FAST PARALLEL SAMPLING FOR MASKED DIFFUSION MODELS VIA DEPENDENCE PARTITIONS PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Masked diffusion models enable parallel generation of discrete data, but simultaneously revealing multiple tokens can introduce error by neglecting their conditional dependencies. Recent methods address this challenge through adaptive unmasking schedules and dependency-aware token selection. However, explicit dependence checks can incur additional model evaluations, while inexpensive dependency estimates need not provide rigorous guarantees on sampling accuracy. In this work, we develop a theoretical framework for sampling with partition prediction. A partition oracle groups masked coordinates into bounded-size blocks, which our sampler processes in parallel while performing successive conditional updates within each block. Under quantitative assumptions on score estimation and partition quality, we establish nonasymptotic KL error bounds that separate initialization error, score estimation error, and residual cross-block dependence and quantify how the block size controls the trade-off between sampling accuracy and parallel query depth. We further construct partition oracles for ferromagnetic Ising models with bounded susceptibility, and masked language-model predictors with stable neural network.

1 INTRODUCTION

Masked diffusion models have emerged as an important class of generative models for discrete data, with applications spanning large-scale language modeling (Nie et al., 2025), code generation (Gong et al., 2025), image generation (Shi et al., 2024), protein design (Wang et al., 2024), molecular design for drug discovery (Lee et al., 2025), and sampling in statistical physics (Zhu et al., 2025). These models learn to reverse a forward process that progressively replaces data tokens with a special mask symbol, generating samples by iteratively predicting and revealing masked tokens. By allowing multiple tokens to be revealed in parallel, this framework can reduce the sequential depth of generation. However, sampling multiple tokens independently from their conditional marginals within the same step neglects their conditional dependencies and can introduce sampling error. A central challenge is therefore to exploit parallelism while controlling the approximation error.

Recent methods exploit additional information to adapt the parallel sampling computation. Token confidence and estimated dependencies guide the selection of parallel updates (Azangulov et al., 2025), auxiliary autoregressive models help determine how many tokens to accept (Israel et al., 2025), and temporal prediction stability identifies opportunities to avoid redundant computation (Oba et al., 2026). These approaches motivate studying how the quality of auxiliary information affects the trade-off between computational cost and sampling accuracy. We focus on information about token dependencies and ask the following question:

How can predicted dependence structure reduce parallel sampling depth?

We address this question by treating predicted partitions as an additional input to masked diffusion sampling. Our contributions are as follows.

1. **A framework for sampling with partition predictions.** We formalize the soundness for partition prediction oracles through bounds on block size and cross-block conditional information (Assumption 2.2). We then design a sampler that uses these predictions and standard score estimation (Algorithm 1). By updating blocks in parallel and coordinates within each block sequentially, the sampler strikes an algorithmic balance between fully parallel and fully sequential

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.

updates. This sampler fits within the broader theme of algorithms-with-predictions, with predicted partitions supplied as an additional input.

2. **Sampling accuracy and parallel query depth.** Under quantitative assumptions on score estimation and partition quality, we establish nonasymptotic KL error bounds that separately account for initialization, score estimation, and residual cross-block dependence (Theorem 1). Together with bounds on parallel query depth in terms of the block-size schedule, these results quantify the trade-off between sampling accuracy and parallel computation (Corollary 1), with provable gains for an Ising example and practical implications for masked language models.
3. **Computable partition-oracle constructions.** To justify our partition-oracle assumptions, we explicitly construct partition oracles for ferromagnetic Ising models and masked language-model predictors using weighted graphs and a fixed merge hierarchy. For Ising models, zero-field susceptibility bounds control cross-block dependence under arbitrary finite external fields (Theorem 2). For language-model predictors, computable sensitivity bounds from network parameters, combined with score accuracy, control conditional information (Theorem 3).

2 Masked Diffusion Models and Partition Prediction

2.1 Masked diffusion models

Notation. Fix integers $d \geq 1$ and $S \geq 2$. Let $\mathcal{V} \coloneqq [S]$ be the token alphabet, let $\texttt{MASK} \notin \mathcal{V}$ be a distinguished mask token, and set $\overline{\mathcal{V}} \coloneqq \mathcal{V} \cup \{\texttt{MASK}\}$. The clean-data space is $\mathcal{X}_0 \coloneqq \mathcal{V}^d$, and the state space of the diffusion is $\overline{\mathcal{V}}^{\,d}$. For $x \in \overline{\mathcal{V}}^{\,d}$, let $U(x) \coloneqq \{i \in [d] : x^i \neq \texttt{MASK}\}$, $M(x) \coloneqq [d] \setminus U(x)$, and $m(x) \coloneqq |M(x)|$. For $A \subseteq [d]$, write $x^A = (x^i)_{i \in A}$, and let $x \odot_i a$ denote the state obtained by replacing coordinate i of x by a, leaving all other coordinates unchanged.

Forward masking process. Let $X_0 \sim q_0 = q_{\mathrm{data}}$, where q_0 is an arbitrary distribution on $\mathcal{V}^d$. Independently of X_0, each coordinate is replaced by $\texttt{MASK}$ at rate one and remains masked thereafter (Campbell et al., 2022; Austin et al., 2021). Write τ for forward noise time and q_τ for the law of X_τ. Each coordinate survives with probability $e^{-\tau}$, independently of the others. Hence

$$q_\tau(x) = e^{-\tau|U(x)|}(1 - e^{-\tau})^{m(x)} q_0^{U(x)}(x^{U(x)}), \qquad \tau > 0, \tag{1}$$

where q_0^A denotes the marginal on A and $q_0^{\varnothing} = 1$.

Reverse process and discrete score. Fix $T > 0$ and let $(W_t)_{t \in [0,T]}$ be the right-continuous time reversal, with $W_t \sim q_{T-t}$. It starts from q_T and ends at q_0. For $q_\tau(x) > 0$ and $y \neq x$, define the discrete score $s_\tau(y, x) \coloneqq q_\tau(y)/q_\tau(x)$ (Lou et al., 2024). The reverse process only reveals masked coordinates. By equation 1, its transition rates satisfy

$$\overleftarrow{Q}_t(x, x \odot_i a) = s_\tau(x \odot_i a, x) = \frac{\Pr(X_0^i = a | X_\tau = x)}{e^\tau - 1}, \qquad \tau = T - t > 0, \tag{2}$$

for $i \in M(x)$ and $a \in \mathcal{V}$. Consequently, each masked coordinate is revealed at total rate $\sum_{a \in \mathcal{V}} s_\tau(x \odot_i a, x) = (e^\tau - 1)^{-1}$, with its token drawn from the current posterior. Let $\widehat{s}_\tau(y, x)$ estimate $s_\tau(y, x)$, and write $\widehat{s}_\tau(x) \coloneqq (\widehat{s}_\tau(x \odot_i a, x))_{(i,a) \in M(x) \times \mathcal{V}}$ for the estimated score vector. The associated token predictor takes (τ, x) as input and estimates $\Pr(X_0^i = a \mid X_\tau = x)$ by the normalized score $\widehat{s}_\tau(x \odot_i a, x) / \sum_{b \in \mathcal{V}} \widehat{s}_\tau(x \odot_i b, x)$, for each $i \in M(x)$ and $a \in \mathcal{V}$.

2.2 Assumptions: Score Estimation and Partition Prediction

Score estimation error. Let $\widehat{s}_\tau$ be a finite, nonnegative score estimator, with positive total estimated rate at each masked coordinate. Following Lou et al. (2024), we measure its accuracy through cost $\mathfrak{d}(u, v) \coloneqq v - u + u \log(\frac{u}{v})$, with $\mathfrak{d}(0, v) = v$ and $\mathfrak{d}(u, 0) = +\infty$ for $u > 0$.

Assumption 2.1 (**Score entropy error**). Let $0 = t_0 < \cdots < t_N = T$, with $h_k = t_{k+1} - t_k$ and $\tau_k = T - t_k$. Define

$$\mathcal{L}_{\mathrm{SE}}(\tau, \widehat{s}_\tau, s_\tau) \coloneqq \mathbb{E}_{Z \sim q_\tau} \sum_{i \in M(Z)} \sum_{a \in \mathcal{V}} \mathfrak{d}(s_\tau(Z \odot_i a, Z), \widehat{s}_\tau(Z \odot_i a, Z)) \,. \tag{3}$$

We assume that $\sum_{k=0}^{N-1} h_k \, \mathcal{L}_{\mathrm{SE}}(\tau_k, \widehat{s}_{\tau_k}, s_{\tau_k}) \leq \varepsilon_{\mathrm{score}} < \infty$.

Score entropy is nonnegative and vanishes at the exact score. Up to an estimator-independent constant, it equals a denoising objective evaluated using clean samples and their masked versions (Lou

et al., 2024, Theorem 3.4). Thus, the assumption specifies the population training accuracy required by our sampling guarantee. The same loss also arises in path-space KL comparisons of jump processes (Dmitriev et al., 2026).

This is the grid-averaged condition of Dmitriev et al. (2026, Assumption 1). Unlike the analysis of Liang et al. (2025, Assumption 2), we impose no additional uniform bound on the estimated scores. Accuracy is required only on average under the true masking marginals, without a pointwise approximation guarantee.

Partition prediction oracle. Motivated by conditional independence tests for parallel decoding (Azangulov et al., 2025) and information-theoretic analyses of parallel unmasking (Chen et al., 2026; Lavenant & Zanella, 2025; Dmitriev et al., 2026), we use a predicted partition to organize the remaining masked coordinates into blocks, with sequential updates within each block and parallel updates across blocks. The partition should retain strongly dependent coordinates within the same block while keeping the blocks small.

For $A \subseteq [d]$, let $\Pi(A)$ denote the collection of partitions of A into nonempty, pairwise disjoint blocks, with $\Pi(\varnothing) = \{\varnothing\}$. For states $x, y \in \overline{\mathcal{V}}^{\,d}$, write $x \preceq y$ if $y^i = x^i$ for every $i \in U(x)$.

At reverse time $0 \le t < T$, a *partition prediction oracle* takes as input the forward noise time $T-t$, the current state $x \in \overline{\mathcal{V}}^{\,d}$, and the score vector $\widehat{s}_{T-t}(x)$, and returns

$$\widehat{\mathcal{P}}_t(x) \coloneqq \mathsf{O}_{\mathrm{part}}\big(T-t, x, \widehat{s}_{T-t}(x)\big) \in \Pi(M(x)).$$

Here, $\widehat{s}_\tau(x)$ contains the score estimates indexed by $(i,a) \in M(x) \times \mathcal{V}$. For $i \in M(x)$, let $\widehat{B}_t(i,x)$ denote the unique block of $\widehat{\mathcal{P}}_t(x)$ containing i.

At a supported state z, $I(X_0^i; X_0^j \mid X_{T-t} = z)$ measures the expected KL change in the posterior of coordinate i after revealing coordinate j. Sequential updates can incorporate this information within each block, whereas parallel updates across blocks omit the other blocks' newly revealed values. We therefore define the *cross-block information rate*

$$\Gamma_{\mathrm{part}}(t) \coloneqq \tfrac{1}{(e^{T-t}-1)^2}\mathbb{E}_{Z\sim q_{T-t}}\Big[\textstyle\sum_{i\in M(Z)}\sum_{j\in M(Z)\setminus \widehat{B}_t(i,Z)} I\big(X_0^i; X_0^j \mid X_{T-t} = Z\big)\Big]. \quad (4)$$

The sum includes ordered pairs of masked coordinates in different blocks and already visible coordinates enter through the conditioning state Z. The expectation averages over both the observed context and its predicted partition, with the integrand set to zero on q_{T-t}-null states.

The normalization reflects the masking process. Writing $\tau = T-t$, two coordinates are simultaneously masked with probability $(1-e^{-\tau})^2$, independently of their clean values. Division by $(e^\tau - 1)^2$ converts this factor to $e^{-2\tau}$, the mutual-information attenuation factor for two independent erasure channels. Thus, the singleton partition recovers the masking information rate of Dmitriev et al. (2026), while general partitions retain only cross-block contributions. This rate differs from posterior block total correlation and we will bound the KL error caused by independent block updates through a weighted time integral of Γ_{part}. The following assumption bounds block size and cross-block information, and requires monotone refinement.

Assumption 2.2 ($(r, \varepsilon_{\mathrm{part}})$**-sound partition prediction oracle).** Let $r : [0,T) \to \{1,\dots,d\}$ and $\varepsilon_{\mathrm{part}} : [0,T) \to [0,\infty)$ be measurable functions satisfying $\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t < \infty$. The oracle is called $(r, \varepsilon_{\mathrm{part}})$-*sound* with respect to $\widehat{s}$ if:

1. Bounded block size. For every $t \in [0,T)$, $x \in \overline{\mathcal{V}}^{\,d}$, and $B \in \widehat{\mathcal{P}}_t(x)$, we have $|B| \le r(t)$.
2. Bounded cross-block information. For almost every $t \in [0,T)$, $\Gamma_{\mathrm{part}}(t) \le \varepsilon_{\mathrm{part}}(t)$.

3. Monotone refinement. For every $0 \le t_1 \le t_2 < T$ and $x, y \in \overline{\mathcal{V}}^{\,d}$ with $x \preceq y$, each $B' \in \widehat{\mathcal{P}}_{t_2}(y)$ is contained in some $B \in \widehat{\mathcal{P}}_{t_1}(x)$.

The first two conditions balance sequential computation against parallelization error. The block-size bound controls the number of sequential coordinate updates within each block, while $\varepsilon_{\mathrm{part}}$ bounds the cross-block information rate used in the error analysis. Grouping dependent coordinates within larger blocks can reduce this rate, at the cost of more sequential work.

The information bound is averaged over the true masking marginal, whereas the size and refinement conditions hold at every state, including partially updated contexts queried by the sampler. Monotone refinement permits blocks to split or shrink as time advances and observations accumulate,

but prevents previously separated coordinates from being merged. It therefore preserves cross-block separation along the unmasking process and keeps later blocks within bounded-size ancestor blocks.

3 Sampling with Partition Prediction

In this section, we introduce a sampler with partition prediction combining parallel updates across blocks with successive first-hitting updates within each block in Section 3.1. Section 3.2 then bounds its KL error, total score queries, and parallel query depth.

3.1 Algorithm

Algorithm 1 combines parallel updates across predicted blocks with successive conditional updates within each block. On each grid interval, a block incorporates its own newly revealed tokens while keeping the other blocks at their interval-start values. The block outputs are combined at the next grid point. Thus the partition determines which conditional updates are performed sequentially and which proceed in parallel. We initialize from $p_0 = \big((1-e^{-T})\delta_{\texttt{MASK}} + e^{-T}\,\mathrm{Unif}(\mathcal{V})\big)^{\otimes d}$ and write $\tau_k = T - t_k$. All sampling operations use fresh randomness, independently across blocks. Write $\tau_k = T - t_k$ and, for $k < N-1$, define $\Delta_k = (e^{\tau_k}-1)\log\frac{1-e^{-\tau_k}}{1-e^{-\tau_{k+1}}}$. Set $\Delta_{N-1} = +\infty$.

Algorithm 1 Inter-Block τ-Leaping with Intra-Block Successive First-Hitting

Input: Grid $0 = t_0 < \cdots < t_N = T$; scores $\widehat{s}_{\tau_k}$; partition oracle $\mathsf{O}_{\mathrm{part}}$.
Output: A sample in $\mathcal{V}^d$.

1: Sample $x_{t_0} \sim p_0$
2: **for** $k = 0, \ldots, N-1$ **do**
3: $\quad z \leftarrow x_{t_k}$; $x_{t_{k+1}} \leftarrow z$
4: $\quad$ **if** $M(z) = \varnothing$ **then return** z
5: $\quad$ Query $\widehat{s}_{\tau_k}(z)$; $\mathcal{P}_k \leftarrow \mathsf{O}_{\mathrm{part}}(\tau_k, z, \widehat{s}_{\tau_k}(z))$
6: $\quad$ **for all** $B \in \mathcal{P}_k$ **in parallel do**
7: $\quad\quad y_B \leftarrow z$; $u_B \leftarrow 0$
8: $\quad\quad$ **while** $B \cap M(y_B) \neq \varnothing$ **do**
9: $\quad\quad\quad \Lambda_B \leftarrow \sum_{i \in B \cap M(y_B)} \sum_{a \in \mathcal{V}} \widehat{s}_{\tau_k}(y_B \odot_i a, y_B)$
10: $\quad\quad\quad$ Draw $\xi_B \sim \mathrm{Exp}(\Lambda_B)$; $u_B \leftarrow u_B + \xi_B$
11: $\quad\quad\quad$ **if** $u_B \geq \Delta_k$ **then break** ▷ Stop evolving block B
12: $\quad\quad\quad$ Draw $(i, a) \in (B \cap M(y_B)) \times \mathcal{V}$ with probability $\frac{\widehat{s}_{\tau_k}(y_B \odot_i a, y_B)}{\Lambda_B}$
13: $\quad\quad\quad y_B \leftarrow y_B \odot_i a$
14: $\quad\quad\quad$ **if** $B \cap M(y_B) \neq \varnothing$ **then**
15: $\quad\quad\quad\quad$ Query $\widehat{s}_{\tau_k}(y_B)$
16: $\quad\quad x^i_{t_{k+1}} \leftarrow y^i_B$ for every $i \in B$
17: **return** x_{t_N}

After initialization from p_0 (Line 1), each reverse-time interval $[t_k, t_{k+1})$ begins with a shared query at x_{t_k} to obtain the partition and initial rates (Line 5). Blocks B evolve in parallel up to the rescaled horizon Δ_k, with coordinates outside each block held fixed (Lines 6–15). Within each block, masked coordinates are updated sequentially, with scores refreshed at the fixed time τ_k (Lines 8–15). Block outputs are merged at t_{k+1} (Line 16).

Remark 1 (**Algorithmic balance between full parallel and sequential updates**). *Algorithm 1 balances τ-leaping sampler (Dmitriev et al., 2026) and first-hitting sampler (Liang et al., 2026) through the choice of partition. Singleton blocks recover modified truncated τ-leaping and updating all masked coordinates independently conditional on the interval-start state. At the other extreme, a single block containing all masked coordinates yields successive first-hitting sampling. Intermediate partitions combine sequential conditioning within blocks with parallel updates across blocks.*

3.2 Theoretical Guarantee of Algorithm 1

We analyze Algorithm 1 and establish bounds on its sampling error, total score queries, and parallel query depth in terms of the score error, partition quality, and block-size budget.

Theorem 1 (**Accuracy and complexity**). *Suppose Assumptions 2.1 and 2.2 hold, and the time grid satisfies $r(t_k)h_k \le 1$ for every $k = 0, \dots, N-1$. Then Algorithm 1, initialized from p_0, terminates almost surely and returns a sample in $\mathcal{V}^d$. Its output law p_{alg} satisfies*

$$\mathrm{KL}(q_{\mathrm{data}} \| p_{\mathrm{alg}}) \lesssim \underbrace{de^{-T} \log S}_{\textit{initialization error}} + \underbrace{\varepsilon_{\mathrm{score}}}_{\textit{score estimation error}} + \underbrace{\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t}_{\textit{partition-wise parallelization error}}. \tag{5}$$

The algorithm terminates with parallel score-query depth at most $\sum_{k=0}^{N-1} r(t_k)$, using at most N partition-oracle queries and $N + d$ score queries.

The bound in equation 5 separates initialization, score, and partition errors. The initialization term decays exponentially with T, while the score term is controlled by score estimation accuracy. The third term captures the cost of inter-block parallelization and reflects the information dependence across blocks. Within each block, successive updates simulate the dynamics exactly and introduce no additional error. Given additional partition predictions, our algorithm achieves provable KL accuracy with parallel score-query depth controlled by block size and residual cross-block dependence.

To obtain explicit score-query depth, we define a certified upper bound on the *partition-wise effective total correlation* and its unweighted analogue as

$$\mathcal{D}_{\mathrm{part}} \coloneqq \int_0^T \min\{1, T-t\}\varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t \ge \int_0^T \min\{1, T-t\}\Gamma_{\mathrm{part}}(t)\,\mathrm{d}t,$$
$$\mathcal{B}_{\mathrm{part}} \coloneqq \int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t \ge \int_0^T \Gamma_{\mathrm{part}}(t)\,\mathrm{d}t.$$

We next give query-depth bounds for specific step-size schedules, with proof in Appendix C.

Corollary 1. *Fix $\varepsilon \in (0,1)$ and $T = \log(4d \log S/\varepsilon)$. Under the assumptions of Theorem 1 it obeys $\mathrm{KL}(q_{\mathrm{data}} \| p_{\mathrm{alg}}) \lesssim \varepsilon_{\mathrm{score}} + \varepsilon$, if*

1. ***Uniform step-size schedule.*** *$t_k - t_{k-1} = \frac{T}{N}$ for all $k = 1, \dots, N$, with $N = \widetilde{O}(r_{\max} + \frac{\mathcal{B}_{\mathrm{part}}}{\varepsilon})$.*
2. ***Exponential-then-constant step-size schedule.*** *$t_{k+1} - t_k \le \kappa \min\{1, T - t_{k+1}\}$ for $k = 0, \dots, N-2$, $T - t_{N-1} = \frac{\varepsilon}{d \log S}$, and $\kappa = \frac{T + \log(\varepsilon^{-1} d \log S)}{N}$, with $N = \widetilde{O}(r_{\max} + \frac{\mathcal{D}_{\mathrm{part}}}{\varepsilon})$.*

Corollary 1 quantifies how the block-size budget and cross-block information affect the query-depth of achieving a prescribed accuracy. Consider the score-query depth $\sum_{k=0}^{N-1} r(t_k)$. For a fixed block-size budget $r(t)$, partitions with lower certified cross-block information, quantified by $\mathcal{D}_{\mathrm{part}}$ or $\mathcal{B}_{\mathrm{part}}$, permit fewer synchronization intervals while preserving the same accuracy guarantee. However, these computational gains must be assessed jointly with the block-size budget. Merging blocks removes nonnegative terms from the cross-block information rate, but larger blocks can increase sequential work within each interval and restrict admissible step sizes. A useful partition must reduce this dependence sufficiently to compensate for the additional within-block computation.

Remark 2 (**Balancing full parallel and sequential updates**). *The partition choice balances the full parallelism of τ-leaping sampler, with query depth $\widetilde{O}\big(\frac{\mathcal{D}(q_{\mathrm{data}})}{\varepsilon}\big)$ (Dmitriev et al., 2026, Corollary 2), against the sequential first-hitting sampler with d query-depth (Liang et al., 2026, Section 4). Here, $\mathcal{D}(q_{\mathrm{data}})$ denotes the effective total correlation of the target distribution, aggregating conditional dependence over all coordinate pairs along the masking process.*

To illustrate Corollary 1 and the benefits of additional partition predictions, we present an Ising example and discuss implications for masked language models.

1. **Ising model.** Consider an Ising model on $d/2$ disjoint edges with fixed coupling $\beta > 0$. Its total, dual total, and effective total correlations are all $\Theta_\beta(d)$. Nevertheless, grouping the endpoints of each edge gives $r_{\max} = 2$ and $\mathcal{B}_{\mathrm{part}} = \mathcal{D}_{\mathrm{part}} = 0$. Even with exact scores and fixed constant accuracy $\varepsilon = \Theta(1)$, Corollary 1 yields parallel query depth $O(\log d)$, whereas modified truncated τ-leaping (Dmitriev et al., 2026) requires $\Omega_\beta(\sqrt{d})$ rounds and first-hitting sampling (Liang et al., 2026) uses d sequential queries. The proof is given in Appendix D. The partition oracle is constructed in Section 4.1.
2. **Masked language models.** Strong local token dependencies can limit the accuracy of independent parallel updates. ReFusion reports that token influence decreases with distance (Li et al., 2025, Appendix A.2), motivating sequential generation within short segments and parallel generation across segments, a strategy also adopted by CoDiLA (Hersche et al., 2026). These observations motivate partition predictions that group strongly dependent tokens into small blocks.

Our sampler handles their dependencies through successive conditioning within each block, leaving only cross-block information in the parallelization error. We further construct the partition prediction oracle for masked language models in Section 4.2.

4 Existence and Construction of Partition Oracles

Motivated by the preceding discussion, we answer this question by constructing partition prediction oracles satisfying Assumption 2.2 for ferromagnetic Ising models with bounded zero-field susceptibilities (Section 4.1) and masked language models with certified network stability and controlled prediction error (Section 4.2). For each application, we present the construction and analyze its soundness and computational cost.

4.1 Ferromagnetic Ising Sampling with External Fields

In this section, we study the ferromagnetic Ising model (Jerrum & Sinclair, 1993) on a finite graph $G = (V, E)$ with $V = [d]$, $m = |E|$ and maximum degree at most Δ:

$$q_0(\sigma) = \tfrac{1}{Z} \exp\Big(\textstyle\sum_{\{i,j\}\in E} J_{ij}\sigma_i\sigma_j + \sum_{i=1}^{d} h_i\sigma_i\Big), \qquad \sigma \in \{-1,+1\}^d, \tag{6}$$

Assume $0 \le J_{ij} \le J$ and where $J_{uv} \ge 0$ and $h_v \in \mathbb{R}$ are finite.

Conditioning on the visible spins gives an Ising model on $G[M(x)]$ with the same internal couplings and fields $h_v^x = h_v + \sum_{u\in U(x):\{u,v\}\in E} J_{uv}x_u$. Indeed, an edge with two unobserved endpoints remains a pair interaction, an edge with one observed endpoint contributes to an external field, and an edge with two observed endpoints contributes a constant (Lemma 12). Thus conditioning does not introduce new interaction edges. The posterior factors exactly over the connected components of $G[M(x)]$, but those components need not be small. Our construction subdivides them using fixed base groups and controls the dependence between the resulting blocks under the original posterior.

Construction of the partition oracle. We preprocess the interaction graph (G, J) once and use the resulting hierarchy to answer queries (t, x).

Offline preprocessing. Algorithm 2 starts with one leaf per vertex $v \in V$, with $S(v) = \{v\}$. It processes edges in decreasing order of J_e, with fixed tie-breaking. When an edge joins distinct components with roots z_1, z_2, it creates a parent z with these children and sets $S(z) = S(z_1)\cup S(z_2)$. Thus, $S(z)$ contains the descendant leaves of z, and $n(z) = |S(z)|$ is its size. We retain the resulting merge forest $\mathcal{H}$, its child pointers, and the original graph G.

For each budget $q \in \{1, \dots, d\}$, select nodes of size at most q whose parents, if present, have size greater than q. Their leaf sets form $\mathcal{C}_q$: each vertex belongs to the block of its highest ancestor of size at most q. These blocks partition V and refine as q decreases.

Online query. Fix a nonincreasing block-size budget schedule $r : [0, T) \to \{1, \dots, d\}$ before observing any query states. Given query (t, x), Algorithm 3 returns a partition $\widehat{\mathcal{P}}_t(x)$ by

$$\widehat{\mathcal{P}}_t(x) = \textstyle\bigcup_{C\in\mathcal{C}_{r(t)}} \mathrm{CC}\big(G[C \cap M(x)]\big), \tag{7}$$

where CC denotes the collection of connected components, with empty intersections contributing no blocks. Group selection uses full leaf counts before masking, so every output block has size at most $r(t)$. Decreasing the budget selects smaller tree nodes, while revealing coordinates deletes vertices. Both operations refine the partition. The query uses the original graph and depends on x only through $M(x)$, independently of any supplied score predictor. Figure 1 illustrates both stages.

