BEYOND LOCAL PREDICTION FREQUENCIES: HIDDEN PATH COUPLING IN MASKED DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion language models (MDLMs) are trained through local masked predictions, but generation composes these predictions along complete reveal schedules. A broad class of training objectives and train–inference alignment methods depends on a schedule only through how often each local prediction problem is used, so frequency-matched schedules induce the same training objective. We prove that this training equivalence does not imply generation equivalence: schedules with the same objective can still produce different output distributions. We characterize exactly what information this frequency-based view loses and how the missing structure changes generation, and further prove that training generically becomes sensitive to it in our shared-representation population model; a controlled Transformer shows the same effect under symmetric masking. Near mutually consistent local predictions, we also identify the minimal additional schedule information whose matching removes every second-order discrepancy and derive a likelihood-based test requiring no retraining. On MDLM-OWT, output distributions differ by a median 43% in total variation across randomly sampled schedule pairs with identical local prediction frequencies.

1 INTRODUCTION

Masked diffusion language models do not generate in a fixed left-to-right order. They start from a fully masked sequence and progressively reveal positions; once a token is filled, it becomes part of the context for later predictions. Training, however, is typically organized as local masked-prediction problems: given the currently visible tokens, predict one of the remaining masked positions. Generation must then compose these local predictions along a complete reveal order. A reveal schedule therefore plays two roles at once: it determines how often different local prediction problems are encountered, and how those predictions are assembled into a complete sample. This raises a basic question: if two schedules use exactly the same local prediction problems with exactly the same frequencies, must they generate the same distribution?

A broad class of any-order, absorbing, and train–inference alignment methods effectively compares schedules through these local usage frequencies (Hoogeboom et al., 2022; Shih et al., 2022; Ou et al., 2025). For objectives that sum one-step prediction losses, this dependence is exact: the complete reveal order does not enter an individual loss term, and the schedule matters only through how often each local prediction problem is used. Hence, two schedules with matching frequencies induce exactly the same training objective, for any model and dataset in this class. Frequency-based train–inference alignment rules see the same information and therefore cannot distinguish the two schedules either. Matching all local prediction frequencies thus gives an exact notion of equivalence at the level of the training objective. What remains unestablished is whether this training equivalence is sufficient for generation.

We prove that it is not. Figure 1 gives the smallest counterexample. With only four positions, two schedules can use the same local prediction problems at the same frequencies and therefore induce the same training objective, yet produce different output distributions. The two schedules contain the same early routes and the same later generation paths; only the pairing between them changes. Because later predictions condition on values produced earlier, changing this pairing can change the final distribution. The missing information is therefore not which local predictions are used, but how they are combined along complete reveal orders. We refer to this missing pairing as hidden path

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.

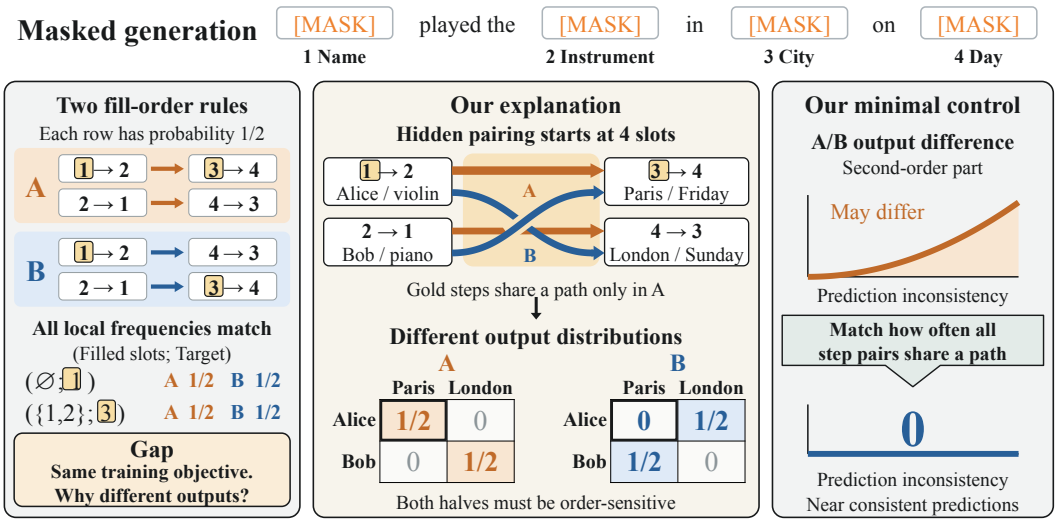


Figure 1: Hidden path coupling and second-order control. With the same predictor, schedules A and B share all local prediction frequencies and the training objective, yet different pairings produce different output distributions. This ambiguity first appears at four positions. Near consistent conditional predictions, also matching all step-pair co-occurrence frequencies eliminates the second-order output difference.

coupling. This observation leads to the three questions studied in the rest of the paper: when does the missing pairing actually change generation, when does training learn sensitivity to it, and what additional information is needed to remove the resulting ambiguity?

Contributions.

(i) **We compute exactly what local prediction frequencies leave unidentified.** They determine the complete schedule law if and only if there are at most three positions. From four positions onward, hidden degrees of freedom appear; at the first threshold, we characterize the six minimum-support four-path switches that span the entire hidden space.

(ii) **We derive the exact generative mechanism.** For any such hidden switch, the change in the terminal distribution factorizes into two terms: the discrepancy created by the early routes and the discrepancy between their later generation paths after the routes first meet. This gives an exact criterion for when hidden path coupling changes generation.

(iii) **We prove that this sensitivity is generically induced by learning in the shared-representation population model we analyze.** Whenever the shared representation leaves predictive structure unexplained, avoiding order dependence requires a special cancellation; outside this exceptional set, training acquires sensitivity to the hidden pairing. A controlled Transformer further shows that symmetric masking alone does not prevent the effect.

(iv) **We characterize the minimal additional path information needed to control the gap and give a no-retraining certificate.** Near mutually consistent local predictions, matching how often any two prediction steps occur on the same complete path, in addition to matching their individual frequencies, removes every second-order difference in the generated distribution, and this extra information is minimal for the corresponding perturbation class. We also derive a four-likelihood certificate that detects the effect without retraining.

2 Background and Motivation

In a broad class of existing methods, local prediction frequencies determine how the schedule enters training and are also used for train–inference alignment. Trajectory generators compose local transitions during generation (Sohl-Dickstein et al., 2015; Ho et al., 2020; Austin et al., 2021; Campbell et al., 2022; Lou et al., 2024); masked, absorbing, and any-order objectives express training as one-step predictions under different visible contexts and aggregate the corresponding losses (Sahoo

et al., 2024; Shi et al., 2025; Zheng et al., 2025; Amin et al., 2025; Hoogeboom et al., 2022). The reveal schedule determines how frequently each prediction problem contributes to this objective; we call these frequencies *local exposure*. Train–inference alignment uses the same information, reweighting training or comparing schedules according to how often those predictions are invoked during generation (Shih et al., 2022; Ou et al., 2025). As reveal orders are randomized, learned, or planned, schedules have also become an explicit generation design variable (Garg et al., 2025; Pang et al., 2025; Yu et al., 2025; Wang et al., 2025; Kim et al., 2025; Peng et al., 2025), with recent work further examining the value of such flexibility (Hong et al., 2026; Ni et al., 2026; Zhang et al., 2026). Thus, the same local-frequency statistics connect what the model practices during training with what it uses at inference. Methods that use richer schedule information are discussed separately in Appendix A (Xue et al., 2025; Du et al., 2026; Luxembourg et al., 2025; Xu et al., 2026).

These training objectives depend only on how often each local prediction is used; generation also depends on how those predictions are combined. In the objectives above, local prediction losses are added independently. After averaging over reveal orders, the schedule affects the objective only through how often each prediction problem occurs. Thus, if all local usage frequencies match, two schedules induce exactly the same training objective. Which other predictions tend to appear with a given prediction in the same complete order does not affect this additive objective. Generation is different: the model executes these predictions sequentially along a complete reveal order, and later predictions condition on values produced earlier. Local frequencies therefore retain everything the training objective needs, but discard how predictions are combined during generation.

Prior work shows both that identical local counts can correspond to different complete orders and that different generation orders can lead to different outcomes, but it does not tell us when the order differences hidden from training actually affect generation. Ranking and combinatorial work represents permutations as complete paths and studies the remaining freedom in path distributions when local transition counts are fixed (Sturmfels & Welker, 2012; Csiszár, 2009; Schrijver, 2003; Dobra, 2003; Takemura & Hara, 2012; Gasanova & Nicklasson, 2024). On the other hand, when the local conditionals across mask contexts all come from a single joint distribution, a *compatible family*, the chain rule guarantees that every reveal order produces the same final distribution (Arnold et al., 1999; Majid et al., 2025). Generative work has also studied schedule-dependent sampling divergence and exact likelihood under specific unmasking orders (Chen et al., 2026; Turok et al., 2026). This leaves three concrete questions: how much of the schedule remains undetermined once all local prediction frequencies are fixed; which of those hidden differences actually change the final distribution; and whether training itself learns sensitivity to them. The first question is model-independent, so we begin there: how tightly do local prediction frequencies determine the complete schedule?

3 What Local Prediction Frequencies Miss

Does local edge exposure determine the complete schedule law over d positions? For $d \leq 3$, yes; from $d = 4$ onward, no. At the first failure, six minimal four-path switches span the hidden space.

Definition 1 (Reveal interface)**.** We study one-time-reveal generators. A schedule draws a complete coordinate order independently of the generated values and reveals one new coordinate per step. At state (S, x_S), the predictor sees the revealed set and values but not the earlier order. Revealed coordinates are never revised, so the terminal assignment retains every intermediate value.

This interface is realized by LLaDA when the number of steps equals the generation length and positions are selected uniformly at random (Nie et al., 2025), and by the MDLM ancestral sampler in the one-reveal-per-step limit (Sahoo et al., 2024).

A complete order π is a path through the Boolean lattice: vertices are revealed sets $S \subseteq [d]$, and edge $e = (S, i)$ reveals $i \notin S$. Let $\mathcal{E}_d$ collect these edges. For schedule law ρ on the permutations $\mathfrak{S}_d$, the edge map records local exposure,

$$T(\rho)_{S,i} = \Pr_{\Pi \sim \rho} \{\{\Pi_1, \ldots, \Pi_{|S|}\} = S,\ \Pi_{|S|+1} = i\}. \tag{1}$$

Thus local prediction problem $e = (S, i)$ has exposure $(T\rho)_e$. Schedules are *edge-equivalent* when $T\rho = T\sigma$; since an edge-additive objective is $\sum_{S,i} T(\rho)_{S,i} L_{S,i}$, this is exactly the training equivalence in Section 2.

Theorem 2 (Edge-observability threshold)**.** Let $\mathcal{K}_d = \ker T$. Then

$$\dim(\mathcal{K}_d) = d! - d2^{d-1} + 2^d - 2. \tag{2}$$

Consequently, the edge map is injective on schedule laws if and only if $d \leq 3$.

Flow conservation and path decomposition identify $T(\Delta(\mathfrak{S}_d))$ with the Boolean-lattice unit-flow polytope. Its dimension is $d2^{d-1} - 2^d + 1$ (Sturmfels & Welker, 2012, Section 4); rank–nullity gives the theorem (Appendix B).

The first failure is the six-dimensional kernel at $d = 4$. For two-element meeting set $S \subset [4]$, take its two prefix orders a, a' and suffix orders b, b'. Writing ab for concatenation, define

$$c_S = \delta_{ab} + \delta_{a'b'} - \delta_{ab'} - \delta_{a'b} \tag{3}$$

This *tail switch* changes the hidden prefix–continuation pairing without changing any edge exposure, so $Tc_S = 0$; its signed vector c_S is a *circuit* (Figure 1).

Corollary 3 (First hidden circuits)**.** At $d = 4$, the six vectors $(c_S)_{|S|=2}$ have disjoint path supports and form a basis of $\mathcal{K}_4$. They are exactly the nonzero kernel circuits with minimum support, up to scale and sign. Edge marginals together with one prefix–suffix joint probability for each S determine all 24 schedule probabilities; no augmentation by fewer than six scalar linear statistics is injective.

Each support is a 2×2 prefix–continuation table whose margins are fixed but interaction is not. The six supports partition $\mathfrak{S}_4$, giving the basis and sharp augmentation (Appendix B).

Revealed-set inputs such as diffusion time leave this map unchanged; block reveal and revision are treated in Appendix C.6. The gap is informational: more independent edge observations cannot recover $\ker T$ (Appendix B.5). We next ask which hidden directions change a fixed model's output.

4 When Hidden Coupling Changes Generation

Hidden schedule directions need not change generation; a fixed model decides which survive. A shared conditional family q induces the path law and schedule mixture

$$Q^q_\pi(x) = \prod_{t=1}^{d} q_{\pi_t}(x_{\pi_t} \mid x_{\pi_{<t}}), \qquad Q^q_\rho = \sum_{\pi \in \mathfrak{S}_d} \rho(\pi) Q^q_\pi. \tag{4}$$

Only the coupling of edges into paths changes across orders. For M_q with columns Q^q_π, define the *edge-fiber susceptibility*

$$\mathrm{Gap}_T(q) = \max_{\rho,\sigma \in \Delta(\mathfrak{S}_d):\, T\rho = T\sigma} \mathrm{TV}(Q^q_\rho, Q^q_\sigma). \tag{5}$$

Equivalently, $\mathrm{Gap}_T(q) = \max\{\frac{1}{2}\|M_q a\|_1 : a \in \ker T, \|a\|_1 \leq 2\}$, and it vanishes exactly when M_q factors through T (Appendix C).

4.1 The two-sided response mechanism

Consider a general tail switch. Let prefixes a, a' meet at S, inducing laws $A_a, A_{a'}$ on x_S, and let suffixes b, b' induce continuations $K_b, K_{b'}$. The straight and crossed laws $\rho_+ = (\delta_{ab} + \delta_{a'b'})/2$ and $\rho_- = (\delta_{ab'} + \delta_{a'b})/2$ differ by $c/2$, so their terminal TV is $\|M_q c\|_1/4$.

Theorem 4 (Two-sided tail-switch response)**.** For every complete assignment x,

$$(M_q c)(x) = [A_a(x_S) - A_{a'}(x_S)][K_b(x_{S^c} \mid x_S) - K_{b'}(x_{S^c} \mid x_S)]. \tag{6}$$

If $\delta_{\mathrm{pre}} = \mathrm{TV}(A_a, A_{a'}) > 0$ and w_c is the normalized absolute prefix difference, then the circuit response is

$$r_q(c) := \tfrac{1}{4}\|M_q c\|_1 = \delta_{\mathrm{pre}}\, \mathbb{E}_{x_S \sim w_c}\, \mathrm{TV}(K_b(\cdot \mid x_S), K_{b'}(\cdot \mid x_S)). \tag{7}$$

Consequently, $\mathrm{Gap}_T(q) > 0$ exactly when some switch has nonzero prefix and suffix discrepancies at the same retained context.

Expanding the four path laws gives $(A_a - A_{a'})(K_b - K_{b'})$ pointwise; retaining x_S prevents averaging across meeting contexts. Since tail switches span $\ker T$, this circuit condition also proves the global equivalence (Appendix C.3).

Exact four-likelihood expectation identities. The four forced-order scores yield bounded rewards with expectations δ_{pre} and $r_q(c)$. When $\delta_{\mathrm{pre}} > 0$, their ratio recovers the aligned continuation factor; the ratio itself need not be unbiased. Since both factors lie in $[0, 1]$,

$$r_q(c) \le \min\left\{\delta_{\mathrm{pre}},\ \mathbb{E}_{x_S \sim w_c} \mathrm{TV}\left(K_b(\cdot \mid x_S), K_{b'}(\cdot \mid x_S)\right)\right\}, \tag{8}$$

so either factor upper-bounds response (Appendix F.2).

At $d = 4$, write $r_S(q) := r_q(c_S)$. Disjoint supports reduce susceptibility to the largest of six circuit responses.

Corollary 5 (Exact susceptibility at $d = 4$)**.** For $d = 4$, $\mathrm{Gap}_T(q) = \max_{S \in \binom{[4]}{2}} r_S(q)$, so the structural threshold yields an exact six-test diagnostic.

A full-support fixed-floor construction shows that response needs neither zero probabilities nor unbounded logits (Corollary 10). We next ask whether edge-only training creates a responsive model.

5 How Learning Creates Susceptibility

Why should training create the fixed-model response above? We answer globally for a shared representation, then locally for symmetric training near compatibility.

5.1 Exact learning results in a shared-representation model

Pair coordinates as $B_1 = \{1, 2\}$ and $B_2 = \{3, 4\}$. Pair j has binary mean m_j and odd-first probability p_j. A shared model uses mask-derived orientation $z_i(S) \in \{-1, 1\}$ and logistic function σ:

$$q_{\beta,\theta,i}(1 \mid x_S) = \sigma(\beta + \theta z_i(S)). \tag{9}$$

Both predictions have $z = -1$ under odd-first reveal and $z = +1$ otherwise, giving orientation probabilities $u = \sigma(\beta - \theta)$ and $v = \sigma(\beta + \theta)$. The model sees the mask, not the schedule identity or cross-pair coupling.

Theorem 6 (Population learning can create susceptibility)**.** There exist full-support product data, full-support edge-equivalent schedules ρ_+, ρ_-, and the model in Equation (9) such that the standard edge-additive population NLL is identical under both schedules, has a unique finite minimizer q^*, and $\mathrm{TV}(Q^{q^*}_{\rho_+}, Q^{q^*}_{\rho_-}) > 0$; an explicit full-support subfamily approaches $4/27$.

Strictly convex weighted cross-entropies give

$$u^* - v^* = \frac{(p_1 - p_2)(m_1 - m_2)}{(p_1 + p_2)(2 - p_1 - p_2)}. \tag{10}$$

Thus $u^* \neq v^*$ exactly when both pair means and precedence probabilities differ. Edge-equivalent schedules preserve each p_j and the objective, yet recouple $R_s := \mathrm{Bern}(s^*)^{\otimes 2}$ for $s \in \{u, v\}$; their terminal difference is proportional to $(R_u - R_v)^{\otimes 2}$ (Appendix D).

For K problems, let representation row j be $\phi_j^\top$ and use $\sigma(\alpha^\top \phi_j + \theta z_i(S))$. At the order-blind optimum α_0, define residual $e_j = m_j - \sigma(\alpha_0^\top \phi_j)$ and precedence bias $b_j = 2p_j - 1$.

Theorem 7 (Residual–precedence criterion)**.** Under full support, interior precedence probabilities, and a full-rank representation matrix, the population NLL has a unique finite minimizer (α^*, θ^*), and $\theta^* = 0$ if and only if $b^\top e = 0$.

The order-direction derivative at $(\alpha_0, 0)$ is $2b^\top e$, and strict convexity turns stationarity into the equivalence. Order-blindness therefore requires exact fit ($e = 0$) or accidental orthogonality ($e \neq 0$ but $b^\top e = 0$).

Corollary 8 (Susceptibility is generic in this model class)**.** Within the shared-representation population model above, let ϱ be the rank of the representation matrix. If ϱ equals the number K of prediction problems, then $e \equiv 0$ and $\theta^* = 0$ always. If $\varrho < K$, then $\{m : e(m) = 0\}$ is a closed ϱ-dimensional submanifold of $(0, 1)^K$, and for every m outside it the precedence profiles with

$b^\top e = 0$ form a null hyperplane; hence $\theta^* \neq 0$ for almost every data distribution and precedence profile. When $K \geq 2$, every such fit has $\mathrm{Gap}_T(q^*) > 0$ by coupling any two pair blocks in an edge-equivalent switch.

