GEOMETRY OF PHASE ORDERING IN DIFFUSION AND FLOW

Anonymous authors
Paper under double-blind review

ABSTRACT

We study same-state conditional prediction contrasts as descriptive measurements, separating the temporal concentration of prediction disagreement from terminal intervention response. The primary statistic is the appearance-minus-structural phase-center gap within a trained evaluating network. For three four-factor diffusion checkpoints, archived model-level summaries report a mean gap of 0.231 (sample standard deviation 0.010) on common trained trajectories, versus -0.0013 for untrained evaluators on the same states; this contrast is descriptive, not a causal estimate of training. A separate Gaussian-block calculation derives variance-dependent phase ordering for unnormalized population contrasts; in a learned-predictor test with five variance ratios, the reported non-null gaps have the predicted signs, whereas their magnitudes, the equal-variance control, and the archived peak positions disagree with the prediction. In SD3.5 Medium, static directions attain the same reported classification accuracy as aligned updates, and text-only features classify 24/24 queries versus 23/24. In an onset audit, eight of 18 cases pass the archived primary gate flags, the geometry-peak error is 1.375 steps versus 1.25 for a fixed-step predictor, and a routing pilot does not establish an advantage. We therefore distinguish the reported phase organization from mechanistic identification, calibrated onset prediction, and verified-budget control utility.

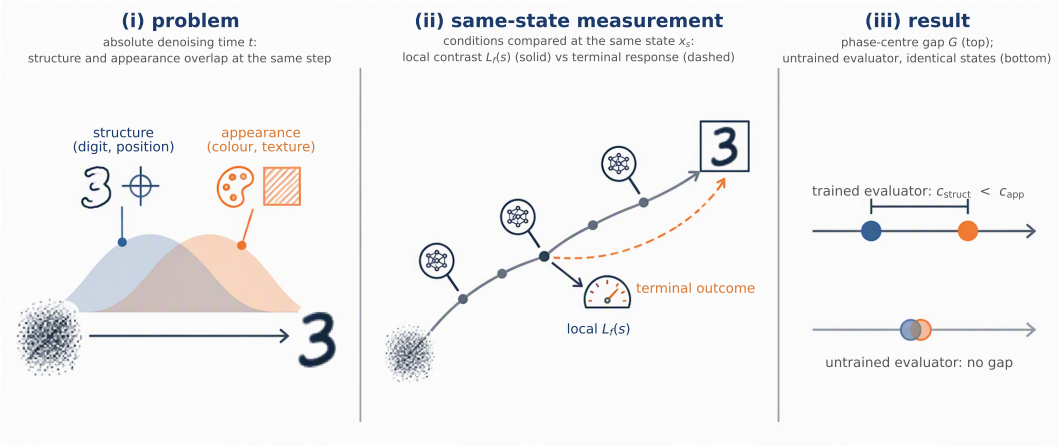
1 PROBLEM AND CONTRIBUTIONS

Diffusion and rectified-flow models admit many useful editing controls, but the time at which a prompt difference is expressed is not, by itself, a prescription for intervention. We study a deliberately narrower question: at the *same* latent state, what does the difference between conditional update fields reveal about functional regime, and what can it constrain? This distinction is important because prior editing methods use noise level, inversion, attention, or guidance to improve an edit, whereas a diagnostic must survive strong static and text-only alternatives before it is interpreted as trajectory-specific. Mistimed interventions can affect target attainment and preservation. We examine same-state conditional contrast as a candidate diagnostic (Figure 1), without claiming that it detects timing errors before generation completes or improves editing efficiency.

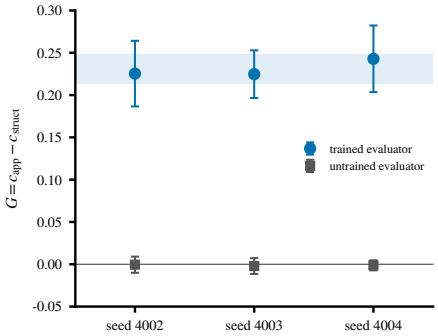
Contributions. The paper makes three limited claims, each tied to the section, figure, and table that carry its evidence.

1. **A descriptive within-network phase-ordering result.** Result 1 in Section 3.1 reports the family gap G directly for three trained checkpoints on common trained trajectories (Figure 1B, Figure 4), with an untrained evaluator on the identical states as the control. The evaluator contrast D and factor-specific contrasts are secondary controls, not substitutes for this within-network quantity.
2. **A separate idealized raw-contrast calculation.** Proposition 1 derives phase ordering for Gaussian population raw contrasts, and Lemma 1 gives a sufficient approximation condition, which the PX-2 estimates do not meet (0/20 non-null models). PX-2 compares learned raw contrasts with this prediction (Figure 5), including quantitative discrepancies. Neither result proves the renderer's normalized-L ordering.
3. **Audits limiting the control interpretation.** The static and onset comparisons in Section 4 do not establish incremental information from temporal alignment or a general onset-prediction

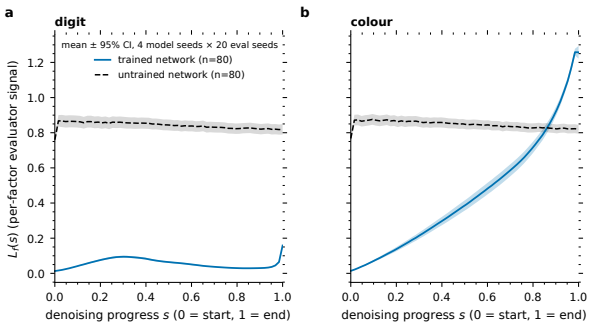
Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping 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.



(A) Problem → same-state measurement → result (schematic)



(B) Direct gap G per independently trained model



(C) Local contrast $L_f(s)$ along the generative process

Figure 1: **At the same sampling state, a trained evaluator places the appearance phase centre later than the structural one in all three independently trained checkpoints ($G \approx 0.23$), while an untrained evaluator on the identical states shows no gap; this is a descriptive within-network measurement, not an estimate of intervention utility.** (A) Three panels: *problem* – absolute denoising time is an incomplete coordinate because structural (digit, position) and appearance (colour, texture) functions overlap at the same step; *measurement design* – conditional predictions are compared at the *same* state x_s and the local contrast $L_f(s)$ is summarised by its phase centre c_f (Equation (8)), which is a local quantity distinct from the terminal response to an intervention; *result* – the family gap $G = c_{\text{app}} - c_{\text{struct}}$. (B) The measured estimand for model seeds 4002–4004 on common trained trajectories (Section 3.1): blue = trained evaluator (error bars: within-model SD over 20 evaluation trajectories; band: 90% CI over the three model seeds), grey = untrained evaluator on the identical states (descriptive: all three model means lie within ± 0.003 of zero; no equivalence test was run on this trained-trajectory series – the TOST in Section 3.1 concerns the untrained network on its *own* trajectories). Plotted from the archived per-model rows by `code/make_fig1_direct_G_models_20260926.py`. (C) The quantity that c_f summarises: on common trained trajectories the trained-evaluator signal $L_f(s)$ for digit peaks early and stays low, whereas colour rises toward the end of denoising; the untrained network (black dashed) is flat in s for both (mean $\pm$ 95% CI over model seeds $\times$ 20 evaluation seeds, Section 3.1). L_f is an evaluator-derived signal, not a latent-state coordinate. Verified range: three checkpoints of one four-factor Colored-MNIST DDPM; nothing here is an SD3.5 result or a controller.

advantage. The recorded primary-case comparison (Figure 6 and Table 7) is descriptive and does not establish calibrated coverage or verified-budget utility.

These are descriptive comparisons rather than equivalence tests or calibrated coverage guarantees. Section 2 fixes the definitions, Section 3 reports the controlled study, Section 4 reports the audits, and Section 6 states the protocol and limitations before the conclusion.

2 SETUP: SAME-STATE ESTIMAND

For a reference condition c_0, an edited condition c_a, and sampling state (x_i, t_i), we examine

$$\Delta v_{a,i} = v_\theta(x_i, t_i; c_a) - v_\theta(x_i, t_i; c_0), \qquad d_{a,i} = \Delta v_{a,i}/(\|\Delta v_{a,i}\| + \epsilon).$$

We use directional disagreement and its trajectory profile as *diagnostic* coordinates. They can identify a candidate interval, and are proposed to restrict a policy space (a utility benefit is not established); they do not identify an optimal onset without an independently measured, task-specific causal response. Same-state conditional contrast is not novel to this work: DiffEdit derives an edit mask from the difference between conditional predictions at a shared noisy state (Couairon et al., 2023), and concept-insertion curves ask when injecting a concept changes the final output (Table 11). What we add is a measured account of what this contrast does and does not identify. Its computational cost and scope are discussed in Section D.1.

The primary same-state quantity is the within-trained-network appearance-minus-structural gap G on common trained trajectory states, rather than the historical comparison of networks on their own trajectories; its positive sign supports the structural-before-appearance description in the abstract. The between-evaluator contrast $D = G(\text{trained}) - G(\text{untrained})$ at fixed trajectory states (Figure 22C) is a secondary control that isolates the evaluating-network contrast from changes in sampled state. The model-level means are 0.232 on trained states and 0.097 on untrained states ($n = 3$ independently retrained models); neither value measures intervention success or by itself identifies a causal effect of training.

2.1 SAMPLING TRAJECTORY AND SAME-STATE PROTOCOL

Let $v_\theta(x, t; c)$ denote the model output selected for a particular experiment. Noise prediction, guided flow prediction, and the actual state increment are different quantities and are not identified by notation alone. The four-factor measurements use noise-prediction disagreement, whereas the SD3.5 definition below uses a classifier-free guided prediction.

A discrete sampler is written abstractly as

$$x_{i+1} = \Phi_i(x_i, v_\theta(x_i, t_i; c)), \qquad i = 0, \ldots, N-1. \tag{1}$$

For text-to-image experiments, v_θ is the classifier-free guided SD3.5 prediction

$$v_\theta(x, t; c) = v_\theta(x, t; \varnothing) + \gamma\big(v_\theta(x, t; c, +) - v_\theta(x, t; \varnothing)\big). \tag{2}$$

Here $\gamma > 0$ is the classifier-free guidance scale, fixed per experiment (see Section I for the value used in the reported SD3.5 experiments). We first sample a source-only calibration trajectory $\{x_i^0\}_{i=0}^N$. Every condition is evaluated at the same source state,

$$v_i^0 = v_\theta(x_i^0, t_i; c_0), \qquad v_i^a = v_\theta(x_i^0, t_i; c_a), \qquad \Delta v_{a,i} = v_i^a - v_i^0. \tag{3}$$

This protocol isolates local differences between fields. Comparing two independently generated trajectories would conflate field geometry with accumulated state drift.

2.2 SCALAR DIAGNOSTICS AND DIRECTION CURVES

For a vector z, define the regularized normalization $\widehat{z} = z/(\|z\| + \epsilon)$, with $\epsilon > 0$. Suppressing the condition index within this subsection, we compute

$$L_i = \|\widehat{v}_i^a - \widehat{v}_i^0\|, \qquad r_i = \|\Delta v_i\|/(\|v_i^0\| + \epsilon), \tag{4}$$

$$p_i = \langle \Delta v_i, v_i^0\rangle/(\|v_i^0\|^2 + \epsilon), \qquad o_i = \|\Delta v_i - p_i v_i^0\|/(\|v_i^0\| + \epsilon). \tag{5}$$

The residual $\Delta v_i - p_i v_i^0$ is a regularized projection residual, not an exactly orthogonal component.

$$\langle \Delta v_i - p_i v_i^0, v_i^0\rangle = \epsilon\langle \Delta v_i, v_i^0\rangle/(\|v_i^0\|^2 + \epsilon). \tag{6}$$

The regularized prompt-delta direction is

$$d_i = \Delta v_i/(\|\Delta v_i\| + \epsilon), \qquad \|d_i\| < 1. \tag{7}$$

The resulting sequence $(d_i)_i$ traces a direction curve over the trajectory. Exact regularized identities and the unregularized comparison are in Section H.4; the temporal, aligned, static, and centered representations used for the SD3.5 controls are defined in Section H.5.

2.3 PHASE COORDINATES

For a nonnegative curve $L(s)$ on normalized denoising progress $s \in [0, 1]$, the center of mass is

$$\bar{s}_L = \frac{\int_0^{s_{\max}} sL(s)\, ds}{\int_0^{s_{\max}} L(s)\, ds}, \tag{8}$$

where $s_{\max} = 0.95$ in Factorized MNIST to avoid the terminal spike. The phase summaries $\bar{s}_L$ and $s_{1/2}$ require an integrable nonnegative curve L with finite, positive total area $\int_0^{s_{\max}} L(s)\, ds$ over the analysis window. The half-area time $s_{1/2}$ is the first progress value at which the cumulative area $\int_0^{s_{1/2}} L(s)\, ds$ reaches half of that total. Both summaries are undefined when the total area is zero (no measurable disagreement over the window). Discrete conventions and the FWHM validity rule are in Section H.4.

These coordinates answer *where disagreement is concentrated*. They do not by themselves answer how a downstream output changes under intervention.

2.4 A SAME-STATE SEPARATION PROPOSITION

Recall from Equation (3) that every comparison is evaluated at a shared calibration state x_i^0. Fix a baseline condition $c = (d, p, c_{\text{col}}, \tau)$ that specifies all four factors of Equation (12), and let c'_S (resp. c'_A) differ from c only in the structural block (d, p) (resp. the appearance block (c_{col}, τ)), as in the single-factor flips already used to define L_f in Equation (13). At a shared state $x = x_i^0$ and step $t = t_i$, define the two per-factor same-state deltas and their difference

$$\begin{aligned} \Delta_S(x,t) &= v_\theta(x,t;c'_S) - v_\theta(x,t;c), \qquad \Delta_A(x,t) = v_\theta(x,t;c'_A) - v_\theta(x,t;c), \\ \Delta(x,t) &= \Delta_S(x,t) - \Delta_A(x,t). \end{aligned} \tag{9}$$

The baseline term $v_\theta(x, t; c)$ cancels in Δ, so $\Delta(x, t) = v_\theta(x, t; c'_S) - v_\theta(x, t; c'_A)$: the same-state update direction compared directly between a structural-only and an appearance-only recondition, matching the abstract's Δ.

A Gaussian block model. As a separate idealization, let each factor block f (structural, $f = S$; appearance, $f = A$) be an independent Gaussian coordinate block of the clean signal, $X_0 \mid c \sim \mathcal{N}\big(\mu(c), \operatorname{diag}(\nu_S^2 I_S, \nu_A^2 I_A)\big)$, and let the forward process be the usual variance-preserving mixture $X_t = a_t X_0 + \sigma_t Z$, $Z \sim \mathcal{N}(0, I)$, with noise level $r_t = \sigma_t / a_t$ decreasing in t's denoising-progress coordinate s (equivalently increasing in noise level, as in Section 3's cosine schedule). Let c, c'_f differ only in block f, with mean shift $\Delta\mu_f = \mu_f(c'_f) - \mu_f(c)$, and let $\epsilon^*(x, t; c) = \mathbb{E}[Z \mid X_t = x, c]$ denote the population-optimal noise predictor that training toward Equation (13)'s objective targets.

Proposition 1 (Gaussian block model: population phase order) *Let $T > 0$. Assume $0 < \nu_S, \nu_A < \infty$, both mean shifts are nonzero, and $r(s) > 0$ is finite and strictly decreasing on the analysis interval $[0, T]$, with $G_f(r(s))$ integrable. Under the model above, for every t the population same-state noise-predictor contrast for block f is state-independent and equal to*

$$\begin{aligned} \Delta\epsilon_f^*(t) &:= \epsilon^*(x,t;c'_f) - \epsilon^*(x,t;c) = -\frac{a_t\sigma_t}{a_t^2\nu_f^2 + \sigma_t^2}\,\Delta\mu_f, \qquad \textit{so} \\ G_f(r) &:= \|\Delta\epsilon_f^*\| = \|\Delta\mu_f\|\,\frac{r}{\nu_f^2 + r^2}, \end{aligned} \tag{10}$$

evaluated at $r = r_t$, a unimodal function of r with a unique maximum at $r_{\text{peak},f} = \nu_f$. The population phase centres $\bar{s}_S^*, \bar{s}_A^*$ of $G_S(r(\cdot)), G_A(r(\cdot))$ satisfy

$$\nu_S > \nu_A \implies \bar{s}_S^* < \bar{s}_A^*, \qquad \nu_S = \nu_A \implies \bar{s}_S^* = \bar{s}_A^*, \qquad \nu_S < \nu_A \implies \bar{s}_S^* > \bar{s}_A^*. \tag{11}$$

The scope of this Gaussian idealization relative to the renderer is discussed in Section C.4. Proofs, the phase-centre stability lemma (Lemma 1), and the scope remark are in Section C.4.

Branch-mixture motivation. Projection-caustic theory describes sharp generative transitions through competition among nearby branches of the data support (Sakamoto & Sakamoto, 2026), and a conditional extension gives a heuristic for prompt-delta rotation (Section C.5). This heuristic suggests a possible source of direction changes, but does not establish that measured direction curves identify switching surfaces or functional regimes in a trained network.

3 CONTROLLED RESULTS IN FACTORIZED DIFFUSION

Large text-to-image models do not expose independent ground-truth semantic factors. We therefore begin with a controlled conditional DDPM whose data are rendered from four factors

$$\begin{aligned} &x = R(d, p, c, \tau), \\ &d \in \{0, \dots, 9\}, \qquad p \in \{1, \dots, 5\}, \\ &c \in \{1, \dots, 4\}, \qquad \tau \in \{\text{solid}, \text{stripes}, \text{dots}\}. \end{aligned} \tag{12}$$

Digit identity d and position p are designated *structural*; foreground color c and stroke texture τ are designated *appearance*. A conditional U-Net is trained with a cosine schedule and EMA weights; deterministic DDIM sampling uses a clipped $\widehat{x}_0$, and $\epsilon_\theta(x, t; c)$ denotes the corresponding noise-prediction network output. For every source trajectory and factor f, the target condition differs in that factor only. In the supplied same-state-control implementation, define $h_\eta(z) = z/\max(\|z\|, \eta)$, with $\eta = 10^{-10}$, and compute

$$L_f(t) = \|h_\eta(\epsilon_\theta(x_t, t; c'_f)) - h_\eta(\epsilon_\theta(x_t, t; c))\| \tag{13}$$

Both predictions are evaluated at the same x_t. This agrees with exact unit normalization only when both prediction norms are at least η. It specifies the supplied implementation, not a retrospective certification of every historical measurement pipeline. The primary statistic is the phase center in Equation (8); 20 source trajectories provide the seed-level statistical units.

Run validity and historical evidence. Generated adherence is 0.8625 for digit and at least 0.998 for every other factor, so digit misses the planned 0.92 stopping target, a limitation we retain in all interpretations; implementation, evaluator accuracies, and budget errors are in Section G. Along the generative process, Figure 3 shows the trained evaluator signal $L_f(s)$ for all four factors with their phase centres: the structural pair centres near $s \approx 0.43$ and the appearance pair near $s \approx 0.66$; the untrained-network panel is descriptive and does not by itself establish absence of ordering. The branching example switches the condition for the whole remaining tail; a pre-fixed single-step impulse at an early, middle, or late position (Figure 8) leaves the target factor unchanged ($O_f \leq 0.0025$, other-factor retention ≥ 0.916 over 36 cases, nominal and budget-matched), so the position of a one-step impulse alone is not what the later measurements attribute. An archived same-protocol replication on three original models exists, but its checkpoints and raw per-trajectory measurements are not recoverable; its numerical summaries are therefore omitted and not used as evidence (Section G). The secondary protocol B is reported only as a sign-consistency check (Section O).

3.1 RESULT 1: WITHIN-NETWORK FAMILY-GAP ORDERING ON COMMON TRAINED TRAJECTORIES

The same-state experiment compares trained and untrained evaluating networks on a common trajectory. For model seed m, trajectory source z, evaluating network e, and factor f, let $\phi(m, z, e, f)$ denote the mean phase center across the 20 analysis seeds. Here $e = T$ denotes the trained network and $e = U$ the untrained network. Define

$$G(m, z, e) = [\phi(m, z, e, c) + \phi(m, z, e, \tau) - \phi(m, z, e, d) - \phi(m, z, e, p)]/2,$$

$$\delta(m, z, f) = \phi(m, z, T, f) - \phi(m, z, U, f), \qquad D(m, z) = G(m, z, T) - G(m, z, U).$$

The factor-specific contrasts reported below are $\delta(m, z, f)$. They are distinct from the within-trained-network family gap $G(m, z, T)$. A positive $D(m, z)$ does not, by itself, establish $G(m, z, T) > 0$.

Holding the trajectory fixed removes sampled-state differences from the evaluating-network comparison. It does not make a within-network family gap interchangeable with its between-network change, and it does not audit the historical models.

We therefore also measure within-network ordering directly on each common trained trajectory state: structural-first means the mean trim-95% phase center of digit and position is strictly less than the mean of foreground color and texture for the same analysis seed and evaluator. Table 1 reports these counts separately from the trained-minus-untrained contrast. All three trained evaluators show structural-first ordering on 20/20 trajectories; untrained evaluators on those identical states show it on 8/20, 8/20, and 11/20. These are neither model replications nor a preregistered pass/fail criterion.