Guarantees under susceptibility bounds. Let $\mu_{G,0}$ denote the Ising measure on G with the same couplings J and external field $h = 0$. This reference measure captures the collective dependence generated by the interactions alone. To control this dependence, we impose the following bounds on the zero-field susceptibilities.

Assumption 4.1 (Susceptibility bounds). The algorithm is given a vector $a \in [1, \infty)^V$ such that $\chi_v := \sum_{u\in V} \mathbb{E}_{\mu_{G,0}}[\sigma_u\sigma_v] \le a_v,$ for every $v \in V$.

Algorithm 2 Preprocess(G, J)

Input: Interaction graph (G, J)
Output: Original graph G and merge forest $\mathcal{H}$

Initialize $\mathcal{H}$ with singleton components
$n(v) \leftarrow 1$ for every $v \in V$
Order E by decreasing J_e with fixed tie-breaking
for each edge $\{u, v\}$ in this order **do**
    $a, b \leftarrow$ the corresponding merge-tree roots
    if $a \neq b$ **then**
        Add a node z with children a, b to $\mathcal{H}$
        $n(z) \leftarrow n(a) + n(b)$
        Merge the two components with root z
return $(G, \mathcal{H})$

Algorithm 3 Query$(G, \mathcal{H}, r; t, x)$

Input: Stored $(G, \mathcal{H}, r)$ and query (t, x)
Output: A partition $\widehat{\mathcal{P}}_t(x)$ of $M(x)$

$q \leftarrow r(t)$; $Q \leftarrow \mathrm{Roots}(\mathcal{H})$
$\mathcal{P} \leftarrow \varnothing$
while $Q \neq \varnothing$ **do**
    Remove a node z from Q
    if $n(z) \leq q$ **then**
        $A \leftarrow S(z) \cap M(x)$
        $\mathcal{P} \leftarrow \mathcal{P} \cup \mathrm{CC}(G[A])$
    else
        Add both children of z to Q
return $\mathcal{P}$

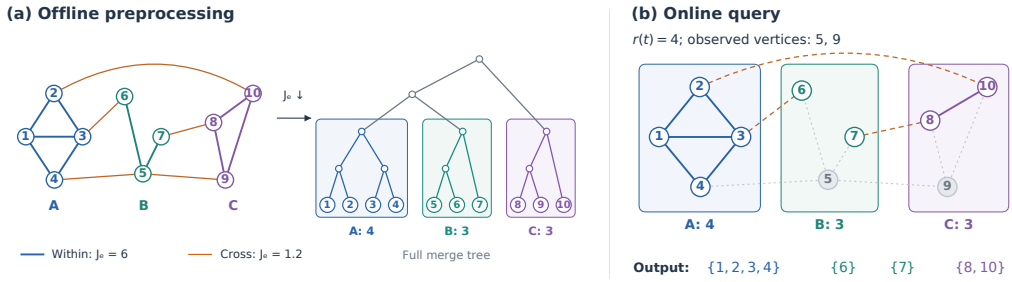

Figure 1: Partition-oracle construction. (a) The full merge tree produced by Algorithm 2, with original vertices as leaves and merges as internal nodes. The colored subtrees define $\mathcal{C}_4$ using full leaf counts. (b) At $r(t) = 4$, observing vertices $5, 9$ leaves connected components within each base group, yielding $\{1, 2, 3, 4\}, \{6\}, \{7\}, \{8, 10\}$. Orange dashed edges represent cross-group interactions that remain in the target model.

The bounds a_v can be obtained from standard high-temperature criteria or graph-dependent spectral estimates (Chen et al., 2023; Saade et al., 2017). Susceptibilities are exactly computable on forests, while size-independent bounds hold for finite free-boundary restrictions of a fixed subcritical transitive infinite model (Duminil-Copin & Tassion, 2016). By Griffiths monotonicity (Griffiths, 1967), the same a_v bound the zero-field susceptibilities of every induced subgraph. Moreover, Ding et al. (2023, Theorem 1.1) bounds single-spin influences under arbitrary finite external fields by the corresponding zero-field correlations. Since revealing spins only deletes vertices and changes the remaining fields, these bounds apply uniformly to all query states and yield the mutual-information estimates used in the partition-error certificate.

For each budget q, let $F(q)$ be the set of original edges joining distinct blocks of $\mathcal{C}_q$, and define the weighted cut cost as $b(q) := \sum_{\{u,v\} \in F(q)} a_u a_v \tanh J_{uv}$. The following theorem shows that our construction yields a sound partition oracle, with block sizes bounded by $r(t)$ and partition error controlled by the weighted cut cost $b(r(t))$.

Theorem 2 (Ferromagnetic Ising partition oracle). *Under Assumption 4.1, for arbitrary finite external fields and any schedule r as above, Algorithms 2 and 3 define a deterministic measurable partition oracle. For every (t, x), its blocks have size at most $r(t)$. Moreover, $\widehat{\mathcal{P}}_{t'}(y) \preceq \widehat{\mathcal{P}}_t(x)$ whenever $t \leq t'$ and $x \preceq y$, and*

$$\Gamma_{\mathrm{part}}(t) \leq \varepsilon_{\mathrm{part}}(t) := 2(\log 2) e^{-2(T-t)} b(r(t)), \tag{8}$$

where $\int_0^T \varepsilon_{\mathrm{part}}(t)\, \mathrm{d}t < \infty$.

Remark 3 (Complexity). *Writing $m = |E|$, preprocessing takes $O(d + m \log(m + 1))$ time and $O(d + m)$ space, and each query takes $O(d + m)$ time. Given the susceptibility bounds, any single budget $b(q)$ can also be evaluated in $O(d + m)$ time.*

The theorem controls the partition error through the total weight of the edges crossing between blocks, with the factors $a_u a_v$ accounting for the propagation of dependence through the graph. Since the partitions $\mathcal{C}_q$ are nested, increasing the block-size budget can only decrease $b(q)$, improving the

error certificate at the cost of more sequential updates within each block. In particular, if no edges cross between blocks, then $\Gamma_{\mathrm{part}}(t) = 0$. The full proof can be found in Appendix E.

Example 1 (High-temperature Ising on a square lattice). *Let G be the $L \times L$ grid with free boundary conditions, where L is a perfect square and $d = L^2$. Fix uniform couplings $J_e = \beta \geq 0$ with $\tanh \beta < \frac{1}{3}$, so that $\chi_v \lesssim_\beta 1$. Take $r(t) \equiv L = \sqrt{d}$ and choose tie-breaking so that $\mathcal{C}_L$ consists of $\sqrt{L} \times \sqrt{L}$ squares. Then $|F(L)| \lesssim d^{3/4}$, and Theorem 2 gives $\Gamma_{\mathrm{part}}(t) \lesssim_\beta e^{-2(T-t)} d^{3/4}$.*

4.2 Masked Language Models

We construct a partition oracle for masked diffusion language models (Sahoo et al., 2024; Nie et al., 2025) using sensitivity bounds computed from network parameters and a fixed merge hierarchy. Soundness also uses the score-error bound in Assumption 2.1.

Predictor and computable certificates. We consider a neural token predictor $\widehat{p}_{k,i}(\cdot \mid x)$, defined as the softmax of logits $\ell_{k,i}(\cdot; x)$ and used to estimate the posterior distribution of each masked token. At time τ_k, the predictor has an embedding $E : \overline{\mathcal{V}} \to \mathbb{R}^H$, hidden layers $F_{k,\ell} : (\mathbb{R}^H)^d \to (\mathbb{R}^H)^d$, $\ell \in [L]$, with $L \geq 1$, and an affine output head $\ell_{k,i}(a; x)$:

$$h_{k,i}^{(0)}(x) = E(x^i) + c_{k,i}, \quad h_k^{(\ell)}(x) = F_{k,\ell}(h_k^{(\ell-1)}(x)), \quad \ell_{k,i}(a; x) = \langle \theta_a, h_{k,i}^{(L)}(x) \rangle + \beta_a, \tag{9}$$

where $h_k^{(\ell)}(x) = (h_{k,i}^{(\ell)}(x))_{i \in [d]}$, $c_{k,i}, \theta_a \in \mathbb{R}^H$, $\beta_a \in \mathbb{R}$. For supported x and $i \in M(x)$, the score vector $(\widehat{s}_{\tau_k}(x \odot_i a, x))_{a \in \mathcal{V}}$ has finite, strictly positive entries and normalizes to the softmax prediction $\widehat{p}_{k,i}(\cdot \mid x)$. Define the maximum embedding change as $D_{\mathrm{emb}} = \max_{a \in \mathcal{V}} \|E(a) - E(\texttt{MASK})\|_2$ and the maximum norm of the output weights as $R_{\mathrm{out}} = \max_{a \in \mathcal{V}} \|\theta_a\|_2$.

Each hidden layer is a finite acyclic computation built from the operations specified in Appendix F. The complete architecture and parameters are fixed independently of x.

Assumption 4.2 (Network regularity). All network parameters are finite. Every scalar activation used in the hidden layers is globally Lipschitz with a known finite Lipschitz constant. Every LayerNorm or RMSNorm operation uses a known positive stabilizer.

Assumption 4.2 accommodates the Transformer backbone of Masked Diffusion Language Models (MDLM) and the original dense Large Language Diffusion with mAsking (LLaDA) predictor in inference mode (Sahoo et al., 2024; Nie et al., 2025). Their Gaussian error linear unit (GELU) and sigmoid linear unit (SiLU) components have bounded derivatives, and positive normalization stabilizers keep denominators bounded away from zero. For the specified computation graphs, Lemma 1 establishes computable sensitivity bounds on domains containing all reachable hidden states. We next quantify how perturbations propagate between token positions through each hidden layer. The following lemma provides computable sensitivity matrices that are valid uniformly over all grid times and masked inputs.

Lemma 1 (Computable network stability). *Under Assumption 4.2, finite nonnegative matrices $A_1, \ldots, A_L \in \mathbb{R}_+^{d \times d}$ can be computed without enumerating masked states such that*

$$\|h_{k,i}^{(\ell)}(x) - h_{k,i}^{(\ell)}(y)\|_2 \leq \textstyle\sum_{j=1}^{d} (A_\ell)_{ij} \|h_{k,j}^{(\ell-1)}(x) - h_{k,j}^{(\ell-1)}(y)\|_2 \tag{10}$$

for every k, $\ell \in [L]$, $i \in [d]$, and $x, y \in \overline{\mathcal{V}}^d$.

The matrices bound propagation of perturbations between positions with details in Appendix F.

Construction of the partition oracle. We preprocess the predictor's architecture and parameters once and use the resulting hierarchy to answer partition queries (t, x).

Offline preprocessing. We first compute the matrices $A_1, \ldots, A_L$ from Lemma 1, with

$$u_0 = v_L = \mathbf{1}, \qquad u_\ell = A_\ell u_{\ell-1}, \qquad v_{\ell-1} = A_\ell^\top v_\ell, \quad \ell \in [L]. \tag{11}$$

The forward and backward vectors account for sensitivity propagation before and after each layer. Set $\Omega_{ii} = 0$ and define the aggregate sensitivity weights by

$$\Omega_{ij} = 2 D_{\mathrm{emb}} R_{\mathrm{out}} \textstyle\sum_{\ell=1}^{L} (v_\ell)_i (A_\ell)_{ij} (u_{\ell-1})_j, \qquad i \neq j. \tag{12}$$

Furthermore, by Lemma 20, summing Ω_{ij} over pairs in different blocks gives an upper bound on the total sensitivity between blocks. Let G be the complete graph on $[d]$ with edge weights $J_{ij} = \Omega_{ij} + \Omega_{ji}$. We then apply Algorithm 2 to (G, J) to obtain the merge tree $\mathcal{H}$, and define its base partitions $(\mathcal{C}_q)_{q=1}^{d}$ as before.

Online query. Fix in advance a deterministic, nonincreasing, right-continuous block size schedule $r : [0, T) \to [d]$. Since the auxiliary graph G is complete, Algorithm 3 returns

$$\widehat{\mathcal{P}}_t(x) = \{C \cap M(x) : C \in \mathcal{C}_{r(t)}\} \setminus \{\varnothing\}. \tag{13}$$

Each nonempty intersection is connected and can be returned directly, without scanning edges. The query therefore uses only $\mathcal{H}$, $r(t)$, and $M(x)$. The block size and refinement guarantees follow as before. To bound the error, we define the weighted boundary b_{LM} as

$$b_{\mathrm{LM}}(q) = \sum_{\substack{i \neq j \\ C_q(i) \neq C_q(j)}} \Omega_{ij}, \tag{14}$$

where $C_q(i)$ is the block of $\mathcal{C}_q$ containing i.

Guarantee and computational cost. Let $\mathcal{L}_{\mathrm{SE},k}$ denote the grid-time score entropy loss in Assumption 2.1, so $\sum_{k=0}^{N-1} h_k \mathcal{L}_{\mathrm{SE},k} \le \varepsilon_{\mathrm{score}}$. Write the maximum grid step size as $h_{\max} = \max_k h_k$ and the maximum number of positions outside a base block at budget q as $\kappa(q) = \max_{C \in \mathcal{C}_q}(d - |C|)$.

Theorem 3 (**Partition oracle for masked language models**). *Under Assumptions 2.1 and 4.2, with the normalized-score parameterization above, Algorithms 2 and 3 define a deterministic measurable $(r, \varepsilon_{\mathrm{part}})$-sound oracle with respect to $\widehat{s}$, where, for $t \in [t_k, t_{k+1})$,*

$$\varepsilon_{\mathrm{part}}(t) = e^{-2(T-t)} b_{\mathrm{LM}}(r(t)) + \kappa(r(t)) e^{d(t-t_k)} \mathcal{L}_{\mathrm{SE},k}. \tag{15}$$

For $q_* = \min_{0 \le t < T} r(t)$,

$$\int_0^T \varepsilon_{\mathrm{part}}(t)\, \mathrm{d}t \le \tfrac{1 - e^{-2T}}{2} b_{\mathrm{LM}}(q_*) + (d-1) e^{d h_{\max}} \varepsilon_{\mathrm{score}} < \infty. \tag{16}$$

Remark 4 (**Complexity**). *Preprocessing takes $O(|\mathcal{V}|H + T_{\mathrm{layer}} + Ld^2 + d^2 \log(d+1))$ operations, where T_{layer} is the cost of computing $A_1, \ldots, A_L$. Each partition query takes $O(d)$ operations plus evaluation of $r(t)$.*

The certificate separates cross-group prediction sensitivity from score approximation error. The structural term is computable from the network parameters. The full proof is given in Appendix F. Corollary 2 gives sufficient conditions for a small structural term through bounds on cross-group sensitivity and its amplification.

Remark 5 (**Limitations of partition construction for mask diffusion models**). *Revealing tokens can change the conditional dependence among the remaining masked positions. Recent samplers exploit this context dependence through online non-influence tests in PUNT (Azangulov et al., 2025) and attention-based dependency graphs in DAPD (Kim et al., 2026). These methods adapt their parallel update sets to the observed context. Our construction accommodates changing dependencies through sensitivity bounds that hold uniformly over masked contexts. However, these bounds may be conservative, and the fixed hierarchy restricts adaptation since monotone refinement prevents previously separated blocks from being merged when new dependencies emerge.*

5 Conclusion and Future Work

We developed a theoretical framework for masked diffusion sampling with partition prediction. Our KL error bounds quantify how partition quality and block size govern the trade-off between sampling accuracy and parallel query depth, while separating initialization, score estimation, and cross-block dependence errors. These results yield provable gains for an Ising example. We further gave explicit partition oracle constructions for ferromagnetic Ising models and masked language-model predictors under quantitative structural assumptions.

Future work could develop adaptive partition oracles that allow both splitting and merging while retaining provable KL guarantees. Another natural direction is to extend our framework to other forms of auxiliary prediction, such as unmasking priorities, joint token proposals, and the stability of previously computed predictions. A central question is how prediction quality translates into guarantees on sampling accuracy and computational cost, and whether these guarantees can remain robust when predictions are inaccurate. Evaluating such methods on trained models would further clarify their practical benefits.

AI USE STATEMENT

In this work, we used generative AI tools to help formulate and refine mathematical statements and assist with the development and writing of proofs. Additionally, we used these tools to identify and summarize related literature and improve the clarity and organization of the manuscript. The authors reviewed all AI-assisted content, checked the mathematical arguments, and verified the accuracy and relevance of the cited references. We take full responsibility for the final content of this work.

REFERENCES

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 34: Annual Conference on Neural Information Processing Systems 2021*, pp. 17981–17993, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/958c530554f78bcd8e97125b70e6973d-Abstract.html. (Cited on page 2.)

Iskander Azangulov, Teodora Pandeva, Niranjani Prasad, Javier Zazo, and Sushrut Karmalkar. Parallel Sampling from Masked Diffusion Models via Conditional Independence Testing. *CoRR*, abs/2510.21961, 2025. doi: 10.48550/ARXIV.2510.21961. URL https://doi.org/10.48550/arXiv.2510.21961. (Cited on pages 1, 3, and 9.)

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet. A Continuous Time Framework for Discrete Denoising Models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b5b528767aa35f5b1a60fe0aaeca0563-Abstract-Conference.html. (Cited on page 2.)

Valérie Castin, Pierre Ablin, and Gabriel Peyré. How Smooth Is Attention? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5817–5840. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/castin24a.html. (Cited on page 37.)

Sitan Chen, Kevin Cong, and Jerry Li. Optimal Inference Schedules for Masked Diffusion Models. In *The Thirty Ninth Annual Conference on Learning Theory, June 29 - July 3, 2026, San Diego, California, USA*, volume 336 of *Proceedings of Machine Learning Research*, pp. 1279–1311. PMLR, 2026. URL https://proceedings.mlr.press/v336/chen26e.html. (Cited on page 3.)

Zongchen Chen, Kuikui Liu, and Eric Vigoda. Rapid Mixing of Glauber Dynamics up to Uniqueness via Contraction. *SIAM J. Comput.*, 52(1):196–237, 2023. doi: 10.1137/20M136685X. URL https://doi.org/10.1137/20m136685x. (Cited on page 7.)

Jian Ding, Jian Song, and Rongfeng Sun. A new correlation inequality for Ising models with external fields. *Probability Theory and Related Fields*, 186(1):477–492, 2023. URL https://link.springer.com/article/10.1007/s00440-022-01132-1. (Cited on pages 7 and 29.)

Daniil Dmitriev, Zhihan Huang, and Yuting Wei. Efficient Sampling with Discrete Diffusion Models: Sharp and Adaptive Guarantees. In Steve Hanneke and Tor Lattimore (eds.), *The Thirty Ninth Annual Conference on Learning Theory, June 29 - July 3, 2026, San Diego, California, USA*, volume 336 of *Proceedings of Machine Learning Research*, pp. 2038–2104. PMLR, 2026. URL https://proceedings.mlr.press/v336/dmitriev26a.html. (Cited on pages 3, 4, 5, 14, and 24.)

Hugo Duminil-Copin and Vincent Tassion. A New Proof of the Sharpness of the Phase Transition for Bernoulli Percolation and the Ising Model. *Communications in Mathematical Physics*, 343(2):725–745, 2016. URL https://link.springer.com/article/10.1007/s00220-015-2480-z. (Cited on page 7.)

Cees M Fortuin, Pieter W Kasteleyn, and Jean Ginibre. Correlation inequalities on some partially ordered sets. *Communications in Mathematical Physics*, 22(2):89–103, 1971. URL https://link.springer.com/article/10.1007/bf01651330. (Cited on page 29.)

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. DiffuCoder: Understanding and Improving Masked Diffusion Models for Code Generation. *CoRR*, abs/2506.20639, 2025. doi: 10.48550/ARXIV.2506.20639. URL https://doi.org/10.48550/arXiv.2506.20639. (Cited on page 1.)

Robert B Griffiths. Correlations in Ising Ferromagnets. II. External Magnetic Fields . *Journal of Mathematical Physics*, 8(3):484–489, 1967. URL https://doi.org/10.1063/1.1705220. (Cited on pages 7 and 31.)

Michael Hersche, Nicolas Menet, Ronan Tanios, and Abbas Rahimi. Locally coherent parallel decoding in diffusion language models. *CoRR*, abs/2603.20216, 2026. URL https://doi.org/10.48550/arXiv.2603.20216. (Cited on page 5.)

Daniel Israel, Guy Van den Broeck, and Aditya Grover. Accelerating Diffusion LLMs via Adaptive Parallel Decoding. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/4c7d31b2da05a17f15101db5b37a1e24-Abstract-Conference.html. (Cited on page 1.)

Mark Jerrum and Alistair Sinclair. Polynomial-Time Approximation Algorithms for the Ising Model. *SIAM J. Comput.*, 22(5):1087–1116, 1993. doi: 10.1137/0222066. URL https://doi.org/10.1137/0222066. (Cited on page 6.)

Bumjun Kim, Dongjae Jeon, Moongyu Jeon, and Albert No. Dependency-Aware Parallel Decoding via Attention for Diffusion LLMs. *CoRR*, abs/2603.12996, 2026. URL https://doi.org/10.48550/arXiv.2603.12996. (Cited on page 9.)

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The Lipschitz Constant of Self-Attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571. PMLR, 2021. URL http://proceedings.mlr.press/v139/kim21i.html. (Cited on page 37.)

Hugo Lavenant and Giacomo Zanella. Error Bounds and Optimal Schedules for Masked Diffusions with Factorized Approximations. *CoRR*, abs/2510.25544, 2025. doi: 10.48550/ARXIV.2510.25544. URL https://doi.org/10.48550/arXiv.2510.25544. (Cited on page 3.)

Seul Lee, Karsten Kreis, Srimukh Prasad Veccham, Meng Liu, Danny Reidenbach, Yuxing Peng, Saee Gopal Paliwal, Weili Nie, and Arash Vahdat. GenMol: A Drug Discovery Generalist with Discrete Diffusion. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/lee25o.html. (Cited on page 1.)

Christian Léonard. Girsanov Theory Under a Finite Entropy Condition. In *Séminaire de Probabilités XLIV*, pp. 429–465. Springer, 2012. URL https://link.springer.com/chapter/10.1007/978-3-642-27461-9_20. (Cited on page 14.)

Jia-Nan Li, Jian Guan, Wei Wu, and Chongxuan Li. Refusion: A diffusion large language model with parallel autoregressive decoding. *CoRR*, abs/2512.13586, 2025. URL https://doi.org/10.48550/arXiv.2512.13586. (Cited on page 5.)

Yuchen Liang, Renxiang Huang, Lifeng Lai, Ness Shroff, and Yingbin Liang. Absorb and Converge: Provable Convergence Guarantee for Absorbing Discrete Diffusion Models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/1d571ea833394630c1dec71664f16cd9-Abstract-Conference.html. (Cited on pages 3 and 25.)

Yuchen Liang, Zhiheng Tan, Ness B. Shroff, and Yingbin Liang. Sharp convergence rates for masked diffusion models. *CoRR*, abs/2602.22505, 2026. doi: 10.48550/ARXIV.2602.22505. URL https://doi.org/10.48550/arXiv.2602.22505. (Cited on pages 4, 5, and 25.)

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution. In *Forty-first International Conference on Machine Learning, ICML 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/lou24a.html. (Cited on page 2.)

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 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/48b383b24230e0e6e649d9c98dae4d8c-Abstract-Conference.html. (Cited on pages 1, 8, and 37.)

Daisuke Oba, Danushka Bollegala, Masahiro Kaneko, and Naoaki Okazaki. Stopping Computation for Converged Tokens in Masked Diffusion-LM Decoding. *CoRR*, abs/2602.06412, 2026. doi: 10.48550/ARXIV.2602.06412. URL https://doi.org/10.48550/arXiv.2602.06412. (Cited on pages 1 and 38.)

Alaa Saade, Florent Krzakala, and Lenka Zdeborová. Spectral bounds for the Ising ferromagnet on an arbitrary given graph. *Journal of Statistical Mechanics: Theory and Experiment*, 2017(5):053403, 2017. URL https://iopscience.iop.org/article/10.1088/1742-5468/aa6f1e/meta?casa_token=V2PQRnJ_8wEAAAAA:KfqUbAsm03id5YJ5_yWxQ97hdmZ0kR6T-aH1FuAhmXR4ZlHQGC-6gwBWGUo-P2FoFPvGPlX_vnyUhkC1WRrmJQgiiU8r. (Cited on page 7.)

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. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/eb0b13cc515724ab8015bc978fdde0ad-Abstract-Conference.html. (Cited on pages 8 and 37.)

Noam Shazeer. GLU Variants Improve Transformer. *CoRR*, abs/2002.05202, 2020. URL https://arxiv.org/abs/2002.05202. (Cited on page 37.)

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis K. Titsias. Simplified and Generalized Masked Diffusion for Discrete Data. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/bad233b9849f019aead5e5cc60cef70f-Abstract-Conference.html. (Cited on page 1.)

Zhouxing Shi, Huan Zhang, Kai-Wei Chang, Minlie Huang, and Cho-Jui Hsieh. Robustness Verification for Transformers. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJxwPJHFwS. (Cited on page 37.)

Zhouxing Shi, Yihan Wang, Huan Zhang, J. Zico Kolter, and Cho-Jui Hsieh. Efficiently Computing Local Lipschitz Constants of Neural Networks via Bound Propagation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/0ff54b4ec4f70b3ae12c8621ca8a49f4-Abstract-Conference.html. (Cited on page 37.)

Jianlin Su, Murtadha H. M. Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/J.NEUCOM.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063. (Cited on page 37.)

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html. (Cited on page 37.)

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 3839–3848, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/d54e99a6c03704e95e6965532dec148b-Abstract.html. (Cited on page 37.)

Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion Language Models Are Versatile Protein Learners. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 52309–52333. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/wang24ct.html. (Cited on page 1.)

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/0cbc5671ae26f67871cb914d81ef8fc1-Abstract.html. (Cited on page 37.)

Nikolay Yudin, Alexander Gaponov, Sergei Kudriashov, and Maxim V. Rakhuba. Pay Attention to Attention Distribution: A New Local Lipschitz Bound for Transformers. *CoRR*, abs/2507.07814, 2025. doi: 10.48550/ARXIV.2507.07814. URL https://doi.org/10.48550/arXiv.2507.07814. (Cited on page 37.)

Biao Zhang and Rico Sennrich. Root Mean Square Layer Normalization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12360–12371, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/1e8a19426224ca89e83cef47f1e7f53b-Abstract.html. (Cited on page 37.)

Huan Zhang, Pengchuan Zhang, and Cho-Jui Hsieh. RecurJac: An Efficient Recursive Algorithm for Bounding Jacobian Matrix of Neural Networks and Its Applications. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 5757–5764. AAAI Press, 2019. doi: 10.1609/AAAI.V33I01.33015757. URL https://doi.org/10.1609/aaai.v33i01.33015757. (Cited on page 37.)

Yuchen Zhu, Wei Guo, Jaemoo Choi, Guan-Horng Liu, Yongxin Chen, and Molei Tao. MDNS: Masked Diffusion Neural Sampler via Stochastic Optimal Control. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/3289cc9cb9fdde172004497c61359544-Abstract-Conference.html. (Cited on page 1.)

