ORDER-ASYMMETRY IN DIFFUSION LANGUAGE MODELS: DISCRETE CURL AND CONDITIONAL COMPATIBILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion language models decode tokens in flexible orders, but their denoisers provide only local conditionals for partially observed contexts, which need not be compatible with a single joint distribution over an unresolved block. As a result, different commitment orders can induce different order-specific pseudo-joints. We call this model-side failure conditional incompatibility, distinct from data-side dependence. For any two unresolved positions, we define curl as the log-ratio between the pseudo-joints induced by reversing the commitment order; it vanishes exactly when the two orders agree. We then prove that the order-dependent part of sequential decoding error is a data-weighted curl path integral, separate from conditional total correlation and marginal model error. Empirically, pairwise Fisher–Rao curl energy distinguishes list-operation regimes, recovers Sudoku 4×4 constraint structure, and identifies a left-to-right preferred direction on GSM8K for both LLaDA-8B-Instruct and Dream-7B. Order sensitivity is therefore a measurable compatibility property of learned local conditionals.

1 INTRODUCTION

Autoregressive language models factorize text generation into a strict left-to-right product of conditional probabilities. This factorization gives clear probabilistic semantics, but it also imposes an inherent serial bottleneck: the ith token cannot be generated before the preceding $i-1$ tokens have been determined. Diffusion language models (DLMs) relax this constraint by recovering tokens through iterative denoising from highly corrupted text states. Since denoising need not follow a fixed absolute position order, DLMs support arbitrary-order, block-wise, and fully parallel generation (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2024; Nie et al., 2025; Ye et al., 2025). This flexibility raises a compatibility question: when unresolved tokens are committed in different orders, do these orders induce the same distribution over the block? A DLM denoiser provides local conditionals under partially observed contexts rather than a guaranteed joint distribution. Thus, different commitment orders can induce different order-specific pseudo-joints, making order sensitivity a possible consequence of model-side conditional incompatibility.

Recent theoretical analyses characterize related but distinct aspects of DLM decoding. Parallel decoding and inference schedules have been connected to conditional total correlation and related dependence measures under oracle or optimal-schedule settings (Kang et al., 2026; Chen et al., 2025), while fixed sequential rollout explains how local model error accumulates along one chosen order (Zhang et al., 2026). Practical decoders often lie between these settings: they choose which tokens can be safely committed together using confidence, stability, or learned dependency signals (Wu et al., 2025; Kim et al., 2025; Ringel et al., 2026). What remains less understood is how the decoding distribution changes across different sequential orders within the same block. We study this missing axis through conditional compatibility, connecting DLM decoding to conditionally specified distributions and scan-order effects in incompatible conditional families (Arnold & Press, 1989; Arnold et al., 2001; Kuo & Wang, 2019).

We formalize this order asymmetry with a discrete curl: the log-ratio between the pseudo-joints induced by the local orders $i \to j$ and $j \to i$. We also use Fisher–Rao curl energy as a symmetric magnitude of local incompatibility on the probability simplex. Theoretically, we prove that the order-

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.

dependent component of sequential decoding error is governed by a data-weighted curl path integral, separating it from data-side conditional total correlation and marginal model error. Empirically, Fisher–Rao curl energy distinguishes Copy, Sort, and Shuffle on ParallelBench list-operation tasks (Kang et al., 2026), recovers Sudoku 4×4 constraint geometry, and exposes model-induced row–column asymmetry. On GSM8K, the relative curl path integral identifies a local left-to-right preference for both LLaDA-8B-Instruct and Dream-7B, aligning with previously observed left-to-right decoding advantages (Israel et al., 2025) while revealing their local pseudo-joint geometry.

Contributions.

- **Compatibility view.** We formulate DLM block decoding as an order-specific pseudo-joint construction induced by local conditionals, and identify conditional incompatibility as a model-side source of order sensitivity.
- **Curl decomposition.** We define discrete curl as the log-ratio between pseudo-joints induced by reversing a pairwise commitment order, and prove that the order-dependent component of sequential decoding error is exactly a data-weighted curl path integral, separated from conditional total correlation and marginal model error.
- **DLM diagnostics.** We use directed curl and Fisher–Rao curl energy as inference-time diagnostics, showing that learned DLM conditionals exhibit structured incompatibility across list operations, Sudoku constraints, and left-to-right preferences on GSM8K.

2 Related Work

Discrete diffusion language models generate text by iteratively denoising corrupted or masked token states rather than following a fixed left-to-right factorization. Foundational and representative models include D3PMs (Austin et al., 2021), masked discrete diffusion in MDLM (Sahoo et al., 2024) and MD4 (Shi et al., 2024), score-entropy discrete diffusion in SEDD (Lou et al., 2024), and recent large-scale DLMs such as LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025). Alongside model development, recent theoretical analyses have examined the statistical limits of DLM decoding. ParallelBench (Kang et al., 2026) shows that one-shot parallel decoding error is lower-bounded by data conditional total correlation under oracle model assumptions, and constructs task suites where this bound is tight. Chen et al. (2025) characterize optimal unmasking schedules using total correlation and dual total correlation. Zhang et al. (2026) decompose forward KL of sequential rollout under model error and distinguish forward from reverse KL via incoherence analysis. Other recent work studies confidence-based adaptive decoding (Cai & Li, 2026), and distribution-agnostic schedules that adapt to unknown dependence structure (Zhao & Cai, 2026). These analyses explain parallel-commitment cost, schedule design, and error accumulation along a chosen rollout. Our work instead studies variation across sequential commitment orders within the same block.

Practical DLM decoders use token- or pair-level proxies to decide which positions can be safely committed together. Fast-dLLM (Wu et al., 2025) uses confidence thresholds; KLASS (Kim et al., 2025) uses token-level KL stability; DEMASK (Ringel et al., 2026) attaches a learned pairwise dependency predictor; PUNT (Azangulov et al., 2025) performs binary conditional-independence tests through anchor-and-test procedures; and CoDD (Li et al., 2026) replaces the factorized output with a tractable probabilistic inference layer. These methods motivate the need to understand dependence and compatibility inside decoding blocks.

A background line studies conditionally specified distributions, Markov random fields, Gibbs-style constructions, and dependency networks (Besag, 1974; Geman & Geman, 1984; Arnold & Press, 1989; Arnold et al., 2001; Heckerman et al., 2000; Hobert & Casella, 1998). Kuo & Wang (2019) show that incompatible conditional families can produce scan-order-dependent stationary distributions in Gibbs samplers. Related consistency testing for language models appears in Wagner et al. (2024).

3 DLMs as Local Conditional Specifications

3.1 Local Conditionals Rather Than a Joint

Consider a token sequence $x = (x_1, \ldots, x_L) \in \mathcal{V}^L$ of length L, where each x_i belongs to a vocabulary $\mathcal{V}$. In masked or discrete diffusion language modeling, the forward process gradually corrupts the original sequence through masks, noisy tokens, or other discrete corruptions, while the reverse process reconstructs text through iterative denoising. Let z_t denote the partially observed state at time step t, and let M_t denote the set of positions that remain masked or corrupted. A typical denoising predictor is written as (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2024; Nie et al., 2025; Ye et al., 2025)

$$p_\theta(x_i \mid z_t, t), \qquad i \in M_t.$$

For our analysis, we freeze the observed part of z_t and focus on the unresolved coordinates. Let $x = (x_1, \ldots, x_m) \in \mathcal{V}^m$ denote an assignment on those active coordinates, with already observed tokens absorbed into the conditioning context. For any current visible or committed position set A, the denoiser therefore induces a family of partial-context local conditionals

$$q_{\theta,t}(x_i \mid x_A), \qquad i \notin A,$$

where x_A contains the tokens currently visible to the denoiser. A DLM denoiser returns local conditionals under partially observed contexts. These conditionals specify one-step token predictions, but they do not automatically define a jointly normalized distribution over the unresolved block. Compatibility with a single conditional joint is therefore a structural property to be tested, not a consequence of the denoising interface.

3.2 Order-Specific Pseudo-Joints

Fix an observed context index set S and a block $B \subseteq [L] \setminus S$ of unresolved positions. For any permutation $\pi = (\pi_1, \ldots, \pi_{|B|})$ of B, the denoiser induces the order-specific pseudo-joint

$$Q^{\pi}_{\theta,t}(x_B \mid x_S) = \prod_{m=1}^{|B|} q_{\theta,t}(x_{\pi_m} \mid x_S, x_{\pi_{<m}}).$$

Different orders induce different pseudo-joints unless the partial-context local conditionals are compatible. If there exists a conditional joint $P_{\theta,t}(x_B \mid x_S)$ such that for every prefix set $A \subseteq B$ and every $i \in B \setminus A$,

$$q_{\theta,t}(x_i \mid x_S, x_A) = P_{\theta,t}(x_i \mid x_S, x_A),$$

then every order-specific pseudo-joint satisfies

$$Q^{\pi}_{\theta,t}(x_B \mid x_S) = P_{\theta,t}(x_B \mid x_S)$$

by the chain rule. In flexible neural denoisers trained through local denoising losses, however, compatibility is a property to be checked rather than an interface guarantee. Block decoding therefore depends not only on which positions are chosen, but also on whether the order-specific pseudo-joints induced by those positions agree.

4 Curl and Order Incompatibility

4.1 Local Curl as a Pseudo-Joint Log-Ratio

Fix a partial context x_S, two unresolved positions $i, j \notin S$, and candidate tokens $a, b \in \mathcal{V}$. Define the two local order-induced pseudo-joints

$$Q^{i\to j}_{\theta,t}(a, b \mid x_S) = q_{\theta,t}(a \mid x_S)\, q_{\theta,t}(b \mid x_S, x_i{=}a),$$
$$Q^{j\to i}_{\theta,t}(a, b \mid x_S) = q_{\theta,t}(b \mid x_S)\, q_{\theta,t}(a \mid x_S, x_j{=}b).$$

Define the discrete curl, abbreviated as curl,

$$\begin{aligned}\mathcal{C}^{a,b}_{ij}(x_S, t) = {} & \log q_{\theta,t}(x_i = a \mid x_S) + \log q_{\theta,t}(x_j = b \mid x_S, x_i = a) \\ & - \log q_{\theta,t}(x_j = b \mid x_S) - \log q_{\theta,t}(x_i = a \mid x_S, x_j = b).\end{aligned}$$

Theorem 1 (Local curl equals a pseudo-joint log-ratio)**.** *For every x_S, t, i, j, a, b,*

$$\mathcal{C}_{ij}^{a,b}(x_S, t) = \log \frac{Q_{\theta,t}^{i\to j}(a, b \mid x_S)}{Q_{\theta,t}^{j\to i}(a, b \mid x_S)}.$$

Consequently,

$$\mathcal{C}_{ij}^{a,b}(x_S, t) = 0 \quad \forall a, b$$

holds if and only if

$$Q_{\theta,t}^{i\to j}(\cdot, \cdot \mid x_S) = Q_{\theta,t}^{j\to i}(\cdot, \cdot \mid x_S)$$

as distributions. Moreover,

$$\mathbb{E}_{(a,b)\sim Q_{\theta,t}^{i\to j}(\cdot,\cdot\mid x_S)}\big[\mathcal{C}_{ij}^{a,b}(x_S, t)\big] = \mathrm{KL}\big(Q_{\theta,t}^{i\to j}(\cdot, \cdot \mid x_S) \,\|\, Q_{\theta,t}^{j\to i}(\cdot, \cdot \mid x_S)\big).$$

The theorem identifies curl with the log discrepancy between the two elementary commitment paths. This local quantity is the building block for comparing arbitrary block orders and is computable from model conditionals alone.