Table 1: Same-state within-network ordering on common trained trajectories. Each count is the number of analysis seeds (out of 20) with structural phase center earlier than appearance phase center. The gap is appearance minus structural phase center, averaged over those 20 trajectories.

Model seed	Trained first	Trained gap	Untrained first	Untrained gap
4002	20/20	0.225	8/20	-0.0005
4003	20/20	0.225	8/20	-0.0021
4004	20/20	0.243	11/20	-0.0012

The directly measured within-trained-network gap $G(m, \text{trained}, T)$ is 0.22535, 0.22470, and 0.24287 for model seeds 4002, 4003, and 4004, respectively (mean 0.23098, sample SD 0.01031 across three model means). On the *same trained trajectory states*, the untrained evaluator has gaps -0.00048, -0.00208, and -0.00123 (mean -0.00127, sample SD 0.00080). These are separate measurements of G, not inferences from positive D; the corresponding D mean is 0.23224 (sample SD 0.01029). A two-one-sided equivalence test on the three untrained model-level gaps measured on the untrained network's *own* trajectories ($z = \text{untrained}$, $e = U$; a different estimand from the trained-trajectory series just quoted) is equivalent to zero at margin $\delta = 0.02$ (mean -0.0002, 90% CI $[-0.0014, 0.0009]$, unit = model seed, $n = 3$), whereas the trained gaps have 90% CI $[0.214, 0.248]$ excluding zero (`results/pgc-r119-untrained-g-equivalence-20260926/`); with $n = 3$ the equivalence test has low power, so it bounds the untrained gap rather than proving its absence. Raw-row provenance and hashes are in Section G.5. This establishes a descriptive within-trained-network separation for these three checkpoints and trajectories, not an impossibility theorem, a prospective classifier, or an intervention benefit. The audit does not validate upstream L generation or checkpoint identity.

Figure 23 shows the trained-minus-untrained net contrast for each factor, evaluated on both trained-trajectory and untrained-trajectory states ($n = 20$ analysis seeds/factor per model seed). On trained-trajectory states, appearance factors (foreground color, texture) show a positive phase-center contrast in all three model seeds (fg color 0.182 ± 0.011, texture 0.192 ± 0.012; mean $\pm$ sample standard deviation, $n = 3$ model seeds) while structural factors (digit, position) show a negative contrast in all three model seeds (digit -0.047 ± 0.012, position -0.043 ± 0.018); on untrained-trajectory states the appearance-factor sign pattern is consistent across all three model seeds but smaller in magnitude (fg color 0.091 ± 0.004, texture 0.107 ± 0.018), while the structural factors are smaller still and not uniformly signed across seeds on untrained-trajectory states (position 0.010 ± 0.008, positive in all three seeds but near zero for one; digit -0.006 ± 0.016, positive in two of three seeds and negative in one). This is a descriptive, three-model-seed contrast; it does not establish a control benefit or generalize beyond the sampled model seeds. Raw-row recomputation and a measurement-convention sensitivity check are in Section G.5.

The prediction and measured gaps are plotted in Figure 5.

4 Audits and controls

The controlled study in Section 3 shows where conditional geometry is organized. This section audits what that organization licenses, using the static and text-only controls, the onset study, and the routing pilot. Each audit is reported at its resolution; none is an equivalence test.

Static and text-only controls. In SD3.5 Medium, aligned directions classify the original 16/16 edit-family decisions. This is not evidence that temporal switching is necessary: a collapsed static direction also reaches 16/16, and a frozen text-only distance reaches 24/24 across six new object–scene contexts (aligned updates: 23/24). The defensible interpretation is that conditional geometry is an auditable coordinate for functional organization: text-only and static controls match or exceed the reported classification accuracy. These controls show that the reported classification accuracy does not establish an incremental benefit from temporal alignment. Matching aggregate accuracies do not imply equivalent representations, shared predictive information, or a common mechanism. The origin of the classification signal is therefore not identified by these comparisons, and the Colored-MNIST same-state control does not resolve that question for SD3.5.

Onset audit and its limit. On a task-local teapot-to-vase onset study, the geometry interval contains the causal peak in 5/5 held-out seeds. The broader frozen multi-object test does not establish a general onset predictor: of 18 attempted cases, 10 have a valid geometry interval, 8 pass every primary quality gate, and 7 of those 8 are covered by the geometry interval; only 7/18 of *all* attempted cases are simultaneously primary-valid and covered, and this fraction includes invalid or abstaining geometry rather than silently excluding it (Figure 17, Section I). Among the eight primary-valid cases, geometry MAE is 1.375. A point predictor that always returns step 4 has MAE 1.25, which can be reconstructed from the listed causal peaks in the appendix. This calculation does not validate the archived comparator intervals or establish their identity as a magnitude-peak baseline. We therefore retain the fixed-step comparison descriptively and withhold a verified magnitude-baseline comparison pending recovery of its per-case predictions and generating code (Section I). Thus the broader test fails the prespecified superiority criterion against simple alternatives; failing to detect an improvement is not evidence that geometry and the baselines are equivalent, and the result does not justify an exact causal-onset claim. The MAE comparison is plotted in Figure 6; the audit sequence is in Figure 17 (Section I).

The SD3.5 Medium renders behind this onset analysis (source, target-only, geometry onset, causal peak k_c, delayed onset $k_c{+}1$; six cases and a three-seed repeat) are shown in Figures 11 and 12 in the appendix; they are PGC's own outputs and add no new measurements.

Directional disagreement does not determine terminal response. An elementary two-step example (Section I.3) shows that local disagreement alone does not determine terminal response without assumptions on the downstream map; it does not rule out predictive associations in a restricted model and task distribution.

A descriptive routing pilot. Averaged over both families, auto/oracle routing reaches target progress 0.4430 and background RMS 0.0760, while a fixed policy that always applies the dense-texture top-5 intervention reaches 0.4583 and 0.0477; routing is therefore not claimed to beat the best fixed policy (full pilot description in Section J).

C1′ implementation pilot. The reported SD3.5 Medium pilot is an implementation test, not evidence for or against susceptibility-based control: its original selectors and endpoint proxies did not implement the intended quantities, and the revised real- generation path has not been executed. In particular, the legacy geometry probe recorded the norm of the *unnormalized* difference between the two conditional predictions (`code/c1prime_real_adapter.py`), not the regularized directional disagreement of Equation (30); its per-step values are therefore a pre-normalization norm-difference pattern and are not measurements of L. The withdrawn tests and the complete audit are reported in Section E.

What is verified and what is not. Table 2 pairs each theoretical statement with the experiment and figure that test it, the range over which the result is verified, and the utility claim that remains unverified.

5 Related work

Two prior lines of work already use closely related measurements, so we make the comparison explicit rather than only tabular. DiffEdit derives an edit mask from the difference between conditional predictions at a *shared* noisy state (Couairon et al., 2023); same-state conditional contrast is therefore not novel to this work, and our contribution is using the resulting scalar/direction curve as a diagnostic for functional regime and policy restriction rather than for mask construction. Görgün et al. (2026) (PCI) measure concept-insertion curves: they ask at which denoising step injecting a concept through the prompt still changes the final output, which targets the same broad timing question as our regime diagnostic. Their measurement is an intervention-and-rollout on the final sample, whereas ours contrasts conditions at a shared noisy state and uses the result to argue for policy restriction. Table 11 (Appendix) compares measurement target, state-sharing method, and the hypothesis each line of work actually tests, rather than only whether a property is claimed. Broader context on guidance and its timing (Dhariwal & Nichol, 2021; Kynkäänniemi et al., 2024; Wang et al., 2024; Sadat et al., 2024), structural conditioning (Zhang et al., 2023; Mou et al., 2023), critical windows in reverse

Table 2: Correspondence between theory, experiment, figure, verified range, and unverified utility. "Verified" means recomputed from archived rows with checked hashes; it does not mean independently regenerated from checkpoints unless stated.

Theory / definition	Experiment	Figure / table	Verified range	Unverified utility
Same-state family gap G (Equation (8), Section 3.1); no theorem	Result 1: three independently trained checkpoints, 20 common trajectories each; untrained evaluator on identical states; TOST	Figure 1B, Figure 4, Table 1	Descriptive $G = 0.231$ (90% CI $[0.214, 0.248]$, $n = 3$ models); untrained G equivalent to 0 at $\delta = 0.02$; structural-first 20/20 per model	Upstream L generation and checkpoint identity not independently regenerated; no intervention or timing utility
Proposition 1: sign of population raw-Q gap vs. ν_S/ν_A	PX-2: five ratios $\times$ five trained predictors; untrained negative control; $4\times$-budget diagnostic	Figure 5	Signs and ordering 20/20 (untrained control 8/20); magnitudes outside the 95% CI at every ratio	Normalized-L ordering of the renderer; training identity vs. specification
Lemma 1: discrete certificate	Certificate recomputation with equal-variance null	Section G.6	Certificate holds in 0/20 non-null models (distinct from the 20/20 sign match); the five equal-variance null cells are assessed separately (consistent 5/5)	None claimed (lemma does not certify the observed ordering)
Local contrast vs. terminal response (two-step example, Section I.3); no theorem	SD3.5 onset audit, 18 cases, 8 primary-valid	Figure 6, Table 7	Geometry covers the causal peak in 7/8 primary-valid cases; MAE 1.375 vs. 1.25 for a fixed step-4 predictor	Onset-prediction advantage, calibrated coverage, verified-budget control
—	Static and text-only controls (SD3.5)	Section 4	Static 16/16 = aligned 16/16; text-only 24/24 vs. 23/24	Incremental information from temporal alignment
—	Routing pilot; C1′ implementation pilot	Section J, Section E	Routing $0.4430/0.0760$ vs. fixed policy $0.4583/0.0477$ (descriptive)	Any controller benefit
—	One-step impulse at early/middle/late, 36 pre-fixed fMNIST cases, nominal and budget-matched	Figure 8	Negative control only: $O_f \leq 0.0025$, retention ≥ 0.916 at every position (three seeds, descriptive)	Any timing effect of a single-step impulse; sustained-switch trade-off curves
—	SD3.5 representation probes and held-out object-replacement onset (teapot→vase, 24-step)	Withheld from the main text (source sections not compiled)	None: image-level outcomes not independently regenerated	Image-level editing success vs. preservation, fixed-onset/magnitude/t_0-sweep baselines

dynamics (Raya & Ambrogioni, 2023; Li & Chen, 2024), and activation steering (Turner et al., 2023; Li et al., 2023) is discussed in Section B.

6 Limitations and conclusion

The independent unit depends on the experiment: within each four-factor model, source trajectories are resampled jointly across factors, and between-model summaries use independently trained models as units. Causal endpoints based on CLIP are operational measures, not ground truth; no blinded human study was conducted. The checkpoints behind the earlier archived replication are not recoverable and its numerical summaries are omitted, so the primary evidence is the same-state control on separately retrained models 4002–4004 (Section 3.1); protocol B is a secondary sign-consistency check, and the historical permutation test is withdrawn as inferential evidence. The two-stage geometry-restricted protocol is a proposal for future evaluation, not a validated controller, and the branch-mixture picture is a hypothesized mechanism. The full protocol statement, the future-evaluation requirements, and the claim ledger (Table 3) are in Section D.1. Appendix Q documents H3 outputs under archived configurations and an exploratory same-state diagnostic, including realized-magnitude differences and semantic-adherence limitations.

Conclusion. The reported experiments describe phase organization in a controlled four-factor model and audit the interpretation of conditional update geometry using static, text-only, and onset comparisons. They do not establish a calibrated controller, a general impossibility of geometry-based onset prediction, or a utility benefit from restricting the policy space. Seven of eight primary-valid onset cases are covered, while only seven of eighteen attempted cases are both primary-valid and covered. The fixed-step comparison is descriptive, and the routing pilot does not establish an advantage over the best reported fixed policy choice. Reproducible measurement contracts and valid terminal-response experiments are required for stronger claims.

REFERENCES

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.

Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. DiffEdit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.11427.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *International Conference on Learning Representations (ICLR)*, 2026.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Marvin Li and Sitan Chen. Critical windows: Non-asymptotic theory for feature emergence in diffusion models. *arXiv preprint arXiv:2403.01633*, 2024.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*, 2023.

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

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)*, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations (ICLR)*, 2024.

Ryosuke Sakamoto and Kotaro Sakamoto. The geometry of phase transitions in generative dynamics via projection caustics. *arXiv preprint arXiv:2606.13191*, 2026. URL https://arxiv.org/abs/2606.13191.

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.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

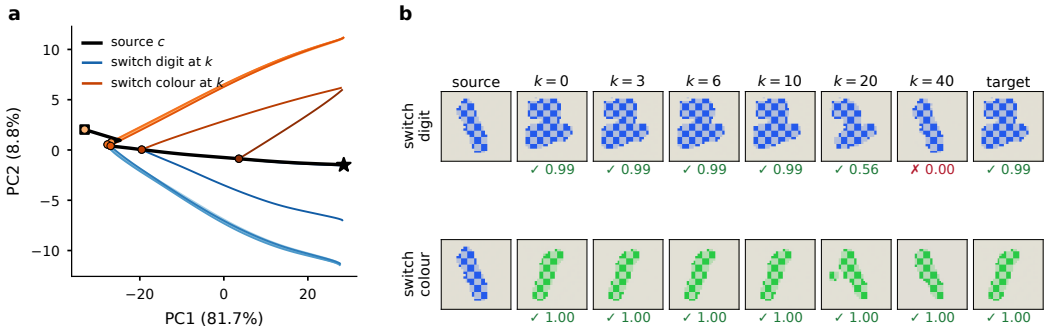


Figure 2: **Switching the digit condition after step 20 no longer changes the endpoint, whereas colour is still switchable at step 40 (one trajectory; illustrative, not an aggregate estimate).** (a) PCA of image-space DDIM states x_t (model seed 4002, one trajectory seed, 64 steps) for the source trajectory (black; square = x_T, star = x_N) and branched trajectories in which the digit (blue) or foreground colour (orange) condition is switched to a target label from step $k \in \{0, 3, 6, 10, 20, 40\}$ (dots mark branch points; light to dark = later k). (b) Endpoint x_N for each factor and k; ✓/× = the evaluator finds the target label and keeps the other factors, number = $P(\text{target})$. This terminal-response example motivates, but is not measured by, the same-state local contrast $L_f(s)$ whose aggregates are reported in Section 3.1.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Additional main-text material

A.1 PX-1: direct model-level gap audit

We recomputed the trim-95% phase centres directly from the archived same-state raw L curves, then formed $G(m, z, e)$ within each evaluator before aggregating across independent models. This is a reproducibility audit of the same experiment, not a second replication. On trained trajectory states, the primary seeds 4002–4004 have $G(m, \text{trained}, T) = 0.22535, 0.22470, 0.24287$; their model mean is 0.23098 with sample SD 0.01031. The paired evaluator contrast, exploratory seed 4005, and input hashes are in Section G.7.

Model-level plots of $G(T)$, $G(U)$, and D and the reason no numeric predicted-versus-observed plot is given are in Section G.7.

A.2 PX-2: direct Gaussian-block raw-Q test

We tested the unfitted Gaussian-block prediction in Proposition 1 with independently trained conditional noise predictors at variance ratios $\nu_S/\nu_A \in \{0.5, 0.75, 1, 1.5, 2\}$, using five model seeds per ratio and a fixed 64-level log-noise grid. The predicted raw-Q phase gaps $S - A$ in $\log r$ are $(-0.375, -0.154, 0, 0.202, 0.328)$; measured means and 95% percentile-bootstrap intervals are respectively $-0.510\,[-0.527, -0.492]$, $-0.134\,[-0.153, -0.119]$, $0.024\,[0.014, 0.034]$, $0.153\,[0.135, 0.178]$, and $0.230\,[0.203, 0.255]$ (5,000 resamples, unit = independently trained model, $n = 5$ per ratio). The reported non-null mean gaps have the predicted signs and ordering, but

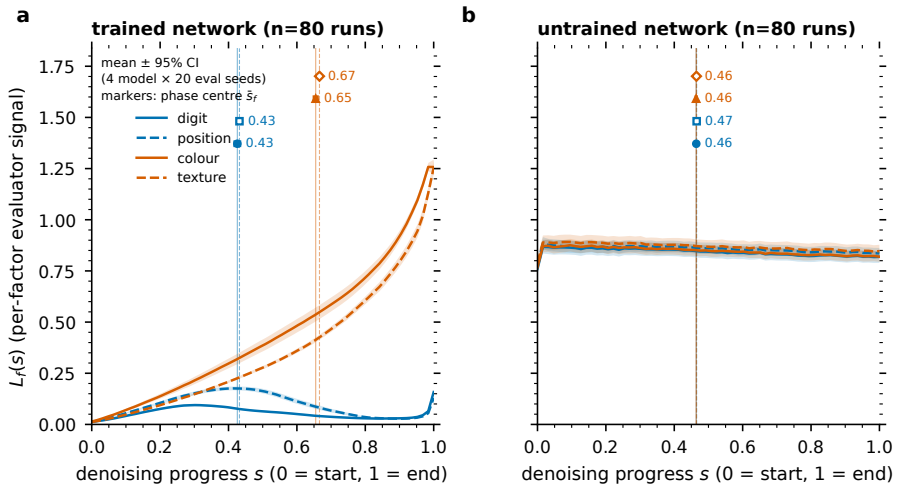


Figure 3: **On its own trajectories, the trained network's factor-wise disagreement profiles place both structural centres (digit, position) before both appearance centres (colour, texture); the untrained-network panel is a descriptive reference only.** The displayed trained-network summaries place the structural factor centres earlier than the appearance factor centres. The panels summarize four model seeds with 20 evaluation trajectories per model; these trajectories are nested within models and are not 80 independent model replications. (a) Trained network on its own trajectories. (b) Untrained network on its own trajectories. The intervals are displayed as archived summaries; no between-model uncertainty claim is made from them. Taking the model as the independent unit instead, the trained-network family gap is 0.232 (model-level 95% CI $[0.218, 0.245]$, four models) and the untrained control is -0.0005 (CI $[-0.0019, 0.0008]$, includes zero); model-level intervals are median $2.3\times$ wider than pooled-run intervals on the same archived L. The untrained panel is descriptive and does not establish a zero family gap, absence of ordering, or equivalence. These own-trajectory panels are distinct from the paired same-state evaluator comparison in Section 3.1. The signal L_f measures conditional prediction disagreement, not a latent-state coordinate or control utility. Blue curves are the structural pair and orange curves the appearance pair (solid/dashed distinguish the two factors of each pair); markers above the curves are the mean trim95 phase centre $\bar{s}_f$. The factor $\times$ step heatmap view of the same means is in Figure 21.

this is not quantitative validation: the fixed theoretical gap lies outside the reported interval at every ratio, the equal-variance control has a nonzero reported gap, and the archived peak-error summaries place the appearance argmax at the upper grid endpoint ($r = 4$) in 25/25 models and the structural argmax at the endpoint in 18/25, rather than at the predicted interior peak $r_{\text{peak},f} = \nu_f$. The causes of these discrepancies are not identified here; they are not a fitted correction or a demonstrated underfitting mechanism. A same-specification diagnostic with a $4\times$ training budget reduces the distance between the measured and predicted gap at ratio 0.5 in 5/5 seeds ($\Delta|\hat{g} - g^\star| = -0.021$, 95% CI $[-0.035, -0.009]$, unit = training model) but not at every ratio (`results/pgc-r119-px2-spec-shift-underfit-20260926/`); this is reported as observed, and hypothesis support is not judged from it. An untrained-predictor negative control matches the predicted sign in only 8/20 non-null models, against 20/20 for the trained predictors (`results/pgc-r119-px2-certificate-negctrl-20260926/`). The separately normalized L_f diagnostic is shown only for comparison: the raw-Q population curve is not a prediction for L_f. Per model, the measured $S - A$ sign matches the prediction in all 20 non-null-ratio models, but the post-hoc discrete certificate of Lemma 1 (Section G.6) holds in 0/20 non-null models ($B_S + B_A \in [1.92, 3.60]$ versus $|S - A| \leq 0.375$ predicted); the equal-variance models have zero theoretical gap and are assessed separately. Sign consistency is observed, but the lemma does not certify the ordering, and two implementation details differ from the archived specification (Section G.6). Sign agreement and the certificate are distinct statements and should not be conflated: sign agreement (20/20) is a per-model binary comparison of the measured and predicted sign of $S - A$, whereas the certificate is the sufficient error-bound condition $B_S + B_A < |S - A|$ of Lemma 1, which no non-null model satisfies (0/20). The earlier count "0/25" includes the five equal-variance ($\nu_S = \nu_A$) models,

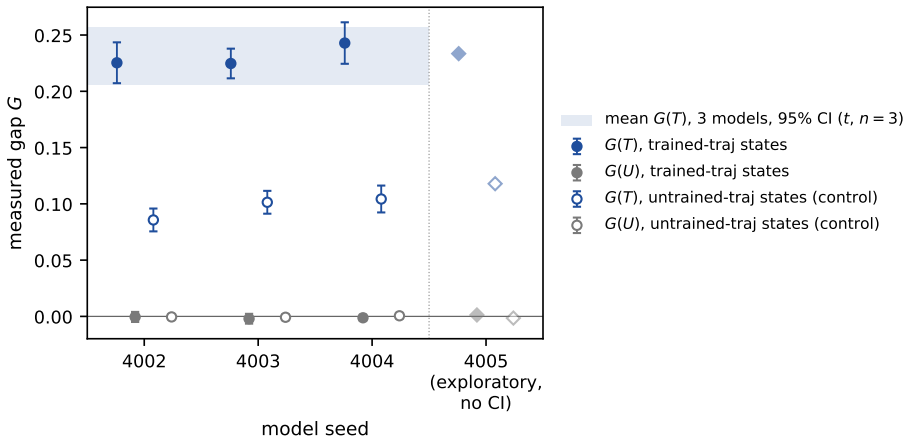


Figure 4: **Direct same-state gap measurements per model: the gap is carried by the trained evaluator, not by the states.** Estimand: the within-network family gap $G(m, z, e)$ (appearance minus structural mean phase centre) for each model seed m, trajectory source z (filled: trained-trajectory states; open: untrained-trajectory control states) and evaluator e (blue: T; grey: U); within each (m, z), $G(T)$ and $G(U)$ evaluate identical states. Error bars are within-model 95% t-confidence intervals over the $n = 20$ analysis seeds of that model (`results/pgc-r119-untrained-g-equivalence-20260926/`); the shaded band is the 95% t-CI of the mean trained-trajectory $G(T)$ over the three primary model seeds (4002–4004, $n = 3$). Seed 4005 is exploratory and has no within-model CI (no per-model dispersion archived). No equivalence (TOST) band is drawn: the TOST in the text applies to the untrained network's own-trajectory series only. Verified range: a reproducibility audit of the archived L curves (script `code/figures/px1_fig_G_v2.py`; input hashes in the sidecar JSON; points agree with `results_pgc_r73_e2_bundle_20260924/` to $< 10^{-15}$); it does not validate upstream L generation or checkpoint identity, and it is not a control-utility estimate.

for which the theoretical gap is zero and a strict-order certificate is undefined; the precise statement is 0/20 non-null models, with the five null cells assessed separately. Finally, no quantitative bridge between the population raw contrast Q_f, the normalized diagnostic L_f, the evaluator gap G, and the downstream task susceptibility $\chi(O)$ has been measured in this work; the PX-2 sign and ordering agreement is reported as an observation, not as a validated mapping from the Gaussian model to the renderer. This toy test also does not certify the Factorized-MNIST result or intervention utility.

