Confidence-Based Decoding is Provably Efficient for Diffusion Language Models under Random Ordering

Changxiao Cai
University of Michigan
Ann Arbor, MI 48105
cxcai@umich.edu

Gen Li
Chinese University of Hong Kong
Hong Kong
genli@cuhk.edu.hk

Abstract

Diffusion language models (DLMs) have emerged as a promising alternative to autoregressive (AR) models for language modeling, allowing flexible generation order and parallel generation of multiple tokens. However, this flexibility introduces a challenge absent in AR models: the *decoding strategy*—which determines the order and number of tokens generated at each iteration—critically affects sampling efficiency. Among decoding strategies explored in practice, confidence-based methods, which adaptively select which and