Directed curl therefore serves two roles at once. Pointwise, it is the signed local preference for the order $i \to j$ over $j \to i$ on the candidate pair (a, b). In expectation under $Q_{\theta,t}^{i\to j}$, it becomes the directed pseudo-joint KL and hence the local penalty for reversing that order.

4.2 Curl Is Not Data Dependence

Any joint distribution is order-compatible by the chain rule:

$$p_{\mathrm{data}}(a, b \mid x_S) = p_{\mathrm{data}}(a \mid x_S)\, p_{\mathrm{data}}(b \mid x_S, a) = p_{\mathrm{data}}(b \mid x_S)\, p_{\mathrm{data}}(a \mid x_S, b).$$

Hence if $q_{\theta,t}$ coincides with p_{data} on all relevant partial contexts, $Q_{\theta,t}^{i\to j} = Q_{\theta,t}^{j\to i} = p_{\mathrm{data}}$ and curl vanishes identically. Curl is therefore a strict model-side quantity, computed entirely from $q_{\theta,t}$ and never sourced from the data: nonzero curl certifies that the learned conditional family does not lift to any single joint over the block.

This distinguishes curl from conditional total correlation. For a fixed context x_S,

$$\mathrm{TC}(X_B \mid x_S) := \mathrm{KL}\left(p(x_B \mid x_S) \,\middle\|\, \textstyle\prod_{i\in B} p(x_i \mid x_S)\right).$$

TC and curl live on different layers: TC measures dependence within a compatible system (the data joint), while curl measures whether the learned conditional family forms a compatible system at all. TC can be large with zero curl; curl can be nonzero with $\mathrm{TC} = 0$.

Separation between dependence and incompatibility.

Proposition 1 (Separation of total correlation and curl)**.** *Data conditional total correlation and curl measure distinct axes. There exist compatible denoisers with $\mathrm{TC}(X_B \mid x_S) > 0$ and zero curl, and there exist incompatible denoisers with $\mathrm{TC}(X_B \mid x_S) = 0$ and nonzero curl.*

Constructive witnesses for both directions are given in Appendix D.1. They show that total correlation belongs to the data conditional, while curl belongs to the learned local conditional specification. The former controls parallel-commitment cost; the latter controls order asymmetry among sequential commitments.

The same curl also admits a discrete exterior-derivative interpretation on the partial-context lattice, which yields the adjacent-swap identity used in the next section.

4.3 From Directed Curl to a Symmetric Magnitude

Directed curl $\mathcal{C}_{ij}^{a,b}$ is signed and depends on the chosen local order. An order-agnostic magnitude should vanish exactly when the two pseudo-joints coincide. Since $Q_{\theta,t}^{i\to j}$ and $Q_{\theta,t}^{j\to i}$ are points in the probability simplex, we measure their discrepancy with the squared Fisher–Rao distance,

$$\Omega_{ij}^{\mathrm{FR}}(x_S, t) = d_{\mathrm{FR}}^2\left(Q_{\theta,t}^{i\to j}(\cdot, \cdot \mid x_S), Q_{\theta,t}^{j\to i}(\cdot, \cdot \mid x_S)\right).$$

This is the standard information-geometric metric on the simplex. Its bounded form is the squared Hellinger distance H^2 used in our experiments.

5 CURL AND SEQUENTIAL ORDER ASYMMETRY

Under Bayes-optimal conditionals, one-shot parallel decoding error is governed by conditional total correlation (Kang et al., 2026; Chen et al., 2025). For sequential decoding of a fixed block, variation across commitment orders enters only through curl.

Theorem 2 (Sequential decoding error decomposition)**.** *Let*

$$p(x_B \mid x_S) = p_{\mathrm{data}}(x_B \mid x_S).$$

For any block B, observed context x_S, and order π of B,

$$\mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\pi}_{\theta,t}(x_B \mid x_S)\big) = \mathrm{TC}(X_B \mid x_S) + \sum_{i \in B} \mathrm{KL}\big(p(x_i \mid x_S) \,\|\, q_{\theta,t}(x_i \mid x_S)\big) - \mathcal{I}^{\pi}_{\theta,t}(x_S),$$

where

$$\mathcal{I}^{\pi}_{\theta,t}(x_S) := \sum_{m=1}^{|B|} \mathbb{E}_p\left[\log \frac{q_{\theta,t}(x_{\pi_m} \mid x_S, x_{\pi_{<m}})}{q_{\theta,t}(x_{\pi_m} \mid x_S)}\right]$$

is the model-perceived intra-block information along order π, evaluated under the data distribution.

Conditional total correlation and the marginal KL terms do not depend on π; only $\mathcal{I}^{\pi}_{\theta,t}$ changes with the commitment order. Curl is the local form of this order-dependent variation.

The summand

$$\log q_{\theta,t}(x_{\pi_m} \mid x_S, x_{\pi_{<m}}) - \log q_{\theta,t}(x_{\pi_m} \mid x_S)$$

is the log-ratio of model conditionals at π_m with and without the earlier commitments along π, and $\mathcal{I}^{\pi}_{\theta,t}$ is the expectation of their sum under p. If $q_{\theta,t}$ agrees with p_{data} on every partial context, the chain rule gives

$$\mathcal{I}^{\pi}_{\theta,t}(x_S) = \mathrm{TC}(X_B \mid x_S)$$

for all π, and Theorem 2 reduces to

$$\mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\pi}_{\theta,t}(x_B \mid x_S)\big) = 0.$$

Bayes-optimal compatibility thus implies vanishing sequential decoding error along every order. Outside this regime, $\mathcal{I}^{\pi}_{\theta,t}$ differs from TC and depends on π, and the change in $\mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\pi}_{\theta,t}(x_B \mid x_S)\big)$ between two orders equals

$$\mathcal{I}^{\pi'}_{\theta,t}(x_S) - \mathcal{I}^{\pi}_{\theta,t}(x_S).$$

Corollary 2.1 (Order asymmetry as a curl path integral)**.** *Let p abbreviate $p(x_B \mid x_S)$. For any two orders π, π' of B,*

$$\mathrm{KL}\big(p \,\|\, Q^{\pi}_{\theta,t}\big) - \mathrm{KL}\big(p \,\|\, Q^{\pi'}_{\theta,t}\big) = -\mathbb{E}_p\left[\sum_{r \in \Gamma} \mathcal{C}^{x_{i_r}, x_{j_r}}_{i_r j_r}(x_{S_r}, t)\right],$$

where Γ is any adjacent-swap path from π to π'.

The two-position adjacent swap is the elementary building block for comparing arbitrary block orders. In this sense, the pairwise curl is an elementary local quantity, but the order asymmetry of an arbitrary multi-token block is obtained by accumulating these local curl terms along any adjacent-swap path between the two decoding orders.

Conditional total correlation determines the parallel-commitment term, whereas the difference between two sequential orders is the curl path integral along an adjacent-swap path. The same identity also yields Pinsker-type bounds on bounded sample-level order statistics.

6 EXPERIMENTS

We compute Fisher–Rao curl energy from the local conditionals of a pretrained instruction-tuned DLM across tasks and constraint geometries. We report two statistics: the squared Hellinger form

$$H^2(P, Q) \;=\; 1 - \sum_{a,b} \sqrt{P_{ab}\, Q_{ab}},$$

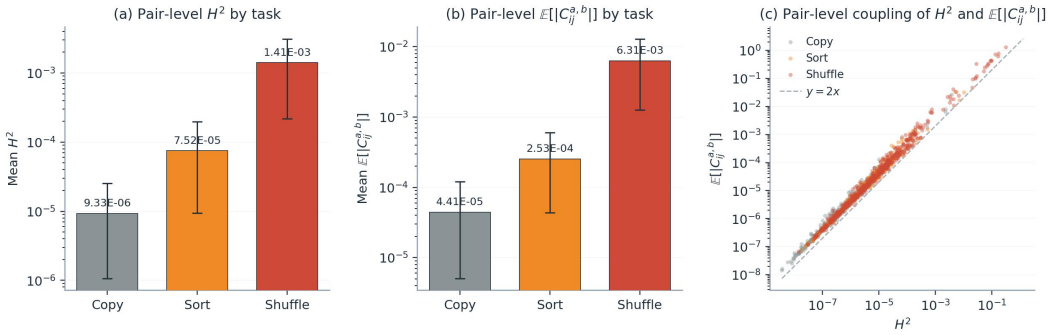

Figure 1: Pair-level incompatibility across Copy, Shuffle, and Sort. (**Left**, **Middle**) Mean H^2 and $\mathbb{E}_{Q_{\theta,t}^{i\to j}}[C_{ij}^{a,b}]$ with 95% bootstrap CI. (**Right**) Per-pair scatter; the two statistics co-vary as predicted by Proposition 3.

a bounded reparameterization of Fisher–Rao curl energy that vanishes if and only if the two pseudo-joints coincide, and

$$\mathrm{KL}\left(Q_{\theta,t}^{i\to j} \,\middle\|\, Q_{\theta,t}^{j\to i}\right) \;=\; \mathbb{E}_{Q_{\theta,t}^{i\to j}}\left[\mathcal{C}_{ij}^{a,b}(x_S,t)\right].$$

The conditional terms $q_{\theta,t}(x_j \mid x_S, x_i{=}a)$ come from intervened forward passes that are cached per candidate token. Estimator implementation and robustness checks are given in Appendix B. All scores are computed from model conditionals at inference time.