The manifold–hyperplane argument and coupling construction are proved in Appendix D.4, together with a near-cancellation bound.

Per-pair extension. With vector order parameter θ acting through u_j, the criterion is $\sum_j b_j e_j u_j = 0$. Tying permits aggregate cancellation; complete untying requires each $b_j e_j = 0$. The full rank statement is Theorem 16; Section 7 tests this distinction.

5.2 Symmetric training near a compatible anchor

At zero bias, we ask whether symmetric training preserves compatibility. Changing a joint distribution induces consistent changes in its conditionals. Fitting projects these onto directions the shared model can represent, potentially breaking their consistency and exposing hidden path coupling.

Formally, for centered joint score h, let $(Dh)_{S,i} := \mathbb{E}_P[h \mid X_{S\cup\{i\}}] - \mathbb{E}_P[h \mid X_S]$ and $\mathcal{C}_P = \mathrm{im}\, D$. Let $\mathcal{M}_q$ be the model tangent and $\Pi^{(w)}_{\mathcal{M}_q}$ projection under $\langle r, s\rangle_w = \sum_e w_e \mathbb{E}_P[r_e s_e]$.

Weighted projection at a compatible anchor. Under Proposition 18's regularity conditions, the fitted conditional-score direction is

$$u_h = \Pi^{(w)}_{\mathcal{M}_q} Dh,$$

and training preserves every compatible first-order direction exactly when

$$\Pi^{(w)}_{\mathcal{M}_q} \mathcal{C}_P \subseteq \mathcal{C}_P.$$

Coordinate-symmetric weights do not force this invariance. A projected direction outside $\mathcal{C}_P$ yields quadratic response for a circuit with nonzero pair-moment contraction (Corollary 19). We next identify the schedule moments controlling these responses.

6 Controlling and Diagnosing the Gap

Path moments minimally control response near compatibility; two certificates bound the gap.

6.1 The minimal path information for response control

Let $E(\pi)$ be the edges visited by order π. An edge set A is *co-visitable* if it lies on some complete path, with moment $H_A(\rho) = \Pr_{\Pi\sim\rho}\{A \subseteq E(\Pi)\}$. Collecting moments through degree k gives Φ_k; in particular, $\Phi_1 = T\rho$ and Φ_2 is represented by $(T\rho, J_\rho)$, where $J_\rho(e,f) = \Pr\{e, f \in E(\Pi)\}$.

Around a full-support compatible joint P, make the legal bounded perturbation

$$q_{e,t}(x_i \mid x_S) = P(x_i \mid x_S)[1 + t\, u_e(x_i, x_S)], \qquad \mathbb{E}_P[u_e(X_i, X_S) \mid X_S] = 0. \tag{11}$$

Expanding the conditionals along each complete path gives the exact polynomial

$$Q_{\rho,t}(x) - Q_{\sigma,t}(x) = P(x) \sum_A [H_A(\rho) - H_A(\sigma)]\, t^{|A|} \prod_{e\in A} u_e(x), \tag{12}$$

pairing degree-j model defects with degree-j schedule moments.

Theorem 9 (Path moments are the minimal response coordinates)**.** For one-time reveal generation over any finite alphabet with at least two symbols and every k, the equality $\Phi_k(\rho) = \Phi_k(\sigma)$ holds if and only if

$$\mathrm{TV}(Q_{\rho,t}, Q_{\sigma,t}) = O(|t|^{k+1}) \quad (t \to 0) \tag{13}$$

for every full-support compatible anchor and every perturbation in Equation (11). If the first mismatching moment has degree j, some legal shared perturbation has exact response $c|t|^j$ with $c > 0$. Hence any schedule statistic whose equality guarantees Equation (13) for the full class must determine Φ_k. Minimality is relative to that class: a model that ties parameters across edges realizes a quotient of it.

Centered contrasts isolate the first mismatch; Appendices E and E.7 give the proof and tying quotient.

The second-order consequence. After edge matching, pairwise exposure is the minimal universal addition removing every quadratic response. If $T\rho = T\sigma$, $J_\rho = J_\sigma$, and every conditional is within multiplicative deviation α of a compatible anchor, then

$$\mathrm{TV}(Q^q_\rho, Q^q_\sigma) \le \min\left\{1,\ \mathrm{TV}(\rho,\sigma)\sum_{j=3}^{d}\binom{d}{j}\alpha^j\right\}. \tag{14}$$

This cubic envelope is $O((d\alpha)^3)$ when $d\alpha \ll 1$. Co-visitation estimates J unbiasedly, with uniform error $\sqrt{\log(2M/\delta)/2N}$ over M prescribed pairs; Section 7.4 compares T-matched and $T + J$-matched contrasts on a frozen checkpoint.

When the hierarchy identifies the entire schedule. Full identification is harder: Φ_k identifies complete-order laws iff $k \ge \lfloor d/2 \rfloor$ (Theorem 22). Even-rank edges recover two positions at a time; a parity construction defeats every lower degree.

6.2 Two complementary certificates

The four-likelihood identity from Section 4 now becomes a finite-sample lower certificate.

A no-retraining lower certificate from four likelihoods. Fix a circuit before confirmation. Let $P^{\pm}_c$ be the equal mixtures of the two path laws on each side, and draw N independent completions from $(P^+_c + P^-_c)/2$. Four frozen-order scores give

$$W_c(X) = \frac{|P^+_c(X) - P^-_c(X)|}{P^+_c(X) + P^-_c(X)} \quad (0/0 := 0), \tag{15}$$

with $\mathbb{E}W_c = r_q(c)$. Hoeffding (1963) then gives

$$L_c(\delta) := \left[\frac{1}{N}\sum_{n=1}^{N} W_c(X_n) - \sqrt{\frac{\log(1/\delta)}{2N}}\right]_+ \le r_q(c) \le \mathrm{Gap}_T(q) \tag{16}$$

with probability at least $1 - \delta$. In log space, $W_c = |\tanh((\ell_+ - \ell_-)/2)|$, where each $\ell_\pm$ is a two-path log-sum-exp. The test needs independent completions but no retraining, discriminator, or schedule/terminal enumeration.

At $d = 4$, the six-circuit maximum gives the simultaneous bracket $\max_S[\hat{r}_S - \varepsilon_S]_+ \le \mathrm{Gap}_T(q) \le \max_S[\hat{r}_S + \varepsilon_S]$, with $\varepsilon_S = \sqrt{\log(12/\delta)/2N_S}$ (Appendix F.3). The two sides certify susceptibility and insensitivity, respectively, with selection separate from confirmation.

A population upper bound from training risk. For common flow $f = T\rho = T\sigma$, let $\kappa_e(q;\mu)$ be the population conditional KL regret and $\mathcal{E}_f = \sum_e f_e\kappa_e$. Telescoping conditional entropies gives

$$\mathcal{E}_f(q;\mu) = \textstyle\sum_e f_e \mathbb{E}_\mu[-\log q_i(X_i \mid X_S)] - H(\mu).$$

For full-support μ,

$$\mathrm{TV}(Q^q_\rho, Q^q_\sigma) \le \min\{1, \mathcal{E}_f(q;\mu)\};$$

the full theorem is Theorem 25. This upper bound covers the fiber but needs μ and $H(\mu)$; the likelihood certificate lower-bounds one circuit from model scores. We now test the claims in order.

7 Experiments

We compare generation under schedules with identical edge frequencies, keeping the predictor fixed, and measure total variation (TV) between output distributions, averaged over prompts. Switch responses use the four-likelihood estimator of Equation (15). Pretrained evaluations use LLaDA-8B (Nie et al., 2025), MDLM-OWT, and a masked-diffusion variant of Qwen3-0.6B (Zhou et al., 2026), generating 16 tokens; fixed-switch and factor comparisons use 32 prompts from CommonsenseQA, MRPC, SNLI, and SST-2, with 4,096 draws per circuit. Section 7.1 tests the headline claim; Sections 7.2 to 7.4 test contributions (ii)–(iv). Appendices G–I give full details.

Table 1: TV between schedules with matched edge frequencies. LCB denotes a lower confidence bound. LLaDA reuses samples drawn for this switch; MDLM and Qwen use new samples.

checkpoint	switch	TV	95% LCB
LLaDA-8B	position pairs $(2, 3), (0, 1)$	4.73%	2.49%
MDLM-OWT	position pairs $(2, 3), (0, 1)$	15.10%	13.18%
Qwen3-0.6B	three fixed switches	all $< 0.05\%$	all 0

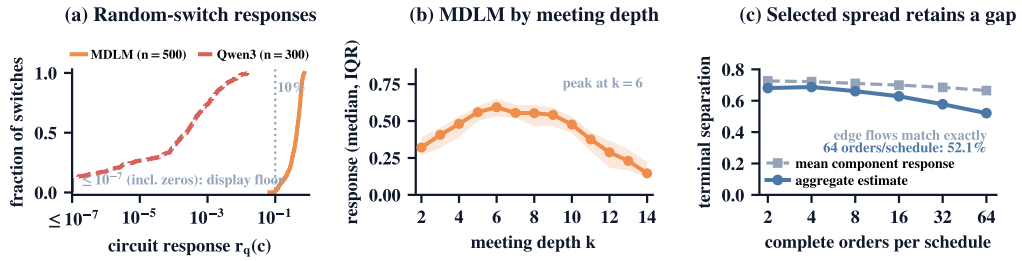


Figure 2: Response across random schedules and mixtures of selected orders. (a) Empirical response distributions; values $\leq 10^{-7}$ use a display floor. (b) MDLM response by meeting depth. (c) Mixture response (solid) and component mean (dashed), evaluated on new samples; both sides have equal edge frequencies.

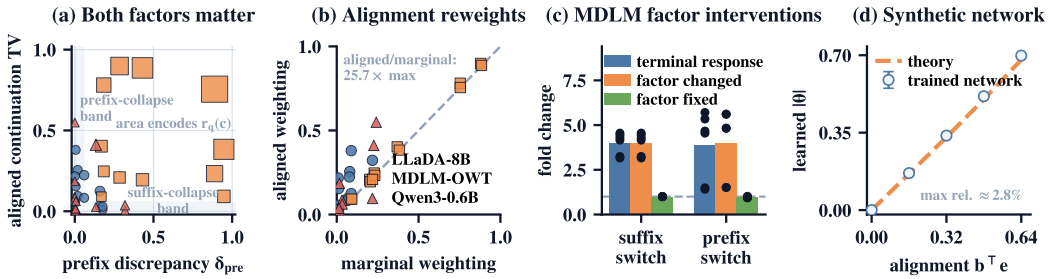


Figure 3: Generation differences and learned dependence. (a) Prefix and aligned continuation discrepancies across 36 settings; area encodes response. (b) Aligned versus marginal continuation discrepancy. (c) MDLM interventions. (d) Learned parameter shifts and the population prediction.

7.1 Frequency-matched schedules can produce different outputs

A fixed switch changes LLaDA and MDLM outputs. We first apply one switch on the first four positions to all three models. LLaDA and MDLM have positive TV lower bounds, whereas Qwen stays below 0.05% on this and two other fixed switches (Table 1).

In MDLM, differences are common across random switches. To test whether this depends on the chosen switch, we sample meeting depths uniformly from 2 to 14, then meeting sets and two distinct orders on each side uniformly; switching their pairing preserves edge frequencies. Over 500 MDLM and 300 Qwen switches, evaluated on eight prompts with 96 samples per circuit, MDLM has median TV 43.0% and positive simultaneous lower bounds for at least 82% of circuits, whereas Qwen's median is 0.02% (Figure 2a). MDLM's median peaks at meeting depth six (panel b).

Mixing many orders does not remove them. We then combine MDLM circuits selected for large response, with disjoint path supports and matched edge frequencies, and evaluate each mixture on new samples. TV decreases only from 68.1% at two orders per schedule to 52.1% at 64 (Figure 2c).

7.2 Both prefix and continuation discrepancies are needed

Either factor alone limits the response. To test the product form of Theorem 4, we vary three meeting depths and two pairs of orders on each side, giving 12 settings per model. A Qwen setting with 54.9% aligned continuation discrepancy but 0.004% prefix discrepancy, and a LLaDA setting with 17.6% prefix but 1.0% aligned continuation discrepancy, both give small responses (Figure 3a).

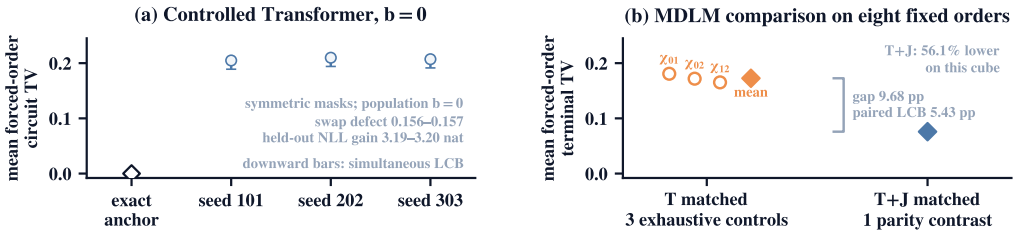


Figure 4: Symmetric training and additional moment matching. (a) Transformer response after training with symmetric masks; lower bounds hold simultaneously across seeds. (b) Three contrasts matching T and one also matching J on the same eight MDLM orders. The bracket shows their mean difference and its lower bound.

Averaging over prefixes can understate the continuation factor. The theorem weights contexts by where the prefixes disagree; we compare this aligned weighting with an average over the two prefix distributions. The aligned discrepancy is a median 3.76 times the averaged one in LLaDA and 1.89 times in Qwen, and the two nearly coincide in MDLM (Figure 3b).

Each factor moves TV on its own. In MDLM we change one stage at a time (Figure 3c). Changing continuations with prefixes fixed multiplies TV by 3.16 to 4.55; changing prefixes multiplies it by 1.41 to 5.71, while the aligned continuation factor stays within 7% of its original value.

Both factors are small in Qwen. Across the random switches of Section 7.1, Qwen's median prefix and aligned continuation discrepancies are 3.5% and 1.3%, versus 84.6% and 65.2% in MDLM.

7.3 Learning creates sensitivity to path coupling

The residual–precedence criterion predicts the learned shifts. To test Theorem 7, we train a network whose shared features cannot fit differences in data means. With bias aligned with the residual, parameter shifts are within 2.8% of the scalar population prediction (Figure 3d). With bias orthogonal to the residual ($b^\top e = 0$), block shifts nearly cancel to a mean of -0.012, but the block used by the switch retains -0.328 and produces positive TV, as the per-pair criterion predicts.

Symmetric masking does not prevent it. To test whether precedence bias is necessary, we train a 13.3M masked Transformer with symmetric masks on GPT-2 samples. Starting from compatible uniform predictions with zero response, all three seeds reach mean TV of 20.5–21.0% (Figure 4a).

7.4 Matching pair co-occurrence removes the quadratic response

Near compatibility, J removes the quadratic term. We compare matching edge frequencies T with also matching pair co-occurrence J, applying small perturbations to compatible conditionals at $d = 6$ and enumerating the output laws. All five random perturbations give quadratic response under T matching and cubic response under $T + J$ matching, as Theorem 9 predicts.

On MDLM, the $T+J$-matched comparison has lower TV. We use eight MDLM orders that flip three disjoint position pairs; each comparison splits them four against four by a flip parity and shares prompts, samples, and likelihoods. The three two-pair parities match T and average 17.3% TV; the three-pair parity also matches J and gives 7.6%, or 56.1% less (Figure 4b). Higher-order moments also differ; the $T + J$ comparison retains a third-order mismatch.

8 Conclusion

Identical local prediction frequencies can yield different output distributions under a fixed predictor. We characterize exactly when hidden path coupling causes this difference. Learning generically creates sensitivity in our shared-representation population model; a Transformer acquires it under symmetric masking. Near compatible predictions, pairwise exposure is the minimal second-order control; four path likelihoods diagnose the effect without retraining. Comparing schedules therefore requires both local frequencies and sensitivity to path coupling.

AI USE STATEMENT

In this work, we used generative AI tools for language polishing and improving readability. The authors reviewed the AI-assisted revisions and take responsibility for the final content of this work.

REFERENCES

Alan N. Amin, Nate Gruver, and Andrew Gordon Wilson. Why masking diffusion works: Condition on the jump schedule for improved discrete diffusion, 2025. URL https://arxiv.org/abs/2506.08316.

Barry C. Arnold, Enrique Castillo, and José María Sarabia. *Conditional Specification of Statistical Models*. Springer Series in Statistics. Springer, 1999. doi: 10.1007/b97592.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2107.03006.

Keith Ball. Cube slicing in $\mathbb{R}^n$. *Proceedings of the American Mathematical Society*, 97(3):465–473, 1986. doi: 10.1090/S0002-9939-1986-0840631-0.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642. Association for Computational Linguistics, 2015. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, and Arnaud Doucet. A continuous time framework for discrete denoising models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.14987.

Sitan Chen, Kevin Cong, and Jerry Li. Optimal inference schedules for masked diffusion models. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1279–1311. PMLR, 2026. URL https://proceedings.mlr.press/v336/chen26e.html.

Villő Csiszár. Markov bases of conditional independence models for permutations. *Kybernetika*, 45 (2):249–260, 2009. URL https://www.kybernetika.cz/content/2009/2/249.

Adrian Dobra. Markov bases for decomposable graphical models. *Bernoulli*, 9(6):1093–1108, 2003. doi: 10.3150/bj/1072215202.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL https://aclanthology.org/I05-5002/.

Tianqi Du, Lizhe Fang, Weijie Yang, Chenheng Zhang, Zeming Wei, Yifei Wang, and Yisen Wang. Autoregressive models rival diffusion models at ANY-ORDER generation, 2026. URL https://arxiv.org/abs/2601.13228.

Prateek Garg, Bhavya Kohli, and Sunita Sarawagi. Masked diffusion models are secretly learned-order autoregressive models. *arXiv preprint arXiv:2511.19152*, 2025. URL https://arxiv.org/abs/2511.19152.

Oleksandra Gasanova and Lisa Nicklasson. Chain algebras of finite distributive lattices. *Journal of Algebraic Combinatorics*, 59:473–494, 2024. doi: 10.1007/s10801-023-01294-8.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. URL https://arxiv.org/abs/2006.11239.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Chunsan Hong, Sanghyun Lee, and Jong Chul Ye. Unifying masked diffusion models with various generation orders and beyond, 2026. URL https://arxiv.org/abs/2602.02112.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.02037.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions, 2025. URL https://arxiv.org/abs/2502.06768.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6980.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. URL https://arxiv.org/abs/2310.16834.

Omer Luxembourg, Haim Permuter, and Eliya Nachmani. Plan for speed: Dilated scheduling for masked diffusion language models, 2025. URL https://arxiv.org/abs/2506.19037.

Mahbod Majid, Rattana Pukdee, Vishwajeet Agrawal, Burak Varıcı, and Pradeep Kumar Ravikumar. On the consistent recovery of joint distributions from conditionals. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 4708–4716. PMLR, 2025. URL https://proceedings.mlr.press/v258/majid25a.html.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, and Gao Huang. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models, 2026. URL https://arxiv.org/abs/2601.15165.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models, 2025. URL https://arxiv.org/abs/2502.09992.

Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.03736.

Ziqi Pang, Tianyuan Zhang, Fujun Luan, Yunze Man, Hao Tan, Kai Zhang, William T. Freeman, and Yu-Xiong Wang. RandAR: Decoder-only autoregressive visual generation in random orders. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2412.01827.