A TECHNICAL LEMMAS

Lemma 2 (Chain rule of KL divergence ((Dmitriev et al., 2026, Lemma 4))). *For $N > 0$, let $a_{0:N}$ and $b_{0:N}$ be the joint distributions of two Markov processes. Then,*

$$\mathrm{KL}(a_{0:N} \| b_{0:N}) = \mathrm{KL}(a_0 \| b_0) + \sum_{k=0}^{N-1} \mathbb{E}_{x \sim a_k} \mathrm{KL}\big(a_{k+1|k}(\cdot \mid x) \,\big\|\, b_{k+1|k}(\cdot \mid x)\big)\,.$$

Lemma 3 (Derivative of KL: an upper bound (Dmitriev et al., 2026, Lemma 5)). *Let (q_t) and (p_t) be the marginals of CTMCs with rate matrices (Q_t) and $(\widehat{Q}_t)$, respectively, and let $\overleftarrow{q}_t \equiv q_{T-t}$ be the marginals of the reverse process, with rate matrix $\overleftarrow{Q}_t$. Then, for $t > t_k$ and any supported initial state z,*

$$\begin{aligned} &\frac{\partial}{\partial t} \mathrm{KL}\big(\overleftarrow{q}_{t|t_k}(\cdot \mid z) \,\big\|\, p_{t|t_k}(\cdot \mid z)\big) \\ &\quad \le \mathbb{E}_{x \sim \overleftarrow{q}_{t|t_k}(\cdot|z)} \sum_{y \neq x} \left[\widehat{Q}_t(x,y) - \overleftarrow{Q}_t(x,y) + \overleftarrow{Q}_t(x,y) \log \frac{\overleftarrow{Q}_t(x,y)}{\widehat{Q}_t(x,y)} \right]. \end{aligned} \tag{17}$$

For measurable time-dependent rates, the derivative inequality is understood almost everywhere.

Lemma 4 (Itô's lemma for a Poisson jump process(Dmitriev et al., 2026, Lemma 6)). *Let $(x_t)_{t \in [0,b]}$ be a finite-state CTMC on $\mathcal{X}$, with measurable, bounded transition rates $R_t(x,y)$ and generator*

$$(L_t g)(x) \coloneqq \sum_{y \neq x} R_t(x,y)\big(g(y) - g(x)\big).$$

Let $f : [0,b] \times \mathcal{X} \to \mathbb{R}$ be continuously differentiable in time at each state. Then, for every $t \in [0,b]$,

$$f(t, x_t) = f(0, x_0) + \int_0^t \big[\partial_s f(s, x_{s-}) + \big(L_s f(s,\cdot)\big)(x_{s-})\big]\, \mathrm{d}s + M_t, \tag{18}$$

where $x_{s-} = \lim_{u \uparrow s} x_u$ and $(M_t)_{t \in [0,b]}$ is an integrable martingale with $M_0 = 0$.

More explicitly, for distinct states $x, y \in \mathcal{X}$, let $N_t^{x,y}$ count transitions from x to y during $(0,t]$. Its predictable intensity is

$$\lambda_s^{x,y} = \mathbf{1}_{\{x_{s-}=x\}} R_s(x,y).$$

The martingale in equation 18 is

$$M_t = \sum_{x \neq y} \int_0^t \big(f(s,y) - f(s,x)\big)\,(\mathrm{d}N_s^{x,y} - \lambda_s^{x,y}\,\mathrm{d}s)\,. \tag{19}$$

Lemma 5 (Path-space KL identity for CTMCs (Léonard, 2012, Theorem 2.9)). *Let $\mathbb{P}$ and $\widehat{\mathbb{P}}$ be the path laws of two CTMCs on a finite state space $\mathcal{X}$ over $[0,b]$, where $0 < b < \infty$. Suppose that both processes start from the same state z and have measurable, bounded transition rates $R_t(x,y)$ and $\widehat{R}_t(x,y)$, respectively. If $\mathrm{KL}(\mathbb{P} \| \widehat{\mathbb{P}}) < \infty$, then*

$$\mathrm{KL}(\mathbb{P} \| \widehat{\mathbb{P}}) = \int_0^b \mathbb{E}_{\mathbb{P}} \left[\sum_{y \neq X_t} \mathfrak{d}\big(R_t(X_t,y), \widehat{R}_t(X_t,y)\big) \right] \mathrm{d}t, \tag{20}$$

where (X_t) is the canonical coordinate process and $\mathfrak{d}(v,w) = v\log(v/w) - v + w$ for $v, w > 0$, with $\mathfrak{d}(0,w) = w$ for $w \ge 0$ and $\mathfrak{d}(v,0) = +\infty$ for $v > 0$.

B FULL PROOF OF THEOREM 1

Throughout, $X_0 \sim q_{\mathrm{data}}$ denotes the clean data. The score estimator is fixed, and its induced partition map is deterministic and measurable. All coordinate rates are zero after the coordinate has been revealed.

Let $E_1, \dots, E_d \overset{\text{iid}}{\sim} \text{Exp}(1)$ be independent of X_0. The right-continuous exact reverse process is

$$\begin{aligned} \overleftarrow{X}_t^i &= \begin{cases} X_0^i, & E_i \geq T-t, \\ \text{MASK}, & E_i < T-t, \end{cases} && \text{with } \text{Law}(\overleftarrow{X}_t) = q_{T-t}, \\ \theta_{s,t} &:= \mathbb{P}(\overleftarrow{X}_t^i = \text{MASK} \mid \overleftarrow{X}_s^i = \text{MASK}) = \frac{1-e^{t-T}}{1-e^{s-T}}, && 0 \leq s \leq t < T. \end{aligned} \tag{21}$$

At every deterministic time, $\overleftarrow{X}_t = X_{T-t}$ almost surely. For $\tau > 0$ and a supported partial state x, define $\pi_i^x(a) = \mathbb{P}(X_0^i = a \mid X_0^{U(x)} = x^{U(x)})$. Then

$$\begin{aligned} q_\tau(x) &= e^{-\tau|U(x)|}(1-e^{-\tau})^{|M(x)|}\mathbb{P}(X_0^{U(x)} = x^{U(x)}), \\ q_\tau(x) > 0 &\iff \mathbb{P}(X_0^{U(x)} = x^{U(x)}) > 0. \end{aligned} \tag{22}$$

In particular, support does not depend on $\tau > 0$. For $i \in M(x)$,

$$\begin{aligned} s_\tau(x \odot_i a, x) &= \frac{q_\tau(x \odot_i a)}{q_\tau(x)} = \frac{e^{-\tau}}{1-e^{-\tau}} \frac{\mathbb{P}(X_0^{U(x)} = x^{U(x)}, X_0^i = a)}{\mathbb{P}(X_0^{U(x)} = x^{U(x)})} = \frac{\pi_i^x(a)}{e^\tau - 1}, \\ \sum_{a \in \mathcal{V}} s_\tau(x \odot_i a, x) &= \frac{1}{e^\tau - 1}. \end{aligned} \tag{23}$$

B.1 CTMC Representation and KL Decomposition

Lemma 6 (Block-frozen CTMC representation). *Let $\widehat{X}$ denote the process generated by Algorithm 1, with finite nonnegative score estimates, and write $\tau_k = T - t_k$. Conditional on $\widehat{X}_{t_k} = z$, the algorithm evolves on $[t_k, t_{k+1})$ as a time-inhomogeneous CTMC. For a current state $x \succeq z$ and $i \in M(x)$, put $B = \widehat{B}_{t_k}(i, z)$ and $y = (x_B, z_{-B})$. The rate of revealing coordinate i as $a \in \mathcal{V}$ is*

$$\widehat{\lambda}_{k,t}^{i,a}(z, x) = \rho_k(t)\widehat{s}_{\tau_k}(y \odot_i a, y), \qquad \rho_k(t) = \frac{e^{\tau_k} - 1}{e^{T-t} - 1}. \tag{24}$$

For $i \notin M(x)$, this rate is zero.

Proof. Fix k and condition on $\widehat{X}_{t_k} = z$. The CTMC with rates equation 24 decomposes into independent processes on the blocks of $\widehat{\mathcal{P}}_{t_k}(z)$. Fix such a block B and a current state x at time t. Write $y = (x_B, z_{-B})$ and $\Lambda_B(y) = \sum_{i \in B \cap M(y)} \sum_{a \in \mathcal{V}} \widehat{s}_{\tau_k}(y \odot_i a, y)$.

First, suppose $\Lambda_B(y) > 0$. The probability that this block makes no jump before $t' \in [t, t_{k+1})$ is

$$\exp\left(-\Lambda_B(y) \int_t^{t'} \rho_k(v)\,\mathrm{d}v\right) = \exp\left(-\Lambda_B(y)\big(u_k(t') - u_k(t)\big)\right),$$

where $u_k(t) = \int_{t_k}^t \rho_k(v)\,\mathrm{d}v = (e^{\tau_k} - 1)\log\frac{1-e^{-\tau_k}}{1-e^{t-T}}$. This matches the exponential waiting time used by the algorithm in its internal clock. Next, since all outgoing rates share the factor $\rho_k(t)$, conditional on a jump, its coordinate and value (i, a) have probability

$$\frac{\widehat{s}_{\tau_k}(y \odot_i a, y)}{\Lambda_B(y)},$$

which also matches the algorithm. If $\Lambda_B(y) = 0$, both constructions leave the block unchanged.

After each jump, both constructions update y and repeat the same rule, with at most $|B|$ jumps. Moreover, $u_k(t_{k+1}) = \Delta_k$ for $k < N-1$, while $u_{N-1}(t) \to \infty$ as $t \uparrow T$. Thus, placing each internal-time event u at real time $u_k^{-1}(u)$ gives the same block process in both constructions. Independence across blocks then establishes the claimed conditional CTMC representation. □

Lemma 7 (Marginal KL comparison). *Suppose Assumption 2.1 holds with $\varepsilon_{\text{score}} < \infty$. Let $\widehat{X}$ be the algorithmic process from Lemma 6, initialized from p_0, and write $p_b = \mathcal{L}(\widehat{X}_b)$. Let $\overleftarrow{X}$ be the*

exact reverse process, with $\overleftarrow{X}_t \sim q_{T-t}$ *and rates* $\lambda_t^{*,i,a}(x) = s_{T-t}(x \odot_i a, x)$ *for* $i \in M(x)$*, and zero otherwise. Then, for every* $b \in (t_{N-1}, T)$*,*

$$\mathrm{KL}(q_{T-b}\|p_b) \le \mathrm{KL}(q_T\|p_0) + \sum_{k=0}^{N-1} \int_{t_k}^{\min\{t_{k+1},b\}} \mathbb{E}\left[\sum_{i=1}^{d}\sum_{a\in\mathcal{V}} \mathfrak{d}\Big(\lambda_t^{*,i,a}(\overleftarrow{X}_t), \widehat{\lambda}_{k,t}^{i,a}(\overleftarrow{X}_{t_k}, \overleftarrow{X}_t)\Big)\right] \mathrm{d}t. \tag{25}$$

The expectation is taken over the exact reverse process, with both rate functions evaluated along its path.

Proof. Fix $b \in (t_{N-1}, T)$ and put $v_k = \min\{t_{k+1}, b\}$. We may assume $\mathrm{KL}(q_T\|p_0) < \infty$, since otherwise the claim is immediate.

The exact and algorithmic states sampled at $t_0, \dots, t_{N-1}, b$ form Markov chains, since the algorithm refreshes its frozen context at each grid point. By data processing and Lemma 2,

$$\mathrm{KL}(q_{T-b}\|p_b) \le \mathrm{KL}(q_T\|p_0) + \sum_{k=0}^{N-1} \mathbb{E}_{Z\sim q_{\tau_k}} \mathrm{KL}\big(\overleftarrow{q}_{v_k|t_k}(\cdot \mid Z) \,\big\|\, p_{v_k|t_k}(\cdot \mid Z)\big).$$

Fix k and z with $q_{\tau_k}(z) > 0$. By Lemma 6, conditional on starting from z, both processes are CTMCs on $[t_k, v_k]$. Their rates are bounded because $b < T$ and the state space and estimated scores are finite. Moreover, if $\lambda_t^{*,i,a}(x) > 0$ at an exact-reachable state x, then $y = (x_B, z_{-B}) \preceq x$ implies $q_{\tau_k}(y) > 0$ and $s_{\tau_k}(y \odot_i a, y) > 0$. Indeed, every data configuration compatible with $x \odot_i a$ is also compatible with $y \odot_i a$. The finite score-error assumption therefore gives $\widehat{s}_{\tau_k}(y \odot_i a, y) > 0$. Thus, the learned rates on positive exact transitions are uniformly positive on this interval.

Both conditional laws start from δ_z. The boundedness and positivity above imply that their KL divergence tends to zero as $t \downarrow t_k$. Integrating Lemma 3 consequently yields

$$\mathrm{KL}\big(\overleftarrow{q}_{v_k|t_k}(\cdot \mid z) \,\big\|\, p_{v_k|t_k}(\cdot \mid z)\big) \le \int_{t_k}^{v_k} \mathbb{E}\left[\sum_{i=1}^{d}\sum_{a\in\mathcal{V}} \mathfrak{d}\Big(\lambda_t^{*,i,a}(\overleftarrow{X}_t), \widehat{\lambda}_{k,t}^{i,a}(z, \overleftarrow{X}_t)\Big) \,\middle|\, \overleftarrow{X}_{t_k} = z\right] \mathrm{d}t.$$

Averaging over $z \sim q_{\tau_k}$ and substituting into the preceding chain-rule bound proves equation 25. □

We now insert the exact block-frozen rate. For $Z = \overleftarrow{X}_{t_k}$ and $i \in M(Z)$, set $B_i = \widehat{B}_{t_k}(i, Z)$ and $Y_i = (\overleftarrow{X}_{t,B_i}, Z_{-B_i})$, and define

$$\lambda_{k,t}^{\mathrm{blk},i,a} := \rho_k(t) s_{\tau_k}(Y_i \odot_i a, Y_i) = \frac{\pi_i^{Y_i}(a)}{e^{T-t}-1}, \qquad i \in M(\overleftarrow{X}_t). \tag{26}$$

All rates and costs below are zero for visible i. Along the exact law, coarse-posterior positivity and finite grid loss imply $\lambda_{k,t}^{\mathrm{blk},i,a} = 0 \Longrightarrow \lambda_t^{*,i,a} = 0$ and $\lambda_{k,t}^{\mathrm{blk},i,a} > 0 \Longrightarrow \widehat{\lambda}_{k,t}^{i,a} > 0$.

For $v, w > 0$ and $u \ge 0$, expanding the three terms gives $\mathfrak{d}(v,w) + (v-u)\log(w/v) + \mathfrak{d}(u,v) = w - v + v\log(v/w) + (v-u)\log(w/v) + v - u + u\log(u/v)$, which simplifies to $w - u + u\log(u/w) = \mathfrak{d}(u,w)$. Thus, defining $T_{1,k,t}^{i,a} = \mathfrak{d}(\lambda_{k,t}^{\mathrm{blk},i,a}, \widehat{\lambda}_{k,t}^{i,a})$, $T_{2,k,t}^{i,a} = (\lambda_{k,t}^{\mathrm{blk},i,a} - \lambda_t^{*,i,a})\log(\widehat{\lambda}_{k,t}^{i,a}/\lambda_{k,t}^{\mathrm{blk},i,a})$, and $T_{3,k,t}^{i,a} = \mathfrak{d}(\lambda_t^{*,i,a}, \lambda_{k,t}^{\mathrm{blk},i,a})$, with $T_{2,k,t}^{i,a} = 0$ when $\lambda_{k,t}^{\mathrm{blk},i,a} = 0$, we obtain

$$\mathfrak{d}(\lambda_t^{*,i,a}, \widehat{\lambda}_{k,t}^{i,a}) = T_{1,k,t}^{i,a} + T_{2,k,t}^{i,a} + T_{3,k,t}^{i,a}. \tag{27}$$

Indeed, in the remaining zero case, both sides equal $\widehat{\lambda}_{k,t}^{i,a}$.

For $j \in \{1,2,3\}$, write $T_{j,k}(t) = \mathbb{E}\sum_{i,a} T_{j,k,t}^{i,a}$ and $T_j(b) = \sum_{k=0}^{N-1} \int_{t_k}^{\min\{t_{k+1},b\}} T_{j,k}(t)\,\mathrm{d}t$. Substituting equation 27 into equation 25 gives

$$\mathrm{KL}(q_{T-b}\|p_b) \le \mathrm{KL}(q_T\|p_0) + T_1(b) + T_2(b) + T_3(b). \tag{28}$$

We next control the three terms in this order.

B.2 Control of T_1 by the Score Entropy Error

The score is evaluated at a partially updated context, and its block is selected from the interval-start state. The following lemma bounds expectations under this partially updated context by those under the exact marginal at the earlier time, allowing the block to depend on the interval-start state.

Lemma 8 (Distribution comparison for adaptive blocks). *Fix k and $t_k \le s \le t < T$. On $i \in M(Z)$, let*

$$Z = \overleftarrow{X}_{t_k}, \qquad B_i = \widehat{B}_{t_k}(i, Z), \qquad Y_i(s,t) = (\overleftarrow{X}_{t,B_i}, \overleftarrow{X}_{s,-B_i}). \tag{29}$$

For every nonnegative F vanishing when i is visible,

$$\mathbb{E}[F(Y_i(s,t)); i \in M(\overleftarrow{X}_t)] \le \theta_{s,t} e^{(r(t_k)-1)(t-s)} \mathbb{E}_{W \sim q_{T-s}} F(W). \tag{30}$$

The left integrand is zero on $i \notin M(Z)$.

Proof. The case $s = t$ is immediate, so assume $s < t$. Set $W = \overleftarrow{X}_s$, $R_i = \widehat{B}_{t_k}(i, \mathrm{MASK}^d)$, and $\theta_{s,t} = (1 - e^{t-T})/(1 - e^{s-T})$. Monotone refinement and bounded block size imply $\widehat{B}_{t_k}(i,z) \subseteq R_i$ whenever $i \in M(z)$, and $|R_i| \le r(t_k)$.

Fix y with $q_{T-s}(y) > 0$ and $i \in M(y)$, and put $H = R_i \cap U(y)$, so $|H| \le r(t_k) - 1$. For $A \subseteq H$, let w_A be y with A masked. Every possible predecessor has this form: on $Y_i(s,t) = y$, we have $W \preceq y$ and $A = U(y) \setminus U(W) \subseteq B_i \subseteq R_i$, hence $W = w_A$. Each w_A is supported.

Write $p_A = \mathbb{P}(X_0^A = y^A \mid W = w_A)$. Conditional on $W = w_A$ and any supported $Z = z$, the block $B = \widehat{B}_{t_k}(i,z)$ is fixed. The reverse Markov property and independent masking clocks give

$$\mathbb{P}(Y_i(s,t) = y \mid W = w_A, Z = z) = \mathbf{1}_{\{A \subseteq B\}} \theta_{s,t}^{|B \cap M(y)|} (1 - \theta_{s,t})^{|A|} p_A.$$

Since $i \in B \cap M(y)$, the right-hand side is at most $\theta_{s,t}(1-\theta_{s,t})^{|A|} p_A$. This bound is independent of z, so averaging over $Z \mid W = w_A$ yields

$$\mathbb{P}(Y_i(s,t) = y \mid W = w_A) \le \theta_{s,t}(1 - \theta_{s,t})^{|A|} p_A. \tag{31}$$

The masking marginal gives $q_{T-s}(w_A) p_A = q_{T-s}(y)(e^{T-s} - 1)^{|A|}$: both sides impose the same clean-sample values, while each coordinate in A changes from visible to masked. Moreover, $(e^{T-s} - 1)(1 - \theta_{s,t}) = (e^{t-T} - e^{s-T})/e^{s-T} = e^{t-s} - 1$. Therefore,

$$\begin{aligned}
&\mathbb{P}(Y_i(s,t) = y,\ i \in M(\overleftarrow{X}_t)) \\
&= \sum_{A \subseteq H} q_{T-s}(w_A) \mathbb{P}(Y_i(s,t) = y \mid W = w_A) \\
&\le \theta_{s,t} \sum_{A \subseteq H} (1 - \theta_{s,t})^{|A|} q_{T-s}(w_A) p_A \\
&= \theta_{s,t} q_{T-s}(y) \sum_{A \subseteq H} (e^{t-s} - 1)^{|A|} \\
&= \theta_{s,t} q_{T-s}(y) e^{(t-s)|H|} \\
&\le \theta_{s,t} e^{(r(t_k)-1)(t-s)} q_{T-s}(y),
\end{aligned} \tag{32}$$

where the subset sum is evaluated by the binomial identity.

Every hybrid state is supported by q_{T-s}, since its visible entries agree with a common clean sample. Multiplying equation 32 by $F(y) \ge 0$ and summing over supported y with $i \in M(y)$ proves the claim. □

Proposition 1 (Score-estimation term). *For $t_k \le t < t_{k+1}$,*

$$T_{1,k}(t) \le e^{r(t_k)(t-t_k)} \mathcal{L}_{\mathrm{SE}}(\tau_k, \widehat{s}_{\tau_k}, s_{\tau_k}). \tag{33}$$

Consequently, if $r(t_k) h_k \le 1$ for every k, then $T_1(b) \le e\varepsilon_{\mathrm{score}}$ for all $t_{N-1} < b < T$.

Proof. For supported y with $i \in M(y)$, set $G_{k,i}(y) = \sum_a \mathfrak{d}(s_{\tau_k}(y \odot_i a, y), \widehat{s}_{\tau_k}(y \odot_i a, y))$; set it to zero otherwise. The score assumption and homogeneity give, for $c > 0$, with zero arguments interpreted using the conventions for $\mathfrak{d}$,

$$\begin{aligned} \sum_i \mathbb{E}_{W \sim q_{\tau_k}} G_{k,i}(W) &= \mathcal{L}_{\mathrm{SE}}(\tau_k, \widehat{s}_{\tau_k}, s_{\tau_k}), \\ \mathfrak{d}(cu, cv) &= cv - cu + cu \log(cu/cv) = c\,\mathfrak{d}(u, v), \\ \rho_k(t)\theta_{t_k,t} &= \frac{e^{\tau_k} - 1}{e^{T-t} - 1} \frac{1 - e^{t-T}}{1 - e^{-\tau_k}} = e^{\tau_k} e^{t-T} = e^{t-t_k}. \end{aligned} \tag{34}$$

Applying Lemma 8 with $s = t_k$,

$$\begin{aligned} T_{1,k}(t) &= \rho_k(t) \sum_i \mathbb{E}[G_{k,i}(Y_i(t_k, t)); i \in M(\overleftarrow{X}_t)] \\ &\le \rho_k(t)\theta_{t_k,t} e^{(r(t_k)-1)(t-t_k)} \sum_i \mathbb{E}_{W \sim q_{\tau_k}} G_{k,i}(W) \\ &= e^{r(t_k)(t-t_k)} \mathcal{L}_{\mathrm{SE}}(\tau_k, \widehat{s}_{\tau_k}, s_{\tau_k}) \\ &\le e\,\mathcal{L}_{\mathrm{SE}}(\tau_k, \widehat{s}_{\tau_k}, s_{\tau_k}), \qquad r(t_k)h_k \le 1. \end{aligned} \tag{35}$$

Thus

$$\begin{aligned} \sup_{t_{N-1} < b < T} T_1(b) &\le e \sum_k h_k \mathcal{L}_{\mathrm{SE}}(\tau_k, \widehat{s}_{\tau_k}, s_{\tau_k}) \\ &\le e\,\varepsilon_{\mathrm{score}}. \end{aligned} \tag{36}$$

□

B.3 Cancellation of T_2 by the Block-Frozen Martingale Property

The logarithmic factor in T_2 depends on Y_i, so we condition on both the interval-start state and the block-frozen context.

We will repeatedly use the following posterior identity. Fix $\ell \le s \le t < T$, a state z, and $B \subseteq M(z)$. Set $Z = \overleftarrow{X}_\ell$ and $Y = (\overleftarrow{X}_{t,B}, \overleftarrow{X}_{s,-B})$. For compatible z, y, independence of the clocks gives

$$\begin{aligned} &\mathbb{P}(Z = z, Y = y \mid X_0 = c) \\ &\quad = e^{(\ell-T)|U(z)|}(e^{t-T} - e^{\ell-T})^{|B \cap U(y)|}(1 - e^{t-T})^{|B \cap M(y)|} \\ &\qquad \cdot (e^{s-T} - e^{\ell-T})^{|U(y) \setminus (U(z) \cup B)|}(1 - e^{s-T})^{|M(y) \setminus B|} \mathbf{1}\{c^{U(y)} = y^{U(y)}\}. \end{aligned} \tag{37}$$

Empty products equal one. If $\mathbb{P}(Z = z, Y = y) > 0$, Bayes' rule yields

$$\begin{aligned} \mathbb{P}(X_0 = c \mid Z = z, Y = y) &= \frac{q_{\mathrm{data}}(c)\mathbb{P}(Z = z, Y = y \mid X_0 = c)}{\sum_{c'} q_{\mathrm{data}}(c')\mathbb{P}(Z = z, Y = y \mid X_0 = c')} \\ &= \frac{q_{\mathrm{data}}(c)\mathbf{1}\{c^{U(y)} = y^{U(y)}\}}{\mathbb{P}(X_0^{U(y)} = y^{U(y)})}, \\ \mathbb{P}(X_0^i = a \mid Z = z, Y = y) &= \pi_i^y(a). \end{aligned} \tag{38}$$

The same calculation applies when B is selected from Z, after conditioning on $Z = z$.

Lemma 9 (Block-frozen score projection). *For $Z = \overleftarrow{X}_{t_k}$ and $Y_i = (\overleftarrow{X}_{t,B_i}, Z_{-B_i})$,*

$$\mathbb{E}[\lambda_t^{*,i,a} \mid Z, Y_i] = \lambda_{k,t}^{\mathrm{blk},i,a}. \tag{39}$$

All rates are zero for visible i.

Proof. Fix supported $Z = z$ with $i \in M(z)$, so B_i is fixed. The posterior identity and the tower property give

$$\begin{aligned} \mathbb{P}(X_0^i = a \mid Z, \overleftarrow{X}_t) &= \pi_i^{\overleftarrow{X}_t}(a), \\ \mathbb{E}[\pi_i^{\overleftarrow{X}_t}(a) \mid Z, Y_i] &= \mathbb{E}[\mathbb{P}(X_0^i = a \mid Z, \overleftarrow{X}_t) \mid Z, Y_i] \\ &= \mathbb{P}(X_0^i = a \mid Z, Y_i) = \pi_i^{Y_i}(a). \end{aligned} \tag{40}$$

Since $i \in B_i$,

$$\begin{aligned}
\mathbf{1}\{i \in M(\overleftarrow{X}_t)\} &= \mathbf{1}\{i \in M(Y_i)\}, \\
\mathbb{E}[\lambda_t^{*,i,a} \mid Z, Y_i] &= \frac{\mathbf{1}\{i \in M(Y_i)\}}{e^{T-t}-1}\mathbb{E}[\pi_i^{\overleftarrow{X}_t}(a) \mid Z, Y_i] \\
&= \frac{\mathbf{1}\{i \in M(Y_i)\}\pi_i^{Y_i}(a)}{e^{T-t}-1} = \lambda_{k,t}^{\mathrm{blk},i,a}.
\end{aligned} \tag{41}$$

The rate expressions on visible coordinates are interpreted as zero without evaluating a score. □

We apply this identity to T_2. If $\lambda_{k,t}^{\mathrm{blk},i,a} > 0$, the logarithmic factor is $\sigma(Z, Y_i)$-measurable and finite:

$$\log \frac{\widehat{\lambda}_{k,t}^{i,a}}{\lambda_{k,t}^{\mathrm{blk},i,a}} = \log \frac{\widehat{s}_{\tau_k}(Y_i \odot_i a, Y_i)}{s_{\tau_k}(Y_i \odot_i a, Y_i)}. \tag{42}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[T_{2,k,t}^{i,a} \mid Z, Y_i] &= \log \frac{\widehat{\lambda}_{k,t}^{i,a}}{\lambda_{k,t}^{\mathrm{blk},i,a}}\big(\lambda_{k,t}^{\mathrm{blk},i,a} - \mathbb{E}[\lambda_t^{*,i,a} \mid Z, Y_i]\big) = 0, \\
T_{2,k}(t) &= \sum_{i,a} \mathbb{E}\big[\mathbb{E}[T_{2,k,t}^{i,a} \mid Z, Y_i]\big] = 0.
\end{aligned} \tag{43}$$

If $\lambda_{k,t}^{\mathrm{blk},i,a} = 0$, the mixed term is zero by definition, so the cancellation holds there as well.