6.1 Experimental Setup

We use `LLaDA-8B-Instruct` (Nie et al., 2025) in `bfloat16` with seed 42 and maximum sequence length 512. Active sets and masked-pair samples are restricted to the assistant response region. All bootstrap 95% CIs use context-level cluster bootstrap with 1000 resamples.

6.2 Incompatibility Across Dependence Regimes

We probe three list-operation tasks from ParallelBench (Kang et al., 2026)—Copy, Shuffle, and Sort. Copy and Sort are deterministic permutations of the input: given X, the output Y is uniquely determined, so the conditional total correlation $TC(Y \mid X) = 0$. Shuffle is a random permutation with $TC(Y \mid X) > 0$: under one-step parallel decoding from Kang et al. (2026), Shuffle accuracy converges to zero as the list length n grows, whereas Copy/Sort accuracy remains bounded away from zero.

For each task we draw 200 contexts at mask ratio 0.3 and sample up to 20 masked-position pairs per context from the assistant response.

Table 1: Pair-level incompatibility on three list-operation tasks under LLaDA-8B-Instruct. Mean H^2 and $\mathbb{E}_{Q_{\theta,t}^{i\to j}}[C_{ij}^{a,b}]$ are our measurements.

Task	$TC(Y \mid X)$	Pairs	Mean H^2	$\mathbb{E}_{Q_{\theta,t}^{i\to j}}\left[\mathcal{C}_{ij}^{a,b}(x_S,t)\right]$
Copy	0	565	9.3×10^{-6}	4.4×10^{-5}
Sort	0	546	7.5×10^{-5}	2.5×10^{-4}
Shuffle	>0	583	1.4×10^{-3}	6.3×10^{-3}

(i) Copy is near the vanishing condition. Copy has H^2 near the numerical floor, so the model's local conditionals compose almost consistently for this position-wise mapping. Shuffle has the largest incompatibility; its partially observed outputs produce less stable order-conditioned predictions.

(ii) FR exposes incompatibility that TC does not predict. Sort and Copy share the same data-side $TC(Y \mid X) = 0$—both are deterministic permutations of the input—yet FR on Sort is 7.5×10^{-5},

roughly $8\times$ that of Copy and well above the numerical floor. The Sort–Copy gap cannot originate from data-side dependence, since by construction there is none. It must reflect a model-side property: the learned local conditionals on Sort exhibit residual order-dependence that disappears on Copy, despite both tasks being equally tractable for a Bayes-optimal denoiser.

Copy and Sort both have zero data-side conditional total correlation, but their curl energies differ by roughly a factor of eight. Data-side conditional total correlation therefore does not determine the compatibility of the learned local conditionals.

6.3 Constraint Structure on Sudoku 4 × 4

We evaluate whether pairwise incompatibility follows the known constraint geometry of Sudoku 4×4 from ParallelBench. Cells must take distinct values within each row, column, and 2×2 box. We evaluate 100 puzzles, providing 8 ground-truth cells and masking the remaining 8; for each puzzle we compute H^2 for every pair among the masked cells and classify each pair into one of six mutually exclusive geometric classes.

Table 2: Pairwise H^2 on Sudoku 4×4 under LLaDA-8B-Instruct. The upper panel reports exclusive geometric classes under row-major serialization; confidence intervals use puzzle-level cluster bootstrap. The lower panel shows that the row–column asymmetry reverses under column-major serialization. "Box only" denotes pairs in the same 2×2 box that share neither row nor column.

Geometric class	n_{pairs}	Mean H^2	95% CI
Independent	1505	0.0071	$[0.0064, 0.0078]$
Column only	379	0.0078	$[0.0068, 0.0087]$
Box only	189	0.0099	$[0.0081, 0.0115]$
Column + box	175	0.0130	$[0.0105, 0.0157]$
Row only	354	0.0210	$[0.0183, 0.0237]$
Row + box	198	0.0448	$[0.0349, 0.0555]$
Serialization	**Row-only H^2**	**Column-only H^2**	**Row/Column**
Row-major	0.02097	0.00778	2.69
Column-major	0.01681	0.02809	0.60

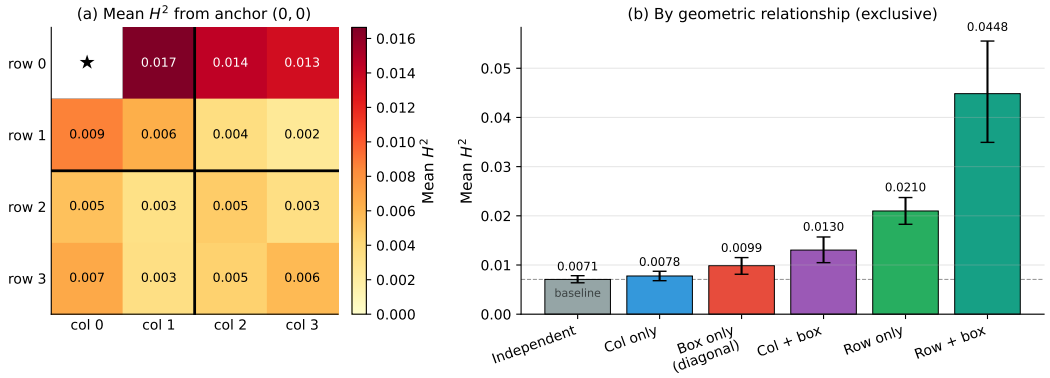

Figure 2: Pair-level incompatibility recovers Sudoku constraint structure with a row–column asymmetry. (**Left**) Mean H^2 between anchor cell $(0, 0)$ and the remaining 15 cells. (**Right**) Mean H^2 by exclusive geometric class with 95% puzzle-level cluster-bootstrap CI; dashed line marks the independent baseline.

(i) The constraint graph is recovered. One-sided Mann–Whitney tests reject equality with the independent baseline at $p < 10^{-3}$ for all three constraint classes.

(ii) Serialization controls the row–column asymmetry. Under row-major serialization, row-only pairs have $2.69\times$ the incompatibility of column-only pairs (0.02097 vs. 0.00778). Under column-major

Table 3: Relative curl path integral $\tilde{I}^\pi$ on LLaDA-8B-Instruct across clean GSM8K mid-decoding blocks. "Kendall to LTR" is the Kendall tau distance to $(1, 2, 3)$. 95% CIs are bootstrap over blocks.

Permutation	Kendall to LTR	Mean $\tilde{I}^\pi$	95% CI
123 (LTR)	0	$\mathbf{0.046}$	$[\ \mathbf{0.033},\ \mathbf{0.059}]$
132	1	-0.001	$[-0.014,\ \ 0.013]$
213	1	0.011	$[-0.002,\ \ 0.024]$
231	2	-0.008	$[-0.020,\ \ 0.005]$
312	2	-0.025	$[-0.038,\ -0.012]$
321 (RTL)	3	-0.023	$[-0.036,\ -0.011]$

serialization, the ratio reverses to 0.60 (0.01681 vs. 0.02809). The reversal ties the asymmetry to sequence serialization rather than to the formally symmetric Sudoku constraints.

(iii) Box is a real increment on row. Box-only pairs are significant but small (0.010); adding the box constraint to a same-row pair raises H^2 from 0.021 to 0.045 (one-sided Mann–Whitney $p = 2.4 \times 10^{-3}$), so box dependence is learned as a genuine increment.

Marginal uncertainty alone does not explain the pattern: the task imposes symmetric row and column restrictions, whereas H^2 is strongly asymmetric. The measured incompatibility therefore depends on how the serialized constraints are represented by the model.

6.4 LTR is Preferred in Pseudo-Joint Geometry

We test for directional structure in local pseudo-joint geometry on two models with different training pipelines: LLaDA-8B-Instruct Nie et al. (2025) (trained from scratch with a masked diffusion objective) and Dream-7B Ye et al. (2025) (initialized from the autoregressive Qwen2.5-7B and adapted via diffusion fine-tuning).

Israel et al. (2025) report that on GSM8K, forced left-to-right (LTR) decoding outperforms the default decoding strategies of both LLaDA and Dream. We use the same generation settings and measure each model's local pseudo-joint preference.

Setup. We run both models on the GSM8K Cobbe et al. (2021) test set with generation length and diffusion steps set to 256, using each model's default decoder (confidence-based for LLaDA, entropy-based for Dream) to generate mid-trajectory states.

We sample size-3 blocks from the mid-trajectory states and enumerate all 6 permutations per block. For each block, we fix the assignment $\hat{x}_B$ to the base-forward argmax tokens at the queried state and evaluate all six permutations on this shared assignment. By Corollary 2.1, the difference $\log Q^\pi_{\theta,t}(\hat{x}_B \mid x_S) - \log Q^{\pi'}_{\theta,t}(\hat{x}_B \mid x_S)$ between any two permutations equals the curl path integral along any adjacent-swap path between them. We define the relative curl path integral as

$$\tilde{I}^\pi = \log Q^\pi_{\theta,t}(\hat{x}_B \mid x_S) - \frac{1}{6}\sum_{\pi'} \log Q^{\pi'}_{\theta,t}(\hat{x}_B \mid x_S), \tag{1}$$

and report the mean and 95% bootstrap CI for each permutation. Blocks whose fixed assignment $\hat{x}_B$ contains special, mask, pad, or non-visible tokens are excluded, as the pseudo-joint is not well-defined for such assignments.

LLaDA: LTR is significantly positive; RTL is significantly negative. For LLaDA, LTR is the unique permutation with significantly positive mean, the two permutations farthest from LTR (including RTL) are significantly negative, and the intermediate permutations cross zero. The mean $\tilde{I}^\pi$ is negatively correlated with Kendall distance from LTR (Spearman $\rho = -0.88$), and the paired LTR-vs-RTL difference is 0.069 nats ($p < 10^{-7}$).

LLaDA: the preference concentrates in order-sensitive regions. LTR is the locally best permutation in 26% of blocks, above the uniform baseline of 17%. When blocks are stratified by order spread $\max_\pi \log Q^\pi - \min_\pi \log Q^\pi$, the mean LTR preference rises monotonically with spread: the bias is

Table 4: Relative curl path integral $\tilde{I}^{\pi}$ on Dream-7B across clean GSM8K mid-decoding blocks sampled from Dream's default entropy-based trajectory.