Fred Zhangzhi Peng, Zachary Bezemek, Sawan Patel, Jarrid Rector-Brooks, Sherwood Yao, Avishek Joey Bose, Alexander Tong, and Pranam Chatterjee. Path planning for masked diffusion model sampling, 2025. URL https://arxiv.org/abs/2502.03540.

Fred Zhangzhi Peng, Zachary Bezemek, Jarrid Rector-Brooks, Shuibai Zhang, Anru R. Zhang, Michael Bronstein, Alexander Tong, and Avishek Joey Bose. Planner aware path learning in diffusion language models training, 2026. URL https://arxiv.org/abs/2509.23405.

Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975. doi: 10.2307/2346567.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language-models.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models, 2024. URL https://arxiv.org/abs/2406.07524.

Alexander Schrijver. *Combinatorial Optimization: Polyhedra and Efficiency*, volume 24 of *Algorithms and Combinatorics*. Springer, 2003. doi: 10.1007/978-3-642-78240-4.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and generalized masked diffusion for discrete data, 2025. URL https://arxiv.org/abs/2406.04329.

Andy Shih, Dorsa Sadigh, and Stefano Ermon. Training and inference on any-order autoregressive models the right way. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/123fd8a56501194823c8e0dca00733df-Abstract-Conference.html.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642. Association for Computational Linguistics, 2013. URL https://aclanthology.org/D13-1170/.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015. URL https://arxiv.org/abs/1503.03585.

Bernd Sturmfels and Volkmar Welker. Commutative algebra of statistical ranking. *Journal of Algebra*, 361:264–286, 2012. doi: 10.1016/j.jalgebra.2012.03.028.

Akimichi Takemura and Hisayuki Hara. Markov chain monte carlo test of toric homogeneous markov chains. *Statistical Methodology*, 9(3):392–406, 2012. doi: 10.1016/j.stamet.2011.10.004.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4149–4158. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1421. URL https://aclanthology.org/N19-1421/.

Gilad Turok, Christopher De Sa, and Volodymyr Kuleshov. Duel: Exact likelihood for masked diffusion via deterministic unmasking, 2026. URL https://arxiv.org/abs/2603.01367.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Zhe Wang, Jiaxin Shi, Nicolas Heess, Arthur Gretton, and Michalis K. Titsias. Learning-order autoregressive models with application to molecular graph generation, 2025. URL https://arxiv.org/abs/2503.05979.

Jiawei Xu, Minghui Liu, Aakriti Agrawal, Yifan Chen, and Furong Huang. Scheduling thoughts: Learning the order of thought in diffusion language models, 2026. URL https://arxiv.org/abs/2606.23567.

Shuchen Xue, Tianyu Xie, Tianyang Hu, Zijin Feng, Jiacheng Sun, Kenji Kawaguchi, Zhenguo Li, and Zhi-Ming Ma. Any-order GPT as masked diffusion model: Decoupling formulation and architecture, 2025. URL https://arxiv.org/abs/2506.19935.

An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Randomized autoregressive visual generation. In *IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2411.00776.

Shaorong Zhang, Longxuan Yu, Rob Brekelmans, Luhan Tang, Salman Asif, and Greg Ver Steeg. Generation order and parallel decoding in masked diffusion models: An information-theoretic perspective, 2026. URL https://arxiv.org/abs/2602.00286.

Yixiu Zhao, Jiaxin Shi, Feng Chen, Shaul Druckmann, Lester Mackey, and Scott Linderman. Informed correctors for discrete diffusion models, 2025. URL https://arxiv.org/abs/2407.21243.

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling, 2025. URL https://arxiv.org/abs/2409.02908.

Zhanhui Zhou, Lingjie Chen, Hanghang Tong, and Dawn Song. dllm: Simple diffusion language modeling, 2026. URL https://arxiv.org/abs/2602.22661.

A Training objectives and schedule information

A.1 Local prediction losses and richer interfaces

For a value-independent schedule law ρ, an edge-additive objective has the form

$$\mathcal{L}(q;\rho) = \sum_{e=(S,i)} (T\rho)_e \, \ell_e(q),$$

where each ℓ_e is fixed independently of the rest of the sampled order. Schedules with equal edge frequencies therefore induce the same objective for every model and dataset. An alignment rule that compares only these frequencies also assigns them the same value. This decomposition provides the objective-level equivalence used in the main text.

Methods using additional schedule information make that information part of the conditioning or transition structure. Order history enters the local prediction index, grouping determines the transition unit, and a policy depending on generated token values changes the action probabilities. Their schedule statistics must account for these inputs.

Order and grouping in local predictions. Xue et al. (2025) distinguish unordered-context encoder predictions from decoder predictions that retain context order. The decoder requires an order-aware local index. Du et al. (2026) predict token groups through two-stream attention with masks determined by the group order. The conditioning interface therefore includes that grouping. The dilated scheduler of Luxembourg et al. (2025) selects spatially separated groups for parallel unmasking; its natural transition unit is a block.

Policies depending on token values. The learned policy of Xu et al. (2026) selects positions from the current partial sequence, making action probabilities depend on token values. Planner Aware Path Learning (Peng et al., 2026) incorporates planner-dependent reverse transitions into a training bound, including planners that use predicted token values. Our comparison fixes the predictor and all local exposure frequencies, isolating how coupling complete reveal paths changes generation within a single training objective.

A.2 Mean gradients and gradient covariance

Let $G_\rho = \sum_{e\in\Pi} g_e(X)$, with X independent of $\Pi \sim \rho$, and let $J_\rho(e,f) = \Pr\{e, f \in \Pi\}$. Expanding the sum gives

$$\mathbb{E}G_\rho = \sum_e T(\rho)_e \mathbb{E}g_e, \qquad \mathbb{E}[G_\rho G_\rho^\top] = \sum_{e,f} J_\rho(e,f)\mathbb{E}[g_e g_f^\top].$$

Equality of the mean for every deterministic local collection is equivalent to equality of T; conditional on this equality, equality of the covariance for every square-integrable local collection is equivalent to equality of J.

For the base circuit in Equation (19), take scalar local gradients

$$g_{(\varnothing,1)} = g_{(\{1,2\},3)} = 1, \qquad g_{(\varnothing,2)} = g_{(\{1,2\},4)} = -1,$$

and set every other g_e to zero. Under $\bar\rho_+$, the full-path sum is 2 or -2 with equal probability; under $\bar\rho_-$, it is always zero. Both means vanish, while the variances are four and zero.

B Identifying schedules from edge frequencies

B.1 Image and dimension of the edge map

The path-flow polytope and prefix–suffix exchanges below specialize the Csiszár-model structure of Sturmfels & Welker (2012, Section 4); see also Csiszár (2009). We give the linear derivation in the notation used for the terminal-response analysis.

Index the Boolean-lattice vertices by $S \subseteq [d]$ and its edges by (S,i) with $i \notin S$. Let B be the vertex–edge incidence matrix, with -1 at the tail and $+1$ at the head of each directed edge. If b is -1 at $\varnothing$, $+1$ at $[d]$, and zero elsewhere, then every permutation incidence vector v_π satisfies $Bv_\pi = b$.

Conversely, let $f \geq 0$ satisfy $Bf = b$. Starting at the source, follow a positive-flow edge. Whenever the path reaches an internal vertex, conservation supplies a positive outgoing edge. Acyclicity forces the walk to reach the sink. Subtract the smallest flow on this path and repeat. At least one positive edge disappears at each iteration, so the procedure terminates and writes f as a nonnegative combination of source-to-sink paths. The source equation makes the coefficients sum to one. Such paths add each coordinate once and are therefore permutations. Hence $T(\Delta(\mathfrak{S}_d))$ is exactly the nonnegative unit-flow polytope.

The underlying undirected Boolean cube is connected, so $\operatorname{rank}(B) = 2^d - 1$. Indeed, the rows sum to zero, while $y^\top B = 0$ forces y to take the same value at the endpoints of every edge and hence at every vertex. The fixed-divergence affine space has dimension

$$d2^{d-1} - (2^d - 1) = d2^{d-1} - 2^d + 1.$$

The nonnegativity constraints do not reduce this affine hull: under the uniform order law,

$$T(\mathrm{Unif}(\mathfrak{S}_d))_{S,i} = \frac{|S|!(d - |S| - 1)!}{d!} > 0,$$

so the polytope contains a relative-interior point. The linear rank of T is one more than the affine dimension of its image. Therefore

$$\dim \ker T = d! - \left(d2^{d-1} - 2^d + 2\right) = d! - d2^{d-1} + 2^d - 2,$$

which proves Theorem 2. The expression is zero for $d = 1, 2, 3$ and equals six for $d = 4$.

B.2 The six circuits at four variables

Partition the 24 permutations by their two-element prefix set S. Within each block, arrange the four path coefficients as a 2×2 matrix indexed by the two prefix orders and two suffix orders. If a signed schedule vector lies in $\ker T$, the two edges entering S force its row sums to vanish, and the two edges leaving S force its column sums to vanish. The block is therefore a scalar multiple of

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}.$$

The six blocks have disjoint path supports. Their circuit vectors are linearly independent, and the dimension calculation shows that they span $\mathcal{K}_4$.

A nonzero kernel vector cannot have support two: two distinct paths have distinct edge incidences. Nor can it have support three. Such a relation would express one 0–1 path-incidence vector as a nontrivial convex combination of two others, which coordinatewise forces all three incidence vectors to agree. Thus support four is minimal, and the block argument classifies every support-four circuit.

Finally, the edge map provides the row and column sums of each 2×2 block. One selected cell probability determines the remaining three, so six prefix–suffix joint probabilities make the map injective. Any scalar linear augmentation that is injective must have rank six on the six-dimensional kernel, proving minimality in the linear class.

B.3 Tail switches on a general DAG

Let $G = (V, E)$ be a finite source-to-sink DAG in which every vertex and edge belongs to an s–t path, and let A_G map path laws to edge-visitation vectors. The same decomposition and incidence-rank arguments show that the image is the nonnegative unit-flow polytope and

$$\dim \ker A_G = |\mathcal{P}_G| - |E| + |V| - 2.$$

We now show that tail switches span this real kernel.

First take an integer kernel vector and represent its positive and negative parts as path multisets $\mathcal{M}_+$ and $\mathcal{M}_-$. Equality of edge incidences makes every edge multiplicity equal on the two sides. Fix $p \in \mathcal{M}_+$. Some path in $\mathcal{M}_-$ shares the first edge of p. Suppose it agrees with p through a vertex v, and let e be the next edge of p. Another negative path contains e. Both paths pass through v, so exchanging their suffixes at v preserves every edge multiplicity and extends the matched prefix by e.

Acyclicity makes this process reach the sink. Remove the resulting copy of p from both multisets and iterate.

Thus every integer kernel vector is an integer combination of tail switches. The path-edge matrix has integer entries, so Gaussian elimination gives a rational kernel basis; scaling makes each basis vector integral, and extension of scalars yields the full real kernel. This proof also clarifies that a tail switch is a relation among complete paths and need not be a local four-cycle in G.

B.4 Identification within restricted schedule families

For a Plackett–Luce schedule (Plackett, 1975) with positive weights w_i, the source-edge probability is $w_i/\sum_j w_j$. The edge map therefore recovers the normalized weights and the complete schedule law. By contrast, in the four-coordinate block family that chooses between $12/21$ and between $34/43$, the edge map recovers the two one-bit marginals but not their coupling. The restricted kernel is the single circuit direction in Equation (19).

B.5 What edge observations cannot identify

Suppose one observes independent edges with law $p_\rho(e) = T(\rho)_e/d$. If $T\rho = T\sigma$, the observation law is identical under the two schedules for every sample size. The triangle inequality therefore gives, for any possibly randomized estimator $\widehat{\rho}$ and any known q,

$$\max_{\tau\in\{\rho,\sigma\}} \mathbb{E}_\tau \,\mathrm{TV}(\widehat{\rho},\tau) \geq \tfrac{1}{2}\,\mathrm{TV}(\rho,\sigma),$$

and

$$\max_{\tau\in\{\rho,\sigma\}} \mathbb{E}_\tau \,\mathrm{TV}(\widehat{Q},Q^q_\tau) \geq \tfrac{1}{2}\,\mathrm{TV}(Q^q_\rho,Q^q_\sigma).$$

C Generation differences at a fixed predictor

C.1 Proof of exact edge-fiber susceptibility

Every difference $a = \rho - \sigma$ of edge-equivalent schedule laws lies in $\mathcal{K}_d$ and satisfies $\|a\|_1 \leq 2$. Conversely, take $a \in \mathcal{K}_d$ with $\|a\|_1 \leq 2$. Summing the source-edge coordinates of $Ta = 0$ gives $\mathbf{1}^\top a = 0$. Hence

$$\|a_+\|_1 = \|a_-\|_1 =: m = \frac{1}{2}\|a\|_1 \leq 1.$$

For any $\nu \in \Delta(\mathfrak{S}_d)$, the measures

$$\rho = a_+ + (1-m)\nu, \qquad \sigma = a_- + (1-m)\nu$$

are probability laws, satisfy $T\rho = T\sigma$, and have difference a. This identifies the feasible difference set and proves the operator form of $\mathrm{Gap}_T(q)$ stated in Section 4. When $\mathcal{K}_d \neq \{0\}$, homogeneity also identifies the same value with the induced ℓ_1 norm of M_q restricted to $\mathcal{K}_d$.

The operator formula gives $\mathrm{Gap}_T(q) = 0$ exactly when $\mathcal{K}_d \subseteq \ker M_q$. In that case define $\widetilde{B}_q(Ta) = M_q a$ on $\mathrm{im}\,T$. If $Ta = Ta'$, then $a - a' \in \mathcal{K}_d \subseteq \ker M_q$, so the definition is independent of the representative. Extending $\widetilde{B}_q$ linearly to the ambient edge space gives $M_q = B_q T$. The converse is immediate. Tail switches span $\mathcal{K}_d$ by Appendix B.3, which proves the last equivalence.

Duality gives an equivalent interpretation. Since $\mathcal{K}_d^\perp = \mathrm{im}\,T^\top$,

$$\mathrm{Gap}_T(q) = \max_{\|f\|_\infty\leq 1} \min_{\lambda\in\mathbb{R}^{\mathcal{E}_d}} \|M_q^\top f - T^\top\lambda\|_\infty. \tag{17}$$

Thus susceptibility is the largest failure of a bounded terminal test to admit an edge-additive path score.

C.2 Proof of exact four-variable susceptibility

By Corollary 3, every $a \in \mathcal{K}_4$ has a unique expansion $a = \sum_S \lambda_S c_S$. The circuit supports are disjoint and every c_S has four unit-magnitude coefficients, so

$$\|a\|_1 = 4 \sum_S |\lambda_S|.$$

For $\|a\|_1 \le 2$,

$$\frac{1}{2}\|M_q a\|_1 \le \frac{1}{2} \sum_S |\lambda_S| \|M_q c_S\|_1 = 2 \sum_S |\lambda_S| r_S(q) \le \max_S r_S(q).$$

Conversely, $c_S/2$ is the difference between the uniform mixtures over the two positive and two negative circuit paths, and its terminal TV is $\|M_q c_S\|_1/4 = r_S(q)$. Maximizing over S gives equality.

C.3 Factorization at a retained meeting state

For a Boolean-lattice tail switch through S, the four path laws factor as

$$Q_{ab} = A_a K_b, \quad Q_{a'b'} = A_{a'} K_{b'}, \quad Q_{ab'} = A_a K_{b'}, \quad Q_{a'b} = A_{a'} K_b.$$

The shared conditional assumption is used here: after reaching (S, x_S), a suffix law depends on that state and the suffix order, not on the prefix order that reached it. Subtracting the four products gives Equation (6).

The terminal assignment retains x_S, so distinct contexts occupy disjoint output cells. Therefore

$$\begin{aligned} \frac{1}{4}\|M_q c\|_1 &= \frac{1}{4} \sum_{x_S} |A_a(x_S) - A_{a'}(x_S)| \sum_{x_{S^c}} |K_b - K_{b'}| \\ &= \frac{1}{2} \sum_{x_S} |A_a(x_S) - A_{a'}(x_S)| \delta_{\mathrm{suf}}(x_S), \end{aligned}$$

which is Equation (7). If the response is positive, at least one nonnegative summand has both factors positive. Conversely, such a context makes the response positive. Tail switches span $\mathcal{K}_d$, so the operator form of $\mathrm{Gap}_T(q)$ in Section 4 turns this circuit condition into the global statement.

C.4 A positive shared-network construction

Corollary 10 (Full-support fixed-floor separation)**.** For any $d \ge 4$, conditional floor $\lambda \in (0, 1/2)$, and schedule-density floor $\gamma \in (0, 1)$, there are positive edge-equivalent schedules and one shared target-and-mask conditional family such that

$$\mathrm{TV}(Q^q_{\rho_+}, Q^q_{\rho_-}) = (1-\gamma)(1-2\lambda)^2, \tag{18}$$

while every logit has magnitude at most $\log((1-\lambda)/\lambda)$.

The four base paths are grouped into the edge-equivalent circuit sides

$$\bar{\rho}_+ = \tfrac{1}{2}(\delta_{1234} + \delta_{2143}), \qquad \bar{\rho}_- = \tfrac{1}{2}(\delta_{1243} + \delta_{2134}). \tag{19}$$

Pair coordinates $B_0 = \{1, 2\}$ and $B_1 = \{3, 4\}$. Let $r_1 = r_3 = 0$ and $r_2 = r_4 = 1$, and let $h_i(S)$ indicate whether the partner of target i is already in S. For $L = \log((1-\lambda)/\lambda)$, define

$$f_L(S, i) = L(2r_i - 1)(1 - 2h_i(S)), \qquad q_i(1 \mid x_S) = \frac{1}{1 + e^{-f_L(S,i)}}.$$

On the binary inputs $r = r_i$ and $h = h_i(S)$,

$$(2r-1)(1-2h) = 2|r-h| - 1,$$

so a width-two ReLU head represents the logit exactly.

It can also be computed from raw target and mask indicators. Let $e_j = \mathbf{1}\{i = j\}$ and $m_j = \mathbf{1}\{j \in S\}$, and set

$$z_1 = \mathrm{ReLU}(e_1 + m_2 - 1), \qquad z_2 = \mathrm{ReLU}(e_2 - m_1),$$
$$z_3 = \mathrm{ReLU}(e_3 + m_4 - 1), \qquad z_4 = \mathrm{ReLU}(e_4 - m_3).$$

For every legal edge, at most one unit is active, and

$$f_L(e, m) = L\left(2\sum_{j=1}^{4} z_j - 1\right)$$

has the desired sign on all eight circuit entries. A dense network with twelve inputs, four hidden units, and one output has $4(12 + 1) + (4 + 1) = 57$ parameters; the observed-value inputs receive zero weight in this construction.

Along the four base paths in Equation (19), the path laws are $L_0 \otimes L_0$, $H_0 \otimes H_0$, $L_0 \otimes H_0$, and $H_0 \otimes L_0$, where $L_0 = \mathrm{Bern}(\lambda)^{\otimes 2}$ and $H_0 = \mathrm{Bern}(1 - \lambda)^{\otimes 2}$. Their signed difference is one half of $(L_0 - H_0) \otimes (L_0 - H_0)$, whose TV is $(1 - 2\lambda)^2$. For $d > 4$, append the same continuation to both unmixed schedule laws and a common conditional factor on the added coordinates. Mixing the resulting d-variable laws with the same $\gamma\,\mathrm{Unif}(\mathfrak{S}_d)$ then preserves edge equality, gives every permutation mass at least $\gamma/d!$, and multiplies the signed terminal difference by $1 - \gamma$. This proves Corollary 10. Off-circuit conditionals can be chosen arbitrarily and positively because the common uniform component cancels from the difference.