B.4 Control of T_3 by Posterior Inter-Block Information

We first identify the path-space meaning of the inter-block cost. Fix k, condition on $Z = \overleftarrow{X}_{t_k} = z$, and freeze $\widehat{\mathcal{P}}_{t_k}(z)$. For $t_k < u \le t_{k+1}$ with $u < T$, let $P_{k,z}^u$ be the exact joint law of the block trajectories on $[t_k, u]$, and let $P_{k,z}^{u,B}$ be its B-th marginal. The following lemma identifies the accumulated T_3 cost with the KL divergence between the exact joint path law and the product of its block marginals.

Lemma 10 (Path-space representation of T_3). *The exact-score block-frozen process has conditional path law $\bigotimes_{B \in \widehat{\mathcal{P}}_{t_k}(z)} P_{k,z}^{u,B}$, and*

$$\int_{t_k}^{u} \mathbb{E}\left[\sum_{i,a} T_{3,k,t}^{i,a} \,\middle|\, Z = z\right] \mathrm{d}t = \mathrm{KL}\left(P_{k,z}^u \,\middle\|\, \bigotimes_{B \in \widehat{\mathcal{P}}_{t_k}(z)} P_{k,z}^{u,B}\right). \tag{44}$$

Proof. Conditional on $Z = z$, each block uses its clean tokens and independent reveal clocks. With $y = (x_B, z_{-B})$ and $i \in B \cap M(x)$,

$$\begin{aligned}
&\mathbb{P}(X_0^i = a \mid Z = z, \overleftarrow{X}_{t,B} = x_B) = \pi_i^y(a), \\
&\text{marginal rate of } (i,a) = \frac{\pi_i^y(a)}{e^{T-t}-1} = \lambda_{k,t}^{\mathrm{blk},i,a}.
\end{aligned} \tag{45}$$

Past block reveal times provide no additional clean-token information since their likelihood is a clock factor independent of compatible clean completions. Thus these are also the intensities conditional on the block history, and independent exact-score block dynamics simulate the product of the exact block marginals. The path-space KL identity (Lemma 5) gives

$$\begin{aligned}
\mathrm{KL}\left(P_{k,z}^u \,\middle\|\, \bigotimes_{B \in \widehat{\mathcal{P}}_{t_k}(z)} P_{k,z}^{u,B}\right) &= \int_{t_k}^{u} \mathbb{E}\left[\sum_{i,a} \mathfrak{d}(\lambda_t^{*,i,a}, \lambda_{k,t}^{\mathrm{blk},i,a}) \,\middle|\, Z = z\right] \mathrm{d}t \\
&= \int_{t_k}^{u} \mathbb{E}\left[\sum_{i,a} T_{3,k,t}^{i,a} \,\middle|\, Z = z\right] \mathrm{d}t.
\end{aligned} \tag{46}$$

□

Assumption 2.2 controls a time-dependent information rate. We therefore bound the rate cost before integrating over the interval.

Lemma 11 (Cross-block information cost). *For $t_k \le t < t_{k+1}$ with $t < T$,*

$$T_{3,k}(t) \le \int_{t_k}^{t} e^{r(t_k)(t-s)} \Gamma_{\mathrm{part}}(s)\, \mathrm{d}s. \tag{47}$$

Proof. At $t = t_k$, both sides vanish. Fix $t_k < t < t_{k+1}$ with $t < T$, and use $Z, B_i, Y_i(s,t)$ from Lemma 8. For supported y with distinct $i, j \in M(y)$, write $I_{ij}(y) = I(X_0^i; X_0^j \mid X_0^{U(y)} = y^{U(y)})$.

For $i \in M(\overleftarrow{X}_t)$, the rate definitions give $\lambda_t^{*,i,a} = \frac{\pi_i^{\overleftarrow{X}_t}(a)}{e^{T-t}-1}$ and $\lambda_{k,t}^{\mathrm{blk},i,a} = \frac{\pi_i^{Y_i(t_k,t)}(a)}{e^{T-t}-1}$. Since both posteriors sum to one, expanding $\mathfrak{d}$ gives

$$\begin{aligned}
&\sum_a \mathfrak{d}\big(\lambda_t^{*,i,a}, \lambda_{k,t}^{\mathrm{blk},i,a}\big) \\
=&\sum_a \left[\lambda_{k,t}^{\mathrm{blk},i,a} - \lambda_t^{*,i,a} + \lambda_t^{*,i,a} \log \frac{\lambda_t^{*,i,a}}{\lambda_{k,t}^{\mathrm{blk},i,a}}\right] \\
=&\frac{1}{e^{T-t}-1} \sum_a \left[\pi_i^{Y_i(t_k,t)}(a) - \pi_i^{\overleftarrow{X}_t}(a) + \pi_i^{\overleftarrow{X}_t}(a) \log \frac{\pi_i^{\overleftarrow{X}_t}(a)}{\pi_i^{Y_i(t_k,t)}(a)}\right] \\
=&\frac{1 - 1 + \sum_a \pi_i^{\overleftarrow{X}_t}(a) \log \frac{\pi_i^{\overleftarrow{X}_t}(a)}{\pi_i^{Y_i(t_k,t)}(a)}}{e^{T-t}-1} \\
=&\frac{\mathrm{KL}\big(\pi_i^{\overleftarrow{X}_t} \| \pi_i^{Y_i(t_k,t)}\big)}{e^{T-t}-1}.
\end{aligned} \tag{48}$$

Fix supported z with $i \in M(z)$, put $B = \widehat{B}_{t_k}(i,z)$, and condition on a positive-probability event $\mathcal{A} = \{Z = z, \overleftarrow{X}_{t,B} = v\}$ with $v^i = \mathrm{MASK}$. Given $Z = z$, the remaining reveal clocks are mutually independent and independent of the clean tokens. Conditioning further on $\overleftarrow{X}_{t,B} = v$ leaves the outside clocks unchanged. Past reveal times provide no additional clean-token information beyond the currently revealed values. Thus, by equation 38, the conditional process $Y_i(s,t) = (v, \overleftarrow{X}_{s,-B})$ is a CTMC with rate $\frac{\pi_j^y(a)}{e^{T-s}-1}$ from y to $y \odot_j a$, for $j \in M(y) \setminus B$. Its total rate on $[t_k, t]$ is bounded by $\frac{\pi_j^y(a)}{e^{T-s}-1}$.

For each such j, the expected entropy change upon revealing j is

$$\begin{aligned}
&\sum_a \pi_j^y(a) \big[H(\pi_i^{y \odot_j a}) - H(\pi_i^y)\big] \\
=&H(X_0^i \mid X_0^j, X_0^{U(y)} = y^{U(y)}) - H(X_0^i \mid X_0^{U(y)} = y^{U(y)}) \\
=&- I_{ij}(y).
\end{aligned} \tag{49}$$

Terms with $\pi_j^y(a) = 0$ are omitted.

Under $\mathcal{A}$, the initial hybrid state $Y_i(t_k, t) = (v, z_{-B})$ is deterministic, and $Y_i(t,t) = \overleftarrow{X}_t$. The posterior identity and the tower property give $\mathbb{E}[\pi_i^{\overleftarrow{X}_t}(a) \mid \mathcal{A}] = \pi_i^{Y_i(t_k,t)}(a)$. If the latter probability is zero, then the former posterior is zero conditionally almost surely, and the corresponding logarithmic term is omitted. Apply Lemma 4 under the conditional law to the bounded, time-independent

function $y \mapsto H(\pi_i^y)$. Taking expectations and using equation 49, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\mathrm{KL}\big(\pi_i^{\overleftarrow{X}_t} \| \pi_i^{Y_i(t_k,t)}\big) \mid \mathcal{A}\right] \\
=& -\mathbb{E}[H(\pi_i^{\overleftarrow{X}_t}) \mid \mathcal{A}] - \sum_a \mathbb{E}[\pi_i^{\overleftarrow{X}_t}(a) \mid \mathcal{A}] \log \pi_i^{Y_i(t_k,t)}(a) \\
=& H(\pi_i^{Y_i(t_k,t)}) - \mathbb{E}[H(\pi_i^{\overleftarrow{X}_t}) \mid \mathcal{A}] \\
=& \int_{t_k}^t \frac{1}{e^{T-s}-1} \mathbb{E}\left[\sum_{j \in M(\overleftarrow{X}_s) \setminus B} I_{ij}(Y_i(s,t)) \;\middle|\; \mathcal{A}\right] \mathrm{d}s.
\end{aligned} \tag{50}$$

Here the martingale term has zero expectation, and $Y_i(s,t)$ agrees with $\overleftarrow{X}_s$ outside B.

For each i, these events $\mathcal{A}$ partition $\{i \in M(\overleftarrow{X}_t)\}$. Multiplying equation 50 by $\mathbb{P}(\mathcal{A})$ and summing over these events and then over i, we obtain from equation 48 and Tonelli's theorem that

$$\begin{aligned}
T_{3,k}(t) &= \frac{1}{e^{T-t}-1} \mathbb{E} \sum_{i \in M(\overleftarrow{X}_t)} \mathrm{KL}\big(\pi_i^{\overleftarrow{X}_t} \| \pi_i^{Y_i(t_k,t)}\big) \\
&= \int_{t_k}^t \frac{\mathbb{E} \sum_{i \in M(\overleftarrow{X}_t)} \sum_{j \in M(\overleftarrow{X}_s) \setminus B_i} I_{ij}(Y_i(s,t))}{(e^{T-t}-1)(e^{T-s}-1)} \mathrm{d}s.
\end{aligned} \tag{51}$$

If $i \in M(\overleftarrow{X}_t)$, $Z \preceq Y_i(s,t)$, so monotone refinement gives $\widehat{B}_s(i, Y_i(s,t)) \subseteq B_i$. Moreover, $M(\overleftarrow{X}_s) \setminus B_i = M(Y_i(s,t)) \setminus B_i$. Since mutual information is nonnegative, the inner sum in equation 51 is therefore at most $\sum_{j \in M(Y_i(s,t)) \setminus \widehat{B}_s(i, Y_i(s,t))} I_{ij}(Y_i(s,t))$.

Apply Lemma 8, for each i, to the nonnegative function $y \mapsto \sum_{j \in M(y) \setminus \widehat{B}_s(i,y)} I_{ij}(y)$ on supported states with i masked, extended by zero otherwise. Using the definition of $\Gamma_{\mathrm{part}}(s)$ and $\frac{\theta_{s,t}}{e^{T-t}-1} = \frac{e^{t-T}}{1-e^{s-T}} = \frac{e^{t-s}}{e^{T-s}-1}$, we conclude that

$$\begin{aligned}
T_{3,k}(t) &\le \int_{t_k}^t \frac{\theta_{s,t} e^{(r(t_k)-1)(t-s)}}{(e^{T-t}-1)(e^{T-s}-1)} \mathbb{E}_{W \sim q_{T-s}} \left[\sum_{i \in M(W)} \sum_{j \in M(W) \setminus \widehat{B}_s(i,W)} I_{ij}(W)\right] \mathrm{d}s \\
&= \int_{t_k}^t \theta_{s,t} e^{(r(t_k)-1)(t-s)} \frac{e^{T-s}-1}{e^{T-t}-1} \Gamma_{\mathrm{part}}(s) \, \mathrm{d}s \\
&= \int_{t_k}^t e^{r(t_k)(t-s)} \Gamma_{\mathrm{part}}(s) \, \mathrm{d}s.
\end{aligned} \tag{52}$$

□

By Tonelli's theorem, $r(t_k)h_k \le 1$, and Assumption 2.2,

$$\begin{aligned} T_3(b) \le T_3 &:= \sum_k \int_{t_k}^{t_{k+1}} T_{3,k}(t)\,\mathrm{d}t \\ &\le \sum_k \int_{t_k}^{t_{k+1}} \int_{t_k}^{t} e^{r(t_k)(t-s)}\Gamma_{\mathrm{part}}(s)\,\mathrm{d}s\,\mathrm{d}t \\ &= \sum_k \int_{t_k}^{t_{k+1}} \Gamma_{\mathrm{part}}(s) \int_{s}^{t_{k+1}} e^{r(t_k)(t-s)}\,\mathrm{d}t\,\mathrm{d}s \\ &\le e \sum_k \int_{t_k}^{t_{k+1}} (t_{k+1}-s)\Gamma_{\mathrm{part}}(s)\,\mathrm{d}s \\ &\le e \sum_k h_k \int_{t_k}^{t_{k+1}} \Gamma_{\mathrm{part}}(s)\,\mathrm{d}s \\ &\le e \sum_k h_k \int_{t_k}^{t_{k+1}} \varepsilon_{\mathrm{part}}(s)\,\mathrm{d}s. \end{aligned} \tag{53}$$

B.5 Proof of Theorem 1

Complexity. For termination, each successful jump decreases the mask count by one; each rejected holding time ends the current block update. In the final interval, a nonempty block satisfies, successively,

$$\begin{aligned} 0 < \Lambda_{B,m} < \infty &\Longrightarrow \mathbb{P}(\xi_{B,m} < \infty \mid \text{past}) = 1, \\ J_B \le |B| &\Longrightarrow \sum_{m=1}^{J_B} \xi_{B,m} < \infty \quad \text{almost surely}, \\ \lim_{t\uparrow T} u_{N-1}(t) = \infty &\Longrightarrow u_{N-1}^{-1}\Big(\sum_{m=1}^{J_B} \xi_{B,m}\Big) < T \quad \text{almost surely}. \end{aligned} \tag{54}$$

There are finitely many blocks; if none remain, completion is immediate. For the exact process,

$$\mathbb{P}\big(\min_i E_i > 0\big) = 1 \quad \Longrightarrow \quad \max_i \max\{0, T - E_i\} < T \quad \text{almost surely}. \tag{55}$$

Both processes are thus eventually equal to their clean outputs. Thus, on the finite state space,

$$p_b \longrightarrow p_{\mathrm{alg}}, \qquad q_{T-b} \longrightarrow q_{\mathrm{data}} \qquad (b \uparrow T). \tag{56}$$

Finally, let $J_{k,B}$ count successful reveals and let $Q_{k,B}$ count score evaluations for block B, including its shared initial one. Then

$$\begin{aligned} &\sum_k \sum_{B\in\mathcal{P}_k} J_{k,B} \le d, \\ &\qquad Q_{k,B} = 1 + \min\{J_{k,B}, |B| - 1\} = \min\{J_{k,B} + 1, |B|\}, \\ &Q_{k,B} - 1 \le J_{k,B}, \qquad\qquad Q_{k,B} \le |B|. \end{aligned} \tag{57}$$

Sharing the initial query and running blocks in parallel gives

$$\begin{aligned} N_{\mathrm{score}} &= \sum_{k:M(x_{t_k})\neq\varnothing} \Big(1 + \sum_{B\in\mathcal{P}_k} (Q_{k,B} - 1)\Big) \\ &\le N + \sum_k \sum_{B\in\mathcal{P}_k} J_{k,B} \le N + d, \\ D_{\mathrm{score}} &\le \sum_k \max_{B\in\mathcal{P}_k} Q_{k,B} \le \sum_k \max_{B\in\mathcal{P}_k} |B| \le \sum_k r(t_k), \\ N_{\mathrm{part}} &\le N. \end{aligned} \tag{58}$$

Empty intervals contribute zero, with $\max \varnothing = 0$.

Correctness. Under both q_T and p_0, the visible set U has independent inclusion probability e^{-T}. Conditional on U, their token laws are q_{data}^U and $\text{Unif}(\mathcal{V}^U)$. Hence

$$\begin{aligned}\text{KL}(q_T \| p_0) &= \mathbb{E}_U \,\text{KL}(q_{\text{data}}^U \| \,\text{Unif}(\mathcal{V}^U)) \\ &= \mathbb{E}_U[|U| \log S - H(X_0^U)] \\ &\le \mathbb{E}|U| \log S = de^{-T} \log S.\end{aligned} \tag{59}$$

For $t_{N-1} < b < T$, the three-term decomposition now yields

$$\begin{aligned}\text{KL}(q_{T-b} \| p_b) &\le \text{KL}(q_T \| p_0) + T_1(b) + T_2(b) + T_3(b) \\ &\le de^{-T} \log S + e\varepsilon_{\text{score}} + e \sum_k h_k \int_{t_k}^{t_{k+1}} \varepsilon_{\text{part}}(s)\, \mathrm{d}s.\end{aligned} \tag{60}$$

C Proof of Corollary 1

Proof. By the choice of T, $de^{-T} \log S = O(\varepsilon)$. It remains to bound the last term in equation 5.

Uniform step-size schedule. Take $N = \lceil T(r_{\max} + 4e\mathcal{B}_{\text{part}}/\varepsilon) \rceil$ and $t_k = kT/N$. Then $h_k r(t_k) \le T r_{\max}/N \le 1$, and

$$e \sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \varepsilon_{\text{part}}(t)\, \mathrm{d}t = \frac{eT}{N} \int_0^T \varepsilon_{\text{part}}(t)\, \mathrm{d}t = \frac{eT\mathcal{B}_{\text{part}}}{N} \le \frac{\varepsilon}{4}.$$

Exponential-then-constant step-size schedule. For every supported state x and distinct $i, j \in M(x)$, $I(X_0^i; X_0^j \mid X_{T-t} = x) \le \log S$. Writing $\tau = T - t$ and using independent masking, we obtain

$$\begin{aligned}\Gamma_{\text{part}}(t) &\le \frac{\log S}{(e^\tau - 1)^2} \sum_{i \ne j} \Pr_{Z \sim q_\tau}(i, j \in M(Z)) \\ &= d(d-1) \log S \, \frac{(1 - e^{-\tau})^2}{(e^\tau - 1)^2} \\ &= d(d-1) \log S \, e^{-2\tau} \le d^2 \log S.\end{aligned}$$

Next, we bound the last term of equation 5 as follows. Writing $h_k = t_{k+1} - t_k$ and $\delta = T - t_{N-1} = \varepsilon/(d \log S)$, we bound the last term of equation 5 by

$$\begin{aligned}\sum_{k=0}^{N-1} h_k \int_{t_k}^{t_{k+1}} \varepsilon_{\text{part}}(t)\, \mathrm{d}t &\le \delta^2 d^2 \log S + \kappa \sum_{k=0}^{N-2} \int_{t_k}^{t_{k+1}} \min\{1, T - t_{k+1}\} \varepsilon_{\text{part}}(t)\, \mathrm{d}t \\ &\le \frac{\varepsilon^2}{\log S} + \kappa \int_0^T \min\{1, T - t\} \varepsilon_{\text{part}}(t)\, \mathrm{d}t \\ &\lesssim \varepsilon + \kappa \mathcal{D}_{\text{part}}.\end{aligned}$$

Here we used $T - t_{k+1} \le T - t$ on each interval, $S \ge 2$, and $\varepsilon < 1$. Such a schedule can be chosen with $\kappa = O((T + \log(1/\delta))/N)$. Thus, taking

$$N = \Theta\left(\left(T + \log(1/\delta)\right) \left(r_{\max} + \frac{\mathcal{D}_{\text{part}}}{\varepsilon} \right) \right) = \widetilde{O}\left(r_{\max} + \frac{\mathcal{D}_{\text{part}}}{\varepsilon} \right)$$

with a sufficiently large absolute constant ensures $\kappa r_{\max} \le 1$ and $\kappa \mathcal{D}_{\text{part}} \lesssim \varepsilon$. □

D Proof for the Ising Example

Proof. Let $d \ge 2$ be even and fix $\beta > 0$. Consider

$$q_{\text{data}}(x) = \prod_{j=1}^{d/2} \pi_\beta(x_{2j-1}, x_{2j}), \qquad \pi_\beta(u, v) = \frac{e^{\beta uv}}{4 \cosh \beta}, \quad u, v \in \{-1, 1\}.$$

Write $\rho = \tanh \beta$. Both marginals of π_β are uniform, and $\pi_\beta(u, v) = (1 + \rho uv)/4$. The mutual information within each pair is $I_\beta = \beta \tanh \beta - \log \cosh \beta > 0$.

Guarantees in Dmitriev et al. (2026). The complexity guarantees for τ-leaping sampler in Dmitriev et al. (2026, Section 3.2) adapt to small total correlation $\mathcal{C}(q)$, dual total correlation $\mathcal{B}(q)$, or effective total correlation $\mathcal{D}(q)$. Their iteration bound is governed by $\mathcal{D}(q)/\varepsilon$ up to logarithmic factors, and $\mathcal{D}(q) \le \min\{\mathcal{B}(q), \mathcal{C}(q)\}$, so smallness of any one of these quantities yields an improved complexity bound.

For $X \sim q$, the total correlation and dual total correlation are defined by

$$\mathcal{C}(q) := \sum_{i=1}^{d} H(X_i) - H(X), \qquad \mathcal{B}(q) := H(X) - \sum_{i=1}^{d} H(X_i \mid X_{-i}).$$

The effective total correlation is

$$\mathcal{D}(q) := \int_0^\infty \min\{1, \tau\} \mathcal{I}_q(\tau)\, \mathrm{d}\tau, \qquad \mathcal{I}_q(\tau) := \sum_{i \neq j} I(X_\tau^i; X_\tau^j \mid X_\tau^{-(i,j)}),$$

where $(X_\tau)_{\tau \ge 0}$ is the rate-one masking process initialized from $X_0 \sim q$, and $X_\tau^{-(i,j)}$ denotes the masked state with coordinates i and j removed. We now compute these three quantities for the disjoint-edge Ising model above and show that all three are $\Theta_\beta(d)$.

Independence across pairs gives $H(X) = d \log 2 - (d/2) I_\beta$. Moreover, conditioning on all other coordinates is equivalent to conditioning on the partner of X_i, so $H(X_i \mid X_{-i}) = \log 2 - I_\beta$. Consequently,

$$\mathcal{C}(q_{\mathrm{data}}) = \sum_{i=1}^{d} H(X_i) - H(X) = \frac{d}{2} I_\beta,$$

$$\mathcal{B}(q_{\mathrm{data}}) = H(X) - \sum_{i=1}^{d} H(X_i \mid X_{-i}) = \frac{d}{2} I_\beta.$$

For the masking information rate, only coordinates belonging to the same pair contribute. There are d ordered partner pairs. Both partners are masked with probability $(1 - e^{-\tau})^2$, in which case their conditional mutual information remains I_β. Thus the all-coordinate information rate is

$$I(\tau) = \frac{d I_\beta (1 - e^{-\tau})^2}{(e^\tau - 1)^2} = d I_\beta e^{-2\tau}.$$

The effective total correlation of Dmitriev et al. (2026) therefore satisfies

$$\mathcal{D}(q_{\mathrm{data}}) = d I_\beta \int_0^\infty \min\{1, \tau\} e^{-2\tau}\, \mathrm{d}\tau = \frac{1 - e^{-2}}{4}\, d I_\beta.$$

All three quantities are $\Theta_\beta(d)$.

Complexity of τ-leaping sampler (Dmitriev et al., 2026). One may wonder whether the preceding complexity guarantees, expressed in terms of total, dual total, and effective total correlation, are tight on this instance. We therefore analyze the sampler directly and establish a lower bound on its parallel query depth, showing that Algorithm 1 of Dmitriev et al. (2026) even requires $\Omega(\sqrt{d})$ sequential score-query rounds to attain a sufficiently small constant KL error.

Consider Algorithm 1 of Dmitriev et al. (2026) with exact scores, its prescribed product initialization, and an arbitrary deterministic grid $0 = t_0 < \cdots < t_N = T$. The initialization independently assigns each coordinate a uniform spin with probability e^{-T} and a mask otherwise.

For each coordinate, let $L \in \{0, \dots, N\}$ denote its revelation round, where $L = 0$ means that it is unmasked at initialization. For $1 \le k < N$, an initially masked coordinate survives interval k with conditional probability

$$\frac{1 - e^{-(T - t_k)}}{1 - e^{-(T - t_{k-1})}},$$

while the final interval reveals every remaining coordinate. These probabilities depend only on the grid. Hence the revelation labels are independent across coordinates and identically distributed. Writing $p_k = \Pr(L = k)$, the probability that two partners are revealed in the same round is

$$a = \sum_{k=0}^{N} p_k^2 \ge \frac{1}{N + 1},$$

where the inequality follows from $(\sum_{k=0}^{N} p_k)^2 \le (N+1)\sum_{k=0}^{N} p_k^2$.

If the partners are revealed in different rounds, the first spin is uniform and the second is sampled from its exact conditional distribution, yielding joint law π_β. If they are revealed in the same round, including round zero, they are sampled independently from their uniform marginals. Thus, writing U_2 for the uniform law on $\{-1,1\}^2$, the output law of each pair is $P_{\mathrm{pair}} = (1-a)\pi_\beta + aU_2$.

The difference between π_β and P_{pair} has absolute value $a\rho/4$ at each of the four configurations, so $\mathrm{TV}(\pi_\beta, P_{\mathrm{pair}}) = a\rho/2$. The algorithm's output factorizes across pairs. By additivity of KL divergence and Pinsker's inequality,

$$\begin{aligned}\mathrm{KL}(q_{\mathrm{data}} \| P_{\mathrm{output}}) &= \frac{d}{2}\,\mathrm{KL}(\pi_\beta \| P_{\mathrm{pair}}) \\ &\ge \frac{d}{2}\cdot 2\,\mathrm{TV}(\pi_\beta, P_{\mathrm{pair}})^2 \\ &= \frac{d\rho^2 a^2}{4} \ge \frac{d\rho^2}{4(N+1)^2}.\end{aligned}$$

Therefore, achieving KL error at most ε requires $N \ge (\rho/2)\sqrt{d/\varepsilon} - 1$.

Complexity of first-hitting sampler (Liang et al., 2025). The first-hitting sampling reveals one coordinate per query and uses exactly d sequential queries, with no error under exact scores (Liang et al., 2026, Section 4). This establishes the claimed separation in parallel query depth.

Complexity of our partition sampler (Algorithm 1). Partition the coordinates into their interacting pairs, restricting each pair to its currently masked coordinates. This partition has block-size budget $r(t) \equiv 2$ and satisfies monotone refinement. Since conditioning on revealed coordinates preserves independence across distinct pairs, $\Gamma_{\mathrm{part}}(t) = 0$ for every $t < T$. We may therefore take $\varepsilon_{\mathrm{part}} \equiv 0$, giving $\mathcal{B}_{\mathrm{part}} = \mathcal{D}_{\mathrm{part}} = 0$.

Assume exact scores and set $T = \log(4d\log 2/\varepsilon)$. The uniform grid with $N = \lceil 2T \rceil$ satisfies $h_k r(t_k) = 2T/N \le 1$. Theorem 1 gives $\mathrm{KL}(q_{\mathrm{data}} \| p_{\mathrm{alg}}) \le de^{-T}\log 2 = \varepsilon/4$ and parallel query depth at most $2N = O(\log(d/\varepsilon))$.