B Extended related work

Guidance and conditional control. Classifier guidance steers sampling with gradients of a noisy-image classifier (Dhariwal & Nichol, 2021), and the score-SDE framework casts sampling and conditional generation as reverse-time stochastic or probability-flow dynamics (Song et al., 2021); latent diffusion and EDM define the latent-space and preconditioning conventions used by many modern samplers (Rombach et al., 2022; Karras et al., 2022). Several works already make guidance strength time-dependent: applying classifier-free guidance only within a limited noise interval (Kynkäänniemi et al., 2024), comparing monotone and other weight schedules (Wang et al., 2024), and annealing the condition to recover diversity (Sadat et al., 2024). Structural conditioning adds trained control branches or adapters (Zhang et al., 2023; Mou et al., 2023), and posterior sampling injects measurement-consistency gradients throughout the trajectory (Chung et al., 2023). Our work differs in what selects the window: rather than a fixed or tuned schedule, the gate is chosen from the measured same-state condition contrast, and the question is whether that measurement predicts where intervention matters.

Critical windows in reverse dynamics. Beyond the regimes discussed in Section 5, reverse diffusion has been analyzed as spontaneous symmetry breaking, with an early phase in which

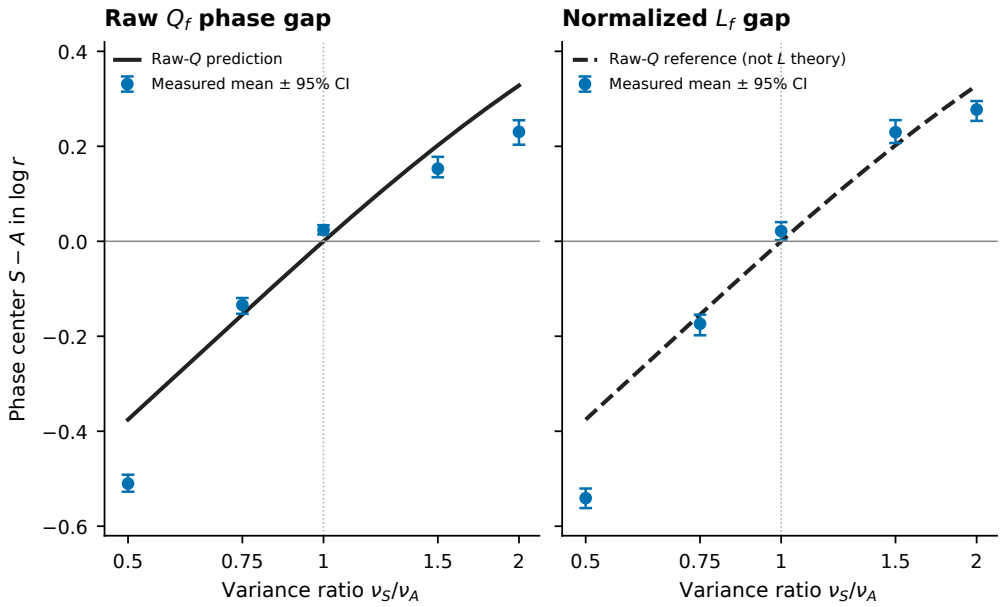


Figure 5: **An unfitted Gaussian raw-Q prediction reproduces the four non-null signs and their ordering (20/20 models) but certifies none of the magnitudes (0/20); an untrained negative control matches only 8/20 signs.** Left: prespecified Gaussian-block curve (Proposition 1) and measured means with model-bootstrap 95% intervals; right: normalized-L gaps, for which the dashed raw-Q curve is a reference, not a prediction. Independent unit: five independently trained models per variance ratio on the same 64-level log-noise grid. Verified range: the run receipt and output hashes were checked and the discrete certificate of Lemma 1 was recomputed with an untrained-predictor negative control (trained: sign match 20/20, certified 0/20; untrained: sign match 8/20; `results/pgc-r119-px2-certificate-negctrl-20260926/summary.json`), but upstream training identity is not independently verified and the phase-centre stability condition holds in 0/20 non-null models (the five equal-variance models have zero theoretical gap and are assessed separately; Section A.2). The population raw-Q result does not validate the renderer's normalized-L ordering.

trajectories commit to a mode (Raya & Ambrogioni, 2023), and non-asymptotic theory characterizes critical windows in which specific features of the sample emerge (Li & Chen, 2024). These results motivate looking for time-localized sensitivity; we do not test their theoretical predictions directly, but use them as context for an empirical condition-contrast diagnostic.

Steering internal representations. Outside image generation, inference-time steering adds directions to intermediate activations: activation addition uses prompt-pair differences (Turner et al., 2023), contrastive activation addition averages such differences over many examples (Panickssery et al., 2023), inference-time intervention shifts selected attention heads (Li et al., 2023), and representation engineering reads and controls population-level representations (Zou et al., 2023). Our controller shares the idea of intervening with a contrast direction at inference time, but acts on the sampling trajectory of a generative flow and is restricted in time by the measured gate.

C Conditional Branch Asymptotics and Switching Structure

This appendix records the mathematical motivation more explicitly. It is a local asymptotic model, not a theorem about the trained SD3.5 network.

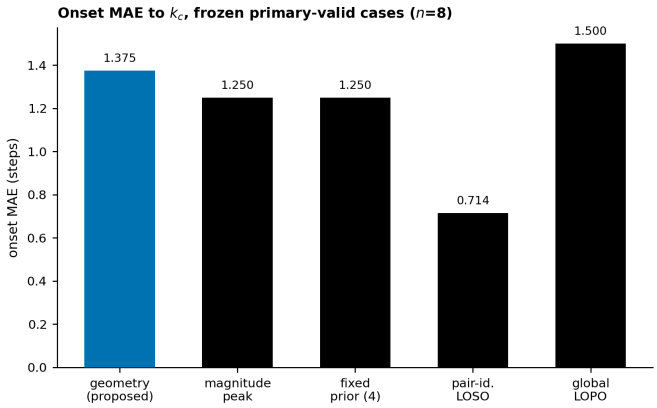


Figure 6: **Geometry covers the observed response peak in seven of the eight primary-valid onset cases but does not beat a fixed-step comparator on them.** Geometry MAE is 1.375, compared with 1.25 for a point predictor fixed at step 4. The case table supports this point-prediction comparison; the archived magnitude comparator remains unverified. This is a descriptive comparison, not an equivalence test.

C.1 Conditional free energy and branch weights

Let a noisy conditional density be represented locally by branchwise contributions

$$p_s(x \mid c) \simeq \sum_{j=1}^{J} A_j(x, s; c) \exp\left[-\frac{S_j(x, s; c)}{\tau(s)}\right]. \tag{14}$$

Its free energy is

$$F_c(x, s) = -\tau(s) \log p_s(x \mid c), \tag{15}$$

and the normalized branch responsibilities are

$$w_j(x, s; c) = \frac{A_j \exp(-S_j/\tau)}{\sum_\ell A_\ell \exp(-S_\ell/\tau)}. \tag{16}$$

Differentiating gives a score or velocity built from branch contributions plus amplitude corrections. Suppressing parameterization-dependent scalars,

$$v_c(x, s) \simeq \sum_j w_j(x, s; c) V_j(x, s; c) + R_A(x, s; c), \tag{17}$$

where R_A collects derivatives of the amplitudes. For a source and target condition at the same state,

$$\Delta v_a \simeq \sum_j (w_j^a - w_j^0) V_j^0 + \sum_j w_j^a (V_j^a - V_j^0) + (R_A^a - R_A^0). \tag{18}$$

The first term is prompt-induced responsibility reweighting; the remaining terms allow the prompt to change branchwise vector fields and amplitudes.

For two shared branches, V_1, V_2, the reweighting term reduces to

$$\Delta v_a \simeq (w_1^a - w_1^0)(V_1 - V_2). \tag{19}$$

In the two-branch approximation, reversing the nonzero scalar responsibility difference reverses the delta direction when the branch difference is held fixed and nonzero. With more than two branches, changing coefficients can rotate the delta direction, but need not do so. Rotation requires the resulting nonzero vector combinations not to remain on the same positive ray. Coefficient changes alone provide no such nondegeneracy guarantee. A broad magnitude profile can coexist with direction rotation.

C.2 SWITCHING SURFACES

For two branches define the effective energy gap

$$\Delta\mathcal{E} = S_1 - S_2 - \tau \log(A_1/A_2). \tag{20}$$

Then $w_1 = (1 + e^{\Delta\mathcal{E}/\tau})^{-1}$ and

$$\nabla_{(x,s)} w_1 = -w_1(1 - w_1)\nabla_{(x,s)}(\Delta\mathcal{E}/\tau). \tag{21}$$

The factor $w_1(1 - w_1)$ is maximal at $w_1 = w_2 = 1/2$. The switching band is a finite-temperature neighborhood of the preimage $\Delta\mathcal{E} = 0$. Projection caustics supply a geometric mechanism by which nearby support branches become simultaneously relevant (Sakamoto & Sakamoto, 2026).

Condition perturbations and state perturbations probe different Jacobian blocks:

$$\underbrace{\nabla_x \widehat{v}(x, s; c)}_{\text{state-local instability}} \qquad \text{versus} \qquad \underbrace{\widehat{v}(x, s; c_a) - \widehat{v}(x, s; c_0)}_{\text{finite condition-local contrast}}. \tag{22}$$

The former motivates state-instability diagnostics. The norm of the finite condition-local contrast is the scalar directional disagreement L_a; the displayed vector contrast is not itself that scalar. These measurements probe different changes and need not peak at the same step.

C.3 WHY REPRESENTATION AND CAUSALITY CAN SEPARATE

For differentiable controlled step maps $F_j(x, u)$, evaluated along the reference trajectory with fixed sampling randomness, define $J_j = \partial_x F_j(x_j^0, 0)$ and $B_i = \partial_u F_i(x_i^0, 0)$. Let $\mathcal{A}_i$ denote the ordered composition of subsequent state Jacobians, from step $N - 1$ down to step $i + 1$, with the identity used when no subsequent step remains. Then

$$\chi_i(O) = \nabla O(x_N^0)^\top \mathcal{A}_i B_i. \tag{23}$$

Exactly normalized directional disagreement depends only on the angle between nonzero local fields. The regularized quantity L_i also depends on their norms. Neither quantity contains the subsequent state Jacobians or the terminal objective gradient. This distinction explains why local disagreement alone does not determine terminal response; it does not prove that L_i identifies a functional regime.

C.4 MOVED FROM THE MAIN TEXT: GAUSSIAN-BLOCK PROOFS AND SCOPE

This subsection analyzes a separate Gaussian idealization of same-state conditional prediction contrasts. It is not an exact representation of the Factorized-MNIST renderer. Independence of semantic rendering factors does not imply independent Gaussian coordinate blocks in image space.

The population calculation concerns the norm of an unnormalized conditional noise-prediction difference. The empirical result concerns a difference between separately normalized predictions. We therefore use the Gaussian model to state a testable result for a distinct estimand, not to derive the observed renderer ordering. The untrained-evaluator comparisons are reported descriptively; a small mean gap does not establish absence or equivalence.

A population-level claim about which sign the same-state gap must take cannot be read off the additive-conditioning, zero-mean-at-initialization argument alone: that argument bounds the *untrained* control, not the *trained* ordering, and stating the trained ordering as an assumption (as an earlier version of this proposition did) merely restates the phenomenon being explained. We instead derive the sign and the location of the same-state gap from an explicit, non-circular generative model, and give a separate, assumption-checkable bound on how far an empirical phase-centre estimate can drift from its population value.

Proof of Proposition 1. Gaussian conditioning gives $\epsilon^*(x, t; c) = \sigma_t(x - a_t\mu(c))/(a_t^2\nu_f^2 + \sigma_t^2)$ in block f, which yields Equation (10) after changing only that block's mean. Differentiating $r/(\nu_f^2 + r^2)$ gives $(\nu_f^2 - r^2)/(\nu_f^2 + r^2)^2$, hence its unique maximum at $r = \nu_f$. For $\nu_S > \nu_A$, the ratio of the two positive curves is

$$R(s) := \frac{G_S(r(s))}{G_A(r(s))} = \frac{\|\Delta\mu_S\|}{\|\Delta\mu_A\|} \frac{\nu_A^2 + r(s)^2}{\nu_S^2 + r(s)^2}. \tag{24}$$

The last fraction is strictly increasing in $r > 0$, so $R(s)$ is strictly decreasing. Let P_A have density $G_A(r(s))/\int_0^T G_A(r(u))\,du$. Then $\bar{s}_S^* - \bar{s}_A^* = \mathrm{Cov}_{P_A}(s, R(s))/\mathbb{E}_{P_A}R(s) < 0$: the covariance is strictly negative because the density is positive almost everywhere and s and $R(s)$ have opposite strict order. Interchanging S, A proves the reverse case, while equal variances make the curves proportional and their centres equal. □ Strict decrease (or nonincrease with genuine variation on a set of positive measure) is essential for a strict centre order: a constant schedule $r(s) \equiv r_0$ makes both curves constant and both centres $T/2$ even when $\nu_S \neq \nu_A$. The peak locations $r = \nu_f$ need not lie inside the observed schedule range; the centre-order result does not require them to be observed. Within this Gaussian model, the variance parameters determine the direction of population raw-contrast phase ordering. These parameters have not been identified with measured quantities of the Factorized-MNIST renderer. The proposition therefore does not explain that renderer's normalized-disagreement ordering without additional assumptions.

Lemma 1 (Phase-centre stability) *Let $l, \hat{l} \geq 0$ be integrable on $[0, T]$ with $\int_0^T l\,ds \geq m > 0$ and $\int_0^T |\hat{l} - l|\,ds \leq \epsilon < m$. Then $\bar{s}(\hat{l})$ is well-defined and*

$$|\bar{s}(\hat{l}) - \bar{s}(l)| \leq \frac{T\epsilon}{m - \epsilon}. \tag{25}$$

Consequently, if factor $f \in \{S, A\}$'s empirical curve $\hat{l}_f = \|\Delta\hat{\epsilon}_f\|$ (from a trained network) satisfies $\int_0^T |\hat{l}_f - G_f(r(\cdot))|\,ds \leq \epsilon_f < m_f \leq \int_0^T G_f(r(s))\,ds$, and $\frac{T\epsilon_S}{m_S - \epsilon_S} + \frac{T\epsilon_A}{m_A - \epsilon_A} < |\bar{s}_S^* - \bar{s}_A^*|$, then $\mathrm{sign}(\bar{s}(\hat{l}_S) - \bar{s}(\hat{l}_A)) = \mathrm{sign}(\bar{s}_S^* - \bar{s}_A^*)$: the population order in Equation (11) survives estimation error, with an explicit, checkable margin, rather than being assumed to transfer.

Proof. Write $M = \int_0^T l(s)\,ds$, $\widehat{M} = \int_0^T \hat{l}(s)\,ds$, and $\mu = \bar{s}(l)$. Since $\widehat{M} \geq M - \epsilon \geq m - \epsilon > 0$,

$$\bar{s}(\hat{l}) - \mu = \frac{\int_0^T (s - \mu)(\hat{l}(s) - l(s))\,ds}{\widehat{M}}. \tag{26}$$

Both s and μ lie in $[0, T]$, so the numerator has absolute value at most $T\epsilon$. The final sign statement follows by applying this bound to both factors and the triangle inequality. □ Proposition 1 predicts the sign and approximate location of the population same-state gap from ν_S, ν_A; Lemma 1 gives a verifiable sufficient condition, stated in estimation error rather than in the conclusion, for that population order to be visible in a finite, trained estimate. The proofs above concern the unnormalized population contrast only; they do not establish the same order for the normalized diagnostic L_f or a causal timing effect.

Remark 1 (Scope: what this does and does not explain) *The proposition concerns the unnormalized population contrast $Q_f = \|\Delta\epsilon_f^*\|$, whereas the renderer experiment uses the difference between two separately normalized predictions. Separate normalization does not normalize the difference vector by its own magnitude.*

A lower bound on prediction norms alone is insufficient to transfer phase ordering. For example, two positive scalar predictions above the normalization floor have identical normalized values even when their raw difference is nonzero. A transfer result would require additional restrictions on directions, relative norms, and their temporal variation. No such result is proved or verified here. The phase-centre stability lemma applies to approximation of the raw population curves, not automatically to normalized L_f. Failure of its sufficient condition means that this certificate is unavailable; it does not by itself refute order preservation. In PX-2 the condition fails in all 25 trained models (Section A.2), so the observed sign agreement there is not certified by this lemma.

A zero-mean matched-untrained control, of the kind the previous version of this proposition proved directly, does not *by itself imply a zero same-state phase gap, because the phase centres in Equation (8) are built from* norms*, and $\mathbb{E}[\|\cdot\|] \neq \|\mathbb{E}[\cdot]\|$ in general. Concretely (a schematic illustration, distinct from Equation (9)'s Δ_S, Δ_A), let $Z \in \{-1, +1\}$ with equal probability and, for $s \in [0, 1]$, $\delta_S(s) = Zse_1$, $\delta_A(s) = Z(1 - s)e_2$ with $e_1 \perp e_2$. Then $\mathbb{E}[\langle \delta_S(s) - \delta_A(s), u\rangle] = 0$ for every fixed direction u and every s — the untrained-control conclusion holds exactly — yet $\|\delta_S(s)\| = s$ and $\|\delta_A(s)\| = 1 - s$ give phase centres $\bar{s}_S = 2/3$, $\bar{s}_A = 1/3$, so the appearance-minus-structural gap $\bar{s}_A - \bar{s}_S = -1/3$ is fixed and nonzero on every realization. This is why the vanishing-expectation*

statement alone, without a population model for the norms as in Proposition 1, has no implication for whether a trained same-state gap should be zero, positive, or negative.

C.5 BRANCH-MIXTURE MOTIVATION

Projection-caustic theory describes sharp generative transitions through competition among nearby branches of the data support (Sakamoto & Sakamoto, 2026). A conditional extension gives a useful heuristic for prompt-delta rotation. For this heuristic, assume a finite collection of corresponding branches across conditions, positive temperature, positive branch amplitudes where logarithms are used, and differentiability of the energies, amplitudes, and branch fields on the region considered. The displayed mixture is an approximation; no residual bound or verification of branch correspondence is supplied for the trained networks. Consequently, the switching calculation is not a guarantee about their measured direction curves or terminal responses. Suppose, under these assumptions, that the conditional field locally admits the following approximation.

$$v_c(x,s) \approx \sum_{j=1}^{J} w_j(x,s;c) V_j(x,s;c), \qquad w_j = \frac{\exp[-E_j(x,s;c)/\tau(s)]}{\sum_\ell \exp[-E_\ell(x,s;c)/\tau(s)]}. \tag{27}$$

Then, at a shared state,

$$\Delta v_a \approx \sum_j \big(w_j(c_a) - w_j(c_0)\big) V_j(c_0) + \sum_j w_j(c_a)\big(V_j(c_a) - V_j(c_0)\big). \tag{28}$$

If a prompt mainly reweights approximately shared branches, the first term dominates. For two branches with $w_1 = \sigma(-\Delta E/\tau)$,

$$\partial_s w_1 = -w_1(1-w_1)\,\partial_s(\Delta E/\tau), \tag{29}$$

so the coefficient $w_1(1-w_1)$ multiplying the responsibility-change rate is largest near a branch tie; the overall rate $\partial_s w_1$ also depends on $\partial_s(\Delta E/\tau)$, which this coefficient does not control, so the actual rate of change need not be largest at the tie. The normalized delta direction can rotate when the vector combination selected by the prompt changes, even if scalar delta magnitude stays broad.