C.5 Compatibility and vanishing response

If $q_i(\cdot \mid x_S) = P(X_i \in \cdot \mid X_S = x_S)$ for one full-support joint P, the chain rule gives $Q^q_\pi = P$ for every order π and hence $\mathrm{Gap}_T(q) = 0$. The converse fails.

Choose $p_i \in (0, 1)$ with at least one $p_i \neq 1/2$, and set

$$q_i(1 \mid \varnothing) = p_i, \qquad q_i(1 \mid x_S) = \frac{1}{2} \quad (S \neq \varnothing).$$

This family is strictly positive and violates source-level adjacent-swap compatibility. Yet every nontrivial tail switch meets after a nonempty revealed set, from which all suffix orders induce the same fair conditional law. Every circuit response is therefore zero, so $\mathrm{Gap}_T(q) = 0$. This proves both that a one-sided defect is insufficient and that edge-fiber insensitivity is strictly weaker than global compatibility.

C.6 Additional inputs, block reveal and revision

Inputs determined by the revealed set and block reveal preserve Theorem 4. Revision can mix contributions from different meeting states and introduces a cancellation factor.

Inputs determined by the revealed set. The shared family is indexed by (S, i, x_S). An additional fixed input $f(S)$, such as a diffusion time or mask count, leaves the conditional a function of the same arguments. The factorization and response identities therefore still apply. Under Definition 1, one coordinate is revealed per step, so the step index is determined by $|S|$. A time or noise embedding set by that index is included in this interface.

Revealing coordinate blocks. Let each step reveal a block of coordinates, so a trajectory is a chain $\varnothing = S_0 \subset S_1 \subset \cdots \subset S_m = [d]$ with $|S_t \setminus S_{t-1}|$ free. A tail switch is defined exactly as before, exchanging the two block-suffixes after two block-prefixes first meet at a common S. Nothing in the derivation of Equation (6) used the block size: the four trajectory laws still factor as $A_a K_b$ and its partners, no coordinate is revised, and the terminal assignment still retains x_S. Hence

Proposition 11 (Block reveal)**.** Under multi-coordinate reveal with a value-independent chain law, Theorem 4 holds verbatim with edges replaced by block transitions (S_{t-1}, S_t), and $\mathrm{Gap}_T(q) > 0$ exactly when some switch has nonzero prefix and continuation discrepancies at a common retained context.

What does change is the combinatorics of the observation map: block transitions form a coarser lattice, so the kernel of the corresponding block-edge map, and the degree at which path moments identify the chain law, must be recomputed for the block structure in question. The response characterization is unaffected.

Revising revealed coordinates. If the terminal assignment no longer determines the meeting state, as when a sampler remasks or corrects an already revealed coordinate (Zhao et al., 2025), linearity still gives the signed-measure identity $M_q c = (A_a - A_{a'})(K_b - K_{b'})$. With $\mu = A_a - A_{a'}$ and $L = K_b - K_{b'}$, the terminal coordinate is $(M_q c)(y) = \sum_x \mu(x) L(y \mid x)$, so contributions from different meeting states can cancel once those states are forgotten. Define

$$\chi_c = \frac{\| \sum_x \mu(x) L(\cdot \mid x) \|_1}{\sum_x |\mu(x)| \|L(\cdot \mid x)\|_1} \in [0, 1],$$

with $\chi_c = 1$ when the denominator is zero. If $\delta_{\mathrm{pre}} = 0$, take w_c to be any probability law on meeting states; the response is then zero.

Proposition 12 (General path systems)**.** For a tail switch on a finite path system whose terminal observation need not retain the meeting state,

$$r_q(c) = \chi_c \, \delta_{\mathrm{pre}} \, \mathbb{E}_{x \sim w_c} \delta_{\mathrm{suf}}(x),$$

and $\chi_c = 1$ whenever the terminal assignment determines the meeting state. In particular $\chi_c = 1$ under Definition 1 and under Proposition 11.

The factor χ_c quantifies cancellation across meeting states in the terminal response. It equals one under Definition 1 and block reveal, and may be smaller when those states are forgotten. For path laws admitting the factorization in Appendix E.8, moment minimality is characterized by the terminal-response quotient defined there.

D How population learning creates order dependence

D.1 Compatible fitting with a well-specified model

Let μ be full support, give every Boolean-lattice edge positive weight λ_e, and consider

$$\mathcal{L}_\lambda(q; \mu) = \sum_{e=(S,i)} \lambda_e \mathbb{E}_\mu[-\log q_i(X_i \mid X_S)]. \tag{20}$$

Every term is its conditional entropy plus $\mathbb{E}_{X_S} \mathrm{KL}(\mu(X_i \in \cdot \mid X_S) \| q_i(\cdot \mid X_S))$. If the model class can represent all conditionals of μ simultaneously, any global minimizer makes every KL term zero. Its conditionals are therefore compatible with μ, and every generation order returns μ. This locates one zero-gap regime; finite capacity, optimization error, or misspecification can move a trained network away from it.

D.2 A full-support four-coordinate construction

Group the coordinates into pairs $B_1 = \{1, 2\}$ and $B_2 = \{3, 4\}$. For target i, let $\bar{i}$ denote its partner and define

$$z_i(S) = \tau_i(1 - 2\mathbf{1}\{\bar{i} \in S\}), \qquad \tau_1 = \tau_3 = -1, \quad \tau_2 = \tau_4 = 1.$$

If the odd coordinate of a pair is revealed first, both predictions in that pair have $z = -1$; if the even coordinate is first, both have $z = 1$.

Let the full-support data law be

$$\mu = \mathrm{Bern}(m_1)^{\otimes 2} \otimes \mathrm{Bern}(m_2)^{\otimes 2}, \qquad m_1, m_2 \in (0, 1),$$

and use the shared two-parameter model in Equation (9). Set

$$u = \sigma(\beta - \theta), \qquad v = \sigma(\beta + \theta).$$

The map $(\beta, \theta) \leftrightarrow (u, v)$ is a bijection between $\mathbb{R}^2$ and $(0, 1)^2$.

Consider the four block paths

$$\pi_{00} = 1234, \quad \pi_{01} = 1243, \quad \pi_{10} = 2134, \quad \pi_{11} = 2143.$$

Choose a strictly positive probability table $\Lambda = (\lambda_{xy})$ and $0 < \kappa < \min_{x,y} \lambda_{xy}$. Define

$$\bar{\rho}_{\pm}(\pi_{xy}) = \lambda_{xy} \pm \kappa s_{xy}, \qquad s_{00} = s_{11} = 1, \quad s_{01} = s_{10} = -1.$$

For $\gamma \in (0, 1)$, set

$$\rho_{\pm} = (1 - \gamma)\bar{\rho}_{\pm} + \gamma \, \mathrm{Unif}(\mathfrak{S}_4).$$

Every permutation has probability at least $\gamma/4!$. Their difference is

$$\rho_+ - \rho_- = 2(1 - \gamma)\kappa(\delta_{1234} + \delta_{2143} - \delta_{1243} - \delta_{2134}),$$

so $T\rho_+ = T\rho_-$.

Let p_j be the common probability that the odd coordinate precedes the even coordinate in pair j:

$$p_1 = (1 - \gamma)(\lambda_{00} + \lambda_{01}) + \gamma/2, \qquad p_2 = (1 - \gamma)(\lambda_{00} + \lambda_{10}) + \gamma/2.$$

The standard population NLL under either schedule is

$$\mathcal{L}(\beta, \theta) = 2 \sum_{j=1}^{2} \left[p_j \, \mathrm{ce}(m_j, u) + (1 - p_j) \, \mathrm{ce}(m_j, v) \right],$$

where $\mathrm{ce}(m, p) = -m \log p - (1 - m) \log(1 - p)$. Minimizing independently in (u, v) gives the unique optimum and difference displayed in Equation (10). Since the map from (u, v) to (β, θ) is bijective, the population minimizer is finite and unique.

Let

$$R_u = \mathrm{Bern}(u^*)^{\otimes 2}, \qquad R_v = \mathrm{Bern}(v^*)^{\otimes 2}, \qquad D = \mathrm{TV}(R_u, R_v).$$

The four block path laws are $R_u \otimes R_u$, $R_u \otimes R_v$, $R_v \otimes R_u$, and $R_v \otimes R_v$. Thus

$$Q_{\rho_+}^{q^*} - Q_{\rho_-}^{q^*} = 2(1 - \gamma)\kappa(R_u - R_v) \otimes (R_u - R_v),$$

and

$$\mathrm{TV}(Q_{\rho_+}^{q^*}, Q_{\rho_-}^{q^*}) = 4(1 - \gamma)\kappa D^2.$$

For $u < v$,

$$\mathrm{TV}(\mathrm{Bern}(u)^{\otimes 2}, \mathrm{Bern}(v)^{\otimes 2}) = (v - u)(1 + |1 - u - v|).$$

Choosing $m_1 = a$, $m_2 = 1 - a$, block precedence probabilities approaching $1 - \tau$ and τ, and $\kappa \uparrow \tau/2$ makes the terminal TV approach $2\tau(1 - 2\tau)^2$ as $a, \gamma \downarrow 0$. At $\tau = 1/6$, this approaches $4/27$. Every prelimit instance has full support and finite logits. This proves Theorem 6.

D.3 The residual and precedence criterion

For pairs $j = 1, \ldots, K$, let $\phi_j \in \mathbb{R}^r$ have full column rank and use $q_{\alpha,\theta,i}(1 \mid x_S) = \sigma(\alpha^\top \phi_j + \theta z_i(S))$ for $i \in B_j$. Let $m_j \in (0, 1)$ be the pair mean and $p_j \in (0, 1)$ its odd-before-even probability. With $\ell(m, t) = \log(1 + e^t) - mt$, the population risk is

$$\mathcal{L}(\alpha, \theta) = 2 \sum_{j=1}^{K} \left[p_j \ell(m_j, \alpha^\top \phi_j - \theta) + (1 - p_j)\ell(m_j, \alpha^\top \phi_j + \theta) \right].$$

The augmented design vectors $(\phi_j, -1)$ and $(\phi_j, 1)$ span $\mathbb{R}^{r+1}$: their differences generate the order coordinate, and their sums generate the feature span. Since every logistic variance and schedule weight is positive, the Hessian is positive definite. The loss is coercive because every unbounded parameter sequence sends at least one full-rank linear predictor to unbounded magnitude, while $\ell(m, t)$ grows linearly at both tails. The minimizer is therefore finite and unique.

At $\theta = 0$, let α_0 minimize the order-blind loss and define

$$\widehat{m}_j = \sigma(\alpha_0^\top \phi_j), \qquad e_j = m_j - \widehat{m}_j, \qquad b_j = 2p_j - 1.$$

The order-blind score equation is $\sum_j \phi_j e_j = 0$. The order-direction derivative at $(\alpha_0, 0)$ is

$$\partial_\theta \mathcal{L}(\alpha_0, 0) = 2 \sum_j b_j e_j.$$

Any joint optimum with $\theta = 0$ must have $\alpha = \alpha_0$. Strict convexity therefore gives

$$\theta^* = 0 \quad \Longleftrightarrow \quad b^\top e = 0,$$

proving Theorem 7. If $e \neq 0$, the zero set is a hyperplane within any feasible precedence-bias family containing an open set.

For small heterogeneous bias $p_j(\varepsilon) = (1 + \varepsilon b_j)/2$, symmetry and implicit differentiation give

$$\alpha^*(\varepsilon) = \alpha_0 + O(\varepsilon^2), \qquad \theta^*(\varepsilon) = -\frac{b^\top e}{\sum_j \widehat{m}_j(1 - \widehat{m}_j)}\varepsilon + O(\varepsilon^3).$$

A two-pair hidden switch then has leading terminal response of order ε^2.

D.4 Genericity and near-cancellation

Throughout, $\Phi \in \mathbb{R}^{K \times r}$ is the representation matrix with rows $\phi_j^\top$, and $\varrho = \operatorname{rank} \Phi$. Write σ for the coordinatewise logistic map and $V = \operatorname{col} \Phi \subseteq \mathbb{R}^K$, so $\dim V = \varrho$.

The residual depends on the data alone. The order-blind optimum solves $\sum_j \phi_j e_j = 0$, that is $\Phi^\top e = 0$ with $e = m - \sigma(\Phi\alpha_0)$. Strict convexity makes $\Phi\alpha_0$, and hence e, unique; and neither depends on the precedence profile p, which enters the objective only through weights that sum to one at $\theta = 0$. We therefore write $e = e(m)$ and set

$$\mathcal{M} = \{\sigma(v) : v \in V\} \subseteq (0,1)^K. \tag{21}$$

Lemma 13 (Exact zero set of the residual)**.** $e(m) = 0$ if and only if $m \in \mathcal{M}$. If $\varrho = K$ then $\mathcal{M} = (0,1)^K$. If $\varrho < K$ then $\mathcal{M}$ is a closed embedded ϱ-dimensional submanifold of $(0,1)^K$; in particular it has Lebesgue measure zero and empty interior.

Proof. If $m = \sigma(\Phi\alpha_\star)$ then $\alpha_\star$ makes every term of the order-blind loss individually minimal, so $\sigma(\Phi\alpha_0) = m$ and $e = 0$. Conversely $e(m) = 0$ gives $m = \sigma(\Phi\alpha_0) \in \mathcal{M}$. For the geometry, $\sigma : \mathbb{R}^K \to (0,1)^K$ is a diffeomorphism with inverse the coordinatewise logit, and V is a closed embedded ϱ-dimensional linear subspace of $\mathbb{R}^K$; a diffeomorphism carries closed embedded submanifolds to closed embedded submanifolds, and $\mathcal{M} = \sigma(V)$. A ϱ-dimensional submanifold with $\varrho < K$ is Lebesgue-null and, being closed with null measure, has empty interior. □

Theorem 14 (Genericity of learned order dependence)**.** Assume the hypotheses of Theorem 7.

(i) If $\varrho = K$, then $\theta^* = 0$ for every (m, p).

(ii) If $\varrho < K$, then $\theta^* \neq 0$ for Lebesgue-almost every $(m, p) \in (0,1)^K \times (0,1)^K$. The exceptional set is contained in $(\mathcal{M} \times (0,1)^K) \cup \{(m,p) : m \notin \mathcal{M},\ b^\top e(m) = 0\}$, a closed-times-full set of measure zero united with a null hyperplane bundle.

When $K \geq 2$, every fit with $\theta^* \neq 0$ also has $\mathrm{Gap}_T(q^*) > 0$, so part (ii) is a generic-susceptibility statement.

Proof. (i) By Lemma 13, $\varrho = K$ forces $e \equiv 0$, hence $b^\top e = 0$ and $\theta^* = 0$ by Theorem 7.

(ii) Fix $\varrho < K$. By Theorem 7, $\theta^* = 0$ exactly on $\{(m,p) : b(p)^\top e(m) = 0\}$, where $b(p) = 2p - 1$. Split on whether $m \in \mathcal{M}$. The slice $\mathcal{M} \times (0,1)^K$ is null by Lemma 13 and Fubini. For $m \notin \mathcal{M}$ the residual $e(m)$ is a fixed nonzero vector, so $b \mapsto b^\top e(m)$ is a nonzero linear functional and its zero set is a hyperplane in $\mathbb{R}^K$, of measure zero; since $p \mapsto 2p - 1$ is an affine diffeomorphism of $(0,1)^K$ onto $(-1,1)^K$, the corresponding p-slice is null. Fubini again gives measure zero for the union.

For the final claim, choose two distinct pair blocks j, k. Let $R_{\ell,\pm} = \mathrm{Bern}(\sigma(\alpha^{*\top}\phi_\ell \pm \theta^*))^{\otimes 2}$, fix the order of every other block, and couple the two orientations with the signed switch $c = \delta_{--} + \delta_{++} - \delta_{-+} - \delta_{+-} \in \ker T$. If R_0 is the terminal law of the fixed blocks, then

$$M_{q^*} c = (R_{j,-} - R_{j,+}) \otimes (R_{k,-} - R_{k,+}) \otimes R_0.$$

The two schedule sides use $c/2$, so their TV is $\frac{1}{4}\|M_{q^*} c\|_1 = \mathrm{TV}(R_{j,-}, R_{j,+})\,\mathrm{TV}(R_{k,-}, R_{k,+}) > 0$ whenever $\theta^* \neq 0$. Thus $\mathrm{Gap}_T(q^*) > 0$ for every nonzero-order fit when $K \geq 2$. □

A dimension-free quantitative form. For uniformly sampled precedence profiles, the following bound controls the probability of near-cancellation independently of K.

Proposition 15 (Anti-concentration of the criterion)**.** Fix $m \notin \mathcal{M}$ and let p be uniform on $(0,1)^K$. For every $\delta > 0$,

$$\Pr\left\{|b^\top e(m)| \leq \delta\, \|e(m)\|_2\right\} \leq \sqrt{2}\,\delta. \tag{22}$$

Proof. Write $u = e/\|e\|_2$, a unit vector, and note $b = 2p - 1$ is uniform on the cube $[-1,1]^K$. By Brunn's theorem the density of $\langle b, u\rangle$ is maximized at the origin, and by the cube-slicing bound of Ball (1986) the central hyperplane section of the unit cube has $(K-1)$-volume at most $\sqrt{2}$. Rescaling from $[-\frac{1}{2}, \frac{1}{2}]^K$ to $[-1,1]^K$ divides the density by two, so $\langle b, u\rangle$ has density at most $\sqrt{2}/2$ everywhere. Integrating over an interval of length 2δ gives Equation (22). □

The constant is close to sharp in our setting: at $K = 5$ with a randomly drawn misspecified representation, the empirical probabilities at $\delta = 0.02, 0.05, 0.10, 0.20$ are 0.0262, 0.0655, 0.1310 and 0.2583 against the bound values 0.0283, 0.0707, 0.1414 and 0.2828.

D.5 Vector-valued order parameters

Replace the shared scalar by $\theta \in \mathbb{R}^G$ acting through per-pair vectors $u_j \in \mathbb{R}^G$, so that $q_{\alpha,\theta,i}(1 \mid x_S) = \sigma(\alpha^\top \phi_j + \theta^\top u_j\, z_i(S))$, and let $U \in \mathbb{R}^{K\times G}$ have rows $u_j^\top$. The population risk is that in Appendix D.3 with θ replaced by $\theta^\top u_j$ in the jth term.

Theorem 16 (The criterion is a per-pair product)**.** If U has full column rank, the population NLL again has a unique finite minimizer, and $\theta^* = 0$ if and only if $\sum_j b_j e_j u_j = 0$. Tying every pair recovers the aggregate $b^\top e$; untying them demands $b_j e_j = 0$ for each pair separately, and the exceptional set among precedence profiles has codimension $\mathrm{rank}(\mathrm{diag}(e)U)$, which untying raises.

Proof. The augmented design vectors $(\phi_j, \pm u_j)$ have sums generating the feature span and differences generating $\mathrm{span}\{u_j\}$, so they span $\mathbb{R}^{r+G}$ exactly when Φ and U both have full column rank, and the coercivity argument of the scalar case applies verbatim. Otherwise the risk is constant along $\ker U$ and θ^* is not identified at all.