□

E Proof of Theorem 2

Throughout, $G = (V, E)$, with $V = [d]$, has couplings $J_{uv} \in [0, \infty)$, and $T > 0$. For $A \subseteq V$, let $E[A]$ be the edge set of $G[A]$, and let $\mu_{G[A],g}$ denote the Ising measure on $G[A]$ with inherited couplings, external field $g \in \mathbb{R}^A$, and free boundary conditions. Write $K_A(u,v) := \mathbb{E}_{\mu_{G[A],0}}[\sigma_u\sigma_v]$ for $u, v \in A$, and $w_{uv} = w_{\{u,v\}} := a_u a_v \tanh J_{uv}$ for $\{u,v\} \in E$. For $A = \varnothing$, the measure is the point mass on the empty configuration. For reverse time $t \in [0,T)$, set $\tau = T - t$ and $W_t = X_\tau$.

We first verify the structural requirements of Assumption 2.2. The key observation is that decreasing the block-size budget and revealing additional coordinates can only refine the partition.

Proposition 2 (Structural properties of the partition oracle). *Let $\mathcal{H}$ be the merge forest returned by Algorithm 2, and let $(\mathcal{C}_q)_{q=1}^{d}$ be its associated base partitions. For every deterministic and nonincreasing block size schedule $r : [0,T) \to \{1,\dots,d\}$, the map*

$$\widehat{\mathcal{P}}_t(x) = \bigcup_{C \in \mathcal{C}_{r(t)}} \mathrm{CC}(G[C \cap M(x)])$$

is a deterministic measurable partition of $M(x)$ satisfying

$$\begin{aligned}|B| &\le r(t) \qquad \big(B \in \widehat{\mathcal{P}}_t(x)\big), \\ \widehat{\mathcal{P}}_{t_2}(y) &\preceq \widehat{\mathcal{P}}_{t_1}(x) \quad \big(t_1 \le t_2,\ x \preceq y\big).\end{aligned}$$

It depends on x only through $M(x)$ and is independent of the supplied score predictor.

Proof. We first establish the properties of the base partitions.

Fact 1 (Properties of the base partitions). *For every $q \in [d]$, $\mathcal{C}_q$ is a partition of V whose blocks have size at most q. Moreover, $\mathcal{C}_{q'} \preceq \mathcal{C}_q$ whenever $1 \le q' \le q \le d$.*

Proof of Fact 1. Recall that $S(z)$ denotes the set of leaves in the subtree rooted at z, and $n(z)$ is the count stored at z. Initially, $S(v) = \{v\}$ and $n(v) = 1$. Each merge joins disjoint components with roots z_1, z_2, so $S(z) = S(z_1) \sqcup S(z_2)$ and $n(z) = n(z_1) + n(z_2)$. Induction therefore gives $n(z) = |S(z)|$ for every node z.

The sets $S(z)$ form a laminar family, meaning that any two are disjoint or one contains the other. By definition, $\mathcal{C}_q$ consists of the inclusion-maximal sets $S(z)$ satisfying $|S(z)| \le q$. Every vertex belongs to such a set, obtained by taking its highest ancestor with count at most q. This ancestor exists because the leaf itself has count one. If two members of $\mathcal{C}_q$ intersect, laminarity and maximality force them to coincide. Thus $\mathcal{C}_q$ partitions V, and each block has size at most q.

Now let $q' \le q$ and $C' = S(z') \in \mathcal{C}_{q'}$. Since $|C'| \le q' \le q$, take the highest ancestor of z' whose count is at most q, allowing z' itself. Its leaf set belongs to $\mathcal{C}_q$ and contains C'. Hence $\mathcal{C}_{q'} \preceq \mathcal{C}_q$. □

We now apply Fact 1 to the online output. The nonempty sets $C \cap M(x)$, with $C \in \mathcal{C}_{r(t)}$, partition $M(x)$. Taking connected components partitions each of these sets, so $\widehat{\mathcal{P}}_t(x)$ also partitions $M(x)$. Every output block B lies in some $C \cap M(x)$ and therefore satisfies $|B| \le |C| \le r(t)$.

To prove refinement, fix $t_1 \le t_2$ and $x \preceq y$. Since r is nonincreasing, Fact 1 gives $\mathcal{C}_{r(t_2)} \preceq \mathcal{C}_{r(t_1)}$. Also, $M(y) \subseteq M(x)$. For any $B' \in \widehat{\mathcal{P}}_{t_2}(y)$, choose $C' \in \mathcal{C}_{r(t_2)}$ containing B' and $C \in \mathcal{C}_{r(t_1)}$ containing C'. Then

$$B' \subseteq C' \cap M(y) \subseteq C \cap M(x).$$

Since B' is connected in $G[C' \cap M(y)]$, it remains connected in $G[C \cap M(x)]$. It is therefore contained in one connected component $B \in \widehat{\mathcal{P}}_{t_1}(x)$. This proves $\widehat{\mathcal{P}}_{t_2}(y) \preceq \widehat{\mathcal{P}}_{t_1}(x)$.

Finally, fixed tie-breaking makes $\mathcal{H}$ deterministic. Since r is nonincreasing, it is Borel measurable. The state space is finite, and the output is a fixed function of the finite-valued pair $(r(t), M(x))$, so the oracle is measurable. Its dependence on x is only through $M(x)$, and its construction does not use the supplied score predictor. □

To bound the partition error, we first identify the conditional law of the masked spins. The following lemma shows that this law is a ferromagnetic Ising measure on the induced subgraph, with inherited couplings and modified external fields.

Lemma 12 (Posterior under independent masking). *Let $X_0 \sim q_0 = \mu_{G,h}$, where $h \in \mathbb{R}^V$ is finite. Suppose the forward masking indicators are mutually independent, independent of X_0, and have masking probability $1 - e^{-\tau}$ at time τ. For every $t \in [0, T)$ and masked state x, $\mathbb{P}(W_t = x) > 0$. Writing $M = M(x)$ and $U = V \setminus M$, define*

$$h_v^x := h_v + \sum_{u \in U : \{u,v\} \in E} J_{uv} x_u, \qquad \mu_x := \mu_{G[M], h^x}.$$

Then

$$\mathcal{L}(X_0^M \mid W_t = x) = \mu_x,$$
$$I(X_0^i; X_0^j \mid W_t = x) = I_{\mu_x}(\sigma_i; \sigma_j) \qquad (i, j \in M,\ i \ne j).$$

Proof. Fix (t, x), and set $M = M(x)$, $U = V \setminus M$, and $p_\tau = 1 - e^{-\tau}$. The masking indicators are mutually independent and independent of X_0. Consequently,

$$\begin{aligned} \mathbb{P}(W_t = x \mid X_0 = \sigma) &= \prod_{v \in M} p_\tau \prod_{u \in U} \big((1 - p_\tau)\mathbf{1}\{\sigma_u = x_u\}\big) \\ &= p_\tau^{|M|}(1 - p_\tau)^{|U|}\mathbf{1}\{\sigma_U = x_U\}. \end{aligned} \tag{61}$$

Since $\tau > 0$ is finite, $p_\tau \in (0,1)$. Finite couplings and fields imply $q_0(\sigma) > 0$ for every σ, so

$$\mathbb{P}(W_t = x) = p_\tau^{|M|}(1-p_\tau)^{|U|} \sum_{\zeta \in \{-1,+1\}^M} q_0(\zeta, x_U) > 0.$$

For every $\eta \in \{-1,+1\}^M$, Bayes' rule therefore gives

$$\begin{aligned}\mathbb{P}(X_0^M = \eta \mid W_t = x) &= \frac{q_0(\eta, x_U) p_\tau^{|M|}(1-p_\tau)^{|U|}}{\sum_\zeta q_0(\zeta, x_U) p_\tau^{|M|}(1-p_\tau)^{|U|}} \\ &= \frac{q_0(\eta, x_U)}{\sum_\zeta q_0(\zeta, x_U)}.\end{aligned}$$

To identify this measure, set

$$h_v^x := h_v + \sum_{u \in U : \{u,v\} \in E} J_{uv} x_u, \qquad c(x_U) := \sum_{\{u,v\} \in E[U]} J_{uv} x_u x_v + \sum_{u \in U} h_u x_u.$$

Splitting the exponent into internal, crossing, and observed terms,

$$\begin{aligned}\log q_0(\eta, x_U) = & -\log Z + c(x_U) + \sum_{\{u,v\} \in E[M]} J_{uv}\eta_u\eta_v \\ & + \sum_{v \in M} h_v \eta_v + \sum_{\substack{v \in M,\, u \in U \\ \{u,v\} \in E}} J_{uv}\eta_v x_u \\ = & -\log Z + c(x_U) + \sum_{\{u,v\} \in E[M]} J_{uv}\eta_u\eta_v + \sum_{v \in M} h_v^x \eta_v.\end{aligned}$$

The factor $e^{c(x_U)}/Z$ cancels in the preceding ratio. Thus

$$\mathcal{L}(X_0^M \mid W_t = x) = \mathcal{L}_{q_0}(\sigma_M \mid \sigma_U = x_U) = \mu_{G[M],h^x} =: \mu_x. \tag{62}$$

All entries of h^x are finite. In particular,

$$I_x(i;j) := I_{\mu_x}(\sigma_i;\sigma_j) = I(X_0^i; X_0^j \mid W_t = x) \qquad (i,j \in M,\ i \neq j). \tag{63}$$

For $M = \varnothing$, the posterior is the point mass on the empty configuration, and all subsequent information sums are zero. □

The posterior field h^x depends on the revealed spins. To control dependence uniformly over all query states, the next lemma bounds pairwise mutual information by zero-field correlations, uniformly over arbitrary external fields.

Lemma 13 (Mutual information under arbitrary external fields). *Fix $M \subseteq V$, $g \in \mathbb{R}^M$, and $\mu = \mu_{G[M],g}$. For distinct $i,j \in M$, set $Q_\pm := \mu(\sigma_i \in \cdot \mid \sigma_j = \pm 1)$. Then*

$$I_\mu(\sigma_i;\sigma_j) \le (\log 2)\,\mathrm{TV}(Q_+, Q_-) \le (\log 2) K_M(i,j).$$

Proof. Fix distinct $i,j \in M$ and write $\mu = \mu_{G[M],h'}$. We first bound mutual information by the total variation distance between the conditional laws, and then bound this distance by the zero-field correlation.

Step 1. Bounding mutual information by total variation.

We use the following fact about binary random variables.

Fact 2 (Binary mutual-information bound). *Let $X, Y \in \{-1,+1\}$ be random variables with $\mathbb{P}(Y = b) > 0$ for $b = \pm 1$. Write $Q_b = \mathcal{L}(X \mid Y = b)$ and $\delta = \mathrm{TV}(Q_+, Q_-)$. Then $I(X;Y) \le \delta H(Y)$.*

Proof of Fact 2. If $\delta = 0$, then $Q_+ = Q_-$, so X and Y are independent and $I(X;Y) = 0$. If $\delta = 1$, then $I(X;Y) = H(Y) - H(Y \mid X) \le H(Y) = \delta H(Y)$. It remains to consider $0 < \delta < 1$.

For $a \in \{-1, +1\}$, define $\nu(a) := \min\{Q_+(a), Q_-(a)\}$. The total mass of this common part is

$$\begin{aligned}\sum_a \nu(a) &= \frac{1}{2}\sum_a \big(Q_+(a) + Q_-(a) - |Q_+(a) - Q_-(a)|\big) \\ &= 1 - \delta.\end{aligned}$$

Define $R(a) := \frac{\nu(a)}{1-\delta}$ and $R_b(a) := \frac{Q_b(a)-\nu(a)}{\delta}$ for $b = \pm 1$. These functions are nonnegative, and

$$\sum_a R(a) = \frac{1-\delta}{1-\delta} = 1,$$

$$\sum_a R_b(a) = \frac{1-(1-\delta)}{\delta} = 1,$$

$$(1-\delta)R(a) + \delta R_b(a) = \nu(a) + Q_b(a) - \nu(a) = Q_b(a).$$

Thus R, R_+, R_- are probability distributions and $Q_b = (1-\delta)R + \delta R_b$.

Write $p_b := \mathbb{P}(Y = b)$. Introduce an auxiliary variable $L \in \{0, 1\}$ through the augmented law

$$\widetilde{\mathbb{P}}(X = a, Y = b, L = \ell) = p_b \begin{cases} (1-\delta)R(a), & \ell = 0, \\ \delta R_b(a), & \ell = 1. \end{cases}$$

Summing over ℓ gives

$$\begin{aligned}\widetilde{\mathbb{P}}(X = a, Y = b) &= p_b\big((1-\delta)R(a) + \delta R_b(a)\big) \\ &= p_b Q_b(a) = \mathbb{P}(X = a, Y = b).\end{aligned}$$

Hence the augmented law preserves the original joint law of (X, Y). Moreover, summing over a yields $\widetilde{\mathbb{P}}(Y = b, L = 0) = (1-\delta)p_b$ and $\widetilde{\mathbb{P}}(Y = b, L = 1) = \delta p_b$. Thus $L \sim \mathrm{Bernoulli}(\delta)$ is independent of Y. Conditional on $L = 0$, we have

$$\widetilde{\mathbb{P}}(X = a, Y = b \mid L = 0) = \frac{p_b(1-\delta)R(a)}{1-\delta} = p_b R(a),$$

so X and Y are independent under this conditioning.

All quantities involving L below are computed under the augmented law. The chain rule and nonnegativity of conditional mutual information give

$$I((X, L); Y) = I(X;Y) + I(L;Y \mid X) \ge I(X;Y).$$

Applying the chain rule in the other order,

$$\begin{aligned}I(X;Y) &\le I((X, L); Y) \\ &= I(L;Y) + I(X;Y \mid L) \\ &= (1-\delta)I(X;Y \mid L = 0) + \delta I(X;Y \mid L = 1) \\ &= \delta I(X;Y \mid L = 1) \\ &= \delta\big(H(Y \mid L = 1) - H(Y \mid X, L = 1)\big) \\ &\le \delta H(Y \mid L = 1) \\ &= \delta H(Y).\end{aligned}$$

Here $I(L;Y) = 0$ follows from independence, $I(X;Y \mid L = 0) = 0$ follows from the conditional factorization, and the inequality uses nonnegativity of conditional entropy. The final equality again uses independence of L and Y. □

Return to the Ising measure μ and set $\delta_{ij} := \mathrm{TV}(Q_+, Q_-)$. Finite couplings and fields give full support, so $\mu(\sigma_j = \pm 1) > 0$ and the fact applies. Also,

$$H_\mu(\sigma_j) = \log 2 - D_{\mathrm{KL}}\big(\mathcal{L}_\mu(\sigma_j) \| \,\mathrm{Unif}\{-1, +1\}\big) \le \log 2.$$

Consequently,

$$I_\mu(\sigma_i;\sigma_j) \le \delta_{ij} H_\mu(\sigma_j) \le (\log 2)\delta_{ij}. \tag{64}$$

Step 2. Bounding total variation by the zero-field correlation.

For any field $\widetilde{h} \in \mathbb{R}^M$, write $\langle f\rangle_{\widetilde{h}} := \mathbb{E}_{\mu_{G[M],\widetilde{h}}}[f]$. For the fixed field h', define $p_\pm := \mu(\sigma_j = \pm 1)$ and $m_\pm := \mathbb{E}_\mu[\sigma_i \mid \sigma_j = \pm 1]$. We first show that $m_+ \ge m_-$ using the following form of the FKG inequality.

Fact 3 (**FKG inequality**; Fortuin et al., 1971)**.** *Let $A \subseteq V$ and $\widetilde{h} \in \mathbb{R}^A$. For $\rho = \mu_{G[A],\widetilde{h}}$ and any coordinatewise nondecreasing functions $f_1, f_2 : \{-1,+1\}^A \to \mathbb{R}$,*

$$\mathbb{E}_\rho[f_1 f_2] \ge \mathbb{E}_\rho[f_1]\mathbb{E}_\rho[f_2]. \tag{65}$$

In particular, $\mathrm{Cov}_\rho(\sigma_u, \sigma_v) \ge 0$ *for every $u, v \in A$.*

Applying Fact 3 to μ, we obtain

$$\begin{aligned} 0 &\le \mathrm{Cov}_\mu(\sigma_i, \sigma_j) \\ &= p_+ m_+ - p_- m_- - (p_+ m_+ + p_- m_-)(p_+ - p_-) \\ &= 2p_+ p_-(m_+ - m_-), \end{aligned}$$

where the last equality uses $p_+ + p_- = 1$. Full support gives $p_+, p_- > 0$, so $m_+ \ge m_-$.

We next compare this difference with its zero-field counterpart using the following influence comparison.

Fact 4 (**Ding–Song–Sun comparison (Ding et al., 2023, Theorem 1.1)**)**.** *Consider a ferromagnetic Ising model on a finite graph with vertex set A and finite nonnegative couplings. Write $\langle\cdot\rangle_f$ for expectation under external field $f \in \mathbb{R}^A$. For any background field $h' \in \mathbb{R}^A$, nonnegative perturbation $\eta \in [0,\infty)^A$, and vertex $u \in A$,*

$$\langle\sigma_u\rangle_{h'+\eta} - \langle\sigma_u\rangle_{h'-\eta} \le \langle\sigma_u\rangle_\eta - \langle\sigma_u\rangle_{-\eta}.$$

Applying Fact 4 to $G[M]$, with background field h' and perturbation $\eta = \lambda e_j$, gives

$$\langle\sigma_i\rangle_{h'+\lambda e_j} - \langle\sigma_i\rangle_{h'-\lambda e_j} \le \langle\sigma_i\rangle_{\lambda e_j} - \langle\sigma_i\rangle_{-\lambda e_j} \qquad (\lambda \ge 0). \tag{66}$$

For $s \in \{-1,+1\}$, changing the field at j multiplies the Gibbs weight by $e^{s\lambda\sigma_j}$. Thus

$$\begin{aligned} \langle\sigma_i\rangle_{h'+s\lambda e_j} &= \frac{\langle\sigma_i e^{s\lambda\sigma_j}\rangle_{h'}}{\langle e^{s\lambda\sigma_j}\rangle_{h'}} \\ &= \frac{e^{s\lambda} p_+ m_+ + e^{-s\lambda} p_- m_-}{e^{s\lambda} p_+ + e^{-s\lambda} p_-}. \end{aligned}$$

For $s = +1$ and $s = -1$, respectively, dividing numerator and denominator by e^λ gives

$$\begin{aligned} \langle\sigma_i\rangle_{h'+\lambda e_j} &= \frac{p_+ m_+ + e^{-2\lambda} p_- m_-}{p_+ + e^{-2\lambda} p_-} \longrightarrow m_+, \\ \langle\sigma_i\rangle_{h'-\lambda e_j} &= \frac{e^{-2\lambda} p_+ m_+ + p_- m_-}{e^{-2\lambda} p_+ + p_-} \longrightarrow m_-. \end{aligned}$$

Both limits hold because $p_+, p_- > 0$.

For the right-hand side of equation 66, zero-field spin-flip symmetry gives $\langle\sigma_i\rangle_0 = \langle\sigma_j\rangle_0 = 0$. Using $e^{s\lambda\sigma_j} = \cosh\lambda + s\sigma_j \sinh\lambda$, we obtain

$$\begin{aligned} \langle\sigma_i\rangle_{s\lambda e_j} &= \frac{\langle\sigma_i(\cosh\lambda + s\sigma_j \sinh\lambda)\rangle_0}{\langle\cosh\lambda + s\sigma_j \sinh\lambda\rangle_0} \\ &= \frac{\cosh\lambda\,\langle\sigma_i\rangle_0 + s\sinh\lambda\,\langle\sigma_i\sigma_j\rangle_0}{\cosh\lambda + s\sinh\lambda\,\langle\sigma_j\rangle_0} \\ &= s K_M(i,j)\tanh\lambda. \end{aligned}$$

Hence $\langle\sigma_i\rangle_{\lambda e_j} - \langle\sigma_i\rangle_{-\lambda e_j} = 2K_M(i,j)\tanh\lambda$. Letting $\lambda \to \infty$ in equation 66, and using $m_+ \geq m_-$, therefore gives $0 \leq m_+ - m_- \leq 2K_M(i,j)$.

Finally, since $\sigma_i \in \{-1,+1\}$, $Q_\pm(+1) = \frac{1+m_\pm}{2}$. The two conditional laws have total mass one, so

$$\begin{aligned}\delta_{ij} &= \frac{1}{2}\sum_{a=\pm 1}|Q_+(a) - Q_-(a)| \\ &= |Q_+(+1) - Q_-(+1)| \\ &= \frac{1}{2}|m_+ - m_-| \\ &= \frac{1}{2}(m_+ - m_-) \\ &\leq K_M(i,j).\end{aligned} \tag{67}$$

Combining equation 64 and equation 67 proves $I_\mu(\sigma_i;\sigma_j) \leq (\log 2)\,\mathrm{TV}(Q_+,Q_-) \leq (\log 2)K_M(i,j)$. □

The preceding lemma bounds pairwise mutual information by zero-field correlations. We next use the susceptibility bounds to control the sum of these correlations across a partition by the total weight of the edges joining distinct blocks.

Lemma 14 (Weighted cut bound). *Under Assumption 4.1, fix $M \subseteq V$, a finite field $g \in \mathbb{R}^M$, and $\mu = \mu_{G[M],g}$. For any partition $\mathcal{P}$ of M, let $E_M(\mathcal{P})$ be the set of original edges whose endpoints lie in distinct blocks of $\mathcal{P}$, and define*

$$\mathcal{D}_\mu(\mathcal{P}) := \sum_{B\in\mathcal{P}}\sum_{i\in B}\sum_{j\in M\setminus B} I_\mu(\sigma_i;\sigma_j).$$

Then

$$\mathcal{D}_\mu(\mathcal{P}) \leq (\log 2)\sum_{B\in\mathcal{P}}\sum_{i\in B}\sum_{j\in M\setminus B} K_M(i,j) \leq 2(\log 2)\sum_{e\in E_M(\mathcal{P})} w_e.$$

Proof. If $|M| \leq 1$, all sums vanish, so assume $|M| \geq 2$. Recall that $w_{uv} = a_u a_v \tanh J_{uv}$.

Step 1. Reduction to zero-field correlations.

For $v \in M$, let $B(v)$ denote its block in $\mathcal{P}$. Each ordered pair of vertices in distinct blocks occurs exactly once in the definition of $\mathcal{D}_\mu(\mathcal{P})$, in the summand corresponding to the block of its first vertex. Applying Lemma 13 to each such pair gives

$$\begin{aligned}\mathcal{D}_\mu(\mathcal{P}) &= \sum_{\substack{i,j\in M\\ B(i)\neq B(j)}} I_\mu(\sigma_i;\sigma_j) \\ &= \sum_{B\in\mathcal{P}}\sum_{i\in B}\sum_{j\in M\setminus B} I_\mu(\sigma_i;\sigma_j) \\ &\leq (\log 2)\sum_{B\in\mathcal{P}}\sum_{i\in B}\sum_{j\in M\setminus B} K_M(i,j).\end{aligned} \tag{68}$$

This proves the first inequality in the lemma. It remains to bound the zero-field correlation sum by the total weight of the crossing edges.

Step 2. Susceptibility bounds on induced subgraphs.

We first verify that the given susceptibility bounds remain valid on every induced subgraph.

Fact 5 (Domain monotonicity and susceptibility). *For $A \subseteq D \subseteq V$ and $i,u \in A$, we have $0 \leq K_A(i,u) \leq K_D(i,u)$. Under Assumption 4.1, $\sum_{i\in A} K_A(i,u) \leq a_u$ for every $A \subseteq V$ and $u \in A$.*

Proof of Fact 5. Fix $A \subseteq D \subseteq V$ and $i, u \in A$. On $G[D]$, retain the couplings on $E[A]$ and multiply each coupling on $E[D] \setminus E[A]$ by $s \in [0,1]$, keeping all external fields zero. Let ρ_s denote the resulting measure.

At $s = 0$, all interactions involving vertices outside A vanish, so the marginal of ρ_0 on A is $\mu_{G[A],0}$. At $s = 1$, we have $\rho_1 = \mu_{G[D],0}$. Differentiating the finite partition sum gives

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}_{\rho_s}[\sigma_i\sigma_u] &= \sum_{\{v,w\}\in E[D]\setminus E[A]} J_{vw}\,\mathrm{Cov}_{\rho_s}(\sigma_i\sigma_u, \sigma_v\sigma_w)\\ &\geq 0.\end{aligned}$$

The inequality follows from nonnegative couplings and the second Griffiths inequality (Griffiths, 1967). Integrating over $s \in [0,1]$ yields

$$K_D(i,u) - K_A(i,u) = \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\mathbb{E}_{\rho_s}[\sigma_i\sigma_u]\,\mathrm{d}s \geq 0.$$

The first Griffiths inequality also gives $K_A(i,u) \geq 0$.

Taking $D = V$ and summing over $i \in A$, we obtain

$$\begin{aligned}\sum_{i\in A} K_A(i,u) &\leq \sum_{i\in A} K_V(i,u)\\ &\leq \sum_{i\in V} K_V(i,u)\\ &= \chi_u \leq a_u.\end{aligned}$$

The first inequality uses domain monotonicity, the second uses nonnegativity, and the last uses Assumption 4.1. □

Step 3. Bounding the contribution of one block.

For $B \subseteq M$, define its oriented boundary by $\partial_M B := \{(u,v) \in B \times (M \setminus B) : \{u,v\} \in E\}$. We use the following inequality to express correlations across B in terms of its boundary edges.

Fact 6 (Modified Simon inequality). *For $B \subseteq M$, $i \in B$, and $j \in M \setminus B$,*

$$K_M(i,j) \leq \sum_{(u,v)\in\partial_M B} K_B(i,u)\tanh J_{uv}\,K_M(v,j). \tag{69}$$

Proof of Fact 6. For an edge set F, let ∂F denote its set of odd-degree vertices. For $A \subseteq M$ and $D \subseteq A$, define

$$\mathcal{Z}_A(D) := \sum_{\substack{F\subseteq E[A]\\ \partial F = D}} \prod_{e\in F} \tanh J_e.$$

Empty products equal 1. Using $e^{J_{uv}\sigma_u\sigma_v} = \cosh J_{uv}(1 + \sigma_u\sigma_v\tanh J_{uv})$, we obtain

$$\begin{aligned}&\sum_{\sigma\in\{-1,+1\}^A} \Big(\prod_{v\in D}\sigma_v\Big)\exp\Big(\sum_{\{u,v\}\in E[A]} J_{uv}\sigma_u\sigma_v\Big)\\ &= \Big(\prod_{e\in E[A]}\cosh J_e\Big)\sum_{F\subseteq E[A]}\Big(\prod_{e\in F}\tanh J_e\Big)\prod_{v\in A}\sum_{s=\pm1} s^{\deg_F(v)+\mathbf{1}_{\{v\in D\}}}\\ &= 2^{|A|}\Big(\prod_{e\in E[A]}\cosh J_e\Big)\mathcal{Z}_A(D).\end{aligned}$$

The last equality holds because $\sum_{s=\pm1} s^k$ equals 2 for even k and 0 for odd k. Thus an edge set contributes precisely when $\partial F = D$. Dividing by the expression with $D = \varnothing$ gives

$$K_A(x,y) = \frac{\mathcal{Z}_A(\{x\}\triangle\{y\})}{\mathcal{Z}_A(\varnothing)}. \tag{70}$$

Here $\triangle$ denotes symmetric difference. The formula includes $x = y$, and the denominator is at least 1 because the empty edge set contributes 1.