This heuristic suggests a possible source of direction changes, but does not establish that measured direction curves identify switching surfaces or functional regimes in a trained network. It does *not* imply that every empirical direction curve requires temporal alignment for classification, nor that a responsibility switch coincides with maximal downstream susceptibility. Equation equation 36 and the static/text representation controls reported in the supplement are therefore necessary companions to the branch-mixture picture. A fuller asymptotic derivation and the limitations of the approximation appear in Section C.

D STATISTICAL PROTOCOLS

D.1 MOVED FROM THE MAIN TEXT: PROTOCOL, LIMITATIONS, AND CLAIM LEDGER

The protocol does not require retraining the generative model. However, evaluating source and target conditions at shared states can require additional model calls. We have not established its incremental utility or total computational advantage over alternative diagnostics. Our contribution is an audit of the reported measurements, not a validated online timing-error detector. We restrict the present study to same-state conditional contrast rather than comparing it with every alternative intervention signal.

The independent unit depends on the experiment. Within each four-factor model, source trajectories are resampled jointly across factors; between-model summaries use independently trained models as units. In the SD3.5 query studies, repeated seeds do not create new prompt pairs. Generalization to new queries requires query- or prompt-pair-level inference. A seed-level routing analysis instead concerns the fixed query set, and the raw paired differences needed to reconstruct that analysis are unavailable. For the causal study, invalid FWHM intervals are counted as abstentions/failed validity, and endpoint and background quantities use the frozen definitions reproduced in the appendix. The reported follow-up analyses were *pre-specified for this follow-up*; we do not call them preregistered because no timestamped immutable protocol evidence is included in this submission. Causal endpoints based on CLIP are operational measures, not ground truth; no blinded human study was conducted.

The checkpoints of the three models behind the earlier archived replication are no longer recoverable; a same-state control was therefore attempted on separately retrained 4-factor models. Initial attempts at seeds 4001 and 4002 failed after three consecutive launch timeouts, and an initial run of seed 4003 diverged (NaN) before a retry (instance 51794853) completed all 30 epochs without the earlier divergence (development loss $1.117 \rightarrow 0.00672$, checkpoint sha256 verified). A subsequent retry of seed 4002 also completed and passed integrity verification. The reported same-state control (Section 3.1) therefore uses all three retrained checkpoints, model seeds 4002, 4003, and 4004, and is a 3-model-seed descriptive contrast on separately retrained checkpoints, not a control of the three original models behind the excluded historical replication. Our primary independent replication is the same-state control (Section 3.1): three independently retrained model seeds (4002, 4003, 4004) evaluated at identical states, with raw per-trajectory reproducibility confirmed (480/480 rows, max abs. error 4.7×10^{-16}). The original same-protocol 3-independent-seed 4-factor rerun, which reused the original 20-trajectory protocol on freshly trained models, is excluded from the reported evidence (Section G): its checkpoints are no longer recoverable, so independent measurement-layer verification is not feasible, and a transcription audit from the committed summary statistics establishes reporting consistency only. The separately versioned three-model Colored-MNIST replication (B) is a secondary sign-consistency check under a different evaluation protocol; it is not a designed negative control. Its mean gap 0.00111 ± 0.00023 is about two orders of magnitude below the within-network gap G of Result 1 (model mean 0.23098), so it supports sign consistency of the ordering across independently trained models but not reproduction of that magnitude; it is not compared with the historical single-model estimate, which is no longer reported. The two protocols differ in conditioning, tail handling, smoothing, and time coordinate, and a full sensitivity analysis reconciling them is deferred. We do not use protocol B as the primary evidence for functional-regime separation. The historical four-seed, four-family permutation test (Section D) permuted labels independently within each seed. We withdraw its use as inferential evidence. A query-shared implementation is supplied, but the frozen prediction table is unavailable and the corrected test has not been applied to the reported experiment. We report no corrected significance claim for that experiment. A complete per-case table for all 18 multi-object trials (selection rationale, interval width, coverage, and the fixed-policy baseline) is maintained in the repository (`C1_allcase_reanalysis_20260908.md`) but is deferred from the main text under the current length budget.

The two-stage protocol is a proposal for future evaluation, not a validated controller. For a future test, the task loss, baseline policy, query distribution, seed distribution, and a positive practical superiority margin must be fixed before outcomes are examined. The estimand is the expected baseline loss minus the expected loss under geometry-restricted selection, evaluated on the same queries and seeds. Superiority requires a confidence interval whose lower endpoint exceeds the pre-fixed practical margin. These choices have not been fixed for a completed valid comparison in the present paper, and the routing pilot does not establish this utility claim. The branch-mixture picture (Section C.5) is a *hypothesized mechanism* for why such a restriction might work, not a demonstrated one: local geometry (L_i, a contrast between conditional fields) and downstream susceptibility (χ_i, a derivative of the task objective through the remaining sampler Jacobian) are linked only through an unverified mixture-model assumption, and our experiments do not establish their equivalence. Stronger claims require a fully crossed, held-out policy experiment with frozen prompt/latent/seed manifests, raw per-case metrics, independent evaluation, and pair-clustered inference. The C1′ outputs described above are retained only as an implementation pilot. Correcting the metric and policy definitions, verifying their end-to-end execution, and freezing a new evaluation protocol are prerequisites for a future hypothesis test. None of the scientific claims in this paper requires C1′ to succeed.

E C1′ IMPLEMENTATION-PILOT AUDIT

The reported SD3.5 Medium pilot comprises 10 development prompt pairs, 5 seeds, and 4 nominal policy labels, yielding 200 reported outputs. An implementation audit found that its original policy selectors and endpoint columns did not implement the intended geometry, susceptibility, target-progress, and background-preservation quantities. The original endpoint proxies were pixel MAE to the source image and the output image's pixel standard deviation; the nominal geometry and susceptibility policies used fixed schedules.

We therefore withdraw the direction tests, the six-contrast significance summary, and the label "Inconclusive" as verdicts on the intended C1′ hypothesis. These outputs document an implementation

Table 3: Claim ledger. "Open" is a result, not a missing-data imputation.

Claim	Status and evidence
Within-trained-network phase organization	Descriptively supported by the reported direct family gaps G for model seeds 4002, 4003, and 4004 (Section 3.1): model mean 0.23098, sample standard deviation 0.01031. The evaluator contrast D is secondary. Aggregation checks do not validate upstream conditional predictions or checkpoint identity. Historical summary-only comparisons are not additional measurement-verified evidence.
Trajectory-specific SD3.5 family information	Open: signal matches text/static controls (text 24/24; aligned 23/24).
General causal-onset superiority	Fails prespecified criterion (8/18 pass all gates; 7 of those 8 covered; MAE 1.375 vs 1.25).
Family routing beats best fixed policy	Open; pilot is not a fully crossed evaluation.
C1′ geometry-versus-susceptibility comparison	Not yet evaluated with valid measurements. The reported generation pilot is an implementation test; no hypothesis-test verdict is reported.

pilot of the generation and prompt-switching path, not evidence for or against susceptibility-based control. A subsequent adapter revision renames the legacy proxies and adds measurement hooks, but its real-generation path has not been executed for the results reported here. The revised adapter is also excluded from scientific evaluation. Its geometry probe records a conditional-delta norm rather than normalized directional disagreement: the logged value is $\|\epsilon_{\mathrm{tgt}} - \epsilon_{\mathrm{src}}\|$, the norm of the unnormalized difference between the two conditional predictions, with no per-field regularized normalization h_η, so it is not the quantity in Equation (30) and its step profile is a pre-normalization norm-difference pattern rather than an L curve. Its response probe compares decoded one-step previews rather than terminal objective responses, and falls back to the geometry norm when the semantic evaluator is unavailable.

Its budget inputs record successive latent increments rather than online source–target field differences. The logged fractional coefficients are not applied by the full prompt-switch operation, so this quantity is not the field-space budget defined in the main text. The semantic-progress call also lacks the source and target reference-image normalization used in the onset definition.

The probe records post-step states with callback timesteps. Their correspondence to pre-forward intervention indices, and the actual conditions used by each forward call, have not been validated end to end. The revised real-generation path has not been executed for any scientific result reported here. No result from that revision is included as scientific evidence. We also exclude the single-step sensitivity pilot from evaluation of terminal task susceptibility. That script records pixel MAE relative to a source image rather than the signed response of a target-specific terminal objective. Its recorded budget is constructed from fixed design constants, and its step-end prompt switch affects a subsequent forward call. Successful generation must therefore be distinguished from measurement validity and hypothesis support.

F Finite-window factor-response diagnostic

The previously reported numerical summaries and figures for this analysis are excluded from the scientific evidence because their measurement provenance and regeneration contracts have not been verified. Archived summaries alone are not treated as completed evidence. No conclusion about predictive accuracy, mechanism, or intervention utility relies on this analysis.

F.1 Blocked permutation tests

The historical four-seed analysis permuted query labels independently within each seed. We do not use that analysis as inferential evidence because repeated evaluations of the same query did not share one label assignment.

The supplied query-shared implementation uses one assignment across seeds. Its interpretation requires fixed predictions, complete query-by-seed observations, identical query labels across seeds, and exchangeability of query-label assignments under the stated null. Construction tests check the algorithm, not these scientific assumptions.

The frozen prediction table is unavailable in the supplied analysis package, and the corrected test has not been applied to the reported experiment. We therefore report no corrected significance claim for that experiment. Without independently verifiable prospective protocol evidence, the multi-context comparisons are presented descriptively rather than certified as confirmatory analyses.

The Holm calculation on archived C1-prime inputs checks numerical arithmetic only. Those inputs do not implement the intended control hypothesis and provide no hypothesis-test verdict.

F.2 Bootstrap and sign tests

Factorized MNIST uncertainty resamples source trajectories jointly across factors. Controller confidence intervals resample seed clusters so that the morphology and texture query from the same seed remain paired. Exact sign-flip tests operate on context- or pair-level mean differences. Wilcoxon $p = 0.00390625$ is the minimum one-sided value attainable with eight nonzero paired differences and occurs because all eight background differences favor auto routing; this value treats the two families within each seed as independent paired differences, which the main text does not assume, so we report it here only as a descriptive fact about the statistic and not as the inferential result.

If four independent seed-level paired signs all favor auto routing, the exact one-sided sign-test probability is 0.0625. The signs and their pairing have not been independently reconstructed from the supplied records. This value is conditional arithmetic, not an inferential result for the archived pilot. The archived bootstrap interval and target-progress significance test are likewise not used for inference.

F.3 Frozen versus exploratory analyses

Frozen analyses include original per-geometry pass rules, the balanced multi-context text distance, residual-aligned linear control, the final causal quality gates and baselines, and the matched controller budget. The following are exploratory or diagnostic: the best single step, individual encoder components in the four-query text pilot, text/update fusion calibration, quality-only post-hoc causal summaries, and uniform-onset tail probabilities. We report them because they explain the data, but they do not replace the frozen decision.

G Factorized MNIST Details

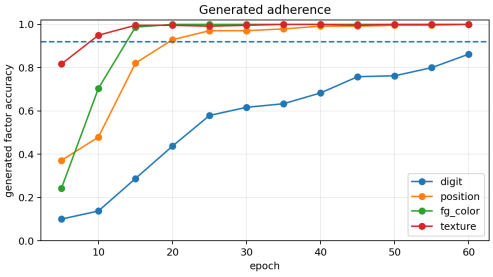
Table 4: Factorized MNIST evaluator quality and generated adherence.

Factor	Evaluator accuracy	Generated adherence
Digit identity	0.9413	0.8625
Position	0.9988	1.0000
Foreground color	1.0000	1.0000
Texture	0.9984	1.0000

The corresponding quantitative table (directional disagreement versus finite-window target-factor gain) is omitted because its measurement and regeneration contract remains unverified.

G.1 Moved from the main text: implementation and run validity

Implementation and reproducibility. *Data and preprocessing.* Each MNIST digit is resized to 26×26 (bilinear), binarized at 0.15 and dilated with a 3×3 window, and placed on a 32×32 RGB canvas at one of five positions; texture modulation is then applied, and foreground/background colors are composited and rescaled to $[-1, 1]$. Renderer randomness is a deterministic function of the sample index and a split-specific seed. The diffusion model uses 60,000 training images, the



(a) Generated factor adherence during training.



(b) Source, early/middle/late equal-budget intervention, and target-only examples.

Figure 7: Factorized MNIST run diagnostics. The image panel is a sanity check; evaluator statistics and 20-trajectory geometry are the primary evidence.

evaluator 50,000 images (independent seed), and evaluation uses the 10,000 MNIST test images. *Network.* The U-Net has 96 base channels, GroupNorm(8) residual blocks, and spatial self-attention (4 heads). The timestep (sinusoidal, 256-d) and the four factor labels (one 256-d learned embedding per factor) are summed into a single embedding that modulates each residual block by FiLM-style scale and shift. *Optimization.* The noise-prediction MSE is minimized with AdamW (learning rate 2×10^{-4}, weight decay 10^{-4}), batch size 160, gradient-norm clipping at 1.0, and mixed-precision training. The diffusion cosine schedule has $T = 1000$ steps; EMA decay is 0.9995; DDIM sampling uses 64 steps. The evaluator uses AdamW (learning rate 10^{-3}, weight decay 10^{-4}) for 8 epochs. *Regularization and hyperparameter selection.* Regularization consists of the AdamW weight decay, gradient clipping, and EMA weights; no dropout is used. Training runs for at least 25 and at most 60 epochs, with generated-adherence checks every 5 epochs on 240 generated samples and early stopping once every factor reaches 0.92 (the planned stopping target of Run validity). The seed is 1234 and the analysis uses 20 source-trajectory seeds (0–19).

G.2 Run validity

An independently trained multi-head evaluator reaches real-data accuracies 0.9413, 0.9988, 1.0000, and 0.9984 for digit, position, color, and texture. Generated adherence is 0.8625 for digit and at least 0.998 for every other factor. Digit therefore misses the planned 0.92 stopping target, a limitation we retain in all interpretations. It cannot explain the high-adherence comparison between position and the two appearance factors. Equal-state-budget interventions have mean relative budget error 5.54×10^{-4} and maximum error 1.25×10^{-3}.

G.3 Historical trained-only factor separation

The historical trained-only factor-level evidence in Figure 20 shows digit and position peaking in an early-to-middle region, whereas color and texture grow toward late denoising. These historical trajectories do not supply the primary controlled result. The primary descriptive family gap G on common trained trajectories is reported in Section 3.1; factor-specific evaluator differences are secondary controls. The corresponding historical phase coordinates are reported in Table 5. Along the generative process itself, Figure 1C shows the trained evaluator signal $L_f(s)$ for digit and colour on common trained trajectories, and Figure 3 shows the same signal for all four factors with their phase centres, in which the structural pair centres near $s \approx 0.43$ and the appearance pair near $s \approx 0.66$ while the untrained network is unordered.

The originally hypothesized strict order digit→position→color→texture is not supported by the phase coordinates in Table 5. The previously reported single-model family-gap estimate, its paired bootstrap interval, the per-measurement separation statement, the pairwise factor differences, and the half-area family gap are excluded from the scientific evidence because their raw-measurement and

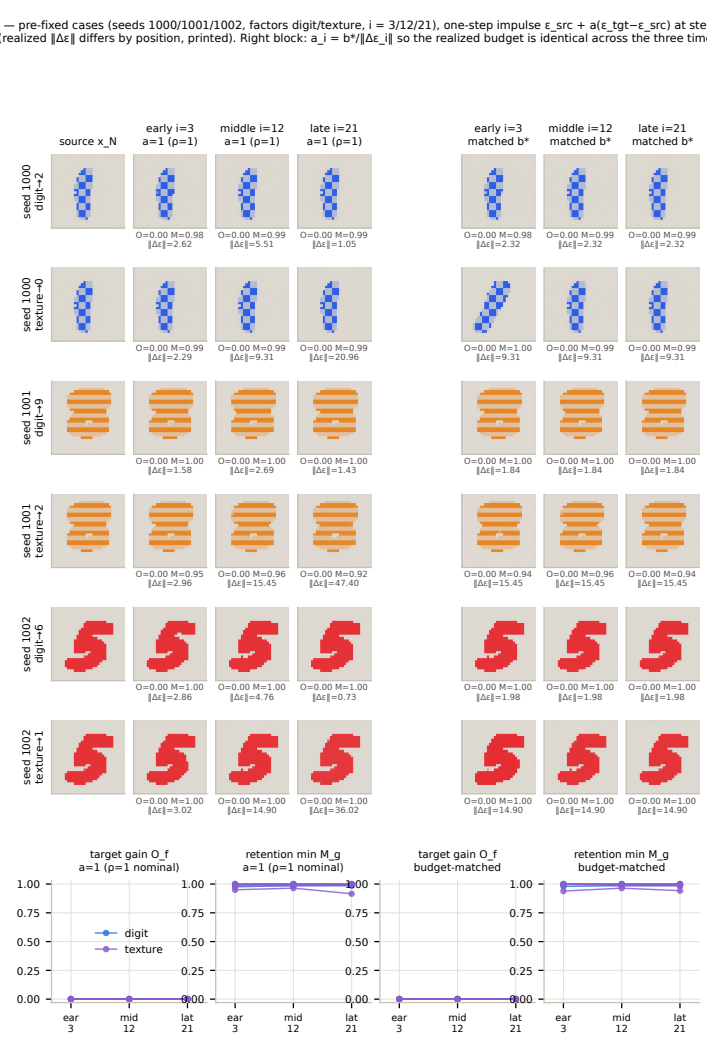


Figure 8: **A single-step impulse leaves the target factor unchanged at every position: over 36 pre-fixed cases (3 seeds $\times$ 2 factors $\times$ early/middle/late $\times$ two budgets) the target gain is $O_f \leq 0.0025$ while the minimum retention of the other factors is ≥ 0.916.** Cases (seeds, factors digit/texture, positions $i \in \{3, 12, 21\}$) were fixed before any one-step result was computed. At step i only, the noise prediction is replaced by $\epsilon_{\text{src}} + a(\epsilon_{\text{tgt}} - \epsilon_{\text{src}})$ and the remaining DDIM tail is left unmodified. Left block: nominal $a = 1$ (realized $\|\Delta\epsilon\|$ printed under each panel; relative budget 0.020–1.29). Right block: $a_i = b^*/\|\Delta\epsilon_i\|$ with b^* the per-(seed, factor) median realized norm, so the three positions receive an identical budget. Bottom row: target gain O_f and minimum retention M_g by position (nominal: $O_f \leq 0.0025$, $M_g \geq 0.916$; matched: $O_f \leq 0.0025$, $M_g \geq 0.939$). This is a negative control for the branching example of Figure 2, where the condition is switched for the whole remaining tail: the position of a one-step impulse alone does not edit the factor at these budgets. It is descriptive (three seeds, CPU, no population claim) and is not the estimand of Section 3.1. Produced by `code/figs/make_gap_pgc_0357_figures.py` (manifest `figures/gap_pgc_0357_manifest.json`); caption numbers are emitted by `code/figs/make_gap05_caption_macros_20260926.py` from the archived case table.

regeneration contracts have not been verified. Archived summaries alone are not treated as completed evidence. The family gap that this paper reports is the within-trained-network gap G on common trained trajectories in Section 3.1.

Panel (b) of Figure 20 presents archived trajectory-level phase centres for the reported trained model; the within-model standard deviations are those of Table 5. Plot-level provenance and

Table 5: Factorized MNIST phase coordinates, mean ± standard deviation across 20 trajectories.

Factor	Phase center $\bar{s}_f$	Half-area $s_{1/2,f}$
Digit identity	0.421 ± 0.046	0.391 ± 0.066
Position	0.428 ± 0.037	0.421 ± 0.047
Foreground color	0.660 ± 0.025	0.707 ± 0.032
Texture	0.663 ± 0.014	0.712 ± 0.019

measurement validity are distinct: tracing a plotted value to a CSV row does not by itself verify the conditioning, model-output parameterization, time coordinate, normalization, or numerical integration that produced that row. The figure therefore summarizes historical factor-level phase coordinates rather than certifying the complete measurement pipeline. It does not establish a control benefit or generalization to a new condition.

G.4 Excluded historical replication summaries

The numerical historical-replication results formerly reported in this subsection (three original four-factor models with 20 analyzed source trajectories each, plus a matched untrained control) are omitted because their measurement-layer provenance and regeneration contract remain unverified: the checkpoints are not archived and no raw per-trajectory measurements for those models exist in this repository. A transcription match between the archived per-factor summary CSVs (`sol/data/e0prime_4run/`) and manuscript values establishes reporting consistency, not reproducibility of the original conditional predictions or trajectories. These summaries are not used as empirical support for Result 1. The retained same-state result is reported separately in Section 3.1, with its own verification limits.

G.5 Moved from the main text: Result 1 provenance and audits

The aggregation is reproducible by `python3 code/audit/same_state_ordering_3model.py` from the three archived `phase_coordinates_per_seed.csv` files; the source `L_curves.csv` files are retained separately.

They are independently recomputed from 61,440 raw L rows for three models, two trajectory sources, four factors, two evaluators, 20 analysis seeds, and 64 steps in `results_pgc_r73_e2_bundle_20260924/gdd_per_model.csv` and `gdd_cross_model.csv`; the manifest records input hashes.

