable from baseline-neutral; all other guided conditions (random and full) have CIs strictly excluding zero on the destructive side.

Condition	hit	$\overline{\Delta\text{Ham}}$	95% CI	pos%	neg%	$\Delta \mid \text{miss}$
baseline	0.000	$+0.000$	—	0.0	0.0	0.000
narrow $(0, 200]$, scale 0.02	0.022	-0.040	$[-0.082, +0.003]$	14.2	15.8	-0.064
narrow $(0, 200]$**, scale** 0.05	0.054	-0.049	$[\mathbf{-0.104}, \mathbf{+0.006}]$	18.8	20.5	-0.129
matched random (scale 0.05)	0.006	-0.115	$[-0.149, -0.080]$	3.1	12.2	-0.124
full $(0, 990]$ **budget-matched,** 0.01	0.054	-0.165	$[-0.233, -0.099]$	19.6	29.6	-0.259
full $(0, 990]$ over-budget, scale 0.02	0.041	-0.461	$[-0.541, -0.381]$	17.5	43.5	-0.545

I.4 BOUNDED DROP CONTROLLABILITY GAP, THREE-PANEL DETAIL FIGURE

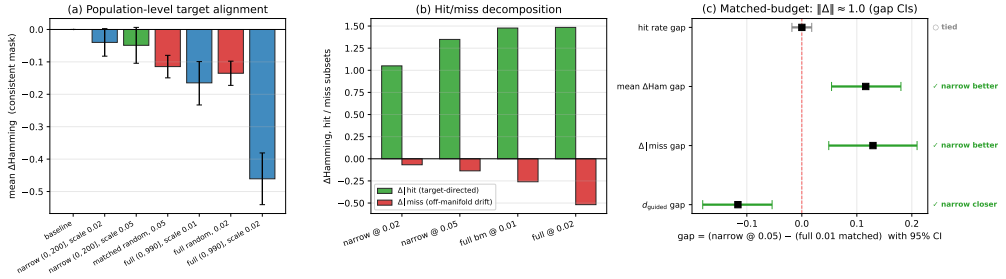

Figure 16: BOUNDED DROP controllability gap detail. (a) Mean ΔHamming under the consistent valid mask, with bootstrap 95% CI error bars. Both narrow late-window conditions (scale 0.02 and 0.05) have CIs that include zero; all other guided conditions have CIs strictly negative. (b) Hit/miss decomposition. (c) Matched-budget head-to-head at total perturbation $\|\Delta\| \approx 1$: narrow $(0, 200]$ at scale 0.05 vs full $(0, 990]$ at scale 0.01, hit-rate statistically tied, narrow significantly less destructive on every other metric. CI bands from a paired bootstrap on the tri-consistent mask, $n = 777$.

I.5 GUIDED-CONDITION PHYSICAL SANITY

For the exact-class guidance conditions we report per-condition physical-validity statistics on BOUNDED DROP. Guidance failure is about *contact targeting*, not physical breakdown of the generated trajectories: the valid-class rate (peak-contact graph lands in a kept class), the integrated penetration count, and the final-state stability score are all comparable to baseline across every guided condition.

Table 16: Per-condition physical sanity on BOUNDED DROP ($n = 1000$). The valid-class rate is the fraction of generated samples whose peak-contact graph lands in a kept (training-frequency) class. Penetration count sums non-trivial body-pair overlaps over the trajectory. Final-state stability is mean ℓ_2 velocity at the last simulator step. Guided conditions retain physical plausibility comparable to baseline, ruling out "guidance simply breaks the generated state" as the explanation for the controllability gap.

Condition	Valid-class rate	Penetration count	Final-state stability
baseline	0.866	359,238	0.002
window $(0, 200]$ @ 0.02 (narrow 0.02)	0.881	361,627	0.002
window $(0, 200]$ @ 0.05 (narrow 0.05)	0.876	363,596	0.002
window $(0, 200]$ @ 0.05 random	0.863	360,574	0.002
full $(0, 990]$ @ 0.004 (budget-matched 0.02)	0.880	354,007	0.002
full $(0, 990]$ @ 0.01 (budget-matched 0.05)	0.876	344,876	0.002
full $(0, 990]$ @ 0.02 (probe, over-budget)	0.843	335,514	0.003
full $(0, 990]$ @ 0.02 random	0.867	363,768	0.002

J CODE AND DATA AVAILABILITY

The supplementary material contains the code for every experiment in this paper. One directory holds the measurement, intervention and policy code: stochastic continuation and the Brier decomposition, paired interventions and their confirmations, label and quality audits, held-out guidance and gating policies, the supplementary sweeps and the post-hoc analyses. A second directory holds the pipeline code for denoiser training, probe training, the exact-class interventions and guidance, and the configuration controls. Result records, trained checkpoints, reverse caches, generated trajectories and simulator datasets are not included. Trial counts are stated in the text: 64 mothers with 16 branches in the branching pilot, 128 mothers with 32 branches per confirmation model, 512 paired proposals per held-out policy task, 32 paired budget blocks per gate task, and 10,000 bootstrap replicates (20,000 for the post-hoc contrasts of Appendix C).