Consider $F \subseteq E[M]$ and $H \subseteq E[B]$ with $\partial F = \{i, j\}$ and $\partial H = \varnothing$. Since degrees add modulo two under symmetric difference,

$$\partial(F \triangle H) = \partial F \triangle \partial H = \{i, j\}.$$

Every connected component has an even number of odd-degree vertices, so $F \triangle H$ contains a path from i to j. Fix an ordering of all simple paths from i to j and choose the first one contained in $F \triangle H$.

Let $e = \{u, v\}$ be its first edge leaving B, with $u \in B$ and $v \in M \setminus B$. Let P be the edge set of its prefix from i to u. Then $P \subseteq E[B]$, $\partial P = \{i\} \triangle \{u\}$, and $e \notin P$. If $u = i$, then $P = \varnothing$. Moreover, $e \notin H$ because $H \subseteq E[B]$, and hence $e \in F$.

Define

$$F' := (F \triangle P) \setminus \{e\}, \qquad H' := H \triangle P.$$

These sets satisfy $F' \subseteq E[M]$ and $H' \subseteq E[B]$. Since $e \in F \triangle P$, removing e toggles the degree parity at u and v, giving

$$\begin{aligned} \partial F' &= \partial F \triangle \partial P \triangle \{u, v\} \\ &= (\{i\} \triangle \{j\}) \triangle (\{i\} \triangle \{u\}) \triangle (\{u\} \triangle \{v\}) \\ &= \{v\} \triangle \{j\}, \\ \partial H' &= \partial H \triangle \partial P \\ &= \{i\} \triangle \{u\}. \end{aligned}$$

Each edge of P belongs to exactly one of F and H. Taking symmetric difference with P in both sets therefore exchanges its membership without changing the product of weights. Consequently,

$$\begin{aligned} \Big(\prod_{f \in F} \tanh J_f\Big)\Big(\prod_{f \in H} \tanh J_f\Big) &= \Big(\prod_{f \in F \triangle P} \tanh J_f\Big)\Big(\prod_{f \in H \triangle P} \tanh J_f\Big) \\ &= \tanh J_{uv} \Big(\prod_{f \in F'} \tanh J_f\Big)\Big(\prod_{f \in H'} \tanh J_f\Big). \end{aligned} \tag{71}$$

The map $(F, H) \mapsto (u, v, F', H')$ is injective. Indeed,

$$\begin{aligned} F' \triangle H' &= (F \triangle P \triangle \{e\}) \triangle (H \triangle P) \\ &= (F \triangle H) \setminus \{e\}. \end{aligned}$$

Thus the output recovers $F \triangle H$ by adding e. The fixed ordering then recovers the chosen path and its prefix P, after which

$$F = (F' \triangle P) \cup \{e\}, \qquad H = H' \triangle P.$$

Hence the original pair is uniquely determined by the output.

Summing equation 71 over all input pairs and using injectivity gives

$$\begin{aligned} &\mathcal{Z}_M(\{i, j\}) \mathcal{Z}_B(\varnothing) \\ &\quad = \sum_{\substack{F \subseteq E[M],\, \partial F = \{i,j\} \\ H \subseteq E[B],\, \partial H = \varnothing}} \Big(\prod_{f \in F} \tanh J_f\Big)\Big(\prod_{f \in H} \tanh J_f\Big) \\ &\quad \leq \sum_{(u,v) \in \partial_M B} \tanh J_{uv} \Big(\sum_{\substack{H' \subseteq E[B] \\ \partial H' = \{i\} \triangle \{u\}}} \prod_{f \in H'} \tanh J_f\Big)\Big(\sum_{\substack{F' \subseteq E[M] \\ \partial F' = \{v\} \triangle \{j\}}} \prod_{f \in F'} \tanh J_f\Big) \\ &\quad = \sum_{(u,v) \in \partial_M B} \tanh J_{uv}\, \mathcal{Z}_B(\{i\} \triangle \{u\}) \mathcal{Z}_M(\{v\} \triangle \{j\}). \end{aligned}$$

The inequality enlarges the image of the injection to all pairs satisfying the displayed parity constraints. All weights are nonnegative, so this enlargement can only increase the sum.

Dividing by $\mathcal{Z}_M(\varnothing)\mathcal{Z}_B(\varnothing) > 0$ and applying equation 70, we conclude

$$K_M(i,j) \le \sum_{(u,v)\in\partial_M B} \tanh J_{uv} \frac{\mathcal{Z}_B(\{i\}\triangle\{u\})}{\mathcal{Z}_B(\varnothing)} \frac{\mathcal{Z}_M(\{v\}\triangle\{j\})}{\mathcal{Z}_M(\varnothing)}$$
$$= \sum_{(u,v)\in\partial_M B} K_B(i,u) \tanh J_{uv}\, K_M(v,j).$$

Finally,

$$K_{\{u,v\}}(u,v) = \frac{2e^{J_{uv}} - 2e^{-J_{uv}}}{2e^{J_{uv}} + 2e^{-J_{uv}}} = \tanh J_{uv},$$

so this also proves the stated bound with $K_{\{u,v\}}(u,v)$ as the middle factor. □

Summing this inequality over both endpoints and applying Fact 5 gives the required bound for a single block.

Fact 7 (Single-block cut bound). *For every $B \subseteq M$,*

$$\sum_{i\in B} \sum_{j\in M\setminus B} K_M(i,j) \le \sum_{(u,v)\in\partial_M B} w_{uv}.$$

Proof of Fact 7. If $B = \varnothing$ or $B = M$, both sides vanish. Otherwise, for each $(u,v) \in \partial_M B$, Fact 5 gives

$$\sum_{i\in B} K_B(i,u) \le a_u,$$
$$\sum_{j\in M\setminus B} K_M(v,j) \le \sum_{j\in M} K_M(v,j)$$
$$= \sum_{j\in M} K_M(j,v) \le a_v.$$

In the second estimate, the first inequality uses nonnegativity, the equality uses symmetry, and the last inequality applies the fact with $A = M$.

Summing equation 69 over $i \in B$ and $j \in M \setminus B$ and interchanging the finite sums, we obtain

$$\begin{aligned}\sum_{i\in B}\sum_{j\in M\setminus B} K_M(i,j) &\le \sum_{i\in B}\sum_{j\in M\setminus B}\sum_{(u,v)\in\partial_M B} K_B(i,u)\tanh J_{uv}\, K_M(v,j)\\ &= \sum_{(u,v)\in\partial_M B} \tanh J_{uv}\Big(\sum_{i\in B} K_B(i,u)\Big)\Big(\sum_{j\in M\setminus B} K_M(v,j)\Big)\\ &\le \sum_{(u,v)\in\partial_M B} a_u a_v \tanh J_{uv}\\ &= \sum_{(u,v)\in\partial_M B} w_{uv}.\end{aligned} \tag{72}$$

The factorization separates the sums over i and j. All factors are nonnegative, so the two endpoint estimates can be multiplied. □

Step 4. Summing over all blocks.

Each edge $\{u,v\} \in E_M(\mathcal{P})$ appears in exactly two oriented boundaries. It appears as (u,v) in $\partial_M B(u)$ and as (v,u) in $\partial_M B(v)$. An edge whose endpoints belong to the same block appears in none of these boundaries. Since $w_{uv} = w_{vu}$, it follows that

$$\sum_{B\in\mathcal{P}}\sum_{(u,v)\in\partial_M B} w_{uv} = \sum_{\{u,v\}\in E_M(\mathcal{P})} (w_{uv} + w_{vu})$$
$$= 2\sum_{e\in E_M(\mathcal{P})} w_e.$$

Combining this identity with equation 68 and Fact 7 yields

$$\begin{aligned}\mathcal{D}_\mu(\mathcal{P}) &\le (\log 2) \sum_{B\in\mathcal{P}} \sum_{i\in B} \sum_{j\in M\setminus B} K_M(i,j) \\ &\le (\log 2) \sum_{B\in\mathcal{P}} \sum_{(u,v)\in\partial_M B} w_{uv} \\ &= 2(\log 2) \sum_{e\in E_M(\mathcal{P})} w_e.\end{aligned} \tag{73}$$

This proves the lemma. □

Proposition 3 (Information-rate certificate). *Under Assumption 4.1, consider the oracle of Proposition 2 and the masking process of Lemma 12. Let $F(q)$ be the set of original edges joining distinct blocks of $\mathcal{C}_q$, and put $b(q) := \sum_{e\in F(q)} w_e$. For every $t \in [0,T)$,*

$$\Gamma_{\mathrm{part}}(t) \le \varepsilon_{\mathrm{part}}(t) := 2(\log 2) e^{-2(T-t)} b(r(t)).$$

The certificate is measurable and satisfies

$$\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t \le (\log 2)(1-e^{-2T}) \sum_{e\in E} w_e < \infty.$$

For $r(t)\equiv q$, its integral equals $(\log 2)(1-e^{-2T})b(q)$.

Proof. Fix (t,x), and write $M = M(x)$, $q = r(t)$, and $\mathcal{P} = \widehat{\mathcal{P}}_t(x)$. Let μ_x be the posterior from Lemma 12, and define

$$\mathcal{D}_t(x) := \mathcal{D}_{\mu_x}(\widehat{\mathcal{P}}_t(x)).$$

For each $v \in M$, let $B(v) \in \mathcal{P}$ and $C(v) \in \mathcal{C}_q$ be its output and base blocks, respectively. For any original edge $\{u,v\}$ with $u,v \in M$,

$$\begin{aligned} C(u) \neq C(v) &\implies B(u) \neq B(v), \\ C(u) = C(v) = C &\implies \{u,v\} \in E(G[C\cap M]) \implies B(u) = B(v). \end{aligned}$$

The first implication uses that each output block lies in one base block and the second uses the original edge as a connecting path. These two implications prove both inclusions in

$$E_M(\widehat{\mathcal{P}}_t(x)) = \{e \in F(r(t)) : e \subseteq M(x)\}.$$

Applying Lemma 14 with $g = h^x$ therefore gives

$$\mathcal{D}_t(x) \le 2(\log 2) \sum_{e\in F(r(t))} w_e\, \mathbf{1}\{e \subseteq M(x)\}. \tag{74}$$

The hierarchy and schedule are fixed before the query, so $F(r(t))$ and all w_e are deterministic. For every $e = \{u,v\} \in E$, the endpoints are distinct, and independence of their masking indicators gives

$$\begin{aligned}\mathbb{P}(e \subseteq M(W_t)) &= \mathbb{P}(u \in M(W_t),\, v \in M(W_t)) \\ &= \mathbb{P}(u \in M(W_t))\mathbb{P}(v \in M(W_t)) \\ &= (1-e^{-\tau})^2.\end{aligned}$$

Taking expectations and using linearity,

$$\begin{aligned}\mathbb{E}[\mathcal{D}_t(W_t)] &\le 2(\log 2)\, \mathbb{E}\Big[\sum_{e\in F(r(t))} w_e \mathbf{1}\{e \subseteq M(W_t)\}\Big] \\ &= 2(\log 2) \sum_{e\in F(r(t))} w_e\, \mathbb{P}(e \subseteq M(W_t)) \\ &= 2(\log 2)(1-e^{-\tau})^2 \sum_{e\in F(r(t))} w_e \\ &= 2(\log 2)(1-e^{-\tau})^2 b(r(t)).\end{aligned}$$

By Lemma 12, $\mathcal{D}_t$ is exactly the ordered information sum in the rate definition. Since $\tau > 0$, its denominator is positive, and

$$\begin{aligned}\Gamma_{\mathrm{part}}(t) &= \frac{\mathbb{E}[\mathcal{D}_t(W_t)]}{(e^\tau - 1)^2} \\ &\le 2(\log 2)\frac{(1-e^{-\tau})^2}{(e^\tau-1)^2} b(r(t)) \\ &= 2(\log 2)\frac{e^{-2\tau}(e^\tau-1)^2}{(e^\tau-1)^2} b(r(t)) \\ &= 2(\log 2) e^{-2(T-t)} b(r(t)) \\ &= \varepsilon_{\mathrm{part}}(t).\end{aligned} \tag{75}$$

This holds for every $t < T$, and hence for almost every t.

Let $W := \sum_{e\in E} w_e < \infty$. Nonnegative weights and $F(q) \subseteq E$ give $0 \le b(q) \le W$. Moreover,

$$\varepsilon_{\mathrm{part}}(t) = 2(\log 2) e^{-2(T-t)} \sum_{q=1}^{d} b(q)\mathbf{1}\{r(t) = q\}$$

is measurable because r is measurable. Substituting $s = T - t$,

$$\begin{aligned}\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t &\le 2(\log 2) W \int_0^T e^{-2(T-t)}\,\mathrm{d}t \\ &= 2(\log 2) W \int_0^T e^{-2s}\,\mathrm{d}s \\ &= 2(\log 2) W \frac{1-e^{-2T}}{2} \\ &= (\log 2)(1-e^{-2T}) W \\ &< \infty.\end{aligned} \tag{76}$$

For $r(t) \equiv q$, the same calculation with $b(q)$ in place of W gives equality: $\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t = (\log 2)(1-e^{-2T})b(q)$. □

Proposition 4 (Computational complexity). *Let $m = |E|$. The merge forest and original graph can be stored using $O(d+m)$ space and constructed in $O(d+m\log(m+1))$ time. If the schedule r of Proposition 2 is supplied by its breakpoint table, each partition query takes $O(d+m)$ time. Given the susceptibility bounds, $b(q)$ can be evaluated for any fixed q in $O(d+m)$ time. These bounds use unit-cost arithmetic, comparisons, and weight evaluation.*

Proof. There are at most $d-1$ successful merges, so $|\mathcal{H}| \le 2d-1$. A disjoint-set component has at most $m+1$ vertices, since connecting s vertices requires at least $s-1$ edges. Union by size therefore gives $O(\log(m+2))$ time per find. Initializing the vertices, sorting the edges, and processing the merges takes $O(d+m\log(m+1))$ time. The graph, forest, and disjoint-set arrays occupy $O(d+m)$ space.

The schedule has at most $d-1$ strict drops, so binary search evaluates $r(t)$ in $O(\log(d+1))$ time. Selecting tree nodes and labeling their leaves takes $O(d)$ time, since the selected subtrees are disjoint. DFS restricted to masked vertices with matching base-block labels computes the output components, with total scan cost

$$O\Big(d + \sum_{v\in M(x)} \deg_G(v)\Big) = O(d+m).$$

This also bounds the total query time.

Finally, compute the base-block labels $\ell_q(v)$ in $O(d)$ time and scan the edges to evaluate

$$b(q) = \sum_{\{u,v\}\in E} w_{uv}\mathbf{1}\{\ell_q(u) \ne \ell_q(v)\}.$$

The total cost is $O(d+m)$. □

Proof of Theorem 2. Proposition 2 gives determinism, measurability, the block-size bound, and monotone refinement. Under Assumption 4.1, Proposition 3 gives $\Gamma_{\mathrm{part}}(t) \le \varepsilon_{\mathrm{part}}(t)$ for every $t < T$, with a measurable integrable certificate. These are precisely the three requirements of Assumption 2.2. Since the oracle is independent of $\widehat{s}$, the conclusion holds for every supplied score predictor. The computational bounds follow from Proposition 4. □

F Proofs for the masked-language-model oracle

We use the network, grid, and oracle notation of the main text. Throughout, $\widehat{p}_{k,i}$ denotes the softmax of the logits for every input state and position. A masked state x is supported when $\mathbb{P}(X_0^{U(x)} = x^{U(x)}) > 0$. For supported x, $i \in M(x)$, and $a \in \mathcal{V}$, define

$$p_i(a \mid x) \coloneqq \mathbb{P}(X_0^i = a \mid X_0^{U(x)} = x^{U(x)}). \tag{77}$$

The posterior prediction error is $e_i^k(x) \coloneqq D_{\mathrm{KL}}(p_i(\cdot \mid x) \| \widehat{p}_{k,i}(\cdot \mid x))$. Independent masking makes the posterior independent of time. Posterior quantities are evaluated only on supported states, and zero-probability revelations are omitted. We extend $e_i^k(x)$ by zero to unsupported states and coordinates outside $M(x)$. For supported x and distinct $i, j \in M(x)$, write $I_x(i;j) \coloneqq I(X_0^i; X_0^j \mid X_0^{U(x)} = x^{U(x)})$. For $\tau > 0$, put $\lambda_\tau \coloneqq (e^\tau - 1)^{-1}$. For the matrices of Lemma 1, put $P \coloneqq A_L \cdots A_1$ and $\gamma \coloneqq 2D_{\mathrm{emb}} R_{\mathrm{out}}$. These posterior quantities and the matrix P enter only the analysis.

For a map F on a prescribed input domain, a radius ρ_F and a nonnegative matrix C_F certify $\max_i \|F(Z)^i\|_2 \le \rho_F$ and $\|F(Z)^i - F(Z')^i\|_2 \le \sum_j (C_F)_{ij} \|Z^j - Z'^j\|_2$ for all inputs Z, Z' in that domain.

F.1 Computable Network Sensitivity Bounds

We prove Lemma 1 for the following class of network architecture.

For each k and ℓ, the map $F_{k,\ell} : (\mathbb{R}^H)^d \to (\mathbb{R}^H)^d$ is represented by a finite directed acyclic graph. Suppressing k, ℓ, number its nodes $0, \dots, s$ in topological order. Node v carries an array $Z^{(v)} \in (\mathbb{R}^{p_v})^d$, where $p_v \ge 1$ and $p_0 = p_s = H$. The computation is

$$Z^{(0)} = Z, \qquad Z^{(v)} = \Phi_v\big((Z^{(u)})_{u \in \mathrm{pa}(v)}\big), \qquad F_{k,\ell}(Z) = Z^{(s)},$$

where $\mathrm{pa}(v)$ is an ordered list of nodes preceding v. Each gate Φ_v is one of the following operations, with compatible input and output dimensions.

1. *Tokenwise affine maps.* For $Z \in (\mathbb{R}^p)^d$,

$$\mathsf{A}(Z)^i = W_i Z^i + b_i, \qquad W_i \in \mathbb{R}^{p' \times p}, \quad b_i \in \mathbb{R}^{p'}, \quad i \in [d].$$

   Shared parameters correspond to $W_i = W$ and $b_i = b$. Constant maps are obtained by taking $W_i = 0$.

2. *Coordinatewise activations.* For a scalar function $\varphi : \mathbb{R} \to \mathbb{R}$ and $Z \in (\mathbb{R}^p)^d$,

$$\mathsf{A}_\varphi(Z)^i = \big(\varphi(Z_1^i), \dots, \varphi(Z_p^i)\big).$$

   As required by Assumption 4.2, φ has a known finite Lipschitz constant L_φ, so $|\varphi(a) - \varphi(b)| \le L_\varphi |a - b|$ for all $a, b \in \mathbb{R}$.

3. *Addition and elementwise multiplication.* For $Z, Y \in (\mathbb{R}^p)^d$,

$$\mathsf{S}(Z,Y)^i = Z^i + Y^i, \qquad \mathsf{P}(Z,Y)^i = Z^i \odot Y^i,$$

   where $(Z^i \odot Y^i)_a = Z_a^i Y_a^i$. Residual connections are instances of addition.

4. *Tokenwise concatenation.* For $Z \in (\mathbb{R}^p)^d$ and $Y \in (\mathbb{R}^{p'})^d$,

$$\mathsf{C}(Z,Y)^i = \begin{pmatrix} Z^i \\ Y^i \end{pmatrix} \in \mathbb{R}^{p+p'}.$$

   This operation is the sum of two fixed affine embeddings into disjoint coordinate sets.

5. *Layer normalization and root mean square normalization.* For $Z \in (\mathbb{R}^p)^d$, define

$$\mathsf{N}(Z)^i = g \odot \frac{\Pi Z^i}{\sqrt{\epsilon + \frac{\|\Pi Z^i\|_2^2}{p}}} + \beta, \qquad g, \beta \in \mathbb{R}^p, \quad \epsilon > 0.$$

Layer normalization uses $\Pi = I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p}$, whereas root mean square normalization uses $\Pi = I_p$.

6. *Dot-product softmax attention.* For an input $Z \in (\mathbb{R}^p)^d$, write $\widetilde{z}_i = ((Z^i)^\top, 1)^\top$. An attention head has projection matrices

$$Q, K \in \mathbb{R}^{h_{\mathrm{head}} \times (p+1)}, \qquad V \in \mathbb{R}^{p' \times (p+1)},$$

whose last columns encode the projection biases. Its queries, keys, and values are

$$q_i = R_i Q \widetilde{z}_i, \qquad k_i = R_i K \widetilde{z}_i, \qquad v_i = V \widetilde{z}_i, \qquad R_i^\top R_i = I_{h_{\mathrm{head}}}.$$

Taking $R_i = I_{h_{\mathrm{head}}}$ gives unrotated projections.

For each i, fix a nonempty support $\mathcal{N}_i \subseteq [d]$ and finite biases $b_{ij} \in \mathbb{R}$ for $j \in \mathcal{N}_i$. Define

$$\zeta_{ij}(Z) = \frac{\langle q_i, k_j \rangle}{\sqrt{h_{\mathrm{head}}}} + b_{ij}, \qquad j \in \mathcal{N}_i,$$

$$\alpha_{ij}(Z) = \frac{\exp(\zeta_{ij}(Z))}{\sum_{a \in \mathcal{N}_i} \exp(\zeta_{ia}(Z))}, \quad j \in \mathcal{N}_i,$$

$$\mathsf{Att}(Z)^i = \sum_{j \in \mathcal{N}_i} \alpha_{ij}(Z) v_j.$$

Set $\alpha_{ij}(Z) = 0$ for $j \notin \mathcal{N}_i$. The denominator is positive because $\mathcal{N}_i$ is nonempty and all scores are finite. The supports are fixed, whereas the attention weights depend on Z.

Multihead attention is obtained by combining these operations. For heads $\mathsf{Att}_1, \ldots, \mathsf{Att}_n$ with output dimensions $p_1, \ldots, p_n$, respectively,

$$\mathsf{MHA}(Z)^i = W_{\mathrm{O}} \begin{pmatrix} \mathsf{Att}_1(Z)^i \\ \vdots \\ \mathsf{Att}_n(Z)^i \end{pmatrix} + b_{\mathrm{O}}, \qquad W_{\mathrm{O}} \in \mathbb{R}^{p_{\mathrm{out}} \times \sum_{a=1}^n p_a}, \quad b_{\mathrm{O}} \in \mathbb{R}^{p_{\mathrm{out}}}.$$

Each head may have its own projections, supports, biases, and rotations.

All computation graphs, gate parameters, attention supports, and rotations are supplied to preprocessing. They may depend on the grid index, layer, and gate, but are fixed independently of x and the intermediate activations. All numerical parameters are finite, and the normalization stabilizers are known and positive. For each scalar activation, preprocessing is supplied with L_φ and a finite upper bound on $|\varphi(0)|$.

Remark 6. *The permitted operations include Transformer self-attention blocks (Vaswani et al., 2017), gated feed-forward layers (Shazeer, 2020), root mean square normalization (Zhang & Sennrich, 2019), and rotary position embeddings (Su et al., 2024). They accommodate the deterministic logit networks of Masked Diffusion Language Models (MDLM) and the original dense Large Language Diffusion with mAsking (LLaDA) in inference mode (Sahoo et al., 2024; Nie et al., 2025). Both full and fixed sparse attention supports are permitted and the attention weights themselves may depend on the input. Input-dependent discrete routing is outside the specified gate class. Coverage of the architecture does not establish the score-error assumption or guarantee a small partition certificate.*

Remark 7. *Our construction uses established ideas from compositional Lipschitz estimation (Virmaux & Scaman, 2018), Jacobian certification (Zhang et al., 2019), and bound propagation on computational graphs (Xu et al., 2020; Shi et al., 2022). Certified analyses of Transformer computations also account for cross-position dependencies (Shi et al., 2020). Since dot-product attention need not be globally Lipschitz on unbounded domains (Kim et al., 2021), we construct bounded domains containing all reachable hidden states. This is consistent with bounded-domain and local analyses of attention (Castin et al., 2024; Yudin et al., 2025).*

For masked diffusion, prediction stability has also been used to motivate token freezing (Oba et al., 2026). Here we construct nonnegative sensitivity matrices uniformly over all masked inputs and grid times. We use these matrices to define graph weights controlling prediction sensitivity across partition cuts, and combine this control with score accuracy to bound the cross-block information rate. The certificates are finite and computable, but their magnitude depends on the network parameters.

We propagate both output radii and positionwise sensitivity bounds through the computation graph. The radii specify bounded domains for subsequent operations, while the sensitivity bounds quantify how perturbations propagate between token positions.

For a map F on a prescribed input domain, a pair (ρ_F, C_F), with $\rho_F \geq 0$ and $C_F \in \mathbb{R}_+^{d\times d}$, is a certificate if

$$\|F(Z)^i\|_2 \leq \rho_F,$$
$$\|F(Z)^i - F(Z')^i\|_2 \leq \sum_{j=1}^{d} (C_F)_{ij} \|Z^j - Z'^j\|_2$$

for all inputs Z, Z' in that domain and all $i \in [d]$. The following lemma supplies certificates for tokenwise operations and rules for combining them.

Lemma 15 (Tokenwise bounds and certificate calculus). *For inputs $z \in \mathbb{R}^p$ with $\|z\|_2 \leq R$, the following maps have the indicated output radius ρ and Euclidean Lipschitz constant c:*

map	ρ	c
$Wz + b$	$\|W\|_{\mathrm{F}} R + \|b\|_2$	$\|W\|_{\mathrm{F}}$
$\varphi(z)$	$c_\varphi R + \sqrt{p}\,\|\varphi(0)\|$	c_φ
$\dfrac{g \odot \Pi z}{\sqrt{\epsilon + \frac{\|\Pi z\|_2^2}{p}}} + \beta$	$\sqrt{p}\,\|g\|_\infty + \|\beta\|_2$	$\dfrac{\|g\|_\infty}{\sqrt{\epsilon}}$

Here φ acts coordinatewise and satisfies $|\varphi(s) - \varphi(t)| \leq c_\varphi |s - t|$; $g, \beta \in \mathbb{R}^p$, $\epsilon > 0$, and $\Pi \in \{I_p, I_p - \frac{1}{p}\mathbf{1}\mathbf{1}^\top\}$. Applying a tokenwise map with Lipschitz constant c at every position gives the positionwise certificate cI_d.

Suppose F and G have certificates (ρ_F, C_F) and (ρ_G, C_G). The following are valid certificates:

map	*output radius*	*positionwise certificate*
$G \circ F$	ρ_G	$C_G C_F$
$F + G$	$\rho_F + \rho_G$	$C_F + C_G$
$F \odot G$	$\rho_F \rho_G$	$\rho_G C_F + \rho_F C_G$

For composition, the image of F must lie in the certified domain of G. For addition and elementwise multiplication, both maps must be certified on a common input domain and have matching output dimensions.

Tokenwise concatenation of F and G on a common input domain has output radius $\sqrt{\rho_F^2 + \rho_G^2}$ and positionwise certificate C with entries $C_{ij} = \sqrt{(C_F)_{ij}^2 + (C_G)_{ij}^2}$.

Proof. For any vector v of the appropriate dimension, Cauchy–Schwarz gives

$$\|Wv\|_2^2 = \sum_a \Big(\sum_b W_{ab} v_b\Big)^2 \leq \sum_a \Big(\sum_b W_{ab}^2\Big) \|v\|_2^2 = \|W\|_{\mathrm{F}}^2 \|v\|_2^2.$$

Applying this bound to $v = z - z'$ yields $\|(Wz + b) - (Wz' + b)\|_2 \leq \|W\|_{\mathrm{F}} \|z - z'\|_2$. For $\|z\|_2 \leq R$, the triangle inequality also gives $\|Wz + b\|_2 \leq \|W\|_{\mathrm{F}} R + \|b\|_2$.