For a separate time-resolved check on the same trained trajectory states, Figure 24 shows the cumulative fraction of each factor's raw L over sampling progress. The structural pair reaches 50% of its cumulative L at mean progress 0.430, before the appearance pair at 0.778 (three model seeds, 20 analysis seeds each). This is a descriptive temporal ordering of conditional-update measurements; it is not a saturation plot of the scalar phase-center gap G or the trained-minus-untrained contrast δ, and does not establish downstream intervention benefit.

Between-model summaries are defined from model-level means:

$$\bar{\delta}(z,f) = \sum_m \delta(m,z,f)/3, \qquad s_\delta(z,f)^2 = \sum_m (\delta(m,z,f) - \bar{\delta}(z,f))^2/2.$$

The sums run over the three reported model seeds. Agreement of trajectory-level paired rows alone does not verify this between-model aggregation. Retraining of model seeds 4002 and 4003 (in addition to the originally reported seed 4004) has since completed (Section 6). The reported comparison uses three independently retrained checkpoints, model seeds 4002, 4003, and 4004. This provides a descriptive comparison across three models; we do not certify a prospective three-model pass/fail criterion. The family-gap definition G is not such a criterion, and the factor-specific contrasts below should not be interpreted as passing one.

An independent CPU recomputation of these paired differences directly from the raw per-trajectory `L_curves.csv` rows (not from the committed `same_state_summary.csv`/`same_state_paired_diff.csv` files) reproduces the

committed values to floating-point precision (480/480 rows joined, max absolute error 4.7×10^{-16}), confirming the reported per-model-seed aggregation. A one-change-at-a-time sensitivity check on the same raw rows (99th-percentile trailing-window trim, smoothing, time coordinate) leaves 7 of 8 (trajectory-source, factor) sign patterns unchanged; the 99th-percentile trim variant flips sign in 1 of 8 cells relative to the reported trim-95% baseline. We report this as a measurement-convention sensitivity finding and do not use it to alter the reported contrasts above. This audit covers the same-state control's model aggregation and measurement convention; it does not validate the excluded historical replication or certify its measurement protocol.

G.6 PX-2 DISCRETE CERTIFICATE AND SPECIFICATION DIFFERENCES

For the discrete certificate of Lemma 1, let $w_j > 0$ be the trapezoidal weights on the 64-level log-noise grid with endpoints $r_a = 0.1$ and $r_b = 4$, $T = \log(r_b/r_a)$, $M_f = \sum_j w_j Q_f(r_j)$, and $E_f = \sum_j w_j |\hat{Q}_f(r_j) - Q_f(r_j)|$, where Q_f is the fixed theory curve and $\hat{Q}_f$ the measured raw contrast of one trained model. When $E_f < M_f$, the phase-centre error bound is $B_f = TE_f/(M_f - E_f)$, and a strict non-null order is certified only when $B_S + B_A < |S - A|$ for the theoretical gap $S - A$. The equal-variance control has zero theoretical gap and is assessed separately rather than counted as a failed strict-order certificate. A grid-based calculation does not certify unsampled continuous curves. The certificate is computed post hoc from the archived per-model curves by `code/queue/pgc_r100_px2_discrete_certificate.py` (output `results/pgc-r116-px2-stability-recert-20260926/summary.json`, input hashes checked against `experiments/px2_qf_direct/full_output/SHA256SUMS`); no new training is involved. Per ratio (0.5, 0.75, 1.5, 2), the sign of the measured $S - A$ matches the prediction in 5/5 models and the certificate holds in 0/5, with $B_S + B_A$ ranging over $[1.92, 2.41]$, $[3.02, 3.41]$, $[2.80, 3.60]$, and $[2.24, 2.59]$. Under the grid-peak definition of the specification, the appearance argmax lies at the upper grid endpoint in 5/5 models at every ratio, and the structural argmax at the endpoint in 0/5, 3/5, 5/5, 5/5, and 5/5 models at ratios 0.5, 0.75, 1, 1.5, and 2. Two implementation details differ from the archived specification (`experiments/px2_qf_direct/SPEC.md`): the conditional mean shift is placed on a single coordinate of each block (`run_px2.py`) rather than spread over the block with unit norm, and the normalization floor for L_f is 10^{-12} rather than 10^{-10}. Equal shift norms preserve the population raw-Q formula under block rotations, but do not certify identical finite-training protocols.

G.7 MOVED FROM THE MAIN TEXT: PX-1 PLOT NOTES

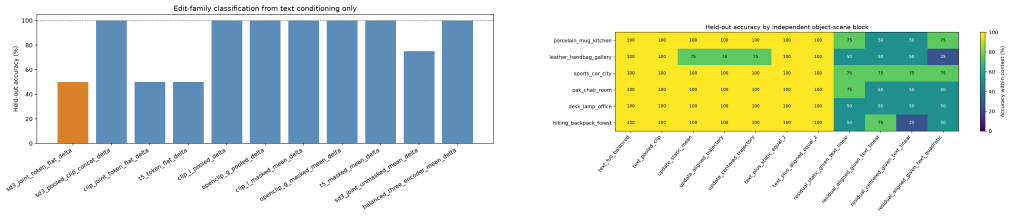
The paired evaluator contrast D has mean 0.23224 and is reported separately. Exploratory seed 4005 has $G(4005, \text{trained}, T) = 0.23346$ but is excluded from the primary three-model estimate. Each archived model has 20 analysis seeds and 20,480 raw L rows; input hashes, model-level rows, and model-resampling percentile intervals are in `results/px1_model_level_G_20260925/`.

Figure 25 plots observed within-evaluator $G(T)$ and $G(U)$ on common trained states, with D in a secondary panel. A numeric predicted-versus-observed plot is not justified: the Gaussian-block proposition concerns an unnormalized population contrast, whereas empirical L uses separately normalized model predictions. Per-model conditional variances and a validated mapping between these estimands are unavailable. This raw-L audit does not verify upstream L generation or checkpoint-load identity.

H SD3.5 REPRESENTATION DETAILS

H.1 ORIGINAL PROMPT DESIGN AND RUN SIGNATURE

All prompts share the template "a centered smooth-glazed blue ceramic teapot on a wooden table, against a light-gray background, with fixed studio lighting, highly detailed." The edit-bearing phrases are: far left, much larger/filling the frame, far right; vase, bowl, goblet; dense fine crackle glaze, dense embossed diamond texture, dense basket-weave relief; one hairline crack, one small rim chip, and one small chipped-glaze patch. The original direction experiment has run signature `ba52774713b2b359`; the static-direction rerun has signature `e3d2589bad454346` with pinned `diffusers 0.40.0`, `transformers 5.15.1`, and `torch 2.11.0+cu128`.



(a) Four-query text representations. Component-level perfect scores are secondary.

(b) Accuracy by context and method in the six-context experiment.

Figure 9: Input-side and context-level controls.

H.2 Text representations on the original four queries

Table 6: Original four-query text-embedding pilot. The primary flattened joint-token delta is not significant; perfect secondary representations do not survive Holm correction over 11 representations.

Representation	Correct	Exact p	Mean true margin	Holm p
SD3 joint token, flattened (primary)	2/4	0.333	0.129	0.750
Pooled CLIP concatenation	4/4	0.0417	0.401	0.458
CLIP-L pooled	4/4	0.0417	0.268	0.458
OpenCLIP-G pooled	4/4	0.0417	0.466	0.458
CLIP-L masked mean	4/4	0.0417	0.047	0.458
OpenCLIP-G masked mean	4/4	0.0417	0.317	0.458
T5 masked mean	4/4	0.0417	0.291	0.458
Balanced three-encoder mean	4/4	0.0417	0.223	0.458

H.3 Six multi-context blocks

Each block uses a frozen seed and retains object, color/material, support surface, background, and lighting across its source and target prompts. Seeds 4101–4106 correspond respectively to the porcelain mug, leather handbag, sports car, oak chair, desk lamp, and hiking backpack contexts. The sole raw update error is the held-out leather handbag→briefcase morphology query, classified as dense texture by both static and aligned geometry. Balanced text is correct for this query and all others.

The corresponding quantitative table (prototype-only linear R^2 per context) is omitted because its measurement and regeneration contract remains unverified.

H.4 Moved from the main text: diagnostics details

Normalized directional disagreement is not generally invariant to parameterization changes that add a state-dependent term. With a fixed positive normalization constant, it is also not generally invariant to rescaling. We therefore interpret phase summaries within the stated output parameterization and numerical conventions, rather than as a parameterization-independent property shared by diffusion and flow samplers.

When both conditional fields are nonzero, let θ_i be their angle, $h_a = \|v_i^a\|/(\|v_i^a\| + \epsilon)$, $h_0 = \|v_i^0\|/(\|v_i^0\| + \epsilon)$, and $q_i = \|v_i^a\|/(\|v_i^0\| + \epsilon)$. The exact regularized identities are

$$L_i^2 = h_a^2 + h_0^2 - 2h_a h_0 \cos\theta_i, \qquad r_i^2 = q_i^2 + h_0^2 - 2q_i h_0 \cos\theta_i. \tag{30}$$

For comparison only, define L_i° and r_i° by exact normalization without regularization, and $q_i^\circ = \|v_i^a\|/\|v_i^0\|$. Then

$$(L_i^\circ)^2 = 2(1 - \cos\theta_i), \tag{31}$$

$$(r_i^\circ)^2 = (q_i^\circ)^2 + 1 - 2q_i^\circ \cos\theta_i. \tag{32}$$

The regularized definitions remain defined at zero vectors, but the angle-based identities require nonzero fields. The current analysis package does not document the regularization value and minimum field norms for every historical result. We therefore make no uniform approximation-error claim.

For discrete results, the time grid, quadrature weights, smoothing boundary rule, and endpoint exclusion must be specified for each experiment; these choices cannot be inferred from the continuous definitions alone. For discrete causal experiments, $\mathrm{FWHM}(L) = [a_L, b_L)$ is the contiguous component around the interior maximum satisfying $L_i \geq \frac{1}{2} \max_j L_j$. In the final frozen test, the interval was valid only if its width was between 2 and 8 analyzed steps and the interior peak exceeded both endpoints by at least 15% of its peak value. Equal-height ties for the interior maximum are broken by the smallest step index; a curve with $\max_i L_i < 10^{-6}$ (no measurable disagreement) yields an invalid, rather than degenerate, interval; and the two endpoint steps excluded from n (Equation (33)) are the first and last sampler steps, where guidance-scale artefacts inflate all four scalar diagnostics.

H.5 Temporal, aligned, static, and centered representations

For the scalar temporal-profile representation defined below, let n be the analyzed length after excluding the first and last sampler steps. Each scalar sequence in this representation is smoothed with a centered three-step moving average and is denoted $g_{a,m} \in \mathbb{R}^n$ for $m \in \{L, r, p, o\}$. This preprocessing statement is not a blanket specification for every phase-coordinate experiment.

The supplied same-state-control implementation instead measures noise-prediction disagreement using $h_\eta(z) = z/\max(\|z\|, \eta)$, with $\eta = 0.0000000001$. Its measured disagreement is $L = \|h_\eta(v^a) - h_\eta(v^0)\|$. This agrees with exact unit normalization only when both field norms are at least η; it is not the additive regularization used in the general definitions above.

In the supplied 64-step configuration, the state is recorded before each sampler step, progress is $s_i = i/63$, and all grid points satisfying $s_i \leq 0.95$ are retained. Thus the initial point is retained and the final retained point is $59/63$; no point is interpolated at 0.95. No moving-average smoothing is applied in this implementation. Both the predicted clean image and the subsequent DDIM state are clipped to the image range.

The phase center uses trapezoidal quadrature for its numerator and denominator. The half-area statistic linearly interpolates cumulative trapezoidal area between adjacent grid points. Total areas at or below 0.000000000001 are returned as undefined. These conventions describe the supplied implementation; their agreement with every historical pipeline has not been certified.

The scalar temporal-profile distance is

$$D_{\mathrm{time}}(a, b) = \frac{1}{4} \sum_{m \in \{L,r,p,o\}} \left(1 - \frac{\langle g_{a,m}, g_{b,m} \rangle}{\|g_{a,m}\| \, \|g_{b,m}\| + \epsilon}\right). \tag{33}$$

It retains the full scalar time series and is therefore stronger than a three-bin early/middle/late baseline. The aligned direction-path distance is

$$D_{\mathrm{dir}}(a, b) = \max\left\{0, 1 - \frac{1}{n} \sum_{i=0}^{n-1} \langle d_{a,i}, d_{b,i} \rangle\right\}. \tag{34}$$

To test whether temporal switching is necessary, define the mean direction and its regularized normalization. The resulting comparison is a dissimilarity, not necessarily a metric: regularization prevents exact unit norm, so self-dissimilarity need not vanish.

$$\mu_a = \frac{1}{n} \sum_{i=0}^{n-1} d_{a,i}, \qquad \bar{d}_a = \frac{\mu_a}{\|\mu_a\| + \epsilon}, \qquad D_{\mathrm{static}}(a, b) = 1 - \langle \bar{d}_a, \bar{d}_b \rangle. \tag{35}$$

The aligned similarity admits the exact decomposition

$$\frac{1}{n} \sum_i \langle d_{a,i}, d_{b,i} \rangle = \langle \mu_a, \mu_b \rangle + \frac{1}{n} \sum_i \langle d_{a,i} - \mu_a, d_{b,i} - \mu_b \rangle. \tag{36}$$

The first term is a time-collapsed component; the second is a centered, trajectory-varying component. Perfect performance of D_{dir} alone cannot establish that the second term is necessary. We separately

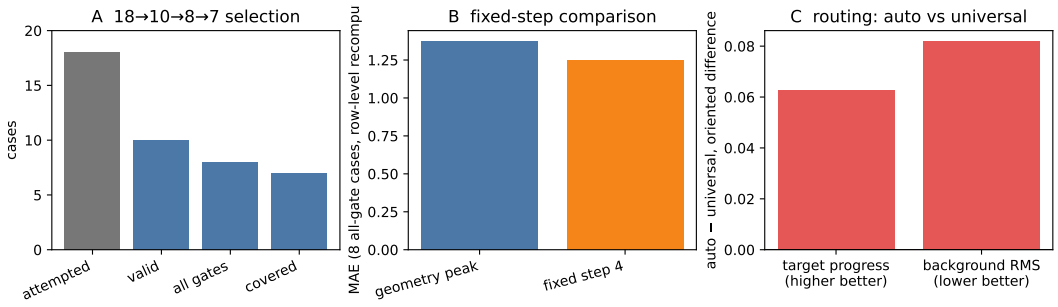


Figure 10: **E1 onset audit and an archived routing comparison.** (A) The archived selection sequence is 18 attempted cases, 10 with valid geometry, 8 passing the primary gates, and 7 covered. (B) On the eight primary cases, the tabulated MAE is 1.375 for the geometry peak and 1.25 for a predictor fixed at step 4. (C) The regeneration script plots archived oriented differences between auto routing and the universal policy. This panel does not compare auto routing with the always-dense-texture policy discussed in the main text. The routing quantities are descriptive and do not establish a verified budget-controlled advantage. Regeneration from summary CSVs does not validate the upstream endpoint measurements or quality gates.

evaluate the centered sequence and independently shuffled time indices. For the centered representation, subtract the temporal mean from each direction vector and divide each nonzero residual by its Euclidean norm; map an exactly zero residual to the zero vector. This convention makes the representation defined at zero, although such entries are not unit vectors. Historical centered-representation results require regeneration under an explicitly documented zero-handling convention before they can certify this definition.

For a distance D, two prototypes $\mathcal{P}_f$ per family define a frozen nearest-family rule

$$S_f(q; D) = \frac{1}{|\mathcal{P}_f|} \sum_{a \in \mathcal{P}_f} D(q, a), \qquad \widehat{f}(q; D) = \arg\min_f S_f(q; D), \tag{37}$$

with margin equal to the second-smallest score minus the smallest score. If multiple families attain the same minimum score, choose the first family in an alphabetical ordering of the frozen family labels, fixed independently of outcomes. A missing or nonfinite required feature yields an abstention, not a nearest-family prediction.

For paired phase comparisons, use a common analysis-seed set across all required factors and both evaluators within each model and trajectory source. Exclude a seed from that paired comparison if any required phase is undefined, and report the retained denominator explicitly. These rules specify a complete analysis convention; they do not retrospectively certify the archived classifications or phase counts, which must be regenerated under their documented rules. Applied to the archived raw L of the same-state control (four models, 20 analysis seeds per cell, trim95 grid phase centres), the paired rule excludes no seed in any of the eight (model, factor-contrast) cells: the retained denominator is 20 in every cell, and the common-seed and all-available-seed family gaps coincide exactly (maximum absolute difference 0; all eight trained-network gaps positive). This is a re-summary of archived L only; the upstream forward passes and checkpoints are not re-verified by it.

I Causal-Onset Details

The panel values and the complete 18-case table in Figure 10 and Table 7 are generated by `code/make_e1_audit.py` from the committed case-level and baseline CSVs. The routing comparison is descriptive: its raw paired observations are not reconstructed by this entry point, so this figure does not add a confidence-interval or superiority claim.

I.1 Final prompts and frozen settings

All source/target pairs use a centered object and hold scene attributes fixed. The full settings are SD3.5 Medium, OpenAI CLIP ViT-L/14 evaluation, 24 steps, 22 analyzed steps, 512×512, CFG

Table 7: **E1 18-case audit regenerated from saved case-level data.** Q: all primary quality gates; V: valid geometry interval; C: geometry interval covers the observed response peak. Observed step and coverage are undefined (–) below the semantic-gap threshold (sol/sections/05_causal_control.tex:50). Fixed-step predictions and errors are reported in the primary comparison only for cases passing the archived all-criteria gate. This reporting restriction is not an additional mathematical definedness condition.

Pair	seed	Q	V	C	observed step	geometry peak	fixed step 4
mug→tumbler	5201	F	T	–	–	6	–
mug→tumbler	5202	F	F	F	3	8	–
mug→tumbler	5203	F	T	–	–	6	–
handbag→briefcase	5201	F	F	T	6	4	–
handbag→briefcase	5202	T	T	F	10	5	4
handbag→briefcase	5203	F	F	T	3	5	–
sports_car→suv	5201	T	T	T	4	4	4
sports_car→suv	5202	T	T	T	5	3	4
sports_car→suv	5203	F	F	F	1	6	–
chair→bench	5201	T	T	T	4	3	4
chair→bench	5202	F	F	T	3	5	–
chair→bench	5203	T	T	T	4	4	4
desk_lamp→lantern	5201	T	T	T	5	5	4
desk_lamp→lantern	5202	T	T	T	6	4	4
desk_lamp→lantern	5203	T	T	T	4	3	4
backpack→messenger_bag	5201	F	F	T	3	5	–
backpack→messenger_bag	5202	F	F	T	6	4	–
backpack→messenger_bag	5203	F	F	T	3	3	–

7, commitment budget $\rho = 2.4$, coefficient cap 12, and fixed baseline onset 4. Geometry validity requires width 2–8 and relative interior bump at least 0.15. Seeds are 5201–5203.

Table 8: Frozen final object–scene pairs. Only the object phrase changes within a row.

Pair	Source → target	Fixed scene
Mug	white porcelain coffee mug → drinking tumbler	black kitchen counter; pale-green tile backsplash; morning side light
Handbag	red leather handbag → briefcase	white marble plinth; cream gallery wall; warm overhead light
Car	glossy yellow sports car → off-road SUV	wet asphalt; blurred city at dusk; low-angle light
Chair	plain oak dining chair → bench	pale rug; minimalist room; large window; natural side light
Lamp	smooth silver desk lamp → lantern	dark office desk; white wall; cool task light
Backpack	green hiking backpack → messenger bag	gray rock; misty pine forest; overcast light

I.2 ALL 18 CASE SUMMARIES

Ten cases have valid geometry intervals, and two of these fail the semantic-gap threshold. The eight primary cases have geometry MAE 1.375 and interval coverage $7/8$. A fixed-onset-4 predictor must predict step 4 in every case; the archived comparator column does not satisfy this requirement. Its interval coverage therefore cannot be attributed to that predictor without a separately specified interval construction. We withdraw the all-18 and geometry-valid-10 onset-risk summaries because they include cases whose endpoints are undefined under the current definition. A continuous selective-risk curve and its uncertainty are not reproducible from the available package.

Table 9: All attempted onset cases under the current endpoint definition. Q denotes the all-criteria gate and G denotes geometry-interval validity. The two cases below the semantic-gap threshold have undefined causal peaks, onset errors, and coverage. Invalid geometry cases are retained for descriptive inspection but excluded from the primary comparison. Archived comparator columns are omitted because their identification as a fixed-onset-4 baseline has not been verified.