Since $\theta = 0$ removes the order term whatever U is, α_0 and e are unchanged. With $\partial_t \ell(m,t) = \sigma(t) - m$,

$$\nabla_\theta \mathcal{L}(\alpha_0, 0) = 2\sum_j u_j(\widehat{m}_j - m_j)(1 - 2p_j) = 2\sum_j b_j e_j\, u_j = 2\,U^\top(b \odot e),$$

writing $b \odot e$ for the vector of per-pair products. Any joint optimum with $\theta = 0$ has $\alpha = \alpha_0$, so strict convexity gives $\theta^* = 0$ if and only if $U^\top(b \odot e) = 0$, which is Theorem 16. With m fixed the map $b \mapsto U^\top(b \odot e)$ has matrix $U^\top \mathrm{diag}(e)$, so its zero set has codimension $\mathrm{rank}(\mathrm{diag}(e)U)$: one under tying ($u_j \equiv 1$), recovering Theorem 7, and $|\{j : e_j \neq 0\}|$ under complete untying ($U = I$). It is positive as soon as one pair has both $e_j \neq 0$ and $u_j \neq 0$, and $\theta^* = 0$ identically when the order features vanish wherever the representation misfits: the order is used only where something is left unexplained *and* an order parameter reaches. □

D.6 Order dependence under weight decay

Give every pair its own intercept and consider

$$q_{\alpha,\theta,i}(1 \mid x_S) = \sigma(\alpha_j + \theta z_i(S)),$$

with regularized population objective

$$\mathcal{L}_\lambda(\alpha,\theta) = 2\sum_j [p_j \ell(m_j, \alpha_j - \theta) + (1-p_j)\ell(m_j, \alpha_j + \theta)] + \frac{\lambda}{2}(\|\alpha\|_2^2 + \theta^2).$$

At $\theta = 0$, the intercepts $\bar{\alpha}_j$ solve

$$2[\sigma(\bar{\alpha}_j) - m_j] + \lambda\bar{\alpha}_j = 0.$$

The order-direction score is

$$\partial_\theta \mathcal{L}_\lambda(\bar{\alpha}, 0) = \lambda \sum_j b_j \bar{\alpha}_j.$$

For $\lambda > 0$, strict convexity gives

$$\theta^*_\lambda = 0 \quad \Longleftrightarrow \quad \sum_j b_j \bar{\alpha}_j = 0.$$

For two pairs with $m_2 = 1 - m_1 \neq 1/2$ and $p_2 = 1 - p_1 \neq 1/2$, odd symmetry gives $\bar{\alpha}_2 = -\bar{\alpha}_1$ and $b_2 = -b_1$, so the sum is nonzero for every $\lambda > 0$.

At $\lambda = 0$, write $a_j = \mathrm{logit}(m_j)$ and $s_j = m_j(1-m_j)$. Profiling out the intercepts gives order-direction curvature

$$H_{\mathrm{prof}} = 8\sum_j p_j(1-p_j)s_j.$$

Implicit differentiation yields

$$\theta^*_\lambda = -\lambda \frac{\sum_j b_j a_j}{H_{\mathrm{prof}}} + O(\lambda^2).$$

Thus a two-pair hidden-switch response is $\Theta(\lambda^2)$ whenever the numerator is nonzero.

A well-specified representation can therefore acquire order dependence under weight decay when $\sum_j b_j \bar{\alpha}_j \neq 0$.

D.7 Finite categorical alphabets

For each coordinate, reserve two symbols $0, 1$ and assign fixed positive probabilities to all additional symbols. Put the remaining mass $r_i > 0$ on the binary contrast. For $i \in B_j$, let the product data law have probabilities $r_i m_j$ and $r_i(1-m_j)$ on 1 and 0, respectively, and define

$$q_i(1 \mid x_S) = r_i \sigma(\alpha^\top \phi_j + \theta z_i(S)), \quad q_i(0 \mid x_S) = r_i[1 - \sigma(\alpha^\top \phi_j + \theta z_i(S))].$$

The model and data assign the same fixed probabilities to the additional symbols. With $w_j = \sum_{i \in B_j} r_i > 0$, the categorical population NLL is

$$\mathcal{L}_{\mathrm{cat}}(\alpha,\theta) = C + \sum_j w_j \left[p_j \ell(m_j, \alpha^\top \phi_j - \theta) + (1-p_j)\ell(m_j, \alpha^\top \phi_j + \theta)\right],$$

where C is parameter independent. Let $\alpha_0^{(w)}$ minimize this risk at $\theta = 0$, and set $e_j^{(w)} = m_j - \sigma(\alpha_0^{(w)\top}\phi_j)$. Its score equations give $\sum_j w_j \phi_j e_j^{(w)} = 0$, while the order-direction derivative is $\sum_j w_j b_j e_j^{(w)}$. Under the conditions of Theorem 7, positive weights preserve strict convexity and coercivity, so

$$\theta^* = 0 \quad \Longleftrightarrow \quad \sum_j w_j b_j e_j^{(w)} = 0.$$

Under the rank condition of Theorem 16, the vector criterion similarly becomes $\sum_j w_j b_j e_j^{(w)} u_j = 0$. Equal pair weights recover the binary criteria.

For any fixed parameters, restrict the signed terminal law to assignments using only $0, 1$. It equals $\prod_i r_i$ times the corresponding binary signed law, so categorical TV is at least $\prod_i r_i$ times binary TV. For two pairs sharing one intercept, the weighted score is $w_1 w_2 (b_1 - b_2)(m_1 - m_2)/(w_1 + w_2)$. Unequal pair means and precedence probabilities therefore yield positive learned separation for arbitrary positive binary masses.

D.8 Compatible conditional-score directions

Let P be a full-support joint distribution on a finite product alphabet. A retained-state edge is $e = (S, i)$ with $i \notin S$, and P_e denotes $P(X_i \mid X_S)$. Define the centered conditional-score space

$$\mathcal{H}_e = \{u_e(X_i, X_S) : \mathbb{E}_P[u_e \mid X_S] = 0\}, \qquad \mathcal{H} = \bigoplus_e \mathcal{H}_e,$$

with positive weighted inner product

$$\langle u, v\rangle_w = \sum_e w_e \, \mathbb{E}_P[u_e v_e], \qquad w_e > 0. \tag{23}$$

Let $\mathcal{G}_P = \{h : \mathbb{E}_P h = 0\}$ be the joint-score space and define $D : \mathcal{G}_P \to \mathcal{H}$ by

$$(Dh)_{S,i} = \mathbb{E}_P[h \mid X_{S\cup\{i\}}] - \mathbb{E}_P[h \mid X_S]. \tag{24}$$

For $u \in \mathcal{H}$, its Boolean-lattice curl is

$$(\mathsf{K}u)_{S,i,j} = u_{S,i} + u_{S\cup\{i\},j} - u_{S,j} - u_{S\cup\{j\},i}, \tag{25}$$

where all four functions are evaluated on the common assignment $X_{S\cup\{i,j\}}$.

Lemma 17 (Compatible conditional-score directions)**.** The tangent space induced by full-support joint perturbations is

$$\mathcal{C}_P := \operatorname{im} D = \ker \mathsf{K}.$$

Proof. For $u = Dh$, the four conditional expectations in every square of the Boolean lattice telescope, so $\mathsf{K}u = 0$. Conversely, take $u \in \ker \mathsf{K}$. Set $g_\emptyset = 0$ and, for any S, define $g_S(X_S)$ by summing $u_{A,i}$ along a path from $\emptyset$ to S. Adjacent swaps connect all such paths, and Equation (25) says that an adjacent swap leaves the sum unchanged; hence g_S is well defined. Moreover,

$$g_{S\cup\{i\}} = g_S + u_{S,i}, \qquad \mathbb{E}_P[g_{S\cup\{i\}} \mid X_S] = g_S,$$

where the second identity is exactly the conditional-centering constraint defining $\mathcal{H}_e$. Put $h = g_{[d]}$. Iterating the last identity down any path from $[d]$ to S gives $g_S = \mathbb{E}_P[h \mid X_S]$, and in particular $\mathbb{E}_P h = g_\emptyset = 0$. Thus $u_{S,i} = g_{S\cup\{i\}} - g_S = (Dh)_{S,i}$. The choice $g_\emptyset = 0$ fixes the otherwise additive, context-independent gauge. □

D.9 Population fitting near a compatible anchor

After quotienting local parameter symmetries, take θ in an open subset of $\mathbb{R}^r$. Let $q_{\theta,e}$ be positive and C^2 in θ, and suppose that an interior parameter θ_0 satisfies $q_{\theta_0,e} = P_e$ for every edge. Its conditional-score derivative is the linear map $J : \mathbb{R}^r \to \mathcal{H}$,

$$(Jv)_e = v^\top \nabla_\theta \log q_{\theta_0,e}(X_i \mid X_S), \qquad \mathcal{M}_q := \operatorname{im} J. \tag{26}$$

Assume regular local identifiability: J is injective in these quotient coordinates, or equivalently its conditional-score Jacobian has full column rank. Hence J^*J is positive definite, where J^* is the adjoint under Equation (23) and the Euclidean parameter inner product.

Proposition 18 (Weighted projection at a compatible anchor)**.** Suppose a positive C^2 shared conditional family contains all conditionals of P at an interior parameter whose conditional-score Jacobian has full column rank after quotienting parameter symmetries, and all population edge weights satisfy $w_e > 0$. Along any C^1 joint perturbation with score h, the local population-NLL minimizer through that anchor has derivative

$$u_h = \Pi^{(w)}_{\mathcal{M}_q} Dh, \tag{27}$$

$$u_h \in \mathcal{C}_P \text{ for every } Dh \in \mathcal{C}_P \quad \Longleftrightarrow \quad \Pi^{(w)}_{\mathcal{M}_q} \mathcal{C}_P \subseteq \mathcal{C}_P. \tag{28}$$

Proof. Fix $h \in \mathcal{G}_P$ and, for sufficiently small $|\varepsilon|$, use the full-support affine representative

$$P_{\varepsilon,h}(x) = P(x)[1 + \varepsilon h(x)].$$

Its conditional log-score derivative at zero is Dh. Every C^1 joint path through P with score h has the same first data derivative below, so the resulting fitted derivative depends only on h. Consider the weighted conditional population NLL

$$L_{\varepsilon,h}(\theta) = \sum_e w_e \, \mathbb{E}_{P_{\varepsilon,h}}[-\log q_{\theta,e}(X_i \mid X_S)]. \tag{29}$$

For fixed v, the conditional KL expansion on the finite state space gives

$$L_{\varepsilon,h}(\theta_0 + \varepsilon v) = C_{\varepsilon,h} + \frac{\varepsilon^2}{2}\|Jv - Dh\|_w^2 + o(\varepsilon^2), \tag{30}$$

where $C_{\varepsilon,h}$ is the weighted sum of the data conditional entropies and is independent of v. Indeed, uniformly over the finitely many contexts,

$$\partial_\varepsilon \log \frac{P_{\varepsilon,h,e}}{P_e}\bigg|_0 = (Dh)_e, \qquad \partial_\varepsilon \log \frac{q_{\theta_0+\varepsilon v,e}}{P_e}\bigg|_0 = (Jv)_e,$$

and the local expansion of conditional KL is one half the conditional second moment of the difference of these two centered scores. Replacing the context marginal under $P_{\varepsilon,h}$ by that under P changes only the $o(\varepsilon^2)$ term.

The implicit-function theorem gives the minimizer branch. Let $F(\varepsilon, \theta) = \nabla_\theta L_{\varepsilon,h}(\theta)$. At $(0, \theta_0)$,

$$F(0, \theta_0) = 0, \qquad \partial_\theta F(0, \theta_0) = J^* J, \qquad \partial_\varepsilon F(0, \theta_0) = -J^* Dh.$$

The Hessian equality is the conditional information identity at $q_{\theta_0,e} = P_e$. For the mixed derivative, conditional centering of every model score gives $\mathbb{E}_P[h(Jv)_e] = \mathbb{E}_P[(Dh)_e(Jv)_e]$, so differentiating the data law produces the stated adjoint term. The implicit-function theorem therefore supplies a unique differentiable stationary branch $\theta_{\varepsilon,h}$ near θ_0; positive definiteness of the Hessian makes it the strict local minimizer branch. Differentiating $F(\varepsilon, \theta_{\varepsilon,h}) = 0$ yields

$$\dot\theta_h = (J^*J)^{-1} J^* Dh, \qquad u_h := J\dot\theta_h = J(J^*J)^{-1}J^*Dh = \Pi_{\mathcal{M}_q}^{(w)} Dh. \tag{31}$$

Together with Lemma 17, this proves both Equation (27) and the equivalence in Equation (28). □

Symmetric weights do not imply invariance. Consider the following symmetric conditional family. Let $d = 2$, $X_1, X_2 \in \{-1, 1\}$, and let P be uniform. Give all four edges equal weight. A one-parameter positive conditional family sets

$$q_{\theta,(\emptyset,i)}(x_i) = \frac{\exp(\theta x_i)}{2\cosh\theta} \quad (i = 1, 2), \qquad q_{\theta,(S,i)}(x_i \mid x_S) = \frac{1}{2} \quad (S \neq \emptyset).$$

It is invariant under exchanging the two coordinates and contains P at $\theta_0 = 0$. Its model tangent is spanned by the field m with $m_{\emptyset,1} = x_1$, $m_{\emptyset,2} = x_2$, and both upper-edge components zero. For the compatible joint score $h(x) = x_1 + x_2$,

$$(Dh)_{\emptyset,1} = x_1, \quad (Dh)_{\emptyset,2} = x_2, \quad (Dh)_{\{1\},2} = x_2, \quad (Dh)_{\{2\},1} = x_1.$$

Under the equal-weight inner product, $\Pi_{\mathcal{M}_q}^{(w)} Dh = m$, but

$$(\mathsf{K}m)_{\emptyset,1,2} = x_1 - x_2 \neq 0.$$

Thus symmetric edge weights, even together with this coordinate symmetry of the family, do not force $\Pi_{\mathcal{M}_q}^{(w)} \mathcal{C}_P \subseteq \mathcal{C}_P$.

D.10 From fitted scores to quadratic path response

Let $c = \rho - \sigma$ be a fixed schedule contrast with $Tc = 0$, and for a co-visitable edge set A write

$$H_A(c) = \sum_\pi c(\pi) \mathbf{1}\{A \subseteq E(\pi)\}.$$

For the fitted branch above, set $q_{\varepsilon,e} = q_{\theta_{\varepsilon,h},e}$ and

$$r_{\varepsilon,e}(x) = \frac{q_{\varepsilon,e}(x_i \mid x_S)}{P_e(x_i \mid x_S)} - 1 = \varepsilon u_{h,e}(x) + o(\varepsilon),$$

uniformly on the finite state space.

Corollary 19 (Conditional quadratic path response)**.** Define, for each terminal assignment x,

$$\mathcal{R}_{c,x}(h) = \sum_{\{e,f\} \text{ co-visitable}} H_{\{e,f\}}(c) u_{h,e}(x) u_{h,f}(x). \tag{32}$$

If $\mathcal{R}_{c,x}(h) \neq 0$ for at least one x, then

$$\mathrm{TV}(Q_\rho^{q_\varepsilon}, Q_\sigma^{q_\varepsilon}) = C_h \varepsilon^2 + o(\varepsilon^2), \qquad C_h = \frac{1}{2} \sum_x P(x) |\mathcal{R}_{c,x}(h)| > 0. \tag{33}$$

Proof. Compatibility of the anchor gives, for every complete path π,

$$Q_\pi^{q_\varepsilon}(x) = P(x) \prod_{e \in E(\pi)} [1 + r_{\varepsilon,e}(x)].$$

Expanding this finite product and summing against c gives the exact identity

$$Q_\rho^{q_\varepsilon}(x) - Q_\sigma^{q_\varepsilon}(x) = P(x) \sum_A H_A(c) \prod_{e \in A} r_{\varepsilon,e}(x),$$

where only co-visitable sets occur. The constant coefficient is $H_\emptyset(c) = \sum_\pi c(\pi) = 0$, and every degree-one coefficient is $H_{\{e\}}(c) = (Tc)_e = 0$. Hence all single-edge terms, including their higher-order dependence on ε, cancel exactly. The degree-two terms equal $P(x)\varepsilon^2 \mathcal{R}_{c,x}(h) + o(\varepsilon^2)$, while all higher-degree terms are $o(\varepsilon^2)$. Summing the absolute pointwise difference proves Equation (33). □

Nonzero curl of u_h, or equivalently $u_h \notin \mathcal{C}_P$, is not enough to satisfy the corollary for an arbitrary c: the additional pair-moment contraction in Equation (32) is essential.

Lemma 20 (Conditional genericity for one fixed circuit)**.** Fix P, w, $\mathcal{M}_q$, and $c \in \ker T$. If $h \mapsto \mathcal{R}_{c,x}(h)$ is not the zero polynomial on $\mathcal{G}_P$ for at least one x, then the joint-score directions with $C_h = 0$ have relative Lebesgue measure zero in $\mathcal{G}_P$.

Proof. The maps $h \mapsto Dh \mapsto u_h$ are linear, so each $\mathcal{R}_{c,x}$ is a quadratic polynomial on the finite-dimensional space $\mathcal{G}_P$. If $C_h = 0$, then every $\mathcal{R}_{c,x}(h)$ is zero. Under the stated premise this exceptional set is contained in the zero set of one nonzero polynomial, which has Lebesgue measure zero. □

E Path moments and response control

E.1 Exact response expansion

Let P be a full-support distribution on a finite product alphabet and let the legal perturbation in Equation (11) satisfy conditional centering. For a path π and complete assignment x, compatibility gives

$$Q_{\pi,t}(x) = P(x) \prod_{e \in E(\pi)} [1 + t u_e(x)].$$

For a general conditionally centered deviation r_e with $q_e = P_e(1 + r_e)$, expanding the product and averaging gives

$$Q_\rho^q(x) - Q_\sigma^q(x) = P(x) \sum_A [H_A(\rho) - H_A(\sigma)] \prod_{e \in A} r_e(x). \tag{34}$$

Substituting $r_e = t u_e$ proves Equation (12).

E.2 Necessity and minimality of the path moments

Suppose $\Phi_k(\rho) = \Phi_k(\sigma)$. Every term through degree k vanishes from the exact expansion. Boundedness of the scores on the finite terminal space gives

$$\mathrm{TV}(Q_{\rho,t}, Q_{\sigma,t}) = O(|t|^{k+1}).$$

Conversely, suppose the moment vectors first differ on a co-visitable set A of minimum cardinality $j \le k$. Every proper-subset moment agrees. Use the uniform independent binary anchor and define

$$u_e(x) = 2x_{i(e)} - 1 \quad (e \in A), \qquad u_f = 0 \quad (f \notin A).$$

A co-visitable set lies on one reveal path, so its edges target distinct coordinates. Each score is conditionally centered. The exact expansion contains only subsets of A, and minimality removes every proper-subset term. Therefore

$$Q_{\rho,t}(x) - Q_{\sigma,t}(x) = P(x)[H_A(\rho) - H_A(\sigma)]t^j \prod_{e\in A}(2x_{i(e)} - 1).$$

The absolute product equals one for every assignment, hence

$$\mathrm{TV}(Q_{\rho,t}, Q_{\sigma,t}) = \frac{1}{2}|H_A(\rho) - H_A(\sigma)|\,|t|^j.$$

This proves the necessity direction and the exact binary coefficient in Theorem 9.

For an arbitrary finite alphabet with at least two symbols, select two symbols a_i, b_i at each target and use the uniform anchor with centered contrast

$$h_i(x_i) = \mathbf{1}\{x_i = a_i\} - \mathbf{1}\{x_i = b_i\}.$$

The same construction gives a strictly positive coefficient

$$\frac{1}{2}|\Delta H_A||t|^j \prod_{e\in A}\mathbb{E}|h_{i(e)}(X_{i(e)})|.$$

Thus binary values are not essential.