Permutation	Kendall to LTR	Mean $\tilde{I}^{\pi}$	95% CI
123 (LTR)	0	$\mathbf{0.643}$	$[\ \mathbf{0.485},\ \mathbf{0.813}]$
132	1	0.219	$[\ 0.060,\ 0.364]$
213	1	-0.146	$[-0.291,\ 0.004]$
231	2	-0.181	$[-0.347, -0.038]$
312	2	0.044	$[-0.121,\ 0.214]$
321 (RTL)	3	$\mathbf{-0.578}$	$[\mathbf{-0.732}, \mathbf{-0.408}]$

weak when the six order-induced pseudo-joints nearly agree, and strongest when the local conditional system is most order-sensitive.

Dream: same direction, substantially larger magnitude. Dream has the same sign pattern: LTR is significantly positive, RTL is significantly negative, and the mean $\tilde{I}^{\pi}$ is negatively correlated with Kendall distance (Spearman $\rho = -0.85$). The paired LTR-vs-RTL difference is 1.22 nats ($p < 10^{-13}$), approximately $18\times$ the value on LLaDA. LTR is the locally best permutation in 27% of blocks. The local LTR preference is therefore present in both models and substantially larger in the AR-initialized Dream.

Relation to forced-LTR accuracy. Israel et al. (2025) reports that forcing LTR improves GSM8K accuracy for both LLaDA and Dream, with a larger gain on LLaDA. Our local measurements show the complementary pattern: Dream has a stronger per-block LTR pseudo-joint preference. These observations are consistent because the gain from forced LTR also depends on how well the default decoder aligns with the model's local preference. If the default decoder already follows that preference, forcing LTR yields a smaller additional gain; if it deviates from it, the gain can be larger. Thus, curl captures a local directional bias that trajectory-level accuracy gaps alone do not identify.

Possible role of AR initialization. Dream inherits Qwen2.5's autoregressive weights (Qwen et al., 2024), which encode strict left-to-right factorization. Its LTR bias is substantially larger than that of the from-scratch trained LLaDA, consistent with directional structure persisting after diffusion fine-tuning. The comparison does not isolate initialization from other differences between the models.

7 Conclusion

DLM denoisers provide local conditionals that need not be compatible with a single global joint. This makes block decoding order-specific: each sequential commitment order induces its own pseudo-joint. Curl explains the signed order asymmetry between such pseudo-joints, and Fisher–Rao curl energy diagnoses the magnitude of the local incompatibility. Theorem 2 places this in a structural picture: sequential block decoding error decomposes into conditional total correlation, marginal model error, and a single order-dependent functional, and curl gives the local form of that functional.

Across list operations, Sudoku, and GSM8K, Fisher–Rao curl energy varies independently of data-side dependence and token confidence. Order sensitivity therefore includes a measurable compatibility failure of the learned local conditionals.

AI use statement

Generative AI assisted with manuscript editing, coding and computational checks.

Reproducibility statement

The code is available at

`https://anonymous.4open.science/r/Order-Asymmetry-in-Diffusion-Language-Models-49DD/`.

REFERENCES

Barry C Arnold and S James Press. Compatible conditional distributions. *Journal of the American Statistical Association*, 84(405):152–156, 1989.

Barry C Arnold, Enrique Castillo, and Jose Maria Sarabia. Conditionally specified distributions: an introduction (with comments and a rejoinder by the authors). *Statistical Science*, 16(3):249–274, 2001.

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*, 2021.

I. Azangulov, T. Pandeva, N. Prasad, J. Zazo, and S. Karmalkar. Parallel sampling from masked diffusion models via conditional independence testing, 2025. Preprint, arXiv:2510.21961.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(2):192–236, 1974.

C. Cai and G. Li. Confidence-based decoding is provably efficient for diffusion language models, 2026. Preprint, arXiv:2603.22248.

S. Chen, K. Cong, and J. Li. Optimal inference schedules for masked diffusion models, 2025. Preprint, arXiv:2511.04647.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6): 721–741, 1984.

David Heckerman, David Maxwell Chickering, Christopher Meek, Robert Rounthwaite, and Carl Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.

James P. Hobert and George Casella. Functional compatibility, markov chains, and gibbs sampling with improper posteriors. *Journal of Computational and Graphical Statistics*, 7(1):42–60, 1998.

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating diffusion LLMs via adaptive parallel decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

W. Kang, K. Galim, S. Oh, M. Lee, Y. Zeng, S. Zhang, C. Hooper, Y. Hu, H. I. Koo, N. I. Cho, et al. Parallelbench: Understanding the trade-offs of parallel decoding in diffusion llms. In *International Conference on Learning Representations*, 2026. arXiv:2510.04767.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. KLASS: KL-guided fast inference in masked diffusion models. In *Advances in Neural Information Processing Systems*, 2025.

K.-L. Kuo and Y. J. Wang. Pseudo-gibbs sampler for discrete conditional distributions. *Annals of the Institute of Statistical Mathematics*, 71(1):93–105, 2019.

I. Li, Z. Shao, B. Wang, R. Yu, G. Van den Broeck, and A. Liu. Breaking the factorization barrier in diffusion language models, 2026. Preprint, arXiv:2603.00045.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

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. In *Advances in Neural Information Processing Systems*, 2025.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc-Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1525–1534, 2016.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL `https://arxiv.org/abs/2412.15115`.

L. Ringel, A. Ali, and Y. Romano. Dependency-guided parallel decoding in discrete diffusion language models, 2026. Preprint, arXiv:2604.02560.

Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37: 103131–103167, 2024.

Eitan Wagner, Yuli Slavutsky, and Omri Abend. CONTESTS: a framework for consistency testing of span probabilities in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15469–15484, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.866.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

S. Zhang, L. Yu, R. Brekelmans, L. Tang, S. Asif, and G. Ver Steeg. Generation order and parallel decoding in masked diffusion models: An information-theoretic perspective, 2026. Preprint, arXiv:2602.00286.

Y. Zhao and C. Cai. Adaptation to intrinsic dependence in diffusion language models, 2026. Preprint, arXiv:2602.20126.

A Additional Experimental Details

A.1 Reverse-Time Trajectory of Incompatibility

We evaluate mask ratios $t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$ along the denoising trajectory. For each t we sample 30 contexts from each of Waiting Line, Sudoku 4×4, and LAMBADA, and report per-context mean pairwise H^2 with 95% bootstrap CIs.

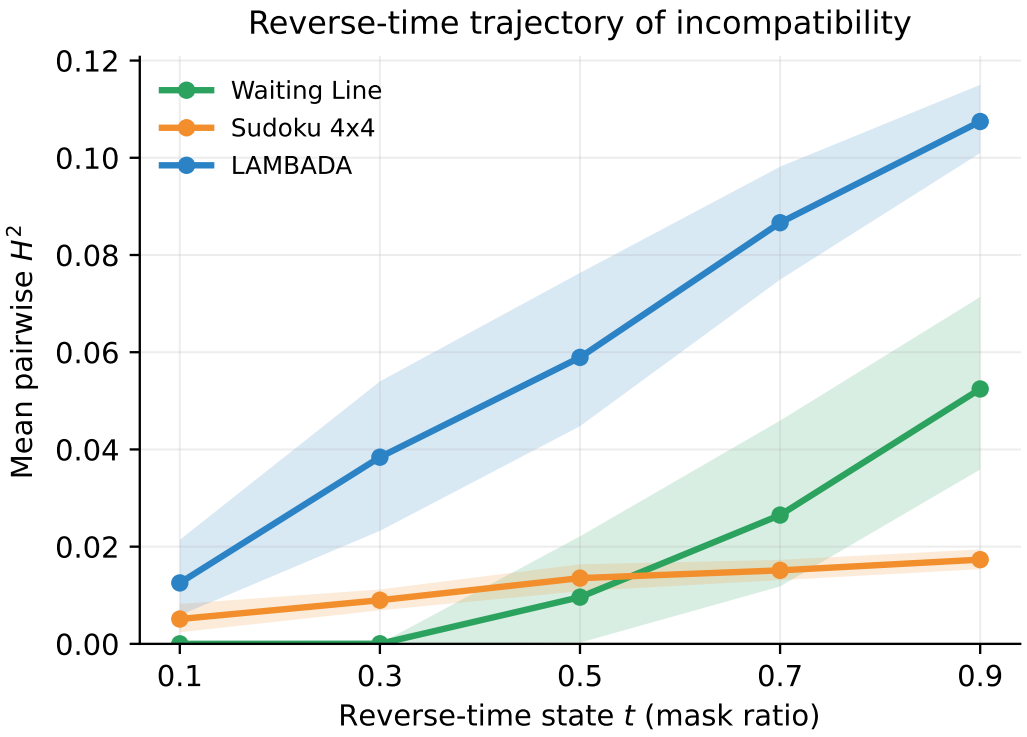

Figure 3: Reverse-time trajectory of pairwise incompatibility on three domains. The three domains exhibit qualitatively distinct trajectory shapes.

The three trajectories differ qualitatively, not merely in magnitude (Figure 3). **LAMBADA** is high and smoothly monotone: $H^2 = 0.012$ already at $t = 0.1$ (when most positions are already resolved) and rises near-linearly to 0.107 at $t = 0.9$. **Sudoku** starts low (0.005), rises gradually, and saturates near 0.017. **Waiting Line** is nearly at the numerical floor for $t \leq 0.3$ ($H^2 \sim 10^{-6}$, the level of Copy in Section 6.2) and then jumps by three orders of magnitude to 0.010 at $t = 0.5$, reaching 0.052 at $t = 0.9$.

LAMBADA remains incompatible even with most positions resolved because neighboring positions admit multiple context-dependent completions. Sudoku rises early and saturates as row, column, and box constraints propagate through its finite solution space. Waiting Line remains near zero while most output tokens are visible, then rises sharply once the visible context falls below roughly half. Thus, the same mask ratio produces different incompatibility levels and trajectory shapes across domains.

B Estimator Details

This appendix collects the implementation details of the Fisher–Rao curl estimator and the robustness checks summarized in Section 6.2. The robustness checks in Sections B.4 and B.5 use an earlier run on `LLaDA-8B-Base` with WikiText-103, GSM8K, and ParallelBench; the main-text experiments use `LLaDA-8B-Instruct` with the prompt templates of ParallelBench.