Since φ acts coordinatewise, its scalar Lipschitz bound implies

$$\|\varphi(z) - \varphi(z')\|_2^2 = \sum_{a=1}^{p} |\varphi(z_a) - \varphi(z'_a)|^2 \leq c_\varphi^2 \|z - z'\|_2^2,$$
$$\begin{aligned} \|\varphi(z)\|_2 &\leq \|\varphi(z) - \varphi(0)\mathbf{1}\|_2 + \|\varphi(0)\mathbf{1}\|_2 \\ &\leq c_\varphi R + \sqrt{p}\,|\varphi(0)|. \end{aligned}$$

This proves the activation bounds.

For the normalization map, put $N(y) = y/\sqrt{a(y)}$, where $a(y) = \epsilon + \frac{\|y\|_2^2}{p}$. Differentiation gives

$$DN(y) = \frac{I_p}{\sqrt{a(y)}} - \frac{yy^\top}{p\,a(y)^{3/2}}.$$

By Cauchy–Schwarz, $0 \preceq yy^\top \preceq \|y\|_2^2 I_p$. Consequently,

$$\frac{\epsilon}{a(y)^{3/2}} I_p = \Big(\frac{1}{\sqrt{a(y)}} - \frac{\|y\|_2^2}{p\,a(y)^{3/2}}\Big) I_p \preceq DN(y) \preceq \frac{I_p}{\sqrt{a(y)}}.$$

Thus $DN(y)$ is symmetric positive semidefinite and $\|DN(y)\|_{\mathrm{op}} \le \frac{1}{\sqrt{\epsilon}}$. Integrating along the segment between y' and y yields

$$\begin{aligned}\|N(y) - N(y')\|_2 &= \Big\| \int_0^1 DN(y' + s(y-y'))(y-y')\,\mathrm{d}s \Big\|_2 \\ &\le \int_0^1 \|DN(y' + s(y-y'))\|_{\mathrm{op}} \|y-y'\|_2\,\mathrm{d}s \\ &\le \frac{1}{\sqrt{\epsilon}} \|y-y'\|_2.\end{aligned}$$

Moreover, $\|N(y)\|_2^2 = \frac{\|y\|_2^2}{\epsilon + \frac{\|y\|_2^2}{p}} \le p$. Both choices of Π are orthogonal projections, so $\|\Pi v\|_2 \le \|v\|_2$. Using also $\|g \odot v\|_2 \le \|g\|_\infty \|v\|_2$, we obtain

$$\begin{aligned}\|g \odot N(\Pi z) + \beta\|_2 &\le \|g\|_\infty \|N(\Pi z)\|_2 + \|\beta\|_2 \\ &\le \sqrt{p}\,\|g\|_\infty + \|\beta\|_2, \\ \|g \odot N(\Pi z) - g \odot N(\Pi z')\|_2 &\le \frac{\|g\|_\infty}{\sqrt{\epsilon}} \|\Pi(z - z')\|_2 \\ &\le \frac{\|g\|_\infty}{\sqrt{\epsilon}} \|z - z'\|_2.\end{aligned}$$

The bias cancels when taking differences. This proves the normalization bounds. Applying any tokenwise Lipschitz bound separately at each position gives the certificate cI_d, since the output at position i depends only on the input at position i.

We now verify the composition rules. Fix admissible inputs Z, Z' and write $\delta_j = \|Z^j - Z'^j\|_2$. For composition, the domain condition permits applying the certificate of G to $F(Z)$ and $F(Z')$. Hence

$$\begin{aligned}\|(G \circ F)(Z)^i - (G \circ F)(Z')^i\|_2 &\le \sum_{a=1}^d (C_G)_{ia} \|F(Z)^a - F(Z')^a\|_2 \\ &\le \sum_{a=1}^d (C_G)_{ia} \sum_{j=1}^d (C_F)_{aj} \delta_j \\ &= \sum_{j=1}^d (C_G C_F)_{ij} \delta_j.\end{aligned}$$

Its output radius is ρ_G, because $F(Z)$ lies in the certified domain of G.

For addition, assume that F and G are certified on a common input domain and have matching output dimensions. The triangle inequality gives

$$\begin{aligned}\|(F+G)(Z)^i\|_2 &\le \|F(Z)^i\|_2 + \|G(Z)^i\|_2 \le \rho_F + \rho_G, \\ \|(F+G)(Z)^i - (F+G)(Z')^i\|_2 &\le \|F(Z)^i - F(Z')^i\|_2 + \|G(Z)^i - G(Z')^i\|_2 \\ &\le \sum_{j=1}^d (C_F + C_G)_{ij} \delta_j.\end{aligned}$$

For elementwise multiplication, observe that $\|a \odot b\|_2 \le \|a\|_2\|b\|_2$, since $\sum_s a_s^2 b_s^2 \le (\sum_s a_s^2)(\sum_s b_s^2)$. It follows that $\|(F \odot G)(Z)^i\|_2 \le \rho_F \rho_G$. For the difference, use the identity

$$\begin{aligned}&(F \odot G)(Z)^i - (F \odot G)(Z')^i \\ &\quad = (F(Z)^i - F(Z')^i) \odot G(Z)^i + F(Z')^i \odot (G(Z)^i - G(Z')^i).\end{aligned}$$

Applying the product bound to each term yields

$$\begin{aligned}&\|(F \odot G)(Z)^i - (F \odot G)(Z')^i\|_2 \\ &\quad \le \|F(Z)^i - F(Z')^i\|_2 \|G(Z)^i\|_2 + \|F(Z')^i\|_2 \|G(Z)^i - G(Z')^i\|_2 \\ &\quad \le \rho_G \sum_{j=1}^{d} (C_F)_{ij}\delta_j + \rho_F \sum_{j=1}^{d} (C_G)_{ij}\delta_j \\ &\quad = \sum_{j=1}^{d} (\rho_G C_F + \rho_F C_G)_{ij}\delta_j.\end{aligned}$$

Finally, let $H(Z)^i = (F(Z)^i, G(Z)^i)$ denote tokenwise concatenation. Its output radius follows from $\|H(Z)^i\|_2^2 = \|F(Z)^i\|_2^2 + \|G(Z)^i\|_2^2 \le \rho_F^2 + \rho_G^2$. For its sensitivity, nonnegativity of the certificate entries and the triangle inequality in $\mathbb{R}^2$ give

$$\begin{aligned}\|H(Z)^i - H(Z')^i\|_2 &= \sqrt{\|F(Z)^i - F(Z')^i\|_2^2 + \|G(Z)^i - G(Z')^i\|_2^2} \\ &\le \Big\| \sum_{j=1}^{d} \begin{pmatrix} (C_F)_{ij} \\ (C_G)_{ij} \end{pmatrix} \delta_j \Big\|_2 \\ &\le \sum_{j=1}^{d} \sqrt{(C_F)_{ij}^2 + (C_G)_{ij}^2}\, \delta_j.\end{aligned}$$

This proves the claimed concatenation certificate. □

We next bound the sensitivity of an attention head using uniform upper bounds on its attention weights.

Lemma 16 (Attention certificate). *Fix $R \ge 0$ and consider the domain $\max_j \|Z^j\|_2 \le R$. Put $\widetilde{R} = \sqrt{R^2+1}$ and $g = \frac{\|Q\|_{\mathrm{F}}\|K\|_{\mathrm{F}}}{\sqrt{h_{\mathrm{head}}}}$. Let $\overline{\alpha}_{ij}$ be finite nonnegative upper bounds on $\alpha_{ij}(Z)$ throughout this domain, with $\overline{\alpha}_{ij} = 0$ outside $\mathcal{N}_i$.*

Then the attention head Att *has output radius $\|V\|_{\mathrm{F}}\widetilde{R}$ and positionwise certificate*

$$C_{ij}^{\mathrm{att}} = \|V\|_{\mathrm{F}}\big((1 + 2g\widetilde{R}^2)\overline{\alpha}_{ij} + g\widetilde{R}^2 \mathbf{1}\{i = j\}\big).$$

Such attention-weight bounds can be computed directly from the head parameters and R.

Proof. We first construct uniform upper bounds on the attention weights. Since $\|\widetilde{z}_j\|_2 \le \widetilde{R}$ and the rotations are orthogonal,

$$\|q_i\|_2 \le \|Q\|_{\mathrm{F}}\widetilde{R}, \qquad \|k_j\|_2 \le \|K\|_{\mathrm{F}}\widetilde{R}, \qquad \|v_j\|_2 \le \|V\|_{\mathrm{F}}\widetilde{R}.$$

Cauchy–Schwarz therefore gives

$$\frac{|\langle q_i, k_j\rangle|}{\sqrt{h_{\mathrm{head}}}} \le \frac{\|Q\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\widetilde{R}^2}{\sqrt{h_{\mathrm{head}}}} = g\widetilde{R}^2.$$

Thus $b_{ij} - g\widetilde{R}^2 \le \zeta_{ij}(Z) \le b_{ij} + g\widetilde{R}^2$. For $j \in \mathcal{N}_i$, it follows that

$$\begin{aligned}\alpha_{ij}(Z) &= \frac{1}{1 + \sum_{a \in \mathcal{N}_i \setminus \{j\}} \exp(\zeta_{ia}(Z) - \zeta_{ij}(Z))} \\ &\le \frac{1}{1 + \sum_{a \in \mathcal{N}_i \setminus \{j\}} \exp(b_{ia} - b_{ij} - 2g\widetilde{R}^2)} =: \overline{\alpha}_{ij}.\end{aligned}$$

Set $\overline{\alpha}_{ij} = 0$ outside $\mathcal{N}_i$. These bounds are finite and computable from the head parameters and R. The remaining argument applies to any uniform upper bounds satisfying the lemma.

The attention weights are nonnegative and sum to one. Consequently,

$$\|\mathsf{Att}(Z)^i\|_2 \le \sum_{j\in\mathcal{N}_i} \alpha_{ij}(Z)\|v_j\|_2 \le \|V\|_{\mathrm{F}}\widetilde{R},$$

which proves the output-radius bound.

To prove the sensitivity bound, fix two admissible inputs Z, Z' and consider $Z(s) = Z' + s(Z - Z')$ for $s \in [0, 1]$. This segment stays in the input domain, since $\|Z(s)^j\|_2 \le (1-s)\|Z'^j\|_2 + s\|Z^j\|_2 \le R$. Write $\delta_j = \|Z^j - Z'^j\|_2$ and use a dot to denote differentiation with respect to s. Augmentation gives $\dot{\widetilde{z}}_j = ((Z^j - Z'^j)^\top, 0)^\top$, so $\|\dot{\widetilde{z}}_j\|_2 = \delta_j$ and

$$\|\dot{q}_i\|_2 \le \|Q\|_{\mathrm{F}}\delta_i, \qquad \|\dot{k}_j\|_2 \le \|K\|_{\mathrm{F}}\delta_j, \qquad \|\dot{v}_j\|_2 \le \|V\|_{\mathrm{F}}\delta_j.$$

Fix an output position i. For the following calculations, suppress the dependence on s, write $f_i = \mathsf{Att}(Z(s))^i$, and take all sums over $\mathcal{N}_i$ unless otherwise indicated. Differentiating the scores and softmax weights gives

$$\dot{\zeta}_{ij} = \frac{\langle \dot{q}_i, k_j\rangle + \langle q_i, \dot{k}_j\rangle}{\sqrt{h_{\mathrm{head}}}},$$
$$\dot{\alpha}_{ij} = \alpha_{ij}\Big(\dot{\zeta}_{ij} - \sum_a \alpha_{ia}\dot{\zeta}_{ia}\Big).$$

The score biases contribute no derivative because they are fixed. Define $\bar{k}_i = \sum_j \alpha_{ij}k_j$ and $C_i^{vk} = \sum_j \alpha_{ij}(v_j - f_i)(k_j - \bar{k}_i)^\top$. Since $\sum_j \alpha_{ij}(v_j - f_i) = 0$, we have $\sum_j \alpha_{ij}(v_j - f_i)k_j^\top = C_i^{vk}$. It follows that

$$\begin{aligned}
\dot{f}_i &= \sum_j \alpha_{ij}\dot{v}_j + \sum_j \dot{\alpha}_{ij}v_j \\
&= \sum_j \alpha_{ij}\dot{v}_j + \sum_j \alpha_{ij}(v_j - f_i)\dot{\zeta}_{ij} \\
&= \sum_j \alpha_{ij}\dot{v}_j + \frac{1}{\sqrt{h_{\mathrm{head}}}}\sum_j \alpha_{ij}(v_j - f_i)\langle q_i, \dot{k}_j\rangle + \frac{C_i^{vk}\dot{q}_i}{\sqrt{h_{\mathrm{head}}}}.
\end{aligned}$$

These three terms account for changes in the values, keys, and query, respectively.

We bound the centered matrix C_i^{vk} using weighted Cauchy–Schwarz. Expanding the squares and using $\sum_j \alpha_{ij}v_j = f_i$ and $\sum_j \alpha_{ij}k_j = \bar{k}_i$ gives

$$\sum_j \alpha_{ij}\|v_j - f_i\|_2^2 = \sum_j \alpha_{ij}\|v_j\|_2^2 - \|f_i\|_2^2 \le \|V\|_{\mathrm{F}}^2\widetilde{R}^2,$$
$$\sum_j \alpha_{ij}\|k_j - \bar{k}_i\|_2^2 = \sum_j \alpha_{ij}\|k_j\|_2^2 - \|\bar{k}_i\|_2^2 \le \|K\|_{\mathrm{F}}^2\widetilde{R}^2.$$

For unit vectors a, b of the appropriate dimensions,

$$\begin{aligned}
|a^\top C_i^{vk} b| &= \Big|\sum_j \alpha_{ij}\langle a, v_j - f_i\rangle\langle b, k_j - \bar{k}_i\rangle\Big| \\
&\le \sqrt{\sum_j \alpha_{ij}\langle a, v_j - f_i\rangle^2}\sqrt{\sum_j \alpha_{ij}\langle b, k_j - \bar{k}_i\rangle^2} \\
&\le \sqrt{\sum_j \alpha_{ij}\|v_j - f_i\|_2^2}\sqrt{\sum_j \alpha_{ij}\|k_j - \bar{k}_i\|_2^2} \\
&\le \|V\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\widetilde{R}^2.
\end{aligned}$$

Taking the supremum over unit vectors proves $\|C_i^{vk}\|_{\mathrm{op}} \le \|V\|_{\mathrm{F}}\|K\|_{\mathrm{F}}\widetilde{R}^2$.

The output-radius bound also implies $\|v_j - f_i\|_2 \le \|v_j\|_2 + \|f_i\|_2 \le 2\|V\|_{\mathrm{F}}\widetilde{R}$. Substituting these estimates into the expression for $\dot{f}_i$ yields

$$\begin{aligned}\|\dot{f}_i\|_2 &\le \sum_j \alpha_{ij}\|\dot{v}_j\|_2 + \frac{\|q_i\|_2}{\sqrt{h_{\mathrm{head}}}}\sum_j \alpha_{ij}\|v_j - f_i\|_2\|\dot{k}_j\|_2 + \frac{\|C_i^{vk}\|_{\mathrm{op}}\|\dot{q}_i\|_2}{\sqrt{h_{\mathrm{head}}}} \\ &\le \|V\|_{\mathrm{F}}\sum_j \alpha_{ij}\delta_j + 2\|V\|_{\mathrm{F}}g\widetilde{R}^2\sum_j \alpha_{ij}\delta_j + \|V\|_{\mathrm{F}}g\widetilde{R}^2\delta_i \\ &\le \sum_{j=1}^{d}\|V\|_{\mathrm{F}}\big((1+2g\widetilde{R}^2)\overline{\alpha}_{ij} + g\widetilde{R}^2\mathbf{1}\{i=j\}\big)\delta_j \\ &= \sum_{j=1}^{d} C_{ij}^{\mathrm{att}}\delta_j.\end{aligned}$$

This bound holds at every point of the segment. Integrating from $s = 0$ to $s = 1$ therefore gives

$$\begin{aligned}\|\mathsf{Att}(Z)^i - \mathsf{Att}(Z')^i\|_2 &\le \int_0^1 \|\dot{f}_i(s)\|_2\,\mathrm{d}s \\ &\le \sum_{j=1}^{d} C_{ij}^{\mathrm{att}}\|Z^j - Z'^j\|_2,\end{aligned}$$

which proves the claimed certificate. □

We now propagate the preceding certificates through each layer's computation graph. Using common input domains across the finite sampling grid gives the following uniform bounds.

Lemma 17 (Computable layerwise sensitivity certificates). *Under Assumption 4.2, finite nonnegative radii $R_0, \dots, R_L$ and matrices $A_1, \dots, A_L \in \mathbb{R}_+^{d\times d}$ can be computed from the supplied network description, the activation Lipschitz constants, and known upper bounds on $|\varphi(0)|$. Define*

$$\mathcal{D}_\ell := \{Z \in (\mathbb{R}^H)^d : \max_i \|Z^i\|_2 \le R_\ell\}, \qquad \ell = 0, \dots, L.$$

For every grid index k and input $x \in \overline{\mathcal{V}}^d$, $h_k^{(0)}(x) \in \mathcal{D}_0$. Moreover, for every k and $\ell \in [L]$, $F_{k,\ell}(\mathcal{D}_{\ell-1}) \subseteq \mathcal{D}_\ell$ and

$$\|F_{k,\ell}(Z)^i - F_{k,\ell}(Z')^i\|_2 \le \sum_{j=1}^{d}(A_\ell)_{ij}\|Z^j - Z'^j\|_2 \tag{78}$$

for all $i \in [d]$ and $Z, Z' \in \mathcal{D}_{\ell-1}$. The construction does not enumerate masked states.

Proof. Set

$$R_0 = \max_{a\in\overline{\mathcal{V}}}\|E(a)\|_2 + \max_{k,i}\|c_{k,i}\|_2 < \infty.$$

Then, for every k, i, x,

$$\|h_{k,i}^{(0)}(x)\|_2 = \|E(x^i) + c_{k,i}\|_2 \le \|E(x^i)\|_2 + \|c_{k,i}\|_2 \le R_0.$$

Fix $\ell \in [L]$ and suppose $R_{\ell-1}$ is known. For each k, process the computation graph of $F_{k,\ell}$ in topological order. Write $H_a(Z)$ for the output at node a, viewed as a function of $Z \in \mathcal{D}_{\ell-1}$. Initialize

$$\begin{aligned}H_{\mathrm{in}}(Z) = Z, &\qquad (\rho_{\mathrm{in}}, C_{\mathrm{in}}) = (R_{\ell-1}, I_d), \\ H_{\mathrm{const}}(Z) = B, &\quad (\rho_{\mathrm{const}}, C_{\mathrm{const}}) = \big(\max_i \|B^i\|_2, 0\big).\end{aligned}$$

For a tokenwise map or attention head f, certify f on $\{Y : \max_i \|Y^i\|_2 \le \rho_a\}$ using Lemmas 15 and 16. If its certificate is (ρ_f, C_f), set

$$H_z = f \circ H_a, \qquad \rho_z = \rho_f, \qquad C_z = C_f C_a.$$

For nodes with two predecessors, use

H_z	ρ_z	$(C_z)_{ij}$
$H_a + H_b$	$\rho_a + \rho_b$	$(C_a)_{ij} + (C_b)_{ij}$
$H_a \odot H_b$	$\rho_a \rho_b$	$\rho_b (C_a)_{ij} + \rho_a (C_b)_{ij}$
$\mathrm{concat}(H_a, H_b)$	$\sqrt{\rho_a^2 + \rho_b^2}$	$\sqrt{(C_a)_{ij}^2 + (C_b)_{ij}^2}$

These rules also cover multi-head attention followed by an affine output map.

The cited lemmas give, inductively at every node a,

$$\max_i \|H_a(Z)^i\|_2 \le \rho_a,$$
$$\|H_a(Z)^i - H_a(Z')^i\|_2 \le \sum_{j=1}^{d} (C_a)_{ij} \|Z^j - Z'^j\|_2, \qquad Z, Z' \in \mathcal{D}_{\ell-1}.$$

In particular, the composition rule is applicable because

$$H_a(\mathcal{D}_{\ell-1}) \subseteq \{Y : \max_i \|Y^i\|_2 \le \rho_a\}.$$

All radii and matrix entries remain finite by Assumption 4.2 and the supplied bounds on $|\varphi(0)|$. At the output node, denote the resulting certificate by $(\rho_{k,\ell}, C_{k,\ell})$.

Since the grid is finite, we may set

$$R_\ell = \max_k \rho_{k,\ell} < \infty, \qquad (A_\ell)_{ij} = \max_k (C_{k,\ell})_{ij} < \infty.$$

For every k and $Z, Z' \in \mathcal{D}_{\ell-1}$,

$$\max_i \|F_{k,\ell}(Z)^i\|_2 \le \rho_{k,\ell} \le R_\ell,$$
$$\|F_{k,\ell}(Z)^i - F_{k,\ell}(Z')^i\|_2 \le \sum_{j=1}^{d} (C_{k,\ell})_{ij} \|Z^j - Z'^j\|_2 \le \sum_{j=1}^{d} (A_\ell)_{ij} \|Z^j - Z'^j\|_2.$$

Thus $F_{k,\ell}(\mathcal{D}_{\ell-1}) \subseteq \mathcal{D}_\ell$ and equation 78 holds. Induction on ℓ completes the construction.

Only vocabulary items, grid points, and computation-graph nodes are traversed and no masked states are enumerated. □

Proof of Lemma 1. Lemma 17 gives $h_k^{(0)}(x) \in \mathcal{D}_0$ and $F_{k,\ell}(\mathcal{D}_{\ell-1}) \subseteq \mathcal{D}_\ell$. Induction therefore gives $h_k^{(\ell)}(x) \in \mathcal{D}_\ell$ for every k, ℓ, x. Apply equation 78 to $Z = h_k^{(\ell-1)}(x)$ and $Z' = h_k^{(\ell-1)}(y)$. The network recursion then gives equation 10 for every pair of masked states x, y. The matrices are finite, nonnegative, and computable without state enumeration by Lemma 17. □

F.2 From Prediction Stability to Conditional Information

We next quantify how revealing one masked coordinate changes the prediction at another position. Write $P = A_L \cdots A_1$ and $\gamma = 2D_{\mathrm{emb}} R_{\mathrm{out}}$. The following lemma converts the layerwise sensitivity bounds into a uniform bound on changes in predicted log probabilities.

Lemma 18 (Revelation stability). *Under Assumption 4.2, let $A_1, \ldots, A_L$ satisfy equation 10. All logits are finite and $\widehat{p}_{k,i}(a \mid x) > 0$ for every k, i, a, x. For distinct $i, j \in M(x)$ and $b \in \mathcal{V}$, set $\Delta_a \coloneqq \ell_{k,i}(a; x \odot_j b) - \ell_{k,i}(a; x)$. Then*

$$\max_{a \in \mathcal{V}} \left| \log \frac{\widehat{p}_{k,i}(a \mid x \odot_j b)}{\widehat{p}_{k,i}(a \mid x)} \right| \le \max_a \Delta_a - \min_a \Delta_a \le \gamma P_{ij}, \tag{79}$$

where $P = A_L \cdots A_1$ and $\gamma = 2D_{\mathrm{emb}} R_{\mathrm{out}}$.

Proof. Lemma 17 and induction give $h_k^{(L)}(x) \in \mathcal{D}_L$ for all k, x. Consequently, $|\ell_{k,i}(a;x)| \le \|\theta_a\|_2 R_L + |\beta_a| < \infty$, and softmax probabilities are strictly positive. For supported x and $i \in M(x)$, finiteness of the alphabet gives $0 \le e_i^k(x) \le \max_{a\in\mathcal{V}} \log(1/\widehat{p}_{k,i}(a \mid x)) < \infty$, where the upper bound uses $\sum_a p_i(a \mid x) \log p_i(a \mid x) \le 0$.

Put $x_b = x \odot_j b$ and $\delta_i^{(\ell)} = \|h_{k,i}^{(\ell)}(x_b) - h_{k,i}^{(\ell)}(x)\|_2$. The state-independent embeddings $c_{k,i}$ cancel, so $\delta^{(0)} \le D_{\mathrm{emb}} \mathbf{e}_j$ coordinatewise. Successive applications of equation 10 and nonnegativity give

$$\begin{aligned}
\delta^{(\ell)} &\le A_\ell \delta^{(\ell-1)}, \\
\delta^{(L)} &\le A_L \cdots A_1 \delta^{(0)} \le D_{\mathrm{emb}} P \mathbf{e}_j, \\
|\Delta_a| &= |\langle \theta_a, h_{k,i}^{(L)}(x_b) - h_{k,i}^{(L)}(x) \rangle| \\
&\le \|\theta_a\|_2 \delta_i^{(L)} \le R_{\mathrm{out}} D_{\mathrm{emb}} P_{ij}.
\end{aligned}$$

The output bias cancels in Δ_a. Therefore $\max_a \Delta_a - \min_a \Delta_a \le 2 \max_a |\Delta_a| \le \gamma P_{ij}$.

For the probability bound, put $c = \log \sum_a \widehat{p}_{k,i}(a \mid x) e^{\Delta_a}$. The weights sum to one, hence $e^{\min_a \Delta_a} \le \sum_a \widehat{p}_{k,i}(a \mid x) e^{\Delta_a} \le e^{\max_a \Delta_a}$. Taking logarithms places c in $[\min_a \Delta_a, \max_a \Delta_a]$. The softmax identity gives

$$\left| \log \frac{\widehat{p}_{k,i}(a \mid x \odot_j b)}{\widehat{p}_{k,i}(a \mid x)} \right| = |\Delta_a - c| \le \max_a \Delta_a - \min_a \Delta_a \le \gamma P_{ij}.$$

Taking the maximum over a proves equation 79. □

We next relate prediction stability to conditional mutual information under the true posterior. The resulting bound includes the expected change in prediction error after revealing one coordinate.

Lemma 19 (Prediction stability and posterior information). *For every supported x, grid index k, and distinct $i, j \in M(x)$,*

$$I_x(i;j) \le \gamma P_{ij} + \sum_{b\in\mathcal{V}} p_j(b \mid x) e_i^k(x \odot_j b) - e_i^k(x).$$

Proof. Put $x_b = x \odot_j b$ and let $\mathbb{E}_{a,b|x}$ denote expectation under the true joint posterior of (X_0^i, X_0^j). At positive joint mass, x_b is supported and both $p_i(a \mid x_b)$ and $p_i(a \mid x)$ are positive. The predicted probabilities are positive by Lemma 18, so inserting them gives

$$\begin{aligned}
I_x(i;j) &= \mathbb{E}_{a,b|x} \log \frac{p_i(a \mid x_b)}{p_i(a \mid x)} \\
&= \mathbb{E}_{a,b|x} \log \frac{\widehat{p}_{k,i}(a \mid x_b)}{\widehat{p}_{k,i}(a \mid x)} + \mathbb{E}_{a,b|x} \log \frac{p_i(a \mid x_b)}{\widehat{p}_{k,i}(a \mid x_b)} \\
&\quad - \mathbb{E}_{a,b|x} \log \frac{p_i(a \mid x)}{\widehat{p}_{k,i}(a \mid x)} \\
&= \mathbb{E}_{a,b|x} \log \frac{\widehat{p}_{k,i}(a \mid x_b)}{\widehat{p}_{k,i}(a \mid x)} + \sum_b p_j(b \mid x) e_i^k(x_b) - e_i^k(x) \\
&\le \gamma P_{ij} + \sum_b p_j(b \mid x) e_i^k(x_b) - e_i^k(x).
\end{aligned}$$

The penultimate equality follows by conditioning on b in the second expectation and using $\sum_b p_j(b \mid x) p_i(a \mid x_b) = p_i(a \mid x)$ in the third. The final inequality is the log-ratio bound in Lemma 18. □

Recall that $\Omega_{ii} = 0$ and

$$\Omega_{ij} = \gamma \sum_{\ell=1}^{L} (v_\ell)_i (A_\ell)_{ij} (u_{\ell-1})_j, \qquad i \ne j,$$

where $u_0 = v_L = \mathbf{1}$, $u_\ell = A_\ell u_{\ell-1}$, and $v_{\ell-1} = A_\ell^\top v_\ell$. These weights aggregate sensitivity through each transition between positions, including propagation before and after that transition. The following lemma connects them to $P = A_L \cdots A_1$, showing that their sum across any partition bounds the corresponding aggregate sensitivity.