Let $S(\rho)$ be any statistic whose equality guarantees $O(|t|^{k+1})$ response for every allowed anchor and perturbation. If S failed to determine Φ_k, the witness above would contradict that guarantee. Hence $S(\rho) = S(\sigma)$ implies $\Phi_k(\rho) = \Phi_k(\sigma)$. This proves minimality up to injective recoding.

E.3 Second-order matching

Theorem 21 (Second-order path matching)**.** If $T\rho = T\sigma$, $J_\rho = J_\sigma$, and each conditional differs multiplicatively from a compatible anchor by at most α, then Equation (14) holds. Pair co-visitation is thus the minimal universal information that removes every quadratic response.

This is the case $k = 2$ of the remainder bound derived next.

E.4 Quantitative bounds for exact and approximate matching

Assume $|u_e(x)| \le \alpha$. For each degree j, define

$$\|\Delta H^{(j)}\|_1 = \sum_{\substack{A \text{ co-visitable}\\ |A|=j}} |H_A(\rho) - H_A(\sigma)|.$$

The exact expansion and triangle inequality give

$$\mathrm{TV}(Q_{\rho,t}, Q_{\sigma,t}) \le \frac{1}{2}\sum_{j=1}^{k}|t|^j\alpha^j\|\Delta H^{(j)}\|_1 + R_{k+1}.$$

For the remainder, let $a_\pi = \rho(\pi) - \sigma(\pi)$. Since each path contains exactly $\binom{d}{j}$ edge subsets of degree j,

$$\sum_{|A|=j}|\Delta H_A| \le \binom{d}{j}\sum_{\pi}|a_\pi| = 2\binom{d}{j}\,\mathrm{TV}(\rho,\sigma).$$

Therefore

$$R_{k+1} \le \mathrm{TV}(\rho,\sigma)\sum_{j=k+1}^{d}\binom{d}{j}(|t|\alpha)^j.$$

When the first k moment tensors agree and $|t| = 1$, the bound becomes

$$\mathrm{TV}(Q^q_\rho, Q^q_\sigma) \le \min\left\{1, \mathrm{TV}(\rho,\sigma) \sum_{j=k+1}^{d} \binom{d}{j} \alpha^j\right\}. \tag{35}$$

For $k = 2$ this is Equation (14), proving Theorem 21.

After T, J matching, the first possible response is cubic. Its coefficient can vanish for a particular model.

E.5 A BOUND FROM UNIFORM CONDITIONAL ERROR

The excess-NLL theorem in Theorem 25 gives a global average-risk guarantee. A complementary result controls a uniform conditional perturbation. For a full-support anchor P, define

$$\begin{aligned} \eta(q;P) &= \max_{(S,i),x_S} \mathrm{TV}(q_i(\cdot \mid x_S), P(X_i \in \cdot \mid X_S = x_S)), \\ m(P) &= \min_{(S,i),x_S,a} P(X_i = a \mid X_S = x_S), \end{aligned} \tag{36}$$

and set $\alpha = \eta(q;P)/m(P)$. Then any edge-equivalent schedules obey

$$\mathrm{TV}(Q^q_\rho, Q^q_\sigma) \le \min\left\{1, \mathrm{TV}(\rho,\sigma)\left[(1+\alpha)^d - 1 - d\alpha\right]\right\}. \tag{37}$$

If at most s visited edges on every relevant path differ from P, replace d by s.

For a complete assignment x and edge $e = (S, i)$, write

$$r_e(x) = \frac{q_i(x_i \mid x_S)}{P(X_i = x_i \mid X_S = x_S)} - 1. \tag{38}$$

Then $|r_e(x)| \le \alpha$ and $Q^q_\pi(x) = P(x)\prod_{e\in\pi}(1 + r_e(x))$. In the difference of two schedule mixtures, the constant term cancels by normalization and the linear coefficient of r_e is $T(\rho)_e - T(\sigma)_e = 0$. Bounding the remaining terms of degree at least two by $\sum_{j=2}^{d}\binom{d}{j}\alpha^j$ proves Equation (37). Keeping the uncancelled linear term gives the approximate-edge version

$$\mathrm{TV}(Q^q_\rho, Q^q_\sigma) \le \min\left\{1, \frac{\alpha}{2}\|T\rho - T\sigma\|_1 + \mathrm{TV}(\rho,\sigma)[(1+\alpha)^d - 1 - d\alpha]\right\}. \tag{39}$$

When $d\alpha \le 1$, the hidden remainder is at most $(e/2)(d\alpha)^2$. This is the $k = 1$ instance of the path-moment hierarchy in Equation (35).

E.6 HALF-PATH IDENTIFICATION

Theorem 22 (Half-path identification)**.** For complete orders of d coordinates, Φ_k identifies the schedule law if and only if $k \ge \lfloor d/2 \rfloor$, and every lower level is strict.

Proof. Let $m = \lfloor d/2 \rfloor$. For a permutation $\pi = (\pi_1, \ldots, \pi_d)$, denote its rank-$t$ edge by

$$e_t(\pi) = \big(\{\pi_1, \ldots, \pi_{t-1}\}, \pi_t\big)$$

and select

$$A_\pi = \{e_2(\pi), e_4(\pi), \ldots, e_{2m}(\pi)\}.$$

A complete path contains A_π only if it equals π. The edge e_2 recovers π_1 from its singleton prefix and π_2 from its target. Inductively, once the first $2r-2$ positions are known, the prefix of e_{2r} differs from the known set by exactly π_{2r-1}, and its target is π_{2r}. If d is odd, the last coordinate is the unique remaining element. Hence

$$H_{A_\pi}(\rho) = \rho(\pi),$$

so moments of degree m identify the complete schedule law.

For the lower bound, group the first $2m$ coordinates into pairs $B_j = \{2j-1, 2j\}$. For orientation bit $x_j \in \{0,1\}$, let

$$o_j(0) = (2j-1, 2j), \qquad o_j(1) = (2j, 2j-1).$$

For $x \in \{0,1\}^m$, concatenate the blocks $o_1(x_1) \cdots o_m(x_m)$ and append the final coordinate when d is odd. Let ρ_{even} and ρ_{odd} be the uniform laws over even- and odd-parity orientation vectors.

On this block family, every nonconstant edge indicator depends on only one orientation bit: earlier prefixes contain complete pairs and are independent of their internal orientation. An event specified by fewer than m edges therefore fixes fewer than all m bits. Flipping a free bit preserves the event and bijects the even and odd parity classes, so all moments below degree m agree. Selecting one orientation-specific edge from every pair gives a co-visitable degree-m set that separates the two parity laws. Common uniform mixing produces full-support examples without changing the moment equalities. This proves Theorem 22 and strictness of every hierarchy level. □

E.7 Response information under parameter tying

Theorem 9 is minimal for the full perturbation class of Equation (11), in which each edge carries its own score. Parameter tying restricts the reachable scores. For fixed conditionally centered features ψ_g and free coefficients θ, suppose

$$u_e(x) = \sum_{g=1}^{G} \theta_g \psi_g(e, x).$$

Write $\Delta H_A = H_A(\rho) - H_A(\sigma)$. For a multi-index $\beta \in \mathbb{N}_0^G$ with $|\beta| = j$, collect the coefficient of the parameter monomial θ^β in the degree-j response:

$$C_{j,\beta}(\Delta H)(x) = \sum_{|A|=j} \Delta H_A \, [\theta^\beta] \prod_{e \in A} \left(\sum_{g=1}^{G} \theta_g \psi_g(e, x) \right).$$

Here $[\theta^\beta]$ denotes coefficient extraction. Substitution into Equation (12) gives the degree-j term $P(x)t^j \sum_{|\beta|=j} \theta^\beta C_{j,\beta}(\Delta H)(x)$. It vanishes for every reachable score exactly when every $C_{j,\beta}(\Delta H)$ is zero as a function of x. Thus the necessary statistic through degree k is the quotient of Φ_k by the joint kernel of these coefficient maps for $1 \le j \le k$.

For $G = 1$, the coefficient at degree j is the function

$$x \longmapsto \sum_{|A|=j} \Delta H_A \prod_{e \in A} \psi_1(e, x).$$

The quotient is determined by linear dependence among these terminal response functions.

E.8 General path systems and the terminal-response quotient

Let $\mathcal{P}$ be a finite path system whose path laws around an anchor admit

$$Q_{\pi,t}(y) = P(y) \prod_{e \in E(\pi)} [1 + t u_e(y)].$$

For degree j, define the response coefficient operator

$$\mathcal{R}_{j,u}(h)(y) = P(y) \sum_{|A|=j} h_A \prod_{e \in A} u_e(y)$$

and universal nullspace

$$\mathcal{N}_j = \bigcap_u \ker \mathcal{R}_{j,u}.$$

Two schedule laws have identical universal response through degree k exactly when

$$\Delta H^{(j)} \in \mathcal{N}_j \qquad (j \le k).$$

Thus the coarsest universal order-k statistic is the sequence of quotient classes

$$H^{(1)} + \mathcal{N}_1, \ldots, H^{(k)} + \mathcal{N}_k.$$

In Boolean/categorical one-time reveal models, independent centered symbol contrasts isolate any nonzero moment coordinate because the terminal assignment retains every target coordinate. Hence $\mathcal{N}_j = \{0\}$ and the quotient reduces to the exact path-moment theorem. For other path systems admitting the displayed factorization, $\mathcal{N}_j$ can be nontrivial. The quotient classes retain exactly the path-moment information detectable in the terminal response.

E.9 ESTIMATING THE HIERARCHY

To estimate pair co-visitation, sample an ordered pair of distinct visited edges uniformly from one complete order. Its population law is

$$\mathsf{P}_\rho^{(2)}(e,f) = \frac{J_\rho(e,f)}{d(d-1)}, \qquad e \neq f. \tag{40}$$

Thus equality of the pair law is equivalent to equality of off-diagonal J and implies equality of T through $\sum_{f\neq e} J_\rho(e,f) = (d-1)T_\rho(e)$. This supplies an unbiased sampling interface to the second-order statistic. Exact population matching gives Equation (14); empirical penalties additionally require a specified discrepancy, finite-sample control, and optimization analysis.

Path moments are estimable from sampled orders without enumerating $\mathfrak{S}_d$. Draw $\Pi_1, \dots, \Pi_N$ i.i.d. from ρ and set

$$\widehat{H}_A = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}\{A \subseteq E(\Pi_n)\}.$$

Each summand is a Bernoulli indicator, so $\widehat{H}_A$ is unbiased and takes values in $[0,1]$.

Proposition 23 (Uniform error grows logarithmically)**.** Let $\mathcal{A}$ be any family of M co-visitable edge sets. With probability at least $1-\delta$,

$$\max_{A\in\mathcal{A}} \left|\widehat{H}_A - H_A(\rho)\right| \le \sqrt{\frac{\log(2M/\delta)}{2N}}.$$

Proof. Hoeffding's inequality for each fixed A, then a union bound over the M sets and both tails. □

The whole level-k hierarchy has $M \le \sum_{j\le k}\binom{d2^{d-1}}{j} \le (d2^{d-1})^k$ sets, so $\log M \le k\log(d2^{d-1}) = O(kd)$ and uniform accuracy ε costs

$$N = O\big(\varepsilon^{-2}\big(kd + \log(1/\delta)\big)\big)$$

sampled orders. The dependence on the number of statistics is logarithmic.

F LIKELIHOOD CERTIFICATES AND POPULATION RISK BOUNDS

F.1 LIKELIHOOD LOWER CERTIFICATE AND CONFIRMATION PROTOCOL

Fix a tail switch c. Let P_c^+ and P_c^- be the terminal laws of its two positive and two negative paths, mixed uniformly on each side, and set $R_c = (P_c^+ + P_c^-)/2$. Define

$$W_c(x) = \begin{cases} \dfrac{|P_c^+(x) - P_c^-(x)|}{P_c^+(x) + P_c^-(x)}, & P_c^+(x) + P_c^-(x) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Then $0 \le W_c \le 1$ and direct summation gives

$$\mathbb{E}_{X\sim R_c} W_c(X) = \mathrm{TV}(P_c^+, P_c^-) = r_q(c) \le \mathrm{Gap}_T(q). \tag{41}$$

For N independent samples, Hoeffding's inequality yields, with probability at least $1-\delta$,

$$\mathrm{Gap}_T(q) \ge \frac{1}{N}\sum_{n=1}^{N} W_c(X_n) - \sqrt{\frac{\log(1/\delta)}{2N}}.$$

Testing several preselected circuits uses the corresponding union-bound correction. A held-out discriminator supplies another lower bound when path likelihoods are unavailable: for any fixed balanced classifier with accuracy A, $2A - 1 \le r_q(c)$.

Each side likelihood is a log-sum-exp of two forced-order log likelihoods, so the estimate requires four path scores per draw and no terminal enumeration.

Selection must be separated from confirmation. A circuit chosen using arbitrary calibration data is fixed conditional on that history, so an untouched confirmation split gives the same one-circuit bound. If L candidates are compared on the confirmation data, replace δ by δ/L. A polynomial candidate family supplies a computationally scalable collection of valid global lower bounds, but its coverage is only family-specific. An arbitrary succinct model can hide its response at an unqueried context, so a negative polynomial scan cannot provide a worst-case global upper bound without symmetry, locality, response-density, or other structural assumptions.

For fixed prompts $z = 1, \dots, m$ with equal allocation, the target is the fixed-set average $m^{-1} \sum_z r_{q_z}(c)$. Stratified sampling remains unbiased for this quantity. Prompt-level sample means are descriptive unless separate prompt-level inference is performed.

F.2 Separating the two factors from the same likelihoods

The certificate of Equation (16) scores four forced orders per draw. We show that the two factors of Theorem 4 are recoverable from exactly those scores, so measuring the mechanism costs no additional inference.

Write $\bar{A} = \frac{1}{2}(A_a + A_{a'})$ and $\bar{K} = \frac{1}{2}(K_b + K_{b'})$. Since

$$P_c^+ + P_c^- = \tfrac{1}{2}(A_a K_b + A_{a'} K_{b'} + A_a K_{b'} + A_{a'} K_b) = \tfrac{1}{2}(A_a + A_{a'})(K_b + K_{b'}),$$

the mixture $R_c = (P_c^+ + P_c^-)/2$ has x_S-marginal $\bar{A}$ and continuation law $\bar{K}$ given x_S. Along a complete order the log likelihood splits at S, so each scored path supplies its prefix and continuation factors separately. Define the bounded rewards

$$W_c^{\mathrm{pre}}(x) = \frac{|A_a(x_S) - A_{a'}(x_S)|}{A_a(x_S) + A_{a'}(x_S)}, \qquad W_c^{\mathrm{suf}}(x) = \frac{|K_b(x_{S^c} \mid x_S) - K_{b'}(x_{S^c} \mid x_S)|}{K_b(x_{S^c} \mid x_S) + K_{b'}(x_{S^c} \mid x_S)},$$

both in $[0, 1]$, with the convention $0/0 := 0$.

Proposition 24 (Factor estimators)**.** Under $X \sim R_c$, with $\delta_{\mathrm{suf}}(x_S) = \mathrm{TV}(K_b(\cdot \mid x_S), K_{b'}(\cdot \mid x_S))$,

$$\mathbb{E}\, W_c^{\mathrm{pre}} = \delta_{\mathrm{pre}}, \qquad \mathbb{E}\, W_c^{\mathrm{suf}} = \mathbb{E}_{x_S \sim \bar{A}}\, \delta_{\mathrm{suf}}(x_S), \qquad \mathbb{E}\big[W_c^{\mathrm{pre}} W_c^{\mathrm{suf}}\big] = r_q(c).$$

Proof. For the first, $\mathbb{E} W_c^{\mathrm{pre}} = \sum_{x_S} \bar{A}(x_S) \frac{|A_a - A_{a'}|}{A_a + A_{a'}}(x_S) = \frac{1}{2} \sum_{x_S} |A_a - A_{a'}|(x_S) = \delta_{\mathrm{pre}}$. For the second, conditioning on x_S gives $\mathbb{E}[W_c^{\mathrm{suf}} \mid x_S] = \sum_{x_{S^c}} \bar{K} \frac{|K_b - K_{b'}|}{K_b + K_{b'}} = \delta_{\mathrm{suf}}(x_S)$, and averaging over $\bar{A}$ gives the claim. For the third, the same conditioning yields

$$\begin{aligned} \mathbb{E}\big[W_c^{\mathrm{pre}} W_c^{\mathrm{suf}}\big] &= \sum_{x_S} \bar{A}(x_S) \frac{|A_a - A_{a'}|}{A_a + A_{a'}}(x_S)\, \delta_{\mathrm{suf}}(x_S) \\ &= \tfrac{1}{2} \sum_{x_S} |A_a - A_{a'}|(x_S)\, \delta_{\mathrm{suf}}(x_S) \\ &= \delta_{\mathrm{pre}}\, \mathbb{E}_{x_S \sim w_c} \delta_{\mathrm{suf}}(x_S), \end{aligned}$$

which is $r_q(c)$ by Equation (7). □

Two consequences are used in Section 7. The pointwise identity $W_c = W_c^{\mathrm{pre}} W_c^{\mathrm{suf}}$ lets the same four likelihoods measure prefix discrepancy and response. When $\delta_{\mathrm{pre}} > 0$, their ratio gives the aligned continuation $r_q(c)/\delta_{\mathrm{pre}} = \mathbb{E}_{w_c} \delta_{\mathrm{suf}}$; the second expectation gives the marginal continuation $\mathbb{E}_{\bar{A}} \delta_{\mathrm{suf}}$. Their contrast measures how continuation discrepancies concentrate on contexts where the prefixes disagree. Since both reward factors lie in $[0, 1]$,

$$r_q(c) = \delta_{\mathrm{pre}}\, \mathbb{E}_{w_c} \delta_{\mathrm{suf}} \le \min\{\delta_{\mathrm{pre}},\ \mathbb{E}_{w_c} \delta_{\mathrm{suf}}\},$$

which is Equation (8); the prefix bound needs only the prefix halves of the four scores. Also, $r_q(c) = \mathbb{E}[W_c^{\mathrm{pre}} W_c^{\mathrm{suf}}] \le \mathbb{E} W_c^{\mathrm{suf}} = \mathbb{E}_{\bar{A}} \delta_{\mathrm{suf}}$. The aligned weighting gives the exact product decomposition; replacing it by $\bar{A}$ in that product would require $\mathbb{E}[W_c^{\mathrm{pre}} W_c^{\mathrm{suf}}] = \mathbb{E} W_c^{\mathrm{pre}}\, \mathbb{E} W_c^{\mathrm{suf}}$.

F.3 A TWO-SIDED CERTIFICATE AT FOUR COORDINATES

At $d = 4$, Corollary 5 expresses $\mathrm{Gap}_T(q)$ as the maximum of six circuit responses, so simultaneous bounds on those responses give both an upper and a lower certificate.

Fix the six circuits c_S, $|S| = 2$, before sampling. For each, draw N_S points from its mixture R_{c_S} and let $\hat{r}_S$ be the mean of W_{c_S}. Since each reward lies in $[0, 1]$ and is unbiased for $r_S(q)$, Hoeffding with a union bound over the six circuits and both tails gives, with probability at least $1 - \delta$ simultaneously,

$$|\hat{r}_S - r_S(q)| \le \varepsilon_S := \sqrt{\frac{\log(12/\delta)}{2N_S}} \qquad \text{for every } S.$$

On that event,

$$\max_S \big[\hat{r}_S - \varepsilon_S\big]_+ \;\le\; \mathrm{Gap}_T(q) \;\le\; \max_S \big[\hat{r}_S + \varepsilon_S\big]. \tag{42}$$

The lower bound holds because $\mathrm{Gap}_T(q) \ge r_S(q)$ for every S; the upper bound because $\mathrm{Gap}_T(q) = \max_S r_S(q)$ and each term is over-estimated by at most ε_S. The bracket has width at most $2 \max_S \varepsilon_S$ and shrinks as $N_S^{-1/2}$.