B.1 Fisher–Rao Curl Energy

Given a partially committed context x_S at reverse-time state z_t and two unresolved positions $i, j \notin S$, the pairwise Fisher–Rao curl energy is

$$\Omega_{ij}^{\mathrm{FR}}(x_S, t) = d_{\mathrm{FR}}^2\Big(Q_{\theta,t}^{i\to j}(\cdot,\cdot \mid x_S),\, Q_{\theta,t}^{j\to i}(\cdot,\cdot \mid x_S)\Big),$$

where on the discrete probability simplex

$$d_{\mathrm{FR}}(P, R) = 2\arccos\left(\sum_{a,b}\sqrt{P_{ab}R_{ab}}\right).$$

By metric definiteness of d_{FR} and Theorem 1,

$$\Omega_{ij}^{\mathrm{FR}}(x_S, t) = 0 \iff \mathcal{C}_{ij}^{a,b}(x_S, t) = 0 \quad \forall a, b.$$

The symmetric score $\Omega_{ij}^{\mathrm{FR}}$ measures the magnitude of local order incompatibility, while the signed pointwise curl $\mathcal{C}_{ij}^{a,b}$ records which local order contributes positively or negatively to a path comparison. The squared Hellinger form H^2 reported in the main text is the bounded reparameterization of d_{FR}^2. Proposition 3 further shows

$$\Omega_{ij}^{\mathrm{FR}} = 2\mathbb{E}_{Q_{\theta,t}^{i\to j}}\left[\mathcal{C}_{ij}^{a,b}\right] + O_m(\|\delta\|^3),$$

so the symmetric and directed scores agree to leading order.

B.2 Top-k Shared-Support Estimation

Computing the full pseudo-joints over $\mathcal{V} \times \mathcal{V}$ is unnecessary. We first run a base forward pass to obtain $q_{\theta,t}(x_i \mid x_S)$ for the active unresolved positions and retain the top-k candidate set $\mathcal{K}_i$ at each position. For a pair (i, j), both pseudo-joints are restricted to the shared support

$$\mathcal{S}_{ij} = \mathcal{K}_i \times \mathcal{K}_j$$

and renormalized before computing $\Omega_{ij}^{\mathrm{FR}}$, the directed KL, or expected curl. Shared-support renormalization limits the influence of incomparable tails and makes scores comparable across positions and reverse-time steps.

B.3 Batched Intervention Cache and Cost

The conditional terms $q_{\theta,t}(x_j \mid x_S, x_i = a)$ and $q_{\theta,t}(x_i \mid x_S, x_j = b)$ are obtained through interventions on the current context. For each active position i and candidate token $a \in \mathcal{K}_i$, we run the model on the intervened context $x_S \cup \{x_i = a\}$ and cache the target distributions at the other active positions. This cache supplies all entries needed to assemble both pseudo-joints for any candidate pair within the active set.

The active set can be restricted by uncertainty, mask geometry, or the experimental block under study. If m active positions are retained and each has k candidates, the estimator requires $O(mk)$ intervened forward contexts and yields $O(m^2k^2)$ cached pseudo-joint entries. All experiments use $k = 16$.

B.4 Top-k Stability

The shared-support top-k approximation gives stable rankings between $k = 8$ and $k = 16$. Pearson correlations range from 0.9335 to 0.999979, and Spearman correlations range from 0.9958 to 0.998486 (Table 5).

Dataset	Pearson	Spearman
WikiText	0.9335	0.9958
GSM8K	0.9985	0.9963
ParallelBench	0.999979	0.998486

Table 5: Top-k stability between $k = 8$ and $k = 16$. Spearman correlation measures stability of the pairwise ranking.

B.5 Cross-State Null Baseline

To verify that the estimated H^2 scores capture coherent within-state structure rather than estimator artifacts, we construct a matched cross-state null baseline. For each real pair, we replace one side with a donor from a different sample, matched by mask ratio, top-k, and distance bucket. This keeps the estimator configuration fixed while breaking the shared masked-state context. Across WikiText, GSM8K, and ParallelBench, cross-state null distances are consistently larger than real within-state distances (Figure 4), confirming that the within-state signal is not an artifact of support truncation or the estimator floor.

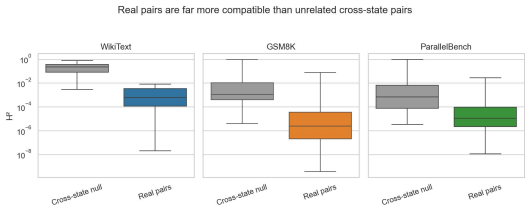

Figure 4: Cross-state null baseline. Real within-state distances are consistently smaller than matched cross-state null distances across WikiText, GSM8K, and ParallelBench, indicating that the reported scores reflect coherent masked-state structure rather than estimator artifacts.

C Order-Sensitivity Prediction on LAMBADA

We test whether block-level pseudo-joint curl predicts sample-level output changes under different decoding orders. The experiment uses LAMBADA (Paperno et al., 2016), where natural-language under-determination produces order-sensitive completions at the level of individual passages.

LAMBADA targets require broad discourse context: humans can often infer the last word from the full passage, but not from the immediately preceding sentence alone. The final block is therefore locally underdetermined but globally constrained, and different sequential orders expose different partial contexts.

Setup. We evaluate the full LAMBADA test set using `LLaDA-8B-Instruct` (Nie et al., 2025). For each passage we select the final four non-special tokens as a target block ($B = 4$), holding the preceding context fixed. To avoid conflating token-level difficulty with order sensitivity—LAMBADA targets are deliberately rare and often outside the model's high-probability region—we replace the gold target with the model's own LTR-decoded output: starting from a fully-masked block, we run forced greedy decoding under the left-to-right ordering $(0, 1, 2, 3)$ to obtain a reference sequence $\mathbf{x}_B = (x_0, x_1, x_2, x_3)$, and all subsequent quantities are computed against this $\mathbf{x}_B$.

We enumerate the $4! = 24$ block permutations. For each permutation π the pseudo-joint log-probability of the reference is

$$\log Q_\theta^\pi(\mathbf{x}_B) = \sum_{m=1}^{4} \log q_\theta(x_{\pi_m} \mid \mathbf{x}_{S_m}),$$
$$S_m = \{\pi_1, \ldots, \pi_{m-1}\},$$

and we define $\pi^+ = \arg\max_\pi \log Q_\theta^\pi(\mathbf{x}_B)$ and $\pi^- = \arg\min_\pi \log Q_\theta^\pi(\mathbf{x}_B)$ as the permutations the model most and least prefers for this sequence.

Order-sensitivity targets. We measure per-sample order sensitivity by two complementary statistics.

Token disagreement Δ_{tok}. Let $\mathbf{y}^+$ and $\mathbf{y}^-$ be the forced greedy outputs under π^+ and π^-. Then

$$\Delta_{\text{tok}} = \tfrac{1}{4} \sum_{k=0}^{3} \mathbf{1}[y_k^+ \neq y_k^-] \in \{0,\ 0.25,\ 0.5,\ 0.75,\ 1\}.$$

Table 6: Order-sensitivity prediction on LAMBADA ($N = 5{,}153$, all $p < 10^{-3}$).

	Spearman ρ		AUC (top-20%)	
Predictor	Δ_{tok}	Δ_{NLL}	Δ_{tok}	Δ_{NLL}
$\bar{p}^{\max}$	-0.418	-0.148	0.695	0.563
$\bar{H}$	$+0.355$	$+0.152$	0.654	0.573
KL_{stab}	$+0.455$	$+0.198$	0.728	0.620
$\bar{H}^2_B$	$\mathbf{+0.507}$	$\mathbf{+0.438}$	0.700	**0.706**

82.8% of samples have $\Delta_{\text{tok}} > 0$, so ordering choice changes the output for the great majority of LAMBADA passages.

NLL gap under preferred ordering Δ_{NLL}. We score both outputs against π^+:

$$\Delta_{\text{NLL}} = \log Q_\theta^{\pi^+}(\mathbf{y}^+) - \log Q_\theta^{\pi^+}(\mathbf{y}^-).$$

This continuous target captures the magnitude—not merely the existence—of the ordering effect.

Predictors. All predictors are computed from the all-mask marginals $p_i = q_\theta(x_i \mid \mathbf{x}_{S=\emptyset})$ and are therefore independent of $\mathbf{x}_B$. This is a strict predictive setup: the predictor has no access to the reference sequence used to define the targets. We compare four position-averaged predictors:

(i) Peak confidence.

$\bar{p}^{\max} = \frac{1}{4}\sum_i \max_v p_i(v)$.

(ii) Marginal entropy.

$\bar{H} = -\frac{1}{4}\sum_i \sum_v p_i(v) \log p_i(v)$.

(iii) Pairwise marginal KL.

$$\text{KL}_{\text{stab}} = \tfrac{1}{B(B-1)} \sum_{i \neq j} \text{KL}\big(q_\theta(x_i \mid \emptyset) \,\big\|\, q_\theta(x_i \mid x_j)\big)\,,$$

measuring how much position-i's marginal shifts when position j is conditioned.

(iv) Block-averaged curl $\bar{H}^2_B$. For each pair (i, j), define pseudo-joints over a shared top-k ($k = 16$) support,

$$Q^{i\to j}(a, b) \propto p_i(a)\, q_\theta(x_j = b \mid x_i = a),$$
$$Q^{j\to i}(a, b) \propto p_j(b)\, q_\theta(x_i = a \mid x_j = b),$$

and average squared Hellinger over the $\binom{4}{2} = 6$ pairs:

$$\bar{H}^2_B = \tfrac{1}{6} \sum_{(i,j)} H^2(Q^{i\to j},\, Q^{j\to i}).$$

This is the block-level aggregation of the H^2 score from Section 6.2.

The first three predictors quantify marginal uncertainty; only $\bar{H}^2_B$ measures pseudo-joint asymmetry, which Theorem 1 identifies as the direct source of order-induced log-probability change.

Results. Table 6 reports Spearman rank correlations and AUC for identifying the top-20% most order-sensitive samples (all $p < 10^{-3}$); Figure 5 visualizes the comparison.

$\bar{H}^2_B$ leads on both targets, with the largest margin on Δ_{NLL}: $\rho = 0.438$ versus the next-best KL_{stab} at 0.198, and AUC 0.706 versus 0.620. To isolate $\bar{H}^2_B$'s unique contribution, we fit a rank regression using all three uncertainty baselines and measure the incremental R^2 from adding $\bar{H}^2_B$ (Table 7). The gain on Δ_{NLL} is $+0.148$, raising R^2 from 0.049 to 0.196—a $4\times$ increase.

Discussion. On Δ_{NLL}, the marginal uncertainty predictors ($\bar{p}^{\max}$, $\bar{H}$) have $|\rho| \leq 0.15$, KL_{stab} reaches $\rho = 0.198$, and $\bar{H}^2_B$ reaches $\rho = 0.438$. The first three statistics measure individual conditionals or their pairwise shifts; $\bar{H}^2_B$ measures the non-commutativity of their composition, which directly determines order-induced log-probability differences by Theorem 1.

Table 7: Incremental rank-R^2 from adding $\bar{H}_B^2$ to the three uncertainty baselines.