Pair	Seed	Q	G	Gap	Geo interval	Geo peak	k_c	Geo error	Coverage
Mug to tumbler	5201	no	yes	-0.008	$[4, 10)$	6	–	–	–
	5202	no	no	0.031	$[4, 13)$	8	3	5	no
	5203	no	yes	0.020	$[2, 10)$	6	–	–	–
Handbag to briefcase	5201	no	no	0.096	$[0, 13)$	4	6	2	yes
	5202	yes	yes	0.099	$[2, 10)$	5	10	5	no
	5203	no	no	0.078	$[2, 12)$	5	3	2	yes
Car to SUV	5201	yes	yes	0.135	$[2, 8)$	4	4	0	yes
	5202	yes	yes	0.142	$[3, 10)$	3	5	2	yes
	5203	no	no	0.131	$[3, 13)$	6	1	5	no
Chair to bench	5201	yes	yes	0.130	$[2, 5)$	3	4	1	yes
	5202	no	no	0.075	$[0, 9)$	5	3	2	yes
	5203	yes	yes	0.079	$[1, 8)$	4	4	0	yes
Lamp to lantern	5201	yes	yes	0.089	$[0, 8)$	5	5	0	yes
	5202	yes	yes	0.128	$[3, 7)$	4	6	2	yes
	5203	yes	yes	0.168	$[0, 6)$	3	4	1	yes
Backpack to messenger	5201	no	no	0.092	$[0, 12)$	5	3	2	yes
	5202	no	no	0.089	$[2, 11)$	4	6	2	yes
	5203	no	no	0.090	$[0, 14)$	3	3	0	yes

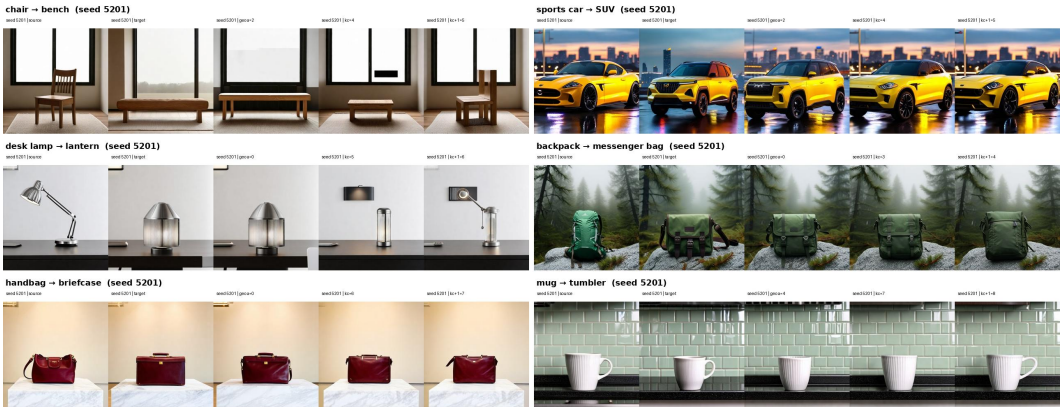


Figure 11: **SD3.5 Medium, six audited object-replacement cases (seed 5201): what the phase-gated intervention looks like.** Columns: source, target-only, geometry onset, causal peak k_c, delayed onset k_c+1; the per-panel labels are the original audit annotations. Renders reproduced from Figures 13 to 15; PGC outputs only, no external editing method is shown. The onset analysis behind these renders is in Section 4.



Figure 12: **Seed dependence of the same intervention.** Chair→bench and backpack→messenger bag across the three audited seeds (rows), same columns as Figure 11. Renders reproduced from Figures 14 and 15.



Figure 13: Complete qualitative audits for mug→tumbler and handbag→briefcase. Columns are source, target-only, geometry onset, causal peak, and delayed onset.



Figure 14: Complete qualitative audits for sports car→SUV and chair→bench.

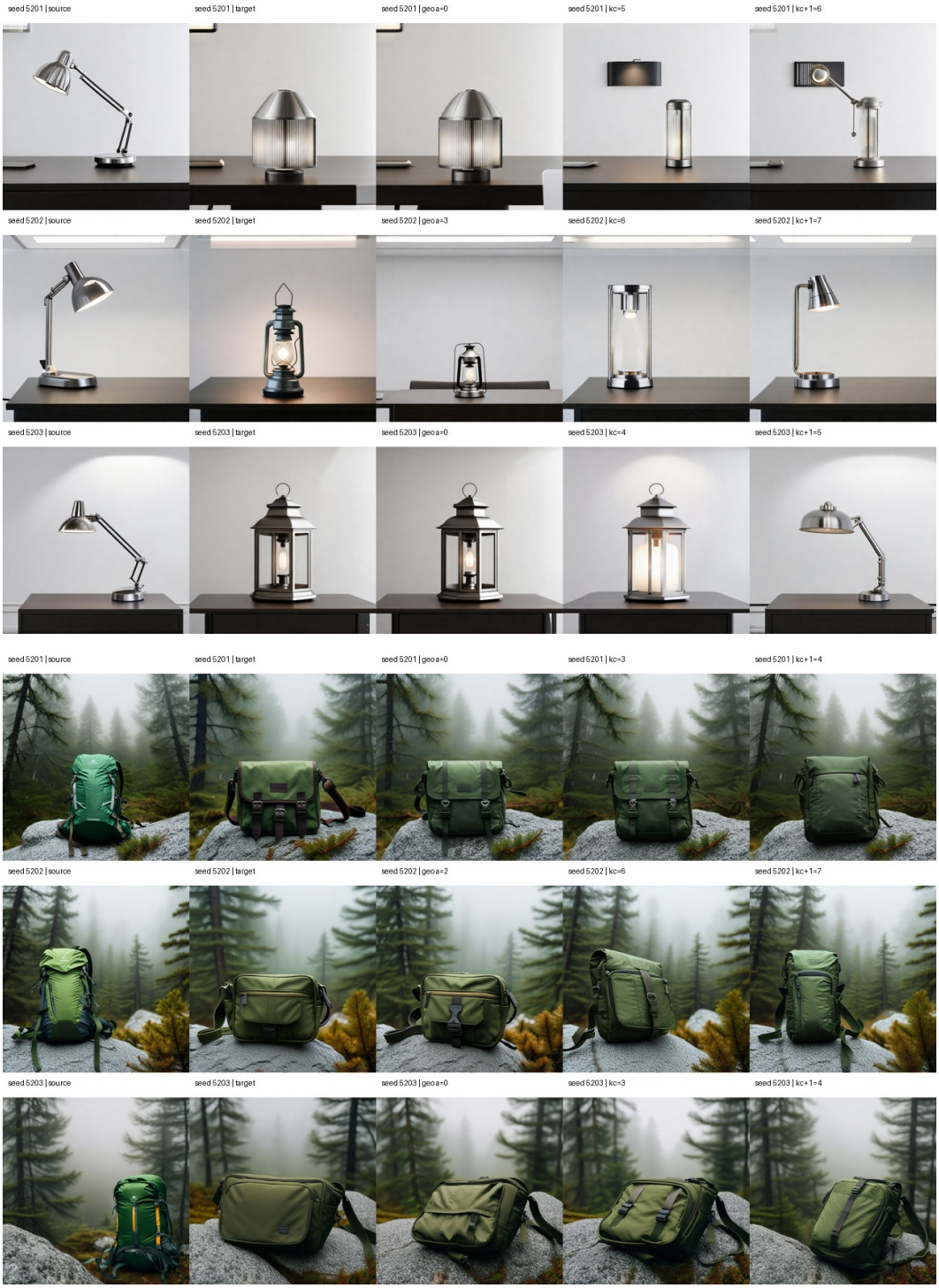


Figure 15: Complete qualitative audits for desk lamp→lantern and backpack→messenger bag.

I.3 CAUSAL INTERVENTIONS AND SUSCEPTIBILITY

For a gate $G \subseteq \{0, \ldots, N-1\}$ and online direction $\mathcal{D}_i(x)$ (calligraphic, distinct from the generic distance D introduced above), controlled sampling uses

$$v_i^{\text{ctrl}} = v_\theta(x_i, t_i; c_0) + \alpha_i \mathbf{1}[i \in G]\mathcal{D}_i(x_i). \tag{38}$$

For full prompt-delta control, $\mathcal{D}_i(x) = v_\theta(x, t_i; c_a) - v_\theta(x, t_i; c_0)$. A normalized intervention budget is

$$B(G, \alpha) = \sum_{i \in G} w_i \alpha_i \frac{\|\mathcal{D}_i(x_i)\|}{\|v_\theta(x_i, t_i; c_0)\| + \epsilon}. \tag{39}$$

For the proposed field-space budget, we specify $w_i = 1$ and $\alpha_i \geq 0$. The target constraint $B = \rho$, where $\rho > 0$ is the target intervention budget fixed per experiment (see Section I for the values used), must be verified from the coefficient actually applied and the online conditional delta at each controlled state; it does not follow from selecting the same number of steps.

The supplied historical package does not establish this measurement contract for every reported comparison. In particular, the Factorized MNIST intervention results are described as using a state-space budget, which is not identified with the field-space quantity defined above. We therefore distinguish nominal budget matching from verified equality of this realized budget. The controller appendix specifies gates and reported outcomes, but does not provide a complete online allocation algorithm.

For sustained-onset experiments, $G_k = \{i : i \geq k\}$ and a fixed α is reused for every k within a case. Let $M(I; c)$ be a CLIP image–text score and $m(I) = M(I; c_a) - M(I; c_0)$ its edit-conditioned margin. Let I_S denote the source image, I_T the fully edited (target) image, and I_k the image produced by controlling with onset index k (i.e. using gate G_k). With source-to-target margin gap

$$g = m(I_T) - m(I_S), \tag{40}$$

we define the source-normalized target progress only when the gap clears a prespecified semantic-separation threshold,

$$P(k) = \begin{cases} \big[m(I_k) - m(I_S)\big]/g, & g \geq 0.03, \\ \text{undefined}, & g < 0.03, \end{cases} \tag{41}$$

For adjacent onset indices, define the raw margin loss

$$A(k) = m(I_k) - m(I_{k+1}). \tag{42}$$

When $g \geq 0.03$, the normalized adjacent loss is $C(k) = A(k)/g$, which agrees with $P(k) - P(k+1)$ since the source-margin term $m(I_S)$ cancels. When $g < 0.03$, the raw loss $A(k)$ remains defined, but the normalized progress P, the normalized causal peak C, its error, and its coverage are not reported.

The current case table does not contain the raw adjacent margin losses or every component needed to recompute the all-criteria gate. Its gate labels are archived outputs, not a complete reproduction specification. Recalculation requires the margin sequence, the causal-effect gate rule, the geometry-validity quantities, and the mapping between sampler indices and analyzed indices.

For cases with defined normalized progress, the adjacent causal loss and causal peak are

$$C(k) = P(k) - P(k+1), \qquad k_c = \arg\max_k C(k), \tag{43}$$

defined only over the range where P is defined; ties in $C(k)$ for k_c are broken by the smallest k, consistent with the tie-break rule used for the interior FWHM maximum above. Conceptually, the task-specific susceptibility of objective O to a local control u_i is

$$\chi_i(O) = \left.\frac{\partial O(x_N(u))}{\partial u_i}\right|_{u=0}. \tag{44}$$

The finite-intervention experiments reported here do not measure this terminal derivative. A sustained-onset contrast changes a finite control schedule and may also change total intervention effort.

Directional disagreement does not determine terminal response. This is an elementary non-identifiability example, not a new theorem or an empirical impossibility result. Consider a two-step scalar sampler with $x_1 = x_0 + 1 + u\delta$, $x_2 = ax_1$, and terminal objective $O(x) = x$, where $\delta > 0$. At the first step, the source and target fields are $v^0 = 1$ and $v^a = 1 + \delta$. Their regularized directional disagreement is $L = \epsilon\delta/((1 + \epsilon)(1 + \delta + \epsilon))$, independently of the downstream coefficient a. However, the terminal susceptibility is $\chi_0(O) = a\delta$. Varying a changes the magnitude and sign of the terminal response while leaving the measured local fields and L unchanged. Local disagreement alone therefore does not determine terminal response without assumptions on the downstream map. This example does not rule out predictive associations in a restricted model and task distribution.

For a differentiable controlled step map $F_j(x, u)$, define the state Jacobian $J_j = \partial_x F_j(x_j^0, 0)$ and the control Jacobian $B_i = \partial_u F_i(x_i^0, 0)$ (partial-derivative notation, to avoid reusing the plain letter D used above for distances and directions; the index is i rather than j so that it matches its use below and the notation in Section C). Let $\mathcal{A}_i$ be the ordered product of subsequent state Jacobians J_j, from step $N - 1$ down to step $i + 1$, or the identity if no step remains (distinct from the raw margin loss $A(k)$ above). Then $\chi_i(O) = \nabla O(x_N^0)^\top \mathcal{A}_i B_i$. Neither a nonzero state response nor a nonzero Jacobian product alone guarantees a nonzero objective response.

J Matched-Budget Controller Details

The policy experiment uses source prompt "a studio photograph of a smooth-glazed blue ceramic teapot centered in the frame on a wooden table, plain light gray background, soft light from the left, highly detailed." Morphology prototypes replace the object with a vase or bowl; dense-texture prototypes retain the teapot and apply fine crackle or embossed diamond texture. Held-out queries are a goblet and dense basket-weave relief. Seeds are 1245–1248.

The morphology gate is the FWHM onset followed by the remaining tail. Because the frozen validity rule is strict, an invalid broad phase can still define a usable onset but is marked invalid diagnostically. The dense-texture policy selects the top five smoothed-L steps at progress at least 0.45; the universal policy uses all 22 endpoint-excluded steps. The historical experiment reports nominal budget matching across different gate sizes. The available package does not establish equality of the field-space budget defined in the main text. Verification requires the coefficient actually applied, the online conditional-delta norm, and the source-field norm at every controlled state. We therefore treat these policy averages as descriptive, not as evidence of a verified budget-controlled advantage. Auto and oracle outputs are identical because all classifications are correct.

Table 10: Controller family means ($\pm$ standard deviation) for principal endpoints.

Family	Policy	Concept progress	CLIP directional	DINO source	Background RMS
Dense texture	Auto/oracle	0.754 ± 0.357	0.290 ± 0.090	0.844 ± 0.025	0.031 ± 0.007
	Universal	0.652 ± 0.253	0.246 ± 0.054	0.881 ± 0.029	0.189 ± 0.046
	Wrong	0.673 ± 0.216	0.248 ± 0.051	0.889 ± 0.029	0.185 ± 0.049
Morphology	Auto/oracle	0.132 ± 0.089	0.166 ± 0.111	0.897 ± 0.058	0.121 ± 0.023
	Universal	0.133 ± 0.084	0.179 ± 0.129	0.916 ± 0.035	0.127 ± 0.021
	Wrong	0.162 ± 0.073	0.212 ± 0.116	0.927 ± 0.023	0.064 ± 0.008

J.1 Moved from the main text: routing pilot

A descriptive routing pilot. The historical pilot compares four nominally budget-matched policies on two edit families and four seeds. The supplied records do not verify equality of the realized field-space budget, so the policy averages are descriptive. The automatically selected policy agrees with a *family-label oracle* in 8/8 pilot cases; this term means access to the experiment's family label, not a per-case optimal-policy oracle. Auto routing has lower background RMS than the universal policy in all eight case-level comparisons. Because the four raw seed-level paired differences needed for a valid cluster-level test are unavailable, we do not treat the previously reported case-level Wilcoxon statistic as inferential; treating the four seeds as four independent sign observations would give an exact one-sided sign test $p = 2^{-4} = 0.0625$ if all four paired signs favor auto routing. Their signs and pairing have not been independently reconstructed from the supplied records; this is conditional

arithmetic, not a reconstructed inferential result. The archived bootstrap interval and target-progress significance test cannot be reconstructed from the available seed-level package, so we do not use them for inference. The policy averages below are descriptive. Averaged over both families, auto/oracle routing reaches target progress 0.4430 and background RMS 0.0760, while a fixed policy that always applies the dense-texture top-5 intervention reaches target progress 0.4583 and background RMS 0.0477 – better than routing on *both* reported averages. The dense-texture policy has a descriptive preservation advantage within its own family; morphology remains a target–preservation trade-off, with the deliberately wrong sparse-texture policy better under the present metrics. The fixed dense-texture policy itself chooses steps using smoothed directional disagreement. Its better reported averages therefore limit the incremental value of family routing, not every use of geometry for step selection. A selector comparison requires a common candidate window, the same number of selected steps, and an independently verified realized budget. Routing is therefore not claimed to beat the best fixed policy, and we report the fixed-policy comparison rather than omitting it.

K FlowEdit and Texture Supplements

The previously reported numerical summaries and figures for this analysis are excluded from the scientific evidence because their measurement provenance and regeneration contracts have not been verified. Archived summaries alone are not treated as completed evidence. No conclusion about predictive accuracy, mechanism, or intervention utility relies on this analysis.

L CelebA-HQ Supplement

The previously reported numerical summaries and figures for this analysis are excluded from the scientific evidence because their measurement provenance and regeneration contracts have not been verified. Archived summaries alone are not treated as completed evidence. No conclusion about predictive accuracy, mechanism, or intervention utility relies on this analysis.

M Compute, Assets, and Build

The final causal experiment records an NVIDIA A100-SXM4-40GB, bfloat16 inference, PyTorch 2.11.0+cu128, Diffusers 0.40.0, Transformers 5.15.1, Accelerate 1.14.0, and Hugging Face Hub 1.28.0. The static, text, and multi-context notebooks use the same pinned core stack. The controller archive records model, resolution, seeds, prompts, prototypes, queries, gates, realized budgets, generated image paths, and NFE. The source package accompanying this manuscript includes compact CSV/JSON result summaries rather than model checkpoints or the full multi-gigabyte intermediate latent tensors.

To compile the manuscript, run `latexmk -pdf main.tex` in the source directory. The project uses the unmodified ICLR 2027 conference style supplied with the manuscript. All plotted and qualitative figures are embedded in `figures/`; compact numerical summaries are in `data/`.

The three-model same-state comparison (Figure 23) is plotted by `code/make_same_state_contrast_multiseed_fig.py` from the committed per-model summaries (model seeds 4002, 4003, 4004; 20 analysis seeds per factor and model), merged into `combined_summary.csv`; the figure reads only these summaries, and raw-trajectory recomputation is reported separately.

N Scope Comparison with Nearby Methods

O Secondary Replication (Protocol B) Details

We additionally ran a separately versioned Colored-MNIST protocol (B) on three independently trained models (seeds 1103, 2207, and 3301), each evaluated on 20 held-out source trajectories. Its pre-specified notebook decision rule is model-level: all three within-model bootstrap 95% intervals for appearance-minus-structural phase centre must exclude zero. The seed-level gaps are 0.000885,

Table 11: Scope of the directly discussed measurement approaches. This is a conceptual comparison, not a new empirical evaluation of the cited methods.

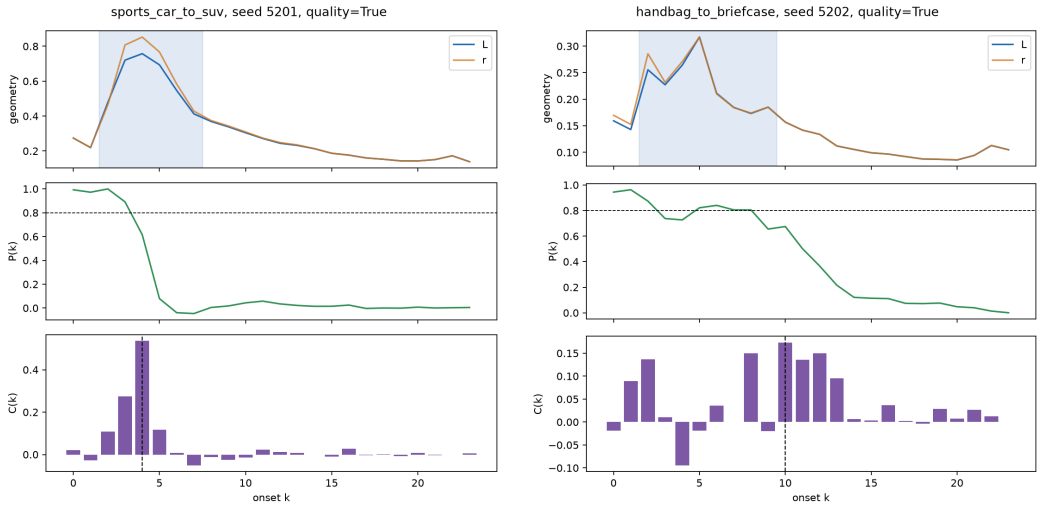
Approach	Measurement target	State or intervention design	Question addressed
DiffEdit	Conditional prediction difference for mask construction	Source and target predictions at a shared noisy state	Where should an edit be spatially applied?
Concept-insertion curves	Final-output change under concept insertion	Injection step is varied across generation schedules	When does concept insertion change the output?
This work	Directional disagreement, phase summaries, and finite-onset response	Conditional predictions share a reference state; onset responses compare separate controlled rollouts	What is described by local geometry, and what remains unidentified about downstream response?

0.001099, and 0.001337, with respective intervals $[0.000287, 0.001508]$, $[0.000780, 0.001419]$, and $[0.000855, 0.001819]$. Thus the rule passes. Across the three trained models, the gap is 0.00111 ± 0.00023 (mean $\pm$ sample standard deviation; $n = 3$), This estimate should not be compared to the earlier unverified single-model family-gap estimate. This is a small, positive replication under the locked sign rule; it neither reproduces the earlier single-model effect magnitude nor establishes a universal ordering of individual factors. Protocol B is a separately versioned supplementary experiment, not a designed negative control and not a replication of the original effect magnitude. Its positive gap satisfies the stated within-protocol sign rule. Differences in conditioning, tail handling, smoothing, and time coordinates require an explicit audit before the scale difference can be interpreted. That sensitivity analysis has not been completed.