The bracket uses only forced-order likelihoods. At $d = 4$, its upper endpoint can certify a small global susceptibility without access to μ or $H(\mu)$.

Prefix screening. By Proposition 24, $r_S(q) \le \delta_{\mathrm{pre},S}$, and $\delta_{\mathrm{pre},S}$ is estimated by the prefix halves of the same scores. Let $\hat{\delta}_S$ be that estimate with radius η_S at the same confidence. Any circuit with $\hat{\delta}_S + \eta_S < \max_{S'}[\hat{r}_{S'} - \varepsilon_{S'}]_+$ cannot attain the maximum and can be dropped before its continuation scores are computed, which is where most of the evaluation cost sits.

F.4 CERTIFICATES FOR MIXTURES OF MANY ORDERS

The same bounded reward applies to schedule mixtures with many complete orders, including those in Section 7.1.

Let $c_1, \dots, c_N$ be tail switches with pairwise disjoint path supports, c_i exchanging suffixes b_i, b_i' after prefixes a_i, a_i' meet at S_i. Define

$$\rho_+ = \frac{1}{2N} \sum_{i=1}^{N} \big(\delta_{a_i b_i} + \delta_{a_i' b_i'}\big), \qquad \rho_- = \frac{1}{2N} \sum_{i=1}^{N} \big(\delta_{a_i b_i'} + \delta_{a_i' b_i}\big), \tag{43}$$

each a probability law on $2N$ complete orders.

Edge equality. $\rho_+ - \rho_- = \frac{1}{2N} \sum_i c_i$ and each $c_i \in \ker T$, so $T\rho_+ = T\rho_-$ exactly, for every N and every choice of the c_i. Disjointness is not needed for this; it is used only to keep the $4N$ orders distinct so that both laws are uniform.

The reward. Write $P_\pm = Q^q_{\rho_\pm}$ for the terminal laws and let $R = (P_+ + P_-)/2$. Put

$$W(x) = \frac{|P_+(x) - P_-(x)|}{P_+(x) + P_-(x)} \quad (0/0 := 0), \tag{44}$$

which is the same functional form as Equation (15) with the two mixtures in place of the two circuit sides. Then $0 \le W \le 1$ and

$$\begin{aligned} \mathbb{E}_{X \sim R} W(X) &= \sum_x \frac{P_+(x) + P_-(x)}{2} \frac{|P_+(x) - P_-(x)|}{P_+(x) + P_-(x)} \\ &= \tfrac{1}{2} \sum_x |P_+(x) - P_-(x)| = \mathrm{TV}(P_+, P_-), \end{aligned}$$

so the Hoeffding bound of Equation (16) applies verbatim and $L(\delta) \le \mathrm{TV}(P_+, P_-) \le \mathrm{Gap}_T(q)$ with probability at least $1 - \delta$.

Computing it. $P_\pm$ are uniform mixtures of $2N$ forced-order path laws, so with $\ell_o(x)$ the log likelihood of order o,

$$\begin{aligned} \log P_\pm(x) &= \operatorname{logsumexp}_{o\in\operatorname{supp}\rho_\pm} \ell_o(x) - \log(2N), \\ W(x) &= \left|\tanh\left(\tfrac{1}{2}(\log P_+(x) - \log P_-(x))\right)\right|, \end{aligned} \tag{45}$$

recovering the four-path expression at $N = 1$. The cost is $4N$ forced-order evaluations per sample, with no enumeration of the terminal space.

Aggregate response. Since $\rho_+ - \rho_- = (2N)^{-1}\sum_i c_i$, the aggregate response satisfies

$$\operatorname{TV}(P_+, P_-) = \frac{1}{4N}\Big\|\sum_i M_q c_i\Big\|_1 \le \frac{1}{N}\sum_i r_q(c_i).$$

Equality holds when the signed responses $M_q c_i(x)$ share a sign at each terminal assignment x. Section 7.1 measures the aggregate response directly with fresh confirmation samples after selecting the circuits.

F.5 Common-flow excess-NLL bound

For an edge $e = (S, i)$, recall the population conditional regret $\kappa_e(q;\mu) = \mathbb{E}_{X_S\sim\mu}\operatorname{KL}(\mu(X_i \in \cdot \mid X_S) \,\|\, q_i(\cdot \mid X_S))$ from Section 6. The chain rule gives, for every complete order π,

$$D_\pi := \operatorname{KL}(\mu\|Q^q_\pi) = \sum_{e\in\pi}\kappa_e(q;\mu). \tag{46}$$

Hence, when $T\rho = T\sigma = f$,

$$\sum_\pi \rho(\pi)D_\pi = \sum_\pi \sigma(\pi)D_\pi = \sum_e f_e\kappa_e(q;\mu) = \mathcal{E}_f(q;\mu). \tag{47}$$

Theorem 25 (Common-flow excess-NLL bound)**.** Let μ have full support, and write $D_\pi = \operatorname{KL}(\mu\|Q^q_\pi)$ for the divergence of the terminal law of a single order π. For any edge-equivalent ρ, σ with common flow f,

$$\operatorname{TV}(Q^q_\rho, Q^q_\sigma) \le \min\left\{1, \tfrac{1}{2}\textstyle\sum_\pi |\rho(\pi) - \sigma(\pi)|\, D_\pi\right\} \le \min\{1, \mathcal{E}_f(q;\mu)\}. \tag{48}$$

The constant one is locally sharp.

Proof of the bound. Write $a_\pi = \rho(\pi) - \sigma(\pi)$, with $0\cdot\infty := 0$. If $\sum_\pi |a_\pi| D_\pi = +\infty$, Equation (48) follows from $\operatorname{TV} \le 1$. Otherwise every path with $a_\pi \ne 0$ has finite D_π; full support of μ then makes $L_\pi = Q^q_\pi/\mu$ positive. For each active edge define $z_e(x) = \log[q_i(x_i \mid x_S)/\mu(x_i \mid x_S)]$. Since $Ta = 0$,

$$\sum_\pi a_\pi \log L_\pi(x) = \sum_e (Ta)_e z_e(x) = 0. \tag{49}$$

Also $\sum_\pi a_\pi = 0$. With $\phi(u) = u - 1 - \log u \ge 0$,

$$\sum_\pi a_\pi L_\pi(x) = \sum_\pi a_\pi \phi(L_\pi(x)). \tag{50}$$

Therefore

$$\begin{aligned} \operatorname{TV}(Q^q_\rho, Q^q_\sigma) &\le \frac{1}{2}\sum_\pi |a_\pi| \mathbb{E}_\mu \phi(L_\pi) && (51)\\ &= \frac{1}{2}\sum_\pi |a_\pi| D_\pi && (52)\\ &\le \frac{1}{2}\left(\sum_\pi \rho(\pi)D_\pi + \sum_\pi \sigma(\pi)D_\pi\right) = \mathcal{E}_f. && (53) \end{aligned}$$

The second line uses $\mathbb{E}_\mu L_\pi = 1$. Taking the minimum with one proves the bound in Theorem 25. The extended-real convention permits zero model probabilities; an infinite weighted KL sum gives the bound one after capping.

Local sharpness. For the local sharpness claim, take μ uniform on four bits and let $R_\theta(z) = \frac{1}{4}[1 + \theta s(z)]$, where $s(z) = 1$ when the two bits agree and -1 otherwise. Use the base circuit from Equation (19) and shared conditionals whose prefix and suffix orders produce R_θ and $R_{-\theta}$. The four path laws are $R_\theta \otimes R_\theta$, $R_{-\theta} \otimes R_{-\theta}$, $R_\theta \otimes R_{-\theta}$, and $R_{-\theta} \otimes R_\theta$. Direct calculation gives

$$\mathrm{TV}(Q^q_\rho, Q^q_\sigma) = \theta^2, \qquad \mathcal{E}_f(q;\mu) = -\log(1-\theta^2), \qquad \frac{\mathrm{TV}}{\mathcal{E}_f} \longrightarrow 1. \tag{54}$$

Thus no uniform constant below one can replace the leading constant in Equation (48). □

Population-NLL interpretation. Finally, the population excess risk equals an NLL difference. If $\mathcal{L}_f(q;\mu) = \sum_e f_e \mathbb{E}_\mu[-\log q_i(X_i \mid X_S)]$, the chain rule implies $\sum_e f_e H_\mu(X_i \mid X_S) = H(\mu)$ and hence

$$\mathcal{E}_f(q;\mu) = \mathcal{L}_f(q;\mu) - H(\mu). \tag{55}$$

The theorem is therefore a population statement, and turning it into a numerical finite-sample upper certificate would additionally require a certified treatment of $H(\mu)$ and the log-loss tails. Appendix F.3 gives an upper bound that avoids this route entirely at $d = 4$, using only forced-order likelihoods and no reference to μ or $H(\mu)$.

G Controlled learning experiments

G.1 Fitting the prescribed conditional table

We fit the conditional table of Appendix C.4. The model receives four mask indicators, four observed values (zero at unrevealed coordinates), and a target indicator with four entries. It has a hidden layer of four $\tanh$ units and one logistic output. Enumerating every legal mask, observed assignment, and target gives $4 \cdot 3^3 = 108$ rows. We minimize equally weighted cross entropy against the conditional probability targets with Adam (Kingma & Ba, 2015) for 3,000 steps at learning rate 0.02. We selected the width using conditional mean absolute error and terminal TV on three development seeds, comparing widths $2, 4, 8, 16$. On ten separate evaluation seeds, 100 to 109, the mean terminal TV is 0.5754 against the analytic value 0.576 at $\lambda = \gamma = 0.1$. Terminal distributions and edge frequencies are computed by enumeration.

For the parameter sweeps, five held-out seeds, 200–204, are fit independently at every point. The learned mean TVs and analytic values are shown in Table 2.

Table 2: Shared-MLP terminal TV across the two parameter sweeps.

Sweep	Curve	Parameter values				
$\lambda, \gamma = 0.1$	values	0.05	0.20	0.30	0.40	0.45
	learned	0.72805	0.32394	0.14411	0.03603	0.00908
	analytic	0.72900	0.32400	0.14400	0.03600	0.00900
$\gamma, \lambda = 0.1$	values	0	0.25	0.50	0.75	0.90
	learned	0.64025	0.48018	0.32012	0.16006	0.06402
	analytic	0.64000	0.48000	0.32000	0.16000	0.06400

The compatible uniform conditional control has zero terminal TV. The analytic construction uses ReLU on the eight circuit entries. The optimized $\tanh$ model fits the complete conditional table, including neutral predictions outside the circuit. Both use the same input encoding, hidden width, and parameter count.

G.2 Learning from aliased pair features

We train an MLP on binary samples to test the aggregate and local criteria in Theorems 7 and 16.

Setup. The data consist of four independent coordinate pairs ($d = 8$) with means $m = (0.80, 0.40, 0.70, 0.30)$. Index the pairs by $j = 0, 1, 2, 3$, with $B_j = \{2j + 1, 2j + 2\}$. The model is an MLP on $[\phi_j, z]$, where $z \in \{-1, +1\}$ is the pair's mask-derived orientation feature. Its single hidden layer has eight tanh units. Using NumPy's PCG64 generator with seed 20260806, we draw

two standard Gaussian vectors in $\mathbb{R}^2$ and assign them to $\phi_0 = \phi_1$ and $\phi_2 = \phi_3$. Within each block, the two pairs thus share the same features but differ in mean. The resulting residual

$$e = (+0.2, -0.2, +0.2, -0.2)$$

cannot be removed by any function of the features. Set $p_j = (1 + \varepsilon a_j)/2$, so $b = \varepsilon a$. The aligned condition uses $a = e/\|e\|_\infty$. For the orthogonal condition, draw the next standard Gaussian vector $g \in \mathbb{R}^4$ from the same generator, set $v = g - e(e^\top g)/\|e\|_2^2$, and use $a = v/\|v\|_\infty$. Thus $b^\top e = 0$ throughout the orthogonal condition.

Each fit uses 20,000 stochastic updates with batch size 512 and learning rate 0.1. Each example independently draws j uniformly, $z = -1$ with probability p_j, and a Bernoulli label with mean m_j. Results average five training seeds, $1234, 5678, 9012, 3456, 7890$. If $f_j(z)$ is the fitted logit at $[\phi_j, z]$, define

$$\theta_j = \tfrac{1}{2}[f_j(+1) - f_j(-1)], \qquad \bar{\theta} = \tfrac{1}{4}\sum_{j=0}^{3} \theta_j.$$

The reported aggregate shift is $\bar{\theta}$, and the switch block has $\theta_{\text{block}} = \theta_0 = \theta_1$ because its two pairs have identical features.

Schedules and normalization. Reveal the pair blocks in order B_0, B_1, B_2, B_3. Let $O_j = 1$ mean odd-first reveal within pair j, with probability p_j, and write $w_j(o) = p_j^o(1 - p_j)^{1-o}$. The two schedules use

$$\rho_\pm(O = o) = \left[w_0(o_0)w_1(o_1) \pm \kappa(-1)^{o_0+o_1}\right] w_2(o_2)w_3(o_3), \qquad \kappa = \tfrac{1}{2}\min_{a,b\in\{0,1\}} w_0(a)w_1(b).$$

The signed perturbation sums to zero in each row and column, preserving every p_j. Each block starts after the same coordinate set, so its edge frequencies depend only on its own orientation marginal. Hence $T\rho_+ = T\rho_-$. The normalized separation $\mathrm{TV}/(2\kappa)$ accounts for the changing feasible coupling range across settings. Terminal laws are enumerated exactly over all 2^8 assignments.

Table 3: Scalar criterion and learned block response. θ_{scalar} is the population reference for a single shared orientation parameter; θ_{block} is the network's learned shift in the block carrying the switch. The orthogonal arm separates aggregate cancellation from local response. Separations are $\mathrm{TV}/(2\kappa)$.

arm	ε	$b^\top e$	θ_{scalar}	θ_{block}	separation
aligned	0.00	0.000	0.0000	−0.00004	0.000040
aligned	0.20	0.160	−0.1637	−0.1672	0.0184
aligned	0.40	0.320	−0.3297	−0.3368	0.0734
aligned	0.60	0.480	−0.5006	−0.5146	0.1667
aligned	0.80	0.640	−0.6796	−0.6983	0.2951
orthogonal	0.20	0.000	0.0000	−0.0628	0.0026
orthogonal	0.40	0.000	0.0000	−0.1368	0.0122
orthogonal	0.60	0.000	0.0000	−0.2195	0.0303
orthogonal	0.80	0.000	0.0000	−0.3280	0.0638

With aligned bias, θ_{block} tracks the scalar reference θ_{scalar} in every setting. The maximum relative deviation is 2.8%, with standard deviations across seeds between 0.007 and 0.009. The normalized separation reaches 0.295 at $\varepsilon = 0.8$.

Cancellation across blocks. With orthogonal bias, $b^\top e = 0$ and the aggregate learned shift ranges from -0.00019 to -0.01181 across ε. The two blocks learn shifts of opposite signs that nearly cancel in the aggregate. The block carrying the switch reaches $\theta_{\text{block}} = -0.328$, producing positive separation despite the small mean shift. This agrees with the per-pair criterion in Theorem 16. The predicted quadratic scaling gives $(0.698/0.328)^2 = 4.53$, close to the measured separation ratio $0.295/0.0638 = 4.63$.

G.3 Learning with zero precedence bias

We use the MLP, data, training settings, and five seeds of Appendix G.2. The baseline uses aligned bias at $\varepsilon = 0.8$; the second condition sets every training precedence probability to $1/2$. Per-pair

logit shifts θ_j are measured as defined there, with $|\theta|_{\max} = \max_j |\theta_j|$ for each fit. Terminal laws are enumerated over all 2^8 assignments using the same biased schedule pair for both conditions. Held-out NLL is also evaluated under the same biased schedule for both conditions.

Zeroing the precedence bias reduces $|\theta|_{\max}$ from 0.698 to 0.010 and the normalized separation $\mathrm{TV}/(2\kappa)$ from 0.2951 to 4×10^{-5}. Held-out NLL rises from 0.62977 to 0.68325. The original model uses the orientation feature to achieve NLL below the order-blind optimum of 0.68308; with zero bias, NLL is 0.00017 above that optimum. Nearly all of the NLL increase is the gain the baseline obtained from the orientation feature.

G.4 Symmetric masked-Transformer training

This controlled experiment asks whether a shared Transformer can acquire an edge-hidden response when the population masking law has no precedence asymmetry. The data joint is obtained by sampling a frozen GPT-2 model (Radford et al., 2019) autoregressively from EOS for exactly 16 tokens at temperature one, without top-k or top-p truncation. Its softmax defines a full-support compatible joint. A single frozen dataset, generated with seed 20260809, contains $65{,}536$ training and $8{,}192$ validation sequences and is shared across student seeds.

The student is a $13{,}315{,}025$-parameter bidirectional masked Transformer (Vaswani et al., 2017) with two encoder layers, hidden size 128, four attention heads, feed-forward size 512, learned absolute positions, and a separate vocabulary output head. Setting that head's weights and bias to zero makes every conditional exactly uniform at initialization, hence compatible and order independent. For each training sequence we draw $m \sim \mathrm{Unif}[0.05, 1]$, mask coordinates independently with probability m, and average cross-entropy over masked targets; an empty mask is replaced by one uniformly selected coordinate. Thus, for each pair $i \neq j$ and fixed m, both observed–masked orientations have the same probability $m(1-m) + (1-m)^{16}/16$. The input contains token identities, positions, and the retained set, but neither m nor a reveal-order label. We use AdamW (Loshchilov & Hutter, 2019) with learning rate 3×10^{-4}, weight decay zero, batch size 64, dropout 0.1, gradient clipping at 1, and $4{,}000$ updates for seeds $101, 202, 303$.

Held-out masked NLL uses $2{,}048$ validation sequences and a separately seeded symmetric mask bank; the uniform anchor has NLL $\log 50257 = 10.8249$. On another $2{,}048$ fixed contexts we record the adjacent-swap compatibility defect

$$\mathcal{C}_{ij,S} = \log q_i(x_i \mid x_S) + \log q_j(x_j \mid x_{S\cup\{i\}}) - \log q_j(x_j \mid x_S) - \log q_i(x_i \mid x_{S\cup\{j\}}),$$

which is zero at the anchor. Four tail-switch circuits with meeting depths $4, 6, 8, 10$ were fixed independently of all student outputs and have exactly equal edge frequencies. For each seed and circuit, 512 terminal samples are drawn from each of its four paths and scored under all four path laws. The bounded likelihood reward of Equation (15) therefore estimates circuit TV. A seed's primary estimate averages the four circuits ($8{,}192$ rewards), and its Hoeffding lower bound uses $\delta = 0.05/3$, simultaneous over the three training seeds.

Table 4: Symmetric-training results. TV is the mean over the four fixed edge-equivalent circuits; its LCB is simultaneous over the three training seeds.

seed	NLL gain (nat)	mean $\lvert\mathcal{C}\rvert$	mean TV	95% LCB
101	3.2043	0.1575	20.48%	18.90%
202	3.2037	0.1559	20.98%	19.40%
303	3.1937	0.1565	20.72%	19.13%

Dropout was disabled during evaluation. All three seeds have positive lower bounds under symmetric masks.