Target	Baselines R^2	$+\,\bar{H}_B^2$ R^2	Gain
Δ_{tok}	0.260	0.337	+0.077
Δ_{NLL}	0.049	0.196	+0.148

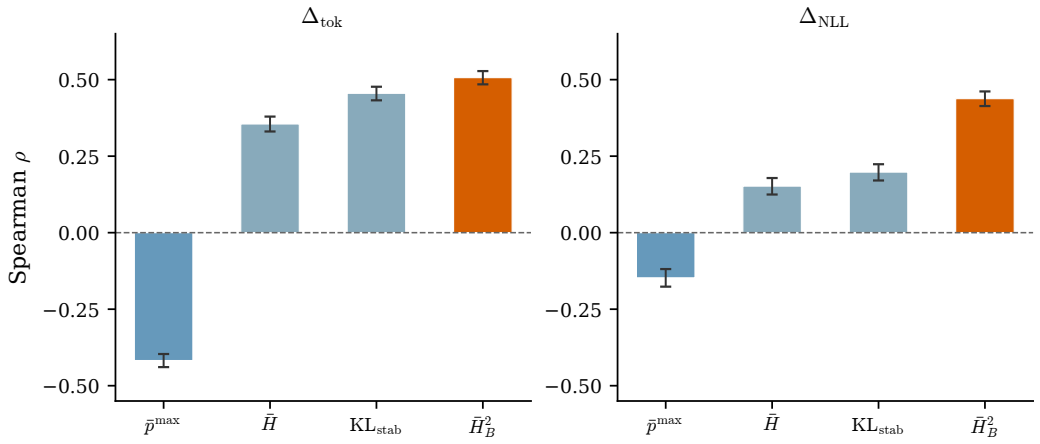

Figure 5: Spearman ρ for each predictor on Δ_{tok} and Δ_{NLL}. $\bar{H}_B^2$ leads on both; the advantage is largest on Δ_{NLL}, where baselines plateau below $\rho = 0.20$.

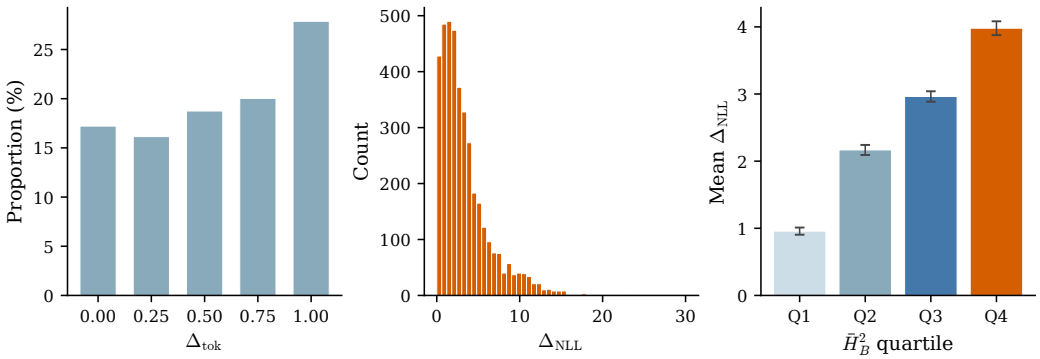

Figure 6: **Left:** distribution of Δ_{tok} (82.8% nonzero). **Center:** distribution of Δ_{NLL} among nonzero samples (right-skewed, mean = 2.52). **Right:** mean Δ_{NLL} by $\bar{H}_B^2$ quartile, showing a monotone increase from 1.0 (Q1) to 4.0 (Q4).

D Proofs and Technical Details

D.1 Constructive Separation Examples

This subsection gives the constructive witnesses for Proposition 1.

Dependent data with zero curl. Let $X_i, X_j \in \{0,1\}$ and suppose

$$p_{\mathrm{data}}(0,0 \mid x_S) = p_{\mathrm{data}}(1,1 \mid x_S) = 1/2.$$

Then $\mathrm{TC}(X_i, X_j \mid x_S) > 0$. If the denoiser is Bayes-optimal on all relevant partial contexts, both local orders recover the same data joint by the chain rule:

$$Q_{\theta,t}^{i\to j}(a,b \mid x_S) = p_{\mathrm{data}}(a,b \mid x_S) = Q_{\theta,t}^{j\to i}(a,b \mid x_S).$$

Hence $\mathcal{C}_{ij}^{a,b}(x_S,t) = 0$ for all a, b despite positive conditional total correlation.

Independent data with nonzero curl. Let X_i and X_j be independent fair bits under $p_{\mathrm{data}}(\cdot,\cdot \mid x_S)$, so $\mathrm{TC}(X_i, X_j \mid x_S) = 0$. Now consider a learned denoiser whose local predictions satisfy

$$q_{\theta,t}(x_i = 1 \mid x_S) = 1/2, \qquad q_{\theta,t}(x_j = 1 \mid x_S) = 1/2,$$
$$q_{\theta,t}(x_j = 1 \mid x_S, x_i = 1) = 0.9, \qquad q_{\theta,t}(x_i = 1 \mid x_S, x_j = 1) = 1/2.$$

Then for $a = b = 1$,

$$\mathcal{C}_{ij}^{1,1}(x_S,t) = \log \frac{(1/2)(0.9)}{(1/2)(1/2)} = \log 1.8 \neq 0.$$

Thus curl can be nonzero even when the data conditional total correlation is zero.

D.2 Proof of Theorem 1

Proof. By definition,

$$Q_{\theta,t}^{i\to j}(a,b \mid x_S) = q_{\theta,t}(a \mid x_S)\, q_{\theta,t}(b \mid x_S, x_i{=}a)$$

and

$$Q_{\theta,t}^{j\to i}(a,b \mid x_S) = q_{\theta,t}(b \mid x_S)\, q_{\theta,t}(a \mid x_S, x_j{=}b).$$

Taking the logarithm of the ratio gives

$$\begin{aligned}\log \frac{Q_{\theta,t}^{i\to j}(a,b \mid x_S)}{Q_{\theta,t}^{j\to i}(a,b \mid x_S)} &= \log q_{\theta,t}(a \mid x_S) + \log q_{\theta,t}(b \mid x_S, x_i{=}a) \\ &\quad - \log q_{\theta,t}(b \mid x_S) - \log q_{\theta,t}(a \mid x_S, x_j{=}b),\end{aligned}$$

which is exactly $\mathcal{C}_{ij}^{a,b}(x_S,t)$.

If $\mathcal{C}_{ij}^{a,b}(x_S,t) = 0$ for all a, b, then the displayed ratio equals 1 for all a, b, so

$$Q_{\theta,t}^{i\to j}(a,b \mid x_S) = Q_{\theta,t}^{j\to i}(a,b \mid x_S)$$

pointwise and hence the two distributions are equal. The converse is immediate from the same identity.

For the expectation claim,

$$\begin{aligned}&\mathbb{E}_{(a,b)\sim Q_{\theta,t}^{i\to j}(\cdot,\cdot \mid x_S)}\big[\mathcal{C}_{ij}^{a,b}(x_S,t)\big] \\ &\quad= \sum_{a,b} Q_{\theta,t}^{i\to j}(a,b \mid x_S) \log \frac{Q_{\theta,t}^{i\to j}(a,b \mid x_S)}{Q_{\theta,t}^{j\to i}(a,b \mid x_S)} \\ &\quad= \mathrm{KL}\big(Q_{\theta,t}^{i\to j}(\cdot,\cdot \mid x_S) \,\|\, Q_{\theta,t}^{j\to i}(\cdot,\cdot \mid x_S)\big),\end{aligned}$$

which is the definition of KL divergence. □

D.3 Discrete Exterior Derivative and Swap-Path Identities

Write $C_{ij}^{a,b}(x_S,t) \equiv \mathcal{C}_{ij}^{a,b}(x_S,t)$ for notational compactness. Define on the partial-context lattice the discrete 1-form

$$\omega_i^a(x_S,t) := \log q_{\theta,t}(x_i = a \mid x_S),$$

where (i,a,x_S) ranges over a position not in S, a candidate token, and a partial context. Its discrete exterior derivative on elementary squares is

$$(d\omega)_{ij}^{a,b}(x_S,t) := \omega_i^a(x_S,t) + \omega_j^b(x_S,x_i{=}a,t) - \omega_j^b(x_S,t) - \omega_i^a(x_S,x_j{=}b,t).$$

Theorem 3. *For every x_S, t, i, j, a, b,*

$$C_{ij}^{a,b}(x_S,t) = (d\omega)_{ij}^{a,b}(x_S,t).$$

The boundary of an elementary context square consists of the two commitment paths $i \to j$ and $j \to i$. Evaluating the log-conditional 1-form ω along this boundary gives the four terms in $C_{ij}^{a,b}$, and hence $C = d\omega$.

Corollary 3.1 (Path independence)**.** *For any two adjacent-swap paths Γ_1, Γ_2 from permutation π to permutation π', the sums*

$$\sum_{r\in\Gamma_1} C_{i_r j_r}^{x_{i_r},x_{j_r}}(x_{S_r^{(1)}},t) = \sum_{r\in\Gamma_2} C_{i_r j_r}^{x_{i_r},x_{j_r}}(x_{S_r^{(2)}},t) = \log \frac{Q_{\theta,t}^{\pi}(x_B \mid x_S)}{Q_{\theta,t}^{\pi'}(x_B \mid x_S)}$$

coincide.