O.1 Moved from the main text: protocol B summary

A separately versioned protocol (B) on three further independently trained models passes its locked sign rule in every within-model bootstrap interval, with a small gap of 0.00111 ± 0.00023 ($n = 3$), about two orders of magnitude below the within-network gap G of Result 1; it supports sign consistency under its own protocol, not magnitude, and is not compared numerically with the historical single-model estimate (details in Appendix).

P Additional figure panels



(a) Car→SUV, seed 5201: geometry peak and causal peak coincide.

(b) Handbag→briefcase, seed 5202: the causal peak falls just outside the geometry interval.

Figure 16: Representative final proof curves. Panels show L, r, target progress P, adjacent causal loss C, and background displacement.

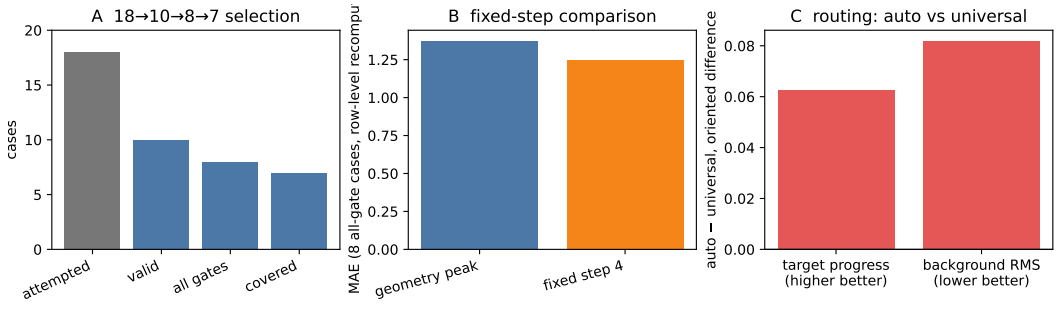
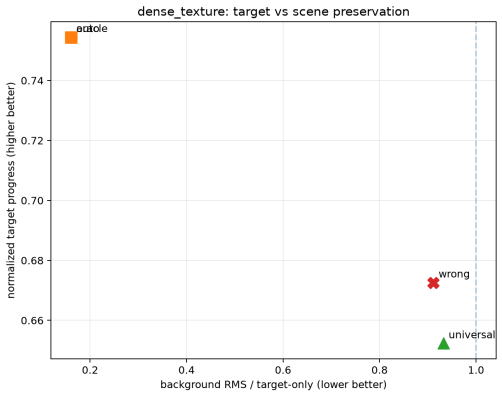
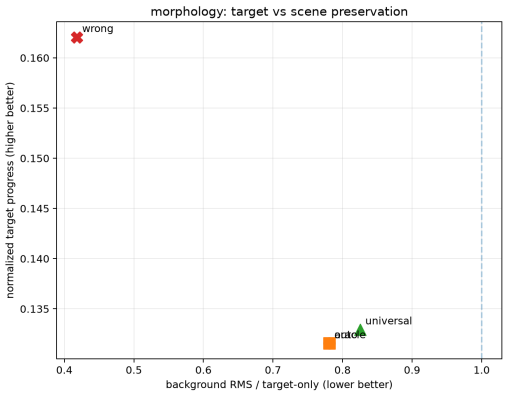


Figure 17: **E1 onset audit and an archived routing comparison.** (A) The archived selection sequence is 18 attempted cases, 10 with valid geometry, 8 passing the primary gates, and 7 covered. (B) On the eight primary cases, the tabulated MAE is 1.375 for the geometry peak and 1.25 for a predictor fixed at step 4. (C) The regeneration script plots archived oriented differences between auto routing and the universal policy. This panel does not compare auto routing with the always-dense-texture policy discussed in the main text. The routing quantities are descriptive and do not establish a verified budget-controlled advantage. Regeneration from summary CSVs does not validate the upstream endpoint measurements or quality gates.



(a) Dense-texture target progress versus background displacement.



(b) Morphology target progress versus background displacement.

Figure 18: Family-specific descriptive controller trade-offs under the reported nominal budget protocol. The dense-texture policy itself uses smoothed directional disagreement for step selection. These comparisons do not isolate the incremental value of geometry or establish an advantage over the best fixed policy choice.

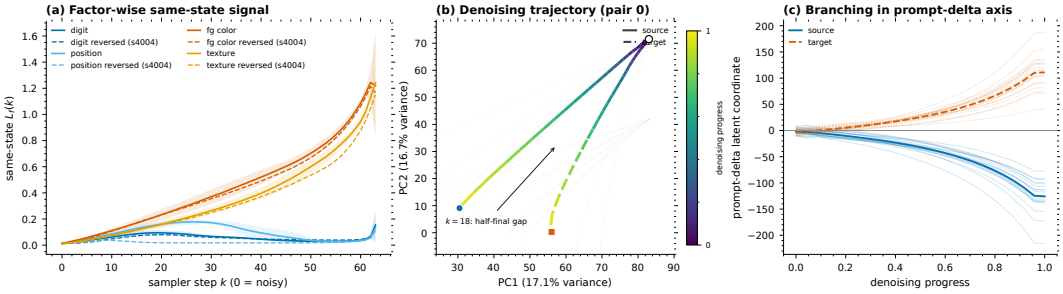
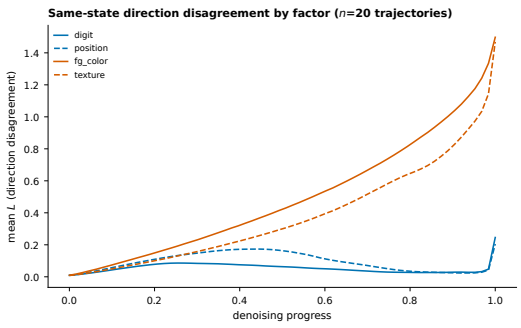
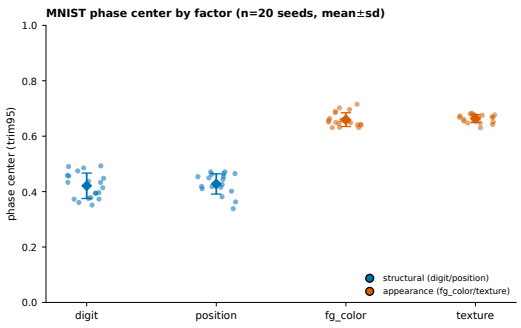


Figure 19: **Overview panels (panels b,c are also shown as Figure 1c,d).** (a) Four-factor same-state signals on trained trajectories (three trained model seeds; 20 trajectories per model; dashed = ν-reversed model 4004); the per-factor view with confidence bands is Figure 3. (b,c) Archived SD3.5 prompt-pair projections (four pairs, five seeds each); the half-final-gap marker is a visualization threshold, not a causal onset estimate.



(a) Mean same-state direction disagreement across 20 trajectories.



(b) Per-trajectory phase centres by factor.

Figure 20: **Structural and appearance regimes separate in the reported trained model.** Structural factor profiles rise earlier than appearance profiles. The terminal spike is excluded from the primary phase statistic; this descriptive result is not a terminal-response or control-utility estimate.

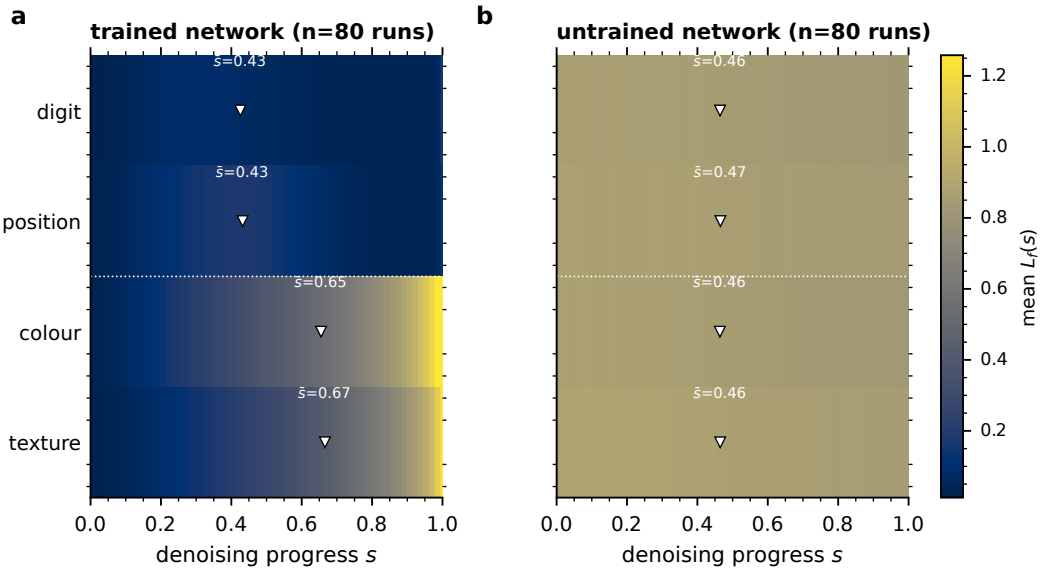


Figure 21: **Factor × step view of Figure 3.** Mean $L_f(s)$ over the same-state control runs (Section 3.1; 4 model seeds × 20 evaluation seeds) as a factor × step heatmap with one shared colour scale; the white marker in each row is the mean trim95 phase centre $\bar{s}_f$ with its value. Panel (b) evaluates the untrained network on its own trajectories. As in Figure 1C, L_f is an evaluator-derived signal along the generative process, not a latent-state coordinate, and no value here is used as a control-utility estimate.

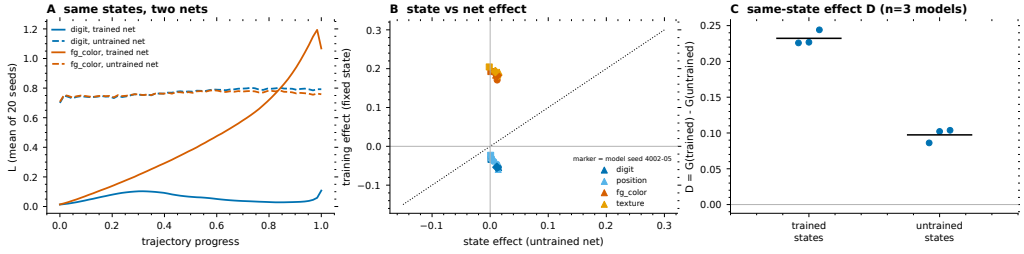


Figure 22: **Same-state decomposition of the four-factor phase signal.** (A) Digit and foreground-color L_f profiles on the same trained-trajectory states, evaluated by trained and untrained networks (model seed 4004; means over 20 analysis seeds). (B) The state effect at a fixed untrained network versus the training effect at fixed trained states, for four factors and model seeds 4002–4005. (C) The secondary between-evaluator contrast $D = G(\text{trained}) - G(\text{untrained})$, where the primary contrast G is the appearance-minus-structural phase-center gap within each evaluator. Across three independently retrained models (seeds 4002–4004), mean D is 0.232 on trained states versus 0.097 on untrained states. Points are model-level values; horizontal bars are their means. These descriptive contrasts do not establish a causal effect of training or downstream intervention benefit.

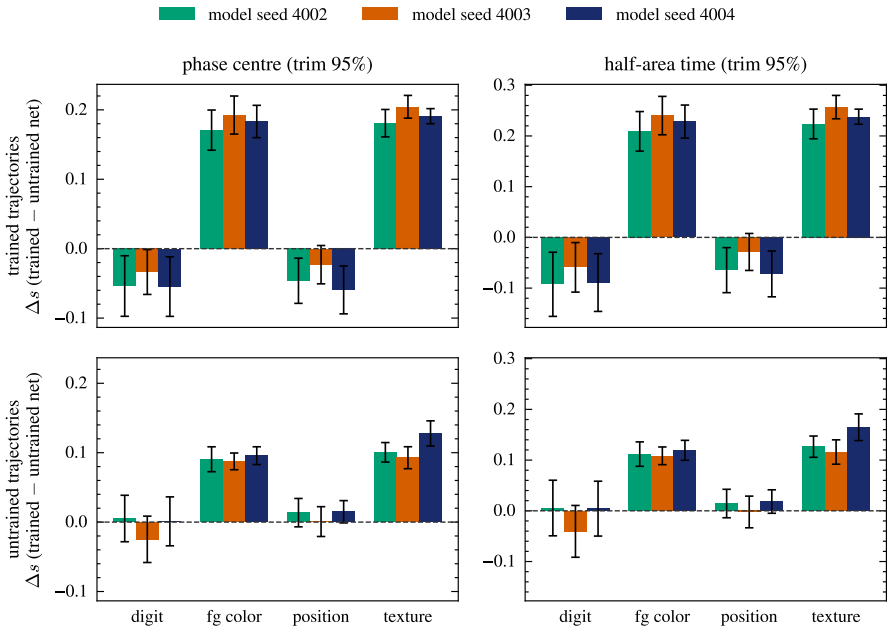


Figure 23: **Same-state controlled comparison on common trained trajectories, three model seeds (4002, 4003, 4004).** The factor-specific trained-minus-untrained contrasts are secondary controls on common trained trajectory states (top row); the bottom row is a separate untrained-trajectory diagnostic. The primary descriptive estimand is the within-trained-network family gap G, reported separately in Figures 4 and 25. A positive evaluator contrast D is not a substitute for a positive within-network gap G. Each bar is the mean difference Δs in denoising progress s (trained minus untrained evaluator network, same states) of the trim-95% phase center $\bar{s}_L$ (left column) or the half-area time $s_{1/2}$ (right column) for one factor and model seed; positive values mean the trained network places the factor later. Error bars are the sample standard deviation across 20 analysis seeds within each model, not between-model uncertainty. The panel is descriptive and does not certify a prospective three-model sign criterion. Summaries are recomputed from committed per-seed phase coordinates and checked against the archived summary table; sources and the plotting script are listed in Section M.

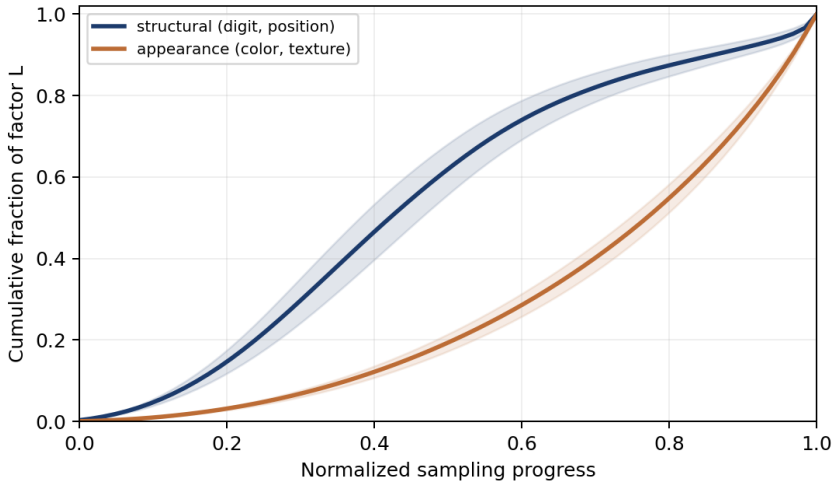


Figure 24: Within-trained-network time-resolved factor ordering on common trained trajectories. Each factor's nonnegative raw L curve is normalized by its own 64-step sum and cumulatively summed; the two structural and two appearance factors are paired within each model and analysis seed before plotting the descriptive mean. Shading is sample SD over 60 paired trajectories, not a model-level confidence interval. The reproducible input is the three archived `L_curves.csv` files; `code/make_same_state_temporal_order_20260924.py` checks complete unique keys before plotting.

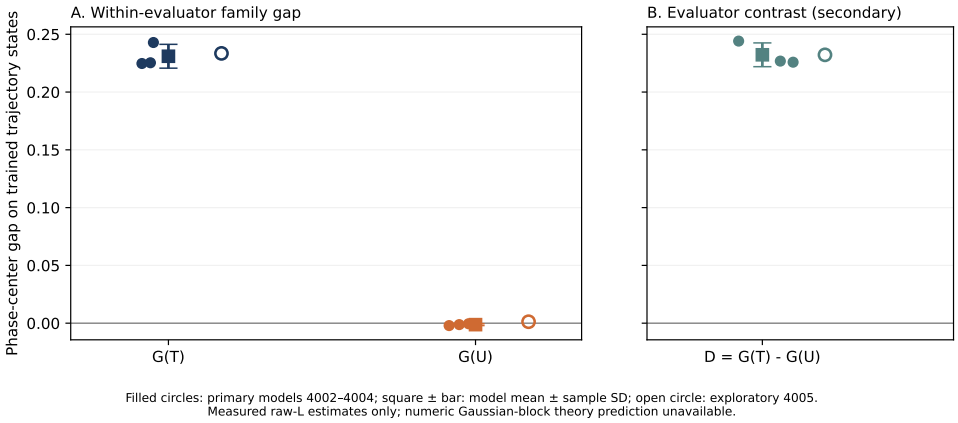


Figure 25: PX-1 measured same-state model-level family gap. Filled points are primary models 4002–4004, squares and bars their mean and sample SD, and open points exploratory model 4005. Panel A shows direct within-evaluator gaps; panel B shows the secondary evaluator contrast D. No numeric Gaussian-block prediction is implied.

Q Recorded H3 Video Outputs and Exploratory Same-State Diagnostics

We report qualitative MiniMax H3 video outputs under archived intervention configurations and an exploratory diagnostic of conditional predictions at common trajectory states. Four prompt pairs were selected after inspection to illustrate changes in clock outline, balloon count, umbrella color, and jacket color. Table 15 provides pair-level qualitative summaries for all 30 prompt pairs, including reference mismatches and accompanying changes. Intervention mode and realized magnitude vary across these comparisons, so the video examples do not identify a family-wide timing effect at a common intervention budget.

Q.1 Protocol and measurement scope

The archive contains 30 source–target prompt pairs and six conditions per pair: SOURCE, TARGET, ALL, EARLY, MIDDLE, and LATE, giving 180 videos. Each condition is generated separately, with one shared sampling seed within each pair and different seeds across pairs. The comparison unit is a prompt pair with one recorded realization per condition. The short-window labels denote adjacent two-step groups within sampler steps 1–6 (Table 12). They refer to the sampling process; playback times identify frames within the completed videos. ALL covers steps 0–7 and therefore includes two endpoints omitted from the union of the short windows. SOURCE and TARGET are separate reference generations under their respective conditions. ALL uses transformed prompt-delta control with online coefficient selection and a cap; it is distinct from the TARGET reference.

Table 12: H3 settings recorded in the saved workflows. Indices are zero-based. The nominal allocation is not an equal-realized-budget guarantee; Section Q.1 defines the realized-magnitude audit.

Setting	Value
Checkpoint	`minimax_h3_fl2va_pruned_int8_convrot.safetensors`
Design	30 prompt pairs $\times$ 6 conditions; one seed per pair
Video	544×288 pixels; 97 frames; 24 fps
Sampling	Euler / simple; 8 steps; CFG 1.0; denoise 1.0
Sigma shift	Video 12.0; audio 3.0
Gates (zero-based)	ALL: 0–7; EARLY: $1, 2$; MIDDLE: $3, 4$; LATE: $5, 6$
Historical allocation	`target_budget` $= 0.5$; `alpha_max` $= 8$
Step weighting	Absolute video-sigma increment divided by its mean over eight steps
Realized magnitude	Unweighted post-cast ratio sums: 0.263381–3.796302 across 120 intervention records
Sampling seeds	P01–P10: 26082211–26082220; P11–P30: 26082311–26082330

The saved workflows identify the generator as MiniMax H3 and specify the checkpoint in Table 12. The direction mode is `fusion_safe` for P01, P12, P13, P16, P17, P19, and P24, and `full` for the remaining pairs. Each pair uses the same mode across its four intervention conditions. In particular, the highlighted shape/count pairs use `full`, whereas the color pairs use `fusion_safe`. The examples therefore compare the outputs of these recorded configurations.

Intervention definition. At the current joint state, let z_i be the free temporal video latents and d_i^0, d_i^a the corresponding source and target denoised predictions. Using the video sigma, the implementation reconstructs Karras ODE update fields

$$u_i^0 = (z_i - d_i^0)/\sigma_i, \qquad u_i^a = (z_i - d_i^a)/\sigma_i, \qquad \Delta_i = u_i^a - u_i^0.$$

Norms and inner products are evaluated on these free video latents. The directions are

$$\mathcal{D}_i^{\text{full}} = \Delta_i, \qquad \mathcal{D}_i^{\text{safe}} = \Delta_i - \min\left(\frac{\langle \Delta_i, u_i^0 \rangle}{\|u_i^0\|^2}, 0\right) u_i^0.$$

For a zero source field, the safe direction is Δ_i. The `fusion_safe` transform removes only the source-opposing parallel component and does not restore the original delta norm.