H Pretrained-model experiments

H.1 Common prompts, reference circuits and TV estimator

The reference and factor comparisons use eight fixed prompts each from CommonsenseQA, MRPC, SNLI, and SST-2 (Talmor et al., 2019; Dolan & Brockett, 2005; Bowman et al., 2015; Socher et al.,

2013). Every circuit consists of two straight-side and two crossed-side orders over a 16-token completion. For each prompt and side, the generator first selects one of that side's two paths with probability $1/2$ and then reveals one masked token at a time along the selected order. We use 64 draws per prompt and side, giving $32 \times 2 \times 64 = 4{,}096$ rewards per circuit. The target is the equally weighted average of the 32 fixed prompt-specific circuit TVs.

For each terminal completion we evaluate all four path log likelihoods under the same conditional law used for generation. If $\ell_{+1}, \ell_{+2}, \ell_{-1}, \ell_{-2}$ are these scores, then

$$\ell_+ = \operatorname{logaddexp}(\ell_{+1}, \ell_{+2}) - \log 2, \tag{56}$$

$$\ell_- = \operatorname{logaddexp}(\ell_{-1}, \ell_{-2}) - \log 2, \tag{57}$$

$$W = \left|\tanh\left((\ell_+ - \ell_-)/2\right)\right|. \tag{58}$$

This evaluates Equation (15) in log space. Temperature is one, the mask token is excluded from both sampling and scoring, and each path has 16 reveal steps. Averaging the rewards gives an unbiased TV estimate.

Reference circuit map. For each row below, let R be the increasing sequence of the other 12 positions. The four complete orders are (a, b, c, d, R) and (b, a, d, c, R) on the straight side, and (a, b, d, c, R) and (b, a, c, d, R) on the crossed side. Together with the table, this specifies all four orders for each circuit.

circuit	(a, b, c, d)
early pairs	$(2, 3, 0, 1)$
interleaved pairs	$(4, 12, 0, 8)$
separated pairs	$(0, 1, 8, 9)$

H.2 Model checkpoints and inputs

LLaDA-8B-Instruct. We use LLaDA-8B-Instruct (Nie et al., 2025) with its instruction chat format. For the three reference circuits, we reuse samples drawn for these circuits before the likelihood estimator was adopted; one-sided bounds are simultaneous with $\delta = 0.05/3$. The design in Table 6 was chosen on MDLM before any LLaDA or Qwen sample was drawn for it, then applied unchanged. Figure 3c reports the MDLM measurements from this design.

MDLM-OWT. For MDLM-OWT (Sahoo et al., 2024), we evaluate the early-pairs circuit, which gave the largest LLaDA response, on newly drawn samples. This checkpoint uses the GPT-2 tokenizer with an added mask token and no chat template. Prompts occupy an observed prefix of fixed length; shorter prompts are padded on the left with EOS tokens visible to the model. Generation and likelihood scoring use the same padded input. The bound for this fixed circuit uses $\delta = 0.05$.

Qwen3-0.6B (masked-diffusion variant). We use the masked-diffusion variant of Qwen3-0.6B released with dLLM (Zhou et al., 2026; Yang et al., 2025). It is evaluated on the same three circuits and prompts with newly drawn samples. Generation and likelihood scoring use the same bidirectional attention mask. Simultaneous lower bounds use $\delta = 0.05/3$ and are clipped at zero as in Equation (16).

Table 5: Complete pretrained-model circuit estimates. Prompt counts above fixed thresholds are descriptive summaries of the 32 prompt-level sample means.

Checkpoint	Circuit	Mean	95% LCB	prompts $> 2\%$	prompts $> 5\%$
LLaDA-8B	early pairs	4.73%	2.49%	$21/32$	$14/32$
	interleaved pairs	0.78%	0	$2/32$	$0/32$
	separated pairs	0.69%	0	$1/32$	$1/32$
MDLM-OWT	early pairs	15.10%	13.18%	$32/32$	$32/32$
Qwen3-0.6B	early pairs	$1.9\times10^{-13}\%$	0	$0/32$	$0/32$
	interleaved pairs	0.04%	0	$0/32$	$0/32$
	separated pairs	$4.8\times10^{-8}\%$	0	$0/32$	$0/32$

For LLaDA on the early pairs, mean TV is 2.72% on CommonsenseQA, 7.21% on MRPC, 6.77% on SNLI, and 2.22% on SST-2. The corresponding MDLM means are 15.74%, 16.26%, 13.26%, and 15.14%. Prompt medians are 4.66% for LLaDA and 14.87% for MDLM; MDLM values range from 6.31% to 24.05%.

H.3 Varying prefixes and continuations

The design combines three meeting depths, two prefix pairs, and two continuation pairs, giving 12 settings per model. Each pair of continuation settings shares the same prefixes, so changing the continuation leaves δ_{pre} unchanged. Figure 3c uses the MDLM measurements to examine these changes. Factors are estimated using Proposition 24. The aligned continuation discrepancy is $r_q(c)/\delta_{\mathrm{pre}}$; the marginal discrepancy averages over $\bar{A}$.

Construction rules. For each $k \in \{4, 8, 12\}$, the construction fixes a meeting set S, a prefix order a on S, and a suffix order b on S^c. Type P1 swaps the last two positions of a to obtain a'; P2 draws a distinct order a' on S. Type S1 swaps the first two positions of b to obtain b'; S2 draws a distinct order b' on S^c. Each cell compares the straight paths $ab, a'b'$ with the crossed paths $ab', a'b$. The same prefix pair is reused across S1/S2, and the same continuation pair across P1/P2.

In the MDLM interventions, changing S1 to S2 leaves the prefix discrepancy unchanged and multiplies response by 3.16 to 4.55. Changing P1 to P2 multiplies prefix discrepancy by 1.51 to 5.61 and response by 1.41 to 5.71. The aligned continuation factor changes by between 0.93 and 1.02 times its original value.

Table 6: Prefix and continuation factors for all 36 settings. P1/S1 swap adjacent positions; P2/S2 use alternative orders, as defined above. Factors and response are percentages; the last column is a ratio.

model	cell	$r_q(c)$	δ_{pre}	aligned	marginal	aligned/marginal
LLaDA-8B	$k = 4$, P1/S1	0.177	1.87	9.46	6.53	1.45
	$k = 4$, P1/S2	0.478	1.87	25.54	7.92	3.22
	$k = 4$, P2/S1	0.495	5.94	8.33	6.54	1.27
	$k = 4$, P2/S2	1.327	5.94	22.34	7.73	2.89
	$k = 8$, P1/S1	0.087	0.57	15.39	1.05	14.63
	$k = 8$, P1/S2	0.216	0.57	37.99	8.85	4.29
	$k = 8$, P2/S1	0.494	16.00	3.09	1.17	2.65
	$k = 8$, P2/S2	1.994	16.00	12.46	8.80	1.42
	$k = 12$, P1/S1	0.006	0.12	5.50	0.29	18.95
	$k = 12$, P1/S2	0.025	0.12	21.73	0.85	25.70
	$k = 12$, P2/S1	0.182	17.64	1.03	0.24	4.34
	$k = 12$, P2/S2	0.497	17.64	2.82	0.63	4.44
MDLM-OWT	$k = 4$, P1/S1	5.951	28.45	20.91	20.12	1.04
	$k = 4$, P1/S2	25.564	28.45	89.84	87.91	1.02
	$k = 4$, P2/S1	8.381	43.01	19.49	20.48	0.95
	$k = 4$, P2/S2	38.148	43.01	88.69	88.42	1.00
	$k = 8$, P1/S1	4.549	18.42	24.70	22.89	1.08
	$k = 8$, P1/S2	14.379	18.42	78.08	75.42	1.04
	$k = 8$, P2/S1	20.693	88.85	23.29	23.29	1.00
	$k = 8$, P2/S2	67.121	88.85	75.54	75.49	1.00
	$k = 12$, P1/S1	1.529	16.89	9.05	8.78	1.03
	$k = 12$, P1/S2	6.800	16.89	40.25	36.94	1.09
	$k = 12$, P2/S1	8.728	94.74	9.21	9.18	1.00
	$k = 12$, P2/S2	36.306	94.74	38.32	38.25	1.00
Qwen3-0.6B	$k = 4$, P1/S1	0.001	0.09	1.62	0.84	1.92
	$k = 4$, P1/S2	0.008	0.09	8.51	3.26	2.61
	$k = 4$, P2/S1	0.007	0.70	1.03	0.84	1.24
	$k = 4$, P2/S2	0.043	0.70	6.20	3.33	1.86
	$k = 8$, P1/S1	0.001	0.004	18.55	1.64	11.34
	$k = 8$, P1/S2	0.002	0.004	54.90	24.13	2.28
	$k = 8$, P2/S1	0.347	13.41	2.59	1.56	1.66
	$k = 8$, P2/S2	5.518	13.41	41.15	22.63	1.82
	$k = 12$, P1/S1	0.001	0.45	0.24	0.20	1.16
	$k = 12$, P1/S2	0.007	0.45	1.55	1.52	1.02
	$k = 12$, P2/S1	0.152	31.90	0.48	0.17	2.88
	$k = 12$, P2/S2	1.178	31.90	3.69	1.36	2.71

Table 7: Factor decomposition of the three LLaDA reference circuits. Values are percentages.

circuit	$r_q(c)$	δ_{pre}	aligned continuation
early pairs	4.58	26.34	17.38
interleaved pairs	0.79	6.92	11.47
separated pairs	0.56	24.86	2.24

This decomposition uses an independent rerun of the three reference circuits. Its small numerical differences from Table 5 arise from Monte Carlo variation between distinct likelihood draws.

H.4 Increasing the prompt set to 512 examples

We repeat the 12 MDLM settings in Appendix H.3 on 128 prompts from each of the same four tasks. Four draws per prompt and side keep the budget at 4,096 rewards per circuit. Relative to the 32-prompt estimates in Table 6, every TV estimate changes by at most 0.77 percentage points.

Table 8: MDLM-OWT breadth check on 512 fixed prompts. Mean TV is a percentage; prompt-level SEs are in percentage points.

cell	mean TV	prompt SE
$k = 4$, P1/S1	6.05	0.166
$k = 4$, P1/S2	25.47	0.377
$k = 4$, P2/S1	7.95	0.193
$k = 4$, P2/S2	38.47	0.483
$k = 8$, P1/S1	4.54	0.123
$k = 8$, P1/S2	14.58	0.280
$k = 8$, P2/S1	20.41	0.325
$k = 8$, P2/S2	67.89	0.524
$k = 12$, P1/S1	1.53	0.058
$k = 12$, P1/S2	7.05	0.180
$k = 12$, P2/S1	8.91	0.236
$k = 12$, P2/S2	36.84	0.504

H.5 Random-switch scans and selected spread schedules

The random-switch measure is fixed as follows. Draw k uniformly from $\{2, \dots, 14\}$ and then draw a k-element set S uniformly among the completion positions. Draw a, a' independently and uniformly from the orders of S, conditioned on $a \neq a'$, and draw b, b' in the same way from the orders of S^c. The straight side mixes $ab, a'b'$ and the crossed side mixes $ab', a'b$; both sides have exactly equal edge frequencies. The MDLM and Qwen scans use 500 and 300 independent draws, respectively, with the eight fixed prompts, length 16, and 12 independent rewards per prompt ($N = 96$ per circuit).

For a scan of L circuits, simultaneous one-sided inference uses the union-bound Hoeffding radius

$$\varepsilon_L = \sqrt{\frac{\log(L/0.05)}{2 \cdot 96}}.$$

All reported lower bounds $[\hat{r}_c - \varepsilon_L]_+$ hold simultaneously with probability at least 0.95. In the MDLM scan, $411/500$ point estimates exceed 25%, above ε_{500}. Hence at least 82% of sampled circuits have positive simultaneous certificates under the sampling rule above.

The median TV is 43.0% for MDLM and 0.02% for Qwen; 1.3% of Qwen circuits exceed 1% TV. Grouping MDLM results by meeting depth gives the largest median at $k = 6$ (Figure 2b).

We build mixtures by ranking the sampled MDLM circuits by response and selecting circuits with disjoint path supports. For $N \in \{1, 2, 4, 8, 16, 32\}$ selected circuits, Equation (43) gives two laws with identical edge frequencies and $2N$ complete orders per side. After selection, each mixture size is evaluated on new terminal samples from its two sides using Equation (45).

components N	orders per side	fresh rewards	TV estimate
1	2	512	0.6806
2	4	512	0.6875
4	8	512	0.6610
8	16	512	0.6284
16	32	1,024	0.5777
32	64	2,048	0.5209

The mixtures use MDLM-OWT, the same eight fixed prompts, and completions of length 16. The last row averages four new draws per component side and prompt.

I Experiments on additional path information

I.1 Identification after matching edge pairs

We compute numerical ranks of the enumerated path matrices to test whether second-order information identifies schedules through $d = 5$ and fails from $d = 6$ (Theorem 22).

For each d we build the edge map T of Equation (1) and the pair co-visitation map J. Rows of J index unordered co-visitable edge pairs $\{e, f\}$, with entry one on permutations containing both. The ranks of T and $[T; J]$ give their kernel dimensions in the $d!$-dimensional path space.

Table 9: Hidden dimensions before and after matching pair co-visitation. The threshold $\lfloor d/2 \rfloor$ predicts closure through $d = 5$.

d	$\lvert\mathfrak{S}_d\rvert$	edges	pairs	$\dim\ker T$	$\dim\ker[T; J]$	closed
3	6	12	18	0	0	all
4	24	32	108	6	0	all
5	120	80	540	70	0	all
6	720	192	2430	590	90	84.7%

Adding J closes the kernel through $d = 5$ and leaves 90 dimensions at $d = 6$, confirming the predicted threshold.

At $d = 4$, each of the six minimal circuits in Equation (3) satisfies $Tc_S = 0$ and $\|Jc_S\|_\infty = 1$, with 16 affected pairs. Matching J removes every direction hidden from T.

I.2 Quadratic and cubic response near compatibility

The conditionals $q_{S,i,t}(x_i \mid x_S) = \frac{1}{2}\left[1 + t\, s_{S,i}(x_S)(2x_i - 1)\right]$ perturb the uniform product anchor on $\{0, 1\}^6$. For score seeds $401, 402, 403, 404, 405$, the scores $s_{S,i}(x_S) \in [0.2, 0.8]$ are fixed across amplitudes and schedules. The contrast $2x_i - 1$ centers each perturbation under the anchor. The T-matched pair assigns equal mass to $123456, 214356$ on one side and $124356, 213456$ on the other. For $T + J$ matching, we project a standard Gaussian vector over all 6! orders onto the degree-two moment kernel, using seed 20260805. Separately normalizing its positive and negative parts gives the two schedule laws.

We evaluate amplitudes $0.02, 0.04, 0.08, 0.12, 0.16, 0.24$, enumerating all 2^6 assignments and all orders in each schedule's support. The T-only tail switch has residuals $\|\Delta T\|_\infty = 0$ and $\|\Delta J\|_\infty = 0.5$. The null-space pair matches both to floating-point precision. Regressing log TV on log amplitude at the four smallest amplitudes gives

Pair	seed 401	402	403	404	405
T only	1.9957	2.0338	2.0040	1.9732	1.9947
$T + J$	2.9993	3.0110	3.0049	2.9945	3.0030

All five score seeds recover the quadratic and cubic response orders predicted by Theorem 9 and Equation (35).

I.3 Moment matching on eight MDLM orders

MDLM-OWT reveals one coordinate at a time over 16 positions, at temperature one without remasking. Prompts use GPT-2 tokenization and left padding with visible EOS tokens in a fixed span, shared by generation and scoring. The 16 prompts comprise four each of arithmetic, factual, copy, and logic queries.

The eight orientations of disjoint pairs $(0,1),(2,3),(4,5)$, followed by $6,\dots,15$, form a cube of complete orders. The degree-two characters $\chi_{01},\chi_{02},\chi_{12}$ and degree-three parity χ_{012} each partition it into equal four-versus-four sides. Exact enumeration shows that the first three match T with nonzero degree-two mismatch, while χ_{012} matches $T+J$ with nonzero degree-three mismatch. The cube, partitions, prompts, and coordinate pairs were fixed independently of every response value.

All four estimates share 32 terminals generated along each order per prompt, giving 256 per prompt and $N = 4{,}096$ overall. Each terminal is scored under all eight path laws. For character χ, log-mean-exp over each side's four paths gives $\ell_{\chi,+}$ and $\ell_{\chi,-}$. The unbiased TV reward is $W_\chi = \left|\tanh\big((\ell_{\chi,+}-\ell_{\chi,-})/2\big)\right|$. The primary shared-sample difference averages

$$\widehat{D}_N = \frac{1}{N}\sum_{s=1}^{N}\left[\frac{W_{01,s}+W_{02,s}+W_{12,s}}{3} - W_{012,s}\right].$$

Its summands lie in $[-1,1]$; the one-sided bound is $\widehat{D}_N - \sqrt{2\log(1/0.025)/N}$. Individual contrast bounds are simultaneous over the four contrasts.

Table 10: All contrasts on the fixed eight-order cube. Individual LCBs are simultaneous over four contrasts; the degree-two mean uses a separate one-sided bound.

contrast	matched moments	mean TV	one-sided LCB
χ_{01}	T	18.08%	15.76%
χ_{02}	T	17.19%	14.87%
χ_{12}	T	16.52%	14.20%
degree-two mean	T	17.26%	15.14%
χ_{012}	$T+J$	7.59%	5.27%

The paired difference is 9.68 percentage points with a one-sided lower bound of 5.43 points. The contrast matching $T+J$ retains 7.59% TV and a nonzero third-order mismatch. The four contrasts also differ in their higher-order moments.

I.4 Conditional deviation from compatible anchors

If a mask-indexed family arises from one joint, then for any revealed set S and unrevealed $i \neq j$,

$$q_i(u \mid x_S)\, q_j(v \mid x_S, u) = q_j(v \mid x_S)\, q_i(u \mid x_S, v),$$

both sides being $P(u,v \mid x_S)$. Writing R for the ratio of the two sides, a family with $q_e = P_e(1+u_e)$ and $|u_e| \le \alpha < 1$ satisfies

$$|\log R| \le 2\log\frac{1+\alpha}{1-\alpha} = 4\,\mathrm{artanh}(\alpha), \qquad \alpha \ge \tanh\left(\frac{|\log R|}{4}\right).$$

The last lower bound also holds when $\alpha \ge 1$ and is measurable without choosing an anchor.

On MDLM-OWT at $d = 16$ we draw $k \sim \mathrm{Unif}\{2,\dots,14\}$, a revealed set of that size uniformly, fill it by sampling from the model along a random order, take $i \neq j$ uniformly from the complement, and draw $u \sim q_i(\cdot \mid x_S)$, $v \sim q_j(\cdot \mid x_S)$: 200 scenarios with eight value pairs each.

Table 11: Square discrepancies on MDLM-OWT, 1,600 measurements. The $d = 16$ envelope $\sum_{j\ge 3}\binom{16}{j}\alpha^j$ reaches one at $\alpha = 0.107$.

	median	p90	p99	max
$\lvert\log R\rvert$	0.161	1.882	5.589	37.11

The square discrepancies have a pronounced upper tail. A uniform multiplicative deviation must dominate the lower bound from every sampled square. The observed maximum therefore implies $\alpha > 0.999$ for every compatible anchor, exceeding the 0.107 threshold of the $d = 16$ cubic envelope.