Proof. Theorem 3 writes each adjacent-swap contribution as a finite difference of log-conditionals. Summing those differences along any adjacent-swap path telescopes to the endpoint pseudo-joint log-ratio $\log Q_{\theta,t}^{\pi} - \log Q_{\theta,t}^{\pi'}$. □

Corollary 3.2 (Path-specific adjacent-swap decomposition)**.** *Let π and π' be two permutations of B, and fix an adjacent-swap path*

$$\Gamma: \qquad \rho^{(0)} = \pi \to \rho^{(1)} \to \cdots \to \rho^{(R)} = \pi'.$$

Suppose the rth step swaps adjacent positions k_r and k_r+1 in the current ordering $\rho^{(r-1)}$. Write

$$i_r = \rho_{k_r}^{(r-1)}, \qquad j_r = \rho_{k_r+1}^{(r-1)},$$

and let

$$S_r = S \cup \{\rho_1^{(r-1)}, \ldots, \rho_{k_r-1}^{(r-1)}\}$$

be the prefix index set before the swapped adjacent pair. Then for every assignment x_B,

$$\log \frac{Q_{\theta,t}^{\pi}(x_B \mid x_S)}{Q_{\theta,t}^{\pi'}(x_B \mid x_S)} = \sum_{r=1}^{R} \mathcal{C}_{i_r j_r}^{x_{i_r},x_{j_r}}(x_{S_r},t).$$

Proof. For each step r, let

$$A_r = (\rho_1^{(r-1)}, \ldots, \rho_{k_r-1}^{(r-1)}) \qquad \text{and} \qquad D_r = (\rho_{k_r+2}^{(r-1)}, \ldots, \rho_{|B|}^{(r-1)}),$$

so that

$$\rho^{(r-1)} = (A_r, i_r, j_r, D_r) \qquad \text{and} \qquad \rho^{(r)} = (A_r, j_r, i_r, D_r).$$

Also write $C_r = S \cup A_r$, which is exactly the prefix index set S_r in the statement. Expanding the sequential products along the two adjacent orders gives

$$Q_{\theta,t}^{\rho^{(r-1)}}(x_B \mid x_S) = Q_{A_r}(x_{A_r} \mid x_S)\, q_{\theta,t}(x_{i_r} \mid x_{C_r}) \cdot q_{\theta,t}(x_{j_r} \mid x_{C_r}, x_{i_r}) \cdot Q_{D_r}(x_{D_r} \mid x_{C_r}, x_{i_r}, x_{j_r}),$$

and

$$Q_{\theta,t}^{\rho^{(r)}}(x_B \mid x_S) = Q_{A_r}(x_{A_r} \mid x_S)\, q_{\theta,t}(x_{j_r} \mid x_{C_r}) \cdot q_{\theta,t}(x_{i_r} \mid x_{C_r}, x_{j_r}) \cdot Q_{D_r}(x_{D_r} \mid x_{C_r}, x_{i_r}, x_{j_r}).$$

The prefix factor Q_{A_r} is identical in both expressions, and the suffix factor Q_{D_r} is also identical because after the adjacent swap both i_r and j_r have entered the conditioning context. Therefore,

$$\begin{aligned}\log \frac{Q_{\theta,t}^{\rho^{(r-1)}}(x_B \mid x_S)}{Q_{\theta,t}^{\rho^{(r)}}(x_B \mid x_S)} &= \log q_{\theta,t}(x_{i_r} \mid x_{C_r}) + \log q_{\theta,t}(x_{j_r} \mid x_{C_r}, x_{i_r}) \\ &\quad - \log q_{\theta,t}(x_{j_r} \mid x_{C_r}) - \log q_{\theta,t}(x_{i_r} \mid x_{C_r}, x_{j_r}) \\ &= \mathcal{C}_{i_r j_r}^{x_{i_r}, x_{j_r}}(x_{C_r}, t).\end{aligned}$$

Summing these consecutive log-ratios along the path Γ telescopes:

$$\begin{aligned}\log \frac{Q_{\theta,t}^{\pi}(x_B \mid x_S)}{Q_{\theta,t}^{\pi'}(x_B \mid x_S)} &= \sum_{r=1}^{R} \log \frac{Q_{\theta,t}^{\rho^{(r-1)}}(x_B \mid x_S)}{Q_{\theta,t}^{\rho^{(r)}}(x_B \mid x_S)} \\ &= \sum_{r=1}^{R} \mathcal{C}_{i_r j_r}^{x_{i_r}, x_{j_r}}(x_{S_r}, t),\end{aligned}$$

which is the claimed decomposition. □

D.4 Bayes-Optimal Compatibility and Conditional Total Correlation

Bayes-optimal compatibility implies zero curl. Suppose the denoiser satisfies

$$q_{\theta,t}(x_i \mid x_S) = p_{\mathrm{data}}(x_i \mid x_S)$$

for every visible context x_S and unresolved coordinate i. Then for every x_S, t, i, j, a, b,

$$Q_{\theta,t}^{i \to j}(a, b \mid x_S) = p_{\mathrm{data}}(a, b \mid x_S) = Q_{\theta,t}^{j \to i}(a, b \mid x_S),$$

and therefore $\mathcal{C}_{ij}^{a,b}(x_S, t) = 0$.

Proof. Fix x_S, t, i, j, a, and b. By the Bayes-optimal assumption,

$$q_{\theta,t}(a \mid x_S) = p_{\mathrm{data}}(a \mid x_S)$$

and, because x_S together with $x_i = a$ is also a visible context for coordinate j,

$$q_{\theta,t}(b \mid x_S, x_i{=}a) = p_{\mathrm{data}}(b \mid x_S, x_i{=}a).$$

Therefore,

$$\begin{aligned}Q_{\theta,t}^{i \to j}(a, b \mid x_S) &= q_{\theta,t}(a \mid x_S) q_{\theta,t}(b \mid x_S, x_i{=}a) \\ &= p_{\mathrm{data}}(a \mid x_S) p_{\mathrm{data}}(b \mid x_S, x_i{=}a) \\ &= p_{\mathrm{data}}(a, b \mid x_S),\end{aligned}$$

where the last step is the chain rule. The same argument with i and j swapped gives

$$\begin{aligned}Q_{\theta,t}^{j \to i}(a, b \mid x_S) &= q_{\theta,t}(b \mid x_S) q_{\theta,t}(a \mid x_S, x_j{=}b) \\ &= p_{\mathrm{data}}(b \mid x_S) p_{\mathrm{data}}(a \mid x_S, x_j{=}b) \\ &= p_{\mathrm{data}}(a, b \mid x_S).\end{aligned}$$

Hence

$$Q_{\theta,t}^{i \to j}(a, b \mid x_S) = p_{\mathrm{data}}(a, b \mid x_S) = Q_{\theta,t}^{j \to i}(a, b \mid x_S).$$

Applying Theorem 1 then gives $\mathcal{C}_{ij}^{a,b}(x_S, t) = 0$. □

Parallel decoding and conditional total correlation. For completeness, define the one-shot independent parallel approximation

$$Q_{\theta,t}^{\mathrm{par}}(x_B \mid x_S) = \prod_{i \in B} q_{\theta,t}(x_i \mid x_S).$$

If $q_{\theta,t} = p_{\mathrm{data}}$, then

$$\mathrm{KL}\Big(p(x_B \mid x_S) \,\Big\|\, Q_{\theta,t}^{\mathrm{par}}(x_B \mid x_S)\Big) = \mathrm{TC}(X_B \mid x_S),$$

which is exactly the conditional-total-correlation identity established by Kang et al. (2026); Chen et al. (2025).

D.5 Proof of Theorem 2

Use the parallel pseudo-joint as a common reference. Its KL error separates into conditional total correlation and marginal estimation error:

$$\mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\mathrm{par}}_{\theta,t}(x_B \mid x_S)\big) = \mathbb{E}_p\left[\log \frac{p(x_B \mid x_S)}{\prod_{i\in B} q_{\theta,t}(x_i \mid x_S)}\right],$$

insert $\log \prod_{i\in B} p(x_i \mid x_S)$ to split into

$$\mathbb{E}_p\left[\log \frac{p(x_B \mid x_S)}{\prod_{i\in B} p(x_i \mid x_S)}\right] + \mathbb{E}_p\left[\log \frac{\prod_{i\in B} p(x_i \mid x_S)}{\prod_{i\in B} q_{\theta,t}(x_i \mid x_S)}\right].$$

The first term is $\mathrm{TC}(X_B \mid x_S)$; the second is $\sum_{i\in B} \mathrm{KL}\big(p(x_i \mid x_S) \,\|\, q_{\theta,t}(x_i \mid x_S)\big)$. This proves

$$\mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\mathrm{par}}_{\theta,t}(x_B \mid x_S)\big) = \mathrm{TC}(X_B \mid x_S) + \sum_{i\in B} \mathrm{KL}\big(p(x_i \mid x_S) \,\|\, q_{\theta,t}(x_i \mid x_S)\big).$$

For sequential order π, decompose

$$\begin{aligned}
&\mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\pi}_{\theta,t}(x_B \mid x_S)\big) \\
&\quad = \mathrm{KL}\big(p(x_B \mid x_S) \,\|\, Q^{\mathrm{par}}_{\theta,t}(x_B \mid x_S)\big) \\
&\qquad + \mathbb{E}_p\big[\log Q^{\mathrm{par}}_{\theta,t}(x_B \mid x_S) - \log Q^{\pi}_{\theta,t}(x_B \mid x_S)\big].
\end{aligned}$$

Expanding the second term and using the definition of $\mathcal{I}^{\pi}_{\theta,t}(x_S)$ gives

$$\begin{aligned}
&\mathbb{E}_p\big[\log Q^{\mathrm{par}}_{\theta,t}(x_B \mid x_S) - \log Q^{\pi}_{\theta,t}(x_B \mid x_S)\big] \\
&\quad = -\sum_{m=1}^{|B|} \mathbb{E}_p\left[\log \frac{q_{\theta,t}(x_{\pi_m} \mid x_S, x_{\pi_{<m}})}{q_{\theta,t}(x_{\pi_m} \mid x_S)}\right] \\
&\quad = -\mathcal{I}^{\pi}_{\theta,t}(x_S).
\end{aligned}$$

Combining gives Theorem 2.

Corollary 2.1 follows by taking the difference for two orders π, π', noting that the total-correlation and marginal-estimation terms cancel, and applying Corollary 3.2 to

$$\mathcal{I}^{\pi}_{\theta,t} - \mathcal{I}^{\pi'}_{\theta,t} = \mathbb{E}_p\big[\log Q^{\pi}_{\theta,t} - \log Q^{\pi'}_{\theta,t}\big].$$

D.6 Cumulative Cross-Entropy Decomposition

For completeness, we record the rollout-weighted KL decomposition used in the proof of Theorem 2.