Lemma 20 (Forward–backward cut certificate). *For every partition $\mathcal{C}$ of $[d]$, writing $C(i)$ for the group containing i, we have*

$$\gamma \sum_{C(i)\neq C(j)} P_{ij} \le \sum_{C(i)\neq C(j)} \Omega_{ij}. \tag{80}$$

The same weights Ω satisfy this bound for all partitions.

Proof. For a path $z = (z_0, \dots, z_L) \in [d]^{L+1}$, define $w(z) = \prod_{\ell=1}^{L} (A_\ell)_{z_\ell z_{\ell-1}} \ge 0$. Expanding matrix products gives

$$P_{ij} = \sum_{z: z_0=j,\ z_L=i} w(z),$$

$$\sum_{z: z_{\ell-1}=j,\ z_\ell=i} w(z) = (v_\ell)_i (A_\ell)_{ij} (u_{\ell-1})_j,$$

$$\Omega_{ij} = \gamma \sum_{\ell=1}^{L} \sum_{z: z_{\ell-1}=j,\ z_\ell=i} w(z), \qquad i \neq j.$$

In the second identity, summing all prefixes ending at j gives $(u_{\ell-1})_j$; summing all suffixes starting at i gives $(v_\ell)_i$. The empty prefix or suffix contributes 1, consistently with $u_0 = v_L = \mathbf{1}$. If consecutive positions belong to the same group at every layer, then the path endpoints belong to the same group. Consequently, $\mathbf{1}\{C(z_0) \neq C(z_L)\} \le \sum_{\ell=1}^{L} \mathbf{1}\{C(z_{\ell-1}) \neq C(z_\ell)\}$, and

$$\begin{aligned}
\gamma \sum_{C(i)\neq C(j)} P_{ij} &= \gamma \sum_z w(z) \mathbf{1}\{C(z_0) \neq C(z_L)\} \\
&\le \gamma \sum_{\ell=1}^{L} \sum_z w(z) \mathbf{1}\{C(z_{\ell-1}) \neq C(z_\ell)\} \\
&= \sum_{C(i)\neq C(j)} \Omega_{ij}.
\end{aligned}$$

This argument bounds the sum across a cut; no entrywise inequality $\gamma P_{ij} \le \Omega_{ij}$ is needed or asserted. □

F.3 Soundness and Computational Complexity

We now prove Theorem 3 and establish the computational cost of the oracle. We begin by averaging the conditional-information bound over independent masking, and then control the resulting prediction error using the grid-time score entropy loss.

Recall that $e_i^k(x) = D_{\mathrm{KL}}(p_i(\cdot \mid x) \| \widehat{p}_{k,i}(\cdot \mid x))$ measures the discrepancy between the true posterior at position i and the predictor at grid time τ_k, given the masked state x.

Lemma 21 (Averaging the revelation error). *Fix a grid index k, $t \in [0, T)$, and $q \in [d]$, and write $\tau = T - t$. Suppose finite constants $\xi_{ij} \ge 0$ satisfy*

$$I_x(i; j) \le \xi_{ij} + \sum_{b \in \mathcal{V}} p_j(b \mid x) e_i^k(x \odot_j b) \tag{81}$$

for every supported x and distinct $i, j \in M(x)$. For the partition $\{C \cap M(x) : C \in \mathcal{C}_q\} \setminus \{\varnothing\}$, the information rate satisfies

$$\Gamma_{\mathrm{part}}(t) \le e^{-2\tau} \sum_{C_q(i)\neq C_q(j)} \xi_{ij} + \kappa(q) \lambda_\tau \mathbb{E}_{Z \sim q_\tau} \sum_{i \in M(Z)} e_i^k(Z). \tag{82}$$

Proof. Independent masking gives

$$q_\tau(x) = (1-e^{-\tau})^{m(x)} e^{-\tau(d-m(x))}\rho(x), \qquad \rho(x) = \mathbb{P}(X_0^{U(x)} = x^{U(x)}). \tag{83}$$

In particular, the support is $\{x : \rho(x) > 0\}$ for every $\tau > 0$. If x is supported, $j \in M(x)$, and $y = x \odot_j b$, then $m(y) = m(x) - 1$ and $\rho(y) = \rho(x)p_j(b \mid x)$. Thus

$$\begin{aligned} q_\tau(x)p_j(b \mid x) &= (1-e^{-\tau})^{m(x)} e^{-\tau(d-m(x))}\rho(y) \\ &= (e^\tau - 1)(1-e^{-\tau})^{m(y)} e^{-\tau(d-m(y))}\rho(y) = \lambda_\tau^{-1} q_\tau(y). \end{aligned} \tag{84}$$

This identity also holds when $p_j(b \mid x) = 0$, since then $\rho(y) = 0$.

Fix distinct i, j. The map $(x, b) \mapsto y = x \odot_j b$ bijects supported pairs with $i, j \in M(x)$ and $p_j(b \mid x) > 0$ onto supported states with $i \in M(y)$ and $j \in U(y)$. Indeed, its inverse masks coordinate j of y and sets $b = y^j$; the resulting x satisfies $\rho(x) \geq \rho(y) > 0$ and $p_j(b \mid x) = \rho(y)/\rho(x) > 0$. Restricting the following state sums to supported states,

$$\begin{aligned} &\mathbb{E}_{Z\sim q_\tau}\left[\mathbf{1}\{i, j \in M(Z)\}\sum_b p_j(b \mid Z) e_i^k(Z \odot_j b)\right] \\ &\quad= \sum_{x: i,j\in M(x)} \sum_{b: p_j(b|x)>0} q_\tau(x)p_j(b \mid x) e_i^k(x \odot_j b) \\ &\quad= \lambda_\tau^{-1} \sum_{y: i\in M(y),\, j\in U(y)} q_\tau(y) e_i^k(y). \end{aligned} \tag{85}$$

Summing over the ordered pairs in different base groups yields

$$\begin{aligned} &\mathbb{E}_{Z\sim q_\tau} \sum_{\substack{i,j\in M(Z)\\ C_q(i)\neq C_q(j)}} \sum_b p_j(b \mid Z) e_i^k(Z \odot_j b) \\ &\quad= \lambda_\tau^{-1}\mathbb{E}_{Z\sim q_\tau} \sum_{i\in M(Z)} |U(Z) \setminus C_q(i)| e_i^k(Z) \\ &\quad\leq \lambda_\tau^{-1}\kappa(q)\mathbb{E}_{Z\sim q_\tau} \sum_{i\in M(Z)} e_i^k(Z). \end{aligned} \tag{86}$$

The inequality uses $e_i^k \geq 0$ and $|U(Z) \setminus C_q(i)| \leq d - |C_q(i)| \leq \kappa(q)$. Two masked positions are in different output blocks exactly when their base groups differ. Their mask indicators are independent, with joint probability $(1-e^{-\tau})^2$; therefore

$$\begin{aligned} \Gamma_{\mathrm{part}}(t) &= \lambda_\tau^2 \mathbb{E}_{Z\sim q_\tau} \sum_{\substack{i,j\in M(Z)\\ C_q(i)\neq C_q(j)}} I_Z(i;j) \\ &\leq \lambda_\tau^2 (1-e^{-\tau})^2 \sum_{C_q(i)\neq C_q(j)} \xi_{ij} \\ &\quad + \lambda_\tau^2 \mathbb{E}_{Z\sim q_\tau} \sum_{\substack{i,j\in M(Z)\\ C_q(i)\neq C_q(j)}} \sum_b p_j(b \mid Z) e_i^k(Z \odot_j b) \\ &\leq e^{-2\tau} \sum_{C_q(i)\neq C_q(j)} \xi_{ij} + \kappa(q)\lambda_\tau \mathbb{E}_{Z\sim q_\tau} \sum_{i\in M(Z)} e_i^k(Z). \end{aligned} \tag{87}$$

The first inequality uses equation 81; the second uses equation 86 and $\lambda_\tau^2(1-e^{-\tau})^2 = e^{-2\tau}$. □

By equation 84, the true score satisfies $s_{\tau_k}(x \odot_i a, x) = q_{\tau_k}(x \odot_i a)/q_{\tau_k}(x) = \lambda_{\tau_k} p_i(a \mid x)$ for supported x and $i \in M(x)$. Hence the grid-time loss in Assumption 2.1 takes the form

$$\mathcal{L}_{\mathrm{SE},k} = \mathbb{E}_{Z\sim q_{\tau_k}} \sum_{i\in M(Z)} \sum_{a\in\mathcal{V}} \mathfrak{d}\big(\lambda_{\tau_k} p_i(a \mid Z), \widehat{s}_{\tau_k}(Z \odot_i a, Z)\big). \tag{88}$$

The normalized-score relation in the main text lets us control posterior error by this loss.

Lemma 22 (**Score normalization and posterior error**). *Fix k, supported x, and $i \in M(x)$. Suppose $\widehat{s}_{\tau_k}(x \odot_i a, x)$ is positive and finite for every $a \in \mathcal{V}$, and its normalization is $\widehat{p}_{k,i}(a \mid x)$. Then*

$$\lambda_{\tau_k} e_i^k(x) \le \sum_{a \in \mathcal{V}} \mathfrak{d}\big(\lambda_{\tau_k} p_i(a \mid x), \widehat{s}_{\tau_k}(x \odot_i a, x)\big). \tag{89}$$

Proof. Put $\widehat{\lambda} = \sum_a \widehat{s}_{\tau_k}(x \odot_i a, x)$. Then $0 < \widehat{\lambda} < \infty$ and $\widehat{s}_{\tau_k}(x \odot_i a, x) = \widehat{\lambda}\widehat{p}_{k,i}(a \mid x)$. Using the unit total masses of p_i and $\widehat{p}_{k,i}$,

$$\begin{aligned}
&\sum_a \mathfrak{d}\big(\lambda_{\tau_k} p_i(a \mid x), \widehat{\lambda}\widehat{p}_{k,i}(a \mid x)\big) \\
&\quad = \widehat{\lambda} - \lambda_{\tau_k} + \lambda_{\tau_k} \sum_a p_i(a \mid x) \log \frac{\lambda_{\tau_k} p_i(a \mid x)}{\widehat{\lambda}\widehat{p}_{k,i}(a \mid x)} \\
&\quad = \widehat{\lambda} - \lambda_{\tau_k} + \lambda_{\tau_k} \log(\lambda_{\tau_k}/\widehat{\lambda}) + \lambda_{\tau_k} e_i^k(x) \\
&\quad = \mathfrak{d}(\lambda_{\tau_k}, \widehat{\lambda}) + \lambda_{\tau_k} e_i^k(x) \ge \lambda_{\tau_k} e_i^k(x).
\end{aligned} \tag{90}$$

Zero posterior masses contribute zero to the logarithmic sums. The last inequality follows from $\log z \le z - 1$ for $z > 0$. □

The score entropy loss controls the weighted prediction error at grid times. We next extend this control to every $t \in [t_k, t_{k+1})$, keeping the predictor index k fixed.

Lemma 23 (**Transfer from grid-time score error**). *Suppose the positive finite learned scores normalize to $\widehat{p}_{k,i}$ on supported states. For $t \in [t_k, t_{k+1})$ and $\tau = T - t$,*

$$\lambda_\tau \mathbb{E}_{Z \sim q_\tau} \sum_{i \in M(Z)} e_i^k(Z) \le e^{d(t - t_k)} \mathcal{L}_{\mathrm{SE},k}. \tag{91}$$

Proof. Since $t \in [t_k, t_{k+1})$, we have $0 < \tau \le \tau_k$ and $\tau_k - \tau = t - t_k$. By equation 83, the two distributions have the same support, and on this support

$$\begin{aligned}
\frac{q_\tau(x)}{q_{\tau_k}(x)} &= e^{(t - t_k)(d - m(x))} \left(\frac{1 - e^{-\tau}}{1 - e^{-\tau_k}}\right)^{m(x)}, \\
\frac{\lambda_\tau}{\lambda_{\tau_k}} &= e^{t - t_k} \frac{1 - e^{-\tau_k}}{1 - e^{-\tau}}.
\end{aligned} \tag{92}$$

For $m(x) \ge 1$, multiplying these identities gives

$$\begin{aligned}
\frac{\lambda_\tau q_\tau(x)}{\lambda_{\tau_k} q_{\tau_k}(x)} &= e^{(t - t_k)(d - m(x) + 1)} \left(\frac{1 - e^{-\tau}}{1 - e^{-\tau_k}}\right)^{m(x) - 1} \\
&\le e^{(t - t_k)(d - m(x) + 1)} \le e^{d(t - t_k)}.
\end{aligned} \tag{93}$$

The inequalities use $1 - e^{-\tau} \le 1 - e^{-\tau_k}$, $m(x) - 1 \ge 0$, and $t - t_k \ge 0$.

The sum $\sum_{i \in M(x)} e_i^k(x)$ is nonnegative and vanishes when $m(x) = 0$. The posterior depends only on x, and the predictor index k is fixed, so this sum is the same function at both times. Summing over supported states and using Lemma 22,

$$\begin{aligned}
\lambda_\tau \mathbb{E}_{Z \sim q_\tau} \sum_{i \in M(Z)} e_i^k(Z) &= \sum_{x : m(x) \ge 1} \lambda_\tau q_\tau(x) \sum_{i \in M(x)} e_i^k(x) \\
&\le e^{d(t - t_k)} \sum_{x : m(x) \ge 1} \lambda_{\tau_k} q_{\tau_k}(x) \sum_{i \in M(x)} e_i^k(x) \\
&\le e^{d(t - t_k)} \mathbb{E}_{Z \sim q_{\tau_k}} \sum_{i \in M(Z)} \sum_a \mathfrak{d}\big(\lambda_{\tau_k} p_i(a \mid Z), \widehat{s}_{\tau_k}(Z \odot_i a, Z)\big) \\
&= e^{d(t - t_k)} \mathcal{L}_{\mathrm{SE},k}.
\end{aligned} \tag{94}$$

□

Lemma 24 (Hierarchy properties). *The partitions $\mathcal{C}_q$ satisfy $\mathcal{C}_{q'} \preceq \mathcal{C}_q$ for $q' \le q$. The oracle equation 13 is deterministic and measurable, has blocks of size at most $r(t)$, and satisfies monotone refinement under increasing time and additional revelations.*

The structural properties follow from Proposition 2, using the same merge construction on the complete graph. Every nonempty $C \cap M(x)$ induces a connected subgraph, so the connected-component query reduces to equation 13. Consequently, the base partitions are nested, and the oracle is deterministic and measurable, has blocks of size at most $r(t)$, and refines under increasing time and additional revelations.

Lemma 25 (Computational cost). *Let $S = |\mathcal{V}|$ and let T_{layer} denote the cost of constructing the layerwise certificates. Preprocessing takes $O(SH + T_{\mathrm{layer}} + Ld^2 + d^2 \log(d+1))$ operations and $O(Ld^2)$ additional workspace beyond the network and its certification procedure. Retaining $(\mathcal{H}, \Omega)$ requires $O(d^2)$ space. Each partition query takes $O(d)$ operations plus evaluation of $r(t)$, and $b_{\mathrm{LM}}(q)$ can be evaluated in $O(d^2)$ operations for any fixed q.*

Proof. We use unit-cost arithmetic, comparisons, and elementary-function evaluations. Computing D_{emb} and R_{out} requires $O(SH)$ operations. By definition, T_{layer} includes the construction of the domains and uniform layerwise matrices, including all grid-dependent costs. The forward and backward recursions require $2L$ matrix–vector products, and each off-diagonal entry of Ω requires a sum over L layers. Thus

$$T_{u,v} = O(Ld^2), \qquad T_{\Omega} = O(Ld(d-1) + d) = O(Ld^2).$$

There are $\binom{d}{2}$ unordered pairs. Forming their weights takes $O(d^2)$ operations, and sorting takes

$$T_{\mathrm{sort}} = O\Big(\binom{d}{2} \log\big(\binom{d}{2} + 1\big)\Big) = O(d^2 \log(d+1)).$$

Implement the merge pass using disjoint sets with union by size, storing each component's merge-tree root separately. Whenever a vertex's disjoint-set depth increases, its component size at least doubles. Hence each find takes $O(\log(d+1))$ operations. There are two finds per pair and at most $d-1$ successful merges, each creating one constant-size tree record, so

$$T_{\mathrm{merge}} = O\Big(d + 2\binom{d}{2} \log(d+1)\Big) = O(d^2 \log(d+1)).$$

Combining these costs gives

$$T_{\mathrm{pre}} = O\big(SH + T_{\mathrm{layer}} + Ld^2 + d^2 \log(d+1)\big).$$

The matrices, forward and backward vectors, weights and sorting records, and tree and disjoint-set arrays use, respectively,

$$\begin{aligned} \mathrm{Space}_{\mathrm{additional}} &= O(Ld^2) + O(Ld) + O(d^2) + O(d) = O(Ld^2), \\ \mathrm{Space}_{\mathrm{retained}} &= O(|\mathcal{H}| + d^2) = O(d^2), \end{aligned}$$

beyond the network and its certification procedure. Here $|\mathcal{H}| \le d + (d-1) = 2d - 1$.

For a query, let $\mathcal{Z}_q$ be the selected tree nodes at budget $q = r(t)$, and let $\mathcal{H}_z$ denote the subtree rooted at z. Top-down selection visits each node at most once. The selected subtrees are disjoint, so

$$N_{\mathrm{selection}} \le |\mathcal{H}|, \qquad \sum_{z \in \mathcal{Z}_q} |\mathcal{H}_z| \le |\mathcal{H}|, \qquad \sum_{z \in \mathcal{Z}_q} |S(z) \cap M(x)| = |M(x)| \le d.$$

On the complete graph, each nonempty masked intersection is one output block. Enumerating the selected subtrees therefore suffices to form the output, without storing descendant leaf sets explicitly. Apart from evaluating $r(t)$, the query cost is

$$T_{\mathrm{query}} = O\Big(|\mathcal{H}| + \sum_{z \in \mathcal{Z}_q} |\mathcal{H}_z| + d\Big) = O(d).$$

Finally, assign each leaf its base-block label $\ell_q(i)$ using the same $O(d)$ traversal. Then evaluate

$$b_{\mathrm{LM}}(q) = \sum_{1 \le i < j \le d} (\Omega_{ij} + \Omega_{ji}) \mathbf{1}\{\ell_q(i) \neq \ell_q(j)\}.$$

Each unordered pair requires constant work, giving $T_{\mathrm{boundary}} = O(d + \binom{d}{2}) = O(d^2)$. □

Proof of Theorem 3. Lemma 1 provides finite uniform layer certificates, and Lemma 24 proves determinism, measurability, bounded block size, and monotone refinement for every query state.

Fix $t \in [t_k, t_{k+1})$, and put $\tau = T - t > 0$ and $q = r(t)$. Since $e_i^k(x) \ge 0$, Lemma 19 gives $I_x(i;j) \le \gamma P_{ij} + \sum_b p_j(b \mid x) e_i^k(x \odot_j b)$. Thus Lemma 21 applies with $\xi_{ij} = \gamma P_{ij}$. Applying that lemma, then Lemmas 20 and 23, yields

$$\begin{aligned}
\Gamma_{\mathrm{part}}(t) &\le e^{-2\tau}\gamma \sum_{C_q(i) \neq C_q(j)} P_{ij} + \kappa(q)\lambda_\tau \mathbb{E}_{Z \sim q_\tau} \sum_{i \in M(Z)} e_i^k(Z) \\
&\le e^{-2\tau} \sum_{C_q(i) \neq C_q(j)} \Omega_{ij} + \kappa(q)\lambda_\tau \mathbb{E}_{Z \sim q_\tau} \sum_{i \in M(Z)} e_i^k(Z) \\
&\le e^{-2\tau} b_{\mathrm{LM}}(q) + \kappa(q) e^{d(t-t_k)} \mathcal{L}_{\mathrm{SE},k} = \varepsilon_{\mathrm{part}}(t).
\end{aligned}$$

The mask indicators are averaged before the cut inequality is applied; no entrywise comparison between γP and Ω is required.

The image of $r : [0,T) \to [d]$ is a nonempty finite set, so $q_* = \min_{0 \le t < T} r(t)$ exists. If $q' \le q$, refinement gives $\mathbf{1}\{C_q(i) \neq C_q(j)\} \le \mathbf{1}\{C_{q'}(i) \neq C_{q'}(j)\}$. Multiplying by $\Omega_{ij} \ge 0$ and summing proves $b_{\mathrm{LM}}(q) \le b_{\mathrm{LM}}(q')$. Hence $b_{\mathrm{LM}}(r(t)) \le b_{\mathrm{LM}}(q_*)$. Also $\kappa(q) \le d-1$, since every base group is nonempty. For each grid interval, $\int_{t_k}^{t_{k+1}} e^{d(t-t_k)}\,\mathrm{d}t = (e^{dh_k} - 1)/d \le h_k e^{dh_k}$; the inequality also follows by bounding the integrand by its value at the right endpoint. Therefore

$$\begin{aligned}
\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t &\le b_{\mathrm{LM}}(q_*) \int_0^T e^{-2(T-t)}\,\mathrm{d}t \\
&\quad + (d-1) \sum_{k=0}^{N-1} \mathcal{L}_{\mathrm{SE},k} \int_{t_k}^{t_{k+1}} e^{d(t-t_k)}\,\mathrm{d}t \\
&\le \frac{1 - e^{-2T}}{2} b_{\mathrm{LM}}(q_*) + (d-1) \sum_{k=0}^{N-1} h_k e^{dh_k} \mathcal{L}_{\mathrm{SE},k} \\
&\le \frac{1 - e^{-2T}}{2} b_{\mathrm{LM}}(q_*) + (d-1) e^{dh_{\max}} \sum_{k=0}^{N-1} h_k \mathcal{L}_{\mathrm{SE},k} \\
&\le \frac{1 - e^{-2T}}{2} b_{\mathrm{LM}}(q_*) + (d-1) e^{dh_{\max}} \varepsilon_{\mathrm{score}}.
\end{aligned}$$

The last step is Assumption 2.1. All weights are finite, and b_{LM} and κ are functions on the finite set $[d]$. Their compositions with the measurable schedule, together with the finitely many grid intervals and exponential factors, make $\varepsilon_{\mathrm{part}}$ measurable. The preceding bound proves integrability. Thus all requirements of Assumption 2.2 hold. The complexity claims follow from Lemma 24. □

F.4 Conditions for a Small Structural Term

We now give sufficient conditions for a small structural term in the rate certificate. These conditions control the total sensitivity across the constructed groups and its amplification through the network.

Corollary 2 (Small certified influence across groups). *Under the assumptions of Theorem 3, fix $q \in [d]$ and let $\eta_\ell(q) := \sum_{C_q(i) \neq C_q(j)} (A_\ell)_{ij}$. Suppose $\alpha_\ell \ge 0$ satisfy $\max\{\|A_\ell\|_1, \|A_\ell\|_\infty\} \le 1 + \alpha_\ell$, $\sum_{\ell=1}^L \alpha_\ell \le a$, and $\sum_{\ell=1}^L \eta_\ell(q) \le \eta$. Then*

$$b_{\mathrm{LM}}(q) \le 2 D_{\mathrm{emb}} R_{\mathrm{out}} e^a \eta. \tag{95}$$

For $r(t) = q$, it follows that

$$\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t \le (1 - e^{-2T}) D_{\mathrm{emb}} R_{\mathrm{out}} e^a \eta + \kappa(q) e^{dh_{\max}} \varepsilon_{\mathrm{score}}.$$

The quantity η bounds the total cross-group sensitivity over all layers, while e^a bounds its amplification through the remaining layers. These conditions are evaluated on the computed certificates and the resulting partition. Assumption 4.2 guarantees finite, computable certificates, but does not guarantee that the resulting structural bound is small.

Proof of Corollary 2. The induced norm inequalities $\|A_\ell z\|_\infty \le \|A_\ell\|_\infty \|z\|_\infty$ and $\|A_\ell^\top z\|_\infty \le \|A_\ell\|_1 \|z\|_\infty$ give, by iteration from $u_0 = v_L = \mathbf{1}$, $\|u_{\ell-1}\|_\infty \le \prod_{h<\ell} \|A_h\|_\infty$ and $\|v_\ell\|_\infty \le \prod_{h>\ell} \|A_h\|_1$. Substituting these bounds and using nonnegativity gives

$$\begin{aligned}
b_{\mathrm{LM}}(q) &= \gamma \sum_{\ell=1}^{L} \sum_{C_q(i) \neq C_q(j)} (v_\ell)_i (A_\ell)_{ij} (u_{\ell-1})_j \\
&\le \gamma \sum_{\ell=1}^{L} \|v_\ell\|_\infty \|u_{\ell-1}\|_\infty \eta_\ell(q) \\
&\le \gamma \sum_{\ell=1}^{L} \prod_{h<\ell} \|A_h\|_\infty \prod_{h>\ell} \|A_h\|_1 \, \eta_\ell(q) \\
&\le \gamma \sum_{\ell=1}^{L} \prod_{h \neq \ell} (1 + \alpha_h) \eta_\ell(q) \\
&\le \gamma e^{\sum_h \alpha_h} \sum_{\ell=1}^{L} \eta_\ell(q) \le \gamma e^a \eta.
\end{aligned}$$

The penultimate inequality uses $1 + u \le e^u$ for $u \ge 0$; the last uses the two assumed sum bounds. For $r(t) = q$, the certificate equation 15 gives

$$\begin{aligned}
\int_0^T \varepsilon_{\mathrm{part}}(t)\,\mathrm{d}t &= \frac{1 - e^{-2T}}{2} b_{\mathrm{LM}}(q) + \kappa(q) \sum_{k=0}^{N-1} \mathcal{L}_{\mathrm{SE},k} \frac{e^{dh_k} - 1}{d} \\
&\le \frac{1 - e^{-2T}}{2} \gamma e^a \eta + \kappa(q) \sum_{k=0}^{N-1} h_k e^{dh_k} \mathcal{L}_{\mathrm{SE},k} \\
&\le \frac{1 - e^{-2T}}{2} \gamma e^a \eta + \kappa(q) e^{dh_{\max}} \sum_{k=0}^{N-1} h_k \mathcal{L}_{\mathrm{SE},k} \\
&\le (1 - e^{-2T}) D_{\mathrm{emb}} R_{\mathrm{out}} e^a \eta + \kappa(q) e^{dh_{\max}} \varepsilon_{\mathrm{score}}.
\end{aligned}$$

Here $(e^{dh_k} - 1)/d = \int_0^{h_k} e^{ds}\,\mathrm{d}s \le h_k e^{dh_k}$, then $h_k \le h_{\max}$; the final line uses Assumption 2.1. □