For gate $\mathcal{G}$, the archived step allocations satisfy $\sum_{i\in\mathcal{G}} b_i = 0.5$ and are nonuniform. The historical rule uses

$$w_i = \frac{|\sigma_{i+1} - \sigma_i|}{\frac{1}{8}\sum_{j=0}^{7} |\sigma_{j+1} - \sigma_j|}, \qquad r_i = \frac{\|\mathcal{D}_i\|}{\max(\|u_i^0\|, 10^{-8})}, \qquad \alpha_i = \min\left(\frac{b_i}{w_i r_i}, 8\right).$$

The coefficient solver rejects $r_i \le 10^{-8}$. Norms are accumulated in FP64 and the ratio passed to the solver is cast to FP32. At an active step, $u_i^{\mathrm{ctrl}} = u_i^0 + \alpha_i \mathcal{D}_i$. The returned video prediction and applied difference are

$$\widetilde{d}_i^{\mathrm{ctrl}} = \mathrm{cast}_{\mathrm{out}}(z_i - \sigma_i u_i^{\mathrm{ctrl}}), \qquad \delta_i^{\mathrm{applied}} = (z_i - \widetilde{d}_i^{\mathrm{ctrl}})/\sigma_i - u_i^0.$$

The reconstruction is evaluated in FP32 and its norms in FP64. Outside the gate, the source prediction is used. Video components outside the free-latent mask and the audio component use the source-conditioned output at each evaluation. Subsequent evaluations depend on the branched state, so completed audio need not match a separate SOURCE trajectory.

Archived budget and realized-magnitude audit. We report two post-cast scalar diagnostics,

$$q_i = \frac{\|\delta_i^{\mathrm{applied}}\|}{\max(\|u_i^0\|, 10^{-8})}, \qquad \widehat{B}_{\mathrm{floor}} = \sum_{i\in\mathcal{G}} q_i, \qquad \widehat{B}_{\mathrm{historical}} = \sum_{i\in\mathcal{G}} w_i q_i.$$

All 120 intervention records were linked to the archived video hashes. The unweighted sums $\widehat{B}_{\mathrm{floor}}$ range from 0.263381 to 3.796302; none is within 10^{-5} of 0.5. Under the historical weighting, 80/120 records are within that tolerance, and 40/120 report coefficient clipping. A common nominal allocation therefore did not yield equal realized intervention magnitude across windows. Table 14 lists the four highlighted pairs.

The unit weights specified for the proposed budget in Appendix I.3 differ from the historical weights. Our post-cast diagnostic also uses a floored denominator, whereas Equation (39) uses additive regularization and the pre-cast direction expression. Both scalar denominator conventions are supplied in the audit CSV; the largest run-level difference is 3.63×10^{-11} here. Recomputing the floor ratios from the stored norms agrees for all 420 active steps. These are retrospective scalar-log checks, not a new tensor-level replay.

The source hashes recorded at custom-node initialization match the five supplied project source files for all 120 records (600/600 comparisons). This checks the recorded file provenance, not the complete historical execution environment. The supplement under `h3_appendix_ordering/audit/` contains the step and run summaries, source hashes, matching implementation, and a scalar verification script.

Visual inspection and prompt design. The archived descriptions inspect frames at playback times 0, 1, 2, 3, and 4 s, recording the requested attribute, scene composition, and visible changes in form. All 180 files decode to the dimensions, frame count, and rate in Table 12, with a stream duration of approximately 4.04 s. The evaluation is descriptive, frame-sampled inspection, not a blinded human study or a video-wide motion or audio assessment. Requested-change categories follow the prompt wording. Material and seasonal requests also modify descriptions of reflection, transmission, vegetation, or ground appearance; P05–P07 additionally change the described acoustic environment.

Q.2 Qualitative observations under recorded configurations

Shape and count. For clock outline (P03), ALL and EARLY produce a square outer frame, whereas MIDDLE and LATE retain the round outline. For balloon count (P04), ALL and EARLY produce three balloons, while MIDDLE and LATE retain one. These examples place the requested shape and count changes in the early-window outputs of the recorded comparisons (Figure 26(a,b)). The accompanying content is retained in the presentation: clock tick marks, size, and frame thickness also vary, and the early balloon output contains a large text-like pattern.

Local color. For umbrella color (P01) and jacket color (P19), LATE produces the requested blue or purple while retaining approximately SOURCE-like person placement (Figure 26(c,d)). ALL also

changes the target color, with differences in person scale and composition. EARLY introduces a strong magenta cast in P01 and a duplicated or ghosted person in P19. These selected outputs show the requested local colors in the recorded LATE conditions. The accompanying scene changes, mode differences, and unequal realized magnitudes limit their interpretation to these configurations.

Variation across the broader set. Mug-to-plant replacement (P18) yields a potted plant under ALL, a plant in a handled container under MIDDLE, and a red mug under LATE. Ceramic-to-glass material (P02) changes both transparency and vase proportions under ALL; LATE retains a white, opaque appearance. For watercolor rendering (P10), EARLY introduces paper-like or granular rendering, while ALL, MIDDLE, and LATE retain volumetric shading and gloss. The TARGET reference has a more pronounced watercolor treatment (Figure 27). These material and rendering changes exhibit temporal behavior beyond the local-color examples. Arm raising (P06), bird takeoff (P22), and reversed image-plane motion (P26) also accompany changes to limb-adjacent patterns, plumage or branch placement, and sculpture form, respectively.

Source and target reference content. The reference generations provide context for every intervention. In lighting direction (P13), the target bust differs in form from the source; in gaze (P17), the source head turns over the video. In pouring (P23), the source already tilts the kettle and directs water toward the cup, although the source prompt specifies a level kettle with zero liquid flow. In depth of field (P15), the source contains blur despite the deep-focus request. For book position (P11), the target contains an additional hand or arm, and ALL places a stack of books on the right (Figure 28). Table 15 retains all 30 pairs, including these reference characteristics and accompanying scene changes.

Q.3 Relation to the structural–appearance phase ordering

The primary result in Section 3.1 is a within-network family gap in the phase centers of conditional predictions evaluated at common trajectory states (Equation (8)). The archived H3 comparisons describe terminal outputs from separately controlled trajectories: they show which requested attributes appear under particular recorded configurations. Their differing intervention modes and realized magnitudes prevent attributing the observed contrast solely to timing or to the structural–appearance grouping.

The exploratory same-state diagnostic in Section Q.4 addresses a closer measurement target, but its small development set and imperfect semantic adherence limit its interpretation. We use this appendix to document qualitative outputs and the limits of transferring the four-factor interpretation to these H3 prompt operations.

Q.4 Exploratory same-state development diagnostic

We evaluated native H3 video predictions under a source condition and four factor-oriented prompt changes at the same stored joint states. Implementation checks confirmed the tested source trajectory, zero difference for identical conditions, agreement under factor-order changes and separate factor evaluation, and replay agreement in a fresh process. These checks establish consistency of the tested measurement path, rather than semantic independence or downstream intervention utility.

The development set contains two scenes (mug/bowl and cube/sphere) and three seeds per scene. A second prompt version was revised after inspecting the first and used the same scene–seed combinations (seeds 2609261001–2609261003). Each version yields 192 factor-by-step measurements and 30 reference videos. The factors are identity and position (structural), and color and texture (appearance). Native video predictions use the floor normalization h_η of Equation (13), with $\eta = 10^{-10}$. Phase centers use progress $i/7$ and trapezoidal integration of numerator and denominator over retained points with progress at most 0.95; the last retained point is $6/7$. This representation differs from the reconstructed Karras update fields used for the archived interventions.

The descriptive mean family gap is $+0.012410$ for v1 and $+0.008747$ for v2 (Table 13). In both versions, identity has a later phase center than both color and texture for every one of the six scene–seed combinations. This is compatible with a positive family-average gap because position has an earlier center. Inspection of reference frames at 0, 2, and 3.5 s found no complete five-condition set satisfying the strict semantic criteria (0/6 in each version). The supplied rubric requires a valid source,

the requested factor change, retention of non-target attributes, and consistency across the inspected frames. Both prompt versions, all seed-level summaries, and the reference judgments are retained in the supplement. These adaptive development results do not constitute independent confirmation or refutation of a general semantic phase ordering in H3.

Table 13: Exploratory H3 development summaries. Each row averages the same two scenes and three seeds per scene. Version 2 was revised after observing version 1. Strict reference sets passed in neither version; all seed-level values and judgments are retained.

Version	Position	Color	Texture	Identity	Mean G	Strict sets
v1	0.157192	0.222335	0.231841	0.272165	$+0.012410$	0/6
v2	0.147017	0.181301	0.230649	0.247439	$+0.008747$	0/6

Table 14: Realized post-cast magnitudes for the highlighted H3 pairs. Values are the unweighted floor-normalized sums $\widehat{B}_{\text{floor}}$. A dagger marks coefficient clipping at one or more steps. SOURCE and TARGET are references, not budget-controlled interventions.

Pair	Mode	ALL	EARLY	MIDDLE	LATE
P03	`full`	1.722444	$3.733026^{\dagger}$	$0.518586^{\dagger}$	$0.272183^{\dagger}$
P04	`full`	2.053055	3.705700	$1.097428^{\dagger}$	$0.411121^{\dagger}$
P01	`fusion_safe`	1.183415	$3.308293^{\dagger}$	$1.464920^{\dagger}$	0.536917
P19	`fusion_safe`	1.261596	$3.287919^{\dagger}$	$1.323437^{\dagger}$	0.542931

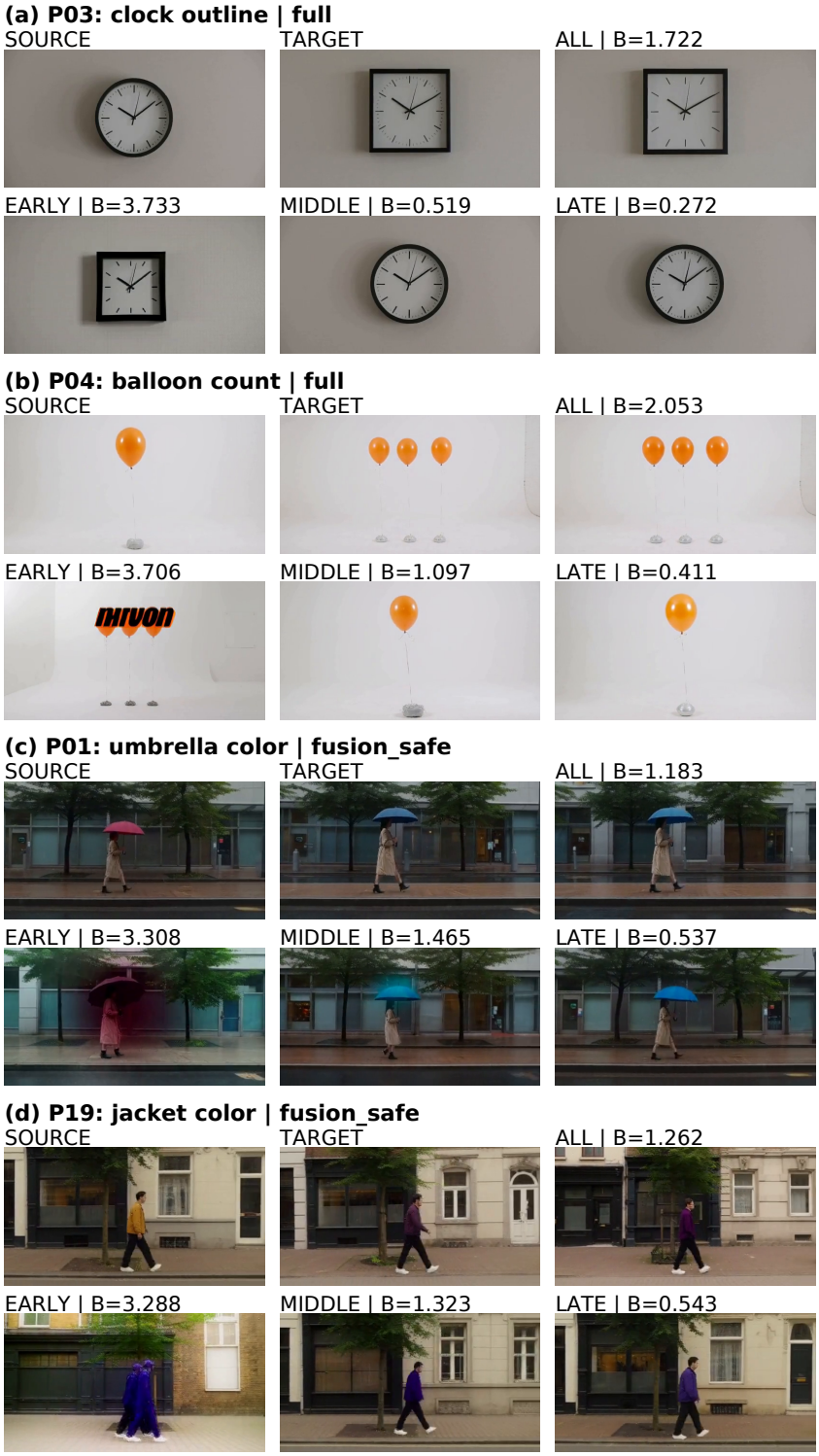
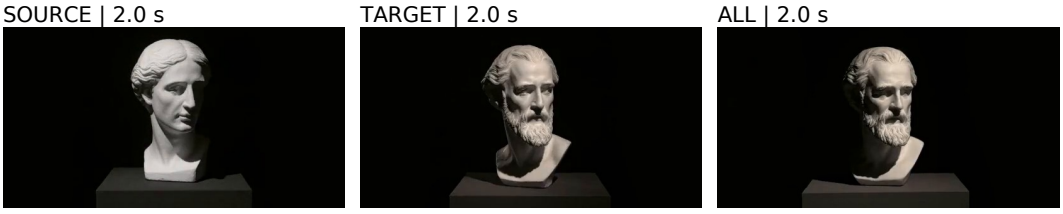


Figure 26: **Selected H3 outputs under archived configurations.** Each block shows six conditions at 4.0 s. In (a,b), ALL and EARLY change clock outline and balloon count; MIDDLE and LATE retain the source attributes. In (c,d), LATE shows the requested local colors. Headings give modes; B labels give $\widehat{B}_{\text{floor}}$ rounded to three decimals (Table 14). This is not an equal-budget family comparison. P04 EARLY also contains a large text-like pattern. The pairs were selected after inspection; Table 15 summarizes all 30 pairs.



Figure 27: **Additional H3 outputs: object replacement and rendering style.** (a) P18, mug-to-plant replacement, shows a potted plant under all-step intervention, a plant with a retained handle under middle-window intervention, and a red mug under late-window intervention. (b) P10, watercolor rendering, shows paper-like or granular rendering under early-window intervention and retained shading and gloss in the all-step, middle, and late outputs. Every panel is the 4.0 s frame from a separate generation. The pair-level qualitative summaries place these outputs alongside the highlighted shape, count, and color examples.

(a) P13: lighting direction and bust form



(b) P17: head direction in the SOURCE reference



(c) P23: pouring already present in SOURCE



(d) P11: book position and scene content



Figure 28: **H3 source and target references and associated scene changes.** (a) P13: source, target-only, and all-step frames at 2.0 s show differences in bust form alongside the lighting request. (b) P17: source frames at 0, 2, and 4 s show changing head direction. (c) P23: source frames at the same times already show pouring. (d) P11: the 0 s source, target-only, and all-step frames show, respectively, a single book, an additional arm, and multiple books. The reference content and intervention outcomes are recorded together for each requested change.

Table 15: Pair-level qualitative summaries for all 30 H3 prompt pairs. Categories follow the prompt wording. Each row summarizes selected observations for one seed, with reference or non-target content listed separately; it is not an exhaustive six-condition rating record. ALL denotes all-step intervention.

ID	Requested change	Observed outcome	Reference or non-target content
P01	Umbrella color	LATE produces a blue umbrella with approximately SOURCE-like person placement. ALL also changes the umbrella to blue.	EARLY has a strong magenta cast; MIDDLE has blue coloration with a halo. ALL also changes the person's image scale.
P02	Vase material	TARGET and ALL have a transparent, glass-like appearance. MIDDLE also changes transmission and reflection. LATE retains a white, opaque appearance.	ALL changes body and neck proportions relative to SOURCE. EARLY has an elongated neck and ribbing; MIDDLE also changes the shoulder shape.
P03	Clock outline	ALL and EARLY produce a square outline; MIDDLE and LATE retain a round outline.	ALL has fewer fine tick marks. EARLY also changes clock size and frame thickness.
P04	Balloon count	ALL and EARLY produce three balloons; MIDDLE and LATE retain one.	EARLY contains large text-like patterns. ALL also changes individual balloon image scales relative to SOURCE.
P05	Weather	ALL produces a snowy scene. EARLY, MIDDLE, and LATE show different partial appearance changes.	ALL changes viewpoint, planting arrangement, and sculpture scale alongside the snowfall.
P06	Person action	ALL, EARLY, and MIDDLE show raised-arm poses. LATE remains closer to stepping in place.	EARLY and MIDDLE contain colored or wing-like patterns around the arms. ALL also changes background and composition.
P07	Dog action	ALL, EARLY, and MIDDLE show a jumping phase; LATE shows walking.	Dog image scale, obstacle orientation and placement, and background change alongside the action.
P08	Bicycle tracking shot	ALL and EARLY show background motion suggestive of tracking. LATE is closer to a fixed shot.	ALL substantially changes viewpoint height and bicycle scale. The tracking interpretation of MIDDLE remains ambiguous.
P09	Camera push-in	TARGET and ALL clearly enlarge the person and workbench in the frame. EARLY also tends toward enlargement.	The visible enlargement is less pronounced in MIDDLE and LATE than in TARGET and ALL. EARLY has whitish, overexposed-looking regions.
P10	Watercolor style	EARLY shows paper-like or granular rendering. TARGET clearly changes outlines and shading toward watercolor.	ALL, MIDDLE, and LATE retain volumetric shading and gloss, with a more three-dimensional appearance than TARGET.
P11	Book position	ALL places a stack of multiple books on the right.	An additional hand or arm appears at the start of TARGET. Book position, book count, and scene content are recorded separately.
P12	Cube scale	ALL produces a large cube; MIDDLE and LATE retain a small cube.	EARLY adds a ring-like floor pattern. Cube size and shadow in ALL differ from TARGET.
P13	Lighting direction	TARGET, ALL, and EARLY show bust forms different from SOURCE.	Features such as the beard change between the reference outputs. The comparison includes both lighting and bust form.
P14	Midday to sunset	ALL produces a sunset scene; EARLY strongly warms the colors. MIDDLE shows partial sunset elements; LATE remains closer to midday.	Lighthouse size and position and coastline shape also differ. EARLY contains an overexposed-looking band.
P15	Depth of field	TARGET, ALL, and EARLY show strong blur.	SOURCE already contains blur despite the deep-focus request. Chess-piece arrangement and viewpoint also change.

Continued on the next page.

Table 15 (continued)

ID	Requested change	Observed outcome	Reference or non-target content
P16	Facial expression	In addition to ALL, MIDDLE and LATE show changes around the mouth consistent with a smile.	Smile-like changes appear across multiple intervention windows and accompany changes in facial appearance.
P17	Gaze direction	The SOURCE head turns sideways over time. ALL and EARLY are more front-facing.	The prompt specifies a fixed head, while the reference output includes changing head direction. Gaze and head orientation vary together.
P18	Object replacement	ALL produces a potted plant. MIDDLE combines a plant with a handled container. LATE retains a red mug.	EARLY produces a plant in a yellow container. MIDDLE retains a handle and combines container and plant features.
P19	Jacket color	ALL, MIDDLE, and LATE produce purple coloration. LATE retains SOURCE-like person placement.	EARLY produces a duplicated or ghosted person and a strong purple cast.
P20	Outerwear type	ALL resembles a long raincoat. EARLY and MIDDLE also change garment length.	EARLY and MIDDLE contain unnatural color bands or gradients. LATE retains a short jacket, contrasting with the requested long raincoat.
P21	Vehicle left turn	ALL, EARLY, and MIDDLE show the main vehicle turning left; LATE shows straight motion.	Intersection appearance and the presence of other vehicles also change. Trajectory change and background preservation are distinct.
P22	Bird takeoff	ALL, EARLY, and MIDDLE show the bird leaving the branch; LATE keeps it perched.	MIDDLE lightens the plumage and changes branch placement alongside takeoff.
P23	Kettle pouring	SOURCE already tilts the kettle and directs a water stream toward the cup. ALL also shows pouring.	Pouring is already present in SOURCE despite the zero-flow request. EARLY has spiky water-stream artifacts.
P24	Candle ignition	ALL and EARLY show a flame; MIDDLE has a small flame. A small glow or flame is also visible in LATE.	Flame-like changes appear in multiple windows. ALL and EARLY also change candle length and its proportions relative to the holder.
P25	Door opening	ALL, EARLY, and MIDDLE show opening. LATE shows partial opening near the end of the video.	Hallway lighting and fixtures also differ across intervention conditions.
P26	Pan direction	In ALL, the sculpture moves in the image plane in the opposite direction to SOURCE.	ALL substantially changes sculpture shape. EARLY has dark regions and text-like patterns; camera behavior is described through image-plane motion.
P27	Tilt direction	ALL shifts the view from the upper building toward the sky, unlike SOURCE.	EARLY resembles an upward translational viewpoint change and has flare. The reported observation is the change in visible field of view.
P28	Arc direction	ALL moves from a more frontward to a more rearward view of the motorcycle, opposite to SOURCE.	EARLY also shows a partial TARGET-like progression, with a different angular range and motorcycle form. Motion is described through successive views.
P29	Running speed	ALL and MIDDLE show longer strides and sprint-like poses. LATE remains closer to SOURCE-like jogging.	EARLY has a glow-like pattern above the head. Background motion and camera tracking also differ; speed is described through pose and apparent motion.
P30	Season	ALL, EARLY, and MIDDLE show red hues associated with autumn foliage. LATE retains a cherry-blossom-like appearance.	EARLY strongly reddens the ground as well. Canopy density, sky, and benches also change.