Lemma 1. *For any order π,*

$$\mathbb{E}_p\big[-\log Q^{\pi}_{\theta,t}(X_B \mid X_S)\big] = H_p(X_B \mid X_S) + \mathrm{KL}\big(p(X_B \mid X_S) \,\|\, Q^{\pi}_{\theta,t}(X_B \mid X_S)\big).$$

Moreover,

$$\begin{aligned}
&\mathrm{KL}\big(p(X_B \mid X_S) \,\|\, Q^{\pi}_{\theta,t}(X_B \mid X_S)\big) \\
&= \sum_{m=1}^{|B|} \mathbb{E}_p\Big[\mathrm{KL}\big(p(X_{\pi_m} \mid X_S, X_{\pi_{<m}}) \,\|\, q_{\theta,t}(X_{\pi_m} \mid X_S, X_{\pi_{<m}})\big)\Big].
\end{aligned}$$

Proof. For the first identity,

$$\begin{aligned}
&\mathrm{KL}\big(p(X_B \mid X_S) \,\|\, Q^{\pi}_{\theta,t}(X_B \mid X_S)\big) \\
&\quad = \mathbb{E}_p\big[\log p(X_B \mid X_S) - \log Q^{\pi}_{\theta,t}(X_B \mid X_S)\big].
\end{aligned}$$

Rearranging gives

$$\begin{aligned}
\mathbb{E}_p\big[-\log Q^{\pi}_{\theta,t}(X_B \mid X_S)\big] &= -\mathbb{E}_p\big[\log p(X_B \mid X_S)\big] + \mathrm{KL}\big(p(X_B \mid X_S) \,\|\, Q^{\pi}_{\theta,t}(X_B \mid X_S)\big) \\
&= H_p(X_B \mid X_S) + \mathrm{KL}\big(p(X_B \mid X_S) \,\|\, Q^{\pi}_{\theta,t}(X_B \mid X_S)\big),
\end{aligned}$$

which is the first claim.

For the second claim, apply the chain rule to both the target conditional and the pseudo-joint in the same order π:

$$p(X_B \mid X_S) = \prod_{m=1}^{|B|} p(X_{\pi_m} \mid X_S, X_{\pi_{<m}})$$

and

$$Q^{\pi}_{\theta,t}(X_B \mid X_S) = \prod_{m=1}^{|B|} q_{\theta,t}(X_{\pi_m} \mid X_S, X_{\pi_{<m}}).$$

Therefore,

$$\begin{aligned}
&\mathrm{KL}\big(p(X_B \mid X_S) \,\|\, Q^{\pi}_{\theta,t}(X_B \mid X_S)\big) \\
&= \mathbb{E}_p\left[\sum_{m=1}^{|B|} \log \frac{p(X_{\pi_m} \mid X_S, X_{\pi_{<m}})}{q_{\theta,t}(X_{\pi_m} \mid X_S, X_{\pi_{<m}})}\right] \\
&= \sum_{m=1}^{|B|} \mathbb{E}_p\left[\log \frac{p(X_{\pi_m} \mid X_S, X_{\pi_{<m}})}{q_{\theta,t}(X_{\pi_m} \mid X_S, X_{\pi_{<m}})}\right].
\end{aligned}$$

Now condition on $(X_S, X_{\pi_{<m}})$ and apply the definition of KL divergence inside each term:

$$\begin{aligned}
&\mathbb{E}_p\left[\log \frac{p(X_{\pi_m} \mid X_S, X_{\pi_{<m}})}{q_{\theta,t}(X_{\pi_m} \mid X_S, X_{\pi_{<m}})}\right] \\
&= \mathbb{E}_p\Big[\mathrm{KL}\big(p(X_{\pi_m} \mid X_S, X_{\pi_{<m}}) \,\|\, q_{\theta,t}(X_{\pi_m} \mid X_S, X_{\pi_{<m}})\big)\Big].
\end{aligned}$$

Summing over m yields the claimed decomposition. □

D.7 Pinsker-Type Bound on Bounded Order Statistics

Theorem 4. *For any bounded function $f : \mathcal{V}^{|B|} \to \mathbb{R}$ and any pair of orders π, π' of B,*

$$\big|\mathbb{E}_{Q^{\pi}_{\theta,t}}[f] - \mathbb{E}_{Q^{\pi'}_{\theta,t}}[f]\big| \le \|f\|_\infty \sqrt{2\mathrm{KL}\big(Q^{\pi}_{\theta,t} \,\|\, Q^{\pi'}_{\theta,t}\big)}.$$

By Corollary 3.2,

$$\mathrm{KL}\big(Q^{\pi}_{\theta,t} \,\|\, Q^{\pi'}_{\theta,t}\big) = \sum_r \mathbb{E}_{Q^{\pi}_{\theta,t}}\big[\mathcal{C}^{x_{i_r}, x_{j_r}}_{i_r j_r}(x_{S_r}, t)\big].$$

Hence

$$\big|\mathbb{E}_{Q^{\pi}_{\theta,t}}[f] - \mathbb{E}_{Q^{\pi'}_{\theta,t}}[f]\big| \le \|f\|_\infty \sqrt{2 \sum_r \mathbb{E}_{Q^{\pi}_{\theta,t}}\big[\mathcal{C}^{x_{i_r}, x_{j_r}}_{i_r j_r}(x_{S_r}, t)\big]}.$$

Proof. Apply Pinsker's inequality to $Q^{\pi}_{\theta,t}$ and $Q^{\pi'}_{\theta,t}$:

$$\big|\mathbb{E}_{Q^{\pi}_{\theta,t}}[f] - \mathbb{E}_{Q^{\pi'}_{\theta,t}}[f]\big| \le \|f\|_\infty \sqrt{2\mathrm{KL}\big(Q^{\pi}_{\theta,t} \,\|\, Q^{\pi'}_{\theta,t}\big)}.$$

By Theorem 1 and Corollary 3.2,

$$\mathrm{KL}\big(Q^{\pi}_{\theta,t} \,\|\, Q^{\pi'}_{\theta,t}\big) = \mathbb{E}_{Q^{\pi}_{\theta,t}}\left[\log \frac{Q^{\pi}_{\theta,t}}{Q^{\pi'}_{\theta,t}}\right] = \sum_r \mathbb{E}_{Q^{\pi}_{\theta,t}}\big[\mathcal{C}^{x_{i_r}, x_{j_r}}_{i_r j_r}(x_{S_r}, t)\big].$$

Substituting gives the claim. □

D.8 Fisher–Rao Energy Details

Proposition 2. *For every x_S, t, i, j,*

$$\Omega_{ij}^{\mathrm{FR}}(x_S, t) = 0 \iff Q_{\theta,t}^{i\to j} = Q_{\theta,t}^{j\to i} \iff \mathcal{C}_{ij}^{a,b}(x_S, t) = 0 \quad \forall a, b.$$

Proof. By definition,

$$\Omega_{ij}^{\mathrm{FR}}(x_S, t) = d_{\mathrm{FR}}^2\Big(Q_{\theta,t}^{i\to j}(\cdot, \cdot \mid x_S), Q_{\theta,t}^{j\to i}(\cdot, \cdot \mid x_S)\Big).$$

Since d_{FR} is a metric on the discrete probability simplex, $\Omega_{ij}^{\mathrm{FR}}(x_S, t) = 0$ holds exactly when

$$Q_{\theta,t}^{i\to j}(\cdot, \cdot \mid x_S) = Q_{\theta,t}^{j\to i}(\cdot, \cdot \mid x_S).$$

Theorem 1 gives the second equivalence: equality of the two local pseudo-joints is equivalent to $\mathcal{C}_{ij}^{a,b}(x_S, t) = 0$ for every a, b. □

Proposition 3. *Fix a shared support $\mathcal{S} \subseteq \mathcal{V} \times \mathcal{V}$. Let*

$$P = Q_{\theta,t}^{i\to j}, \qquad R = Q_{\theta,t}^{j\to i} = P + \delta,$$

where the sums below are over $\mathcal{S}$, $\sum_{(a,b)\in\mathcal{S}} \delta_{ab} = 0$, $P_{ab} \geq m > 0$, and $R_{ab} > 0$ on $\mathcal{S}$. For top-k use, both pseudo-joints are first restricted to a shared support and renormalized on that support. Then

$$d_{\mathrm{FR}}^2(P, P + \delta) = \sum_{a,b} \frac{\delta_{ab}^2}{P_{ab}} + O_m(\|\delta\|^3),$$

and

$$\mathrm{KL}(P \| P + \delta) = \frac{1}{2} \sum_{a,b} \frac{\delta_{ab}^2}{P_{ab}} + O_m(\|\delta\|^3).$$

Consequently,

$$\Omega_{ij}^{\mathrm{FR}} = 2\mathrm{KL}\big(Q_{\theta,t}^{i\to j} \,\|\, Q_{\theta,t}^{j\to i}\big) + O_m(\|\delta\|^3) = 2\mathbb{E}_{Q_{\theta,t}^{i\to j}}\big[\mathcal{C}_{ij}^{a,b}\big] + O_m(\|\delta\|^3).$$

Proof sketch. Write

$$\rho(P, P + \delta) := \sum_{a,b} \sqrt{P_{ab}(P_{ab} + \delta_{ab})}.$$

Using

$$\sqrt{P_{ab}(P_{ab} + \delta_{ab})} = P_{ab}\sqrt{1 + \delta_{ab}/P_{ab}} = P_{ab} + \tfrac{1}{2}\delta_{ab} - \tfrac{1}{8}\frac{\delta_{ab}^2}{P_{ab}} + O_m(\|\delta\|^3),$$

and summing over (a, b) gives

$$\rho(P, P + \delta) = 1 - \tfrac{1}{8} \sum_{a,b} \frac{\delta_{ab}^2}{P_{ab}} + O_m(\|\delta\|^3),$$

because $\sum_{a,b} P_{ab} = 1$ and $\sum_{a,b} \delta_{ab} = 0$. For small u, the expansion

$$4 \arccos^2(1 - u) = 8u + O(u^2)$$

then yields

$$d_{\mathrm{FR}}^2(P, P + \delta) = \sum_{a,b} \frac{\delta_{ab}^2}{P_{ab}} + O_m(\|\delta\|^3).$$

For the KL term,

$$\mathrm{KL}(P \| P + \delta) = -\sum_{a,b} P_{ab} \log\left(1 + \frac{\delta_{ab}}{P_{ab}}\right) = \frac{1}{2} \sum_{a,b} \frac{\delta_{ab}^2}{P_{ab}} + O_m(\|\delta\|^3),$$

using the Taylor series of $\log(1+x)$ and $\sum_{a,b} \delta_{ab} = 0$. Substituting $P = Q_{\theta,t}^{i\to j}$ and $P + \delta = Q_{\theta,t}^{j\to i}$ gives

$$\Omega_{ij}^{\mathrm{FR}} = 2\mathrm{KL}\big(Q_{\theta,t}^{i\to j} \,\|\, Q_{\theta,t}^{j\to i}\big) + O_m(\|\delta\|^3).$$

Finally, Theorem 1 implies

$$\mathrm{KL}\big(Q_{\theta,t}^{i\to j} \,\|\, Q_{\theta,t}^{j\to i}\big) = \mathbb{E}_{Q_{\theta,t}^{i\to j}}\big[\mathcal{C}_{ij}^{a,b}\big],$$

which gives the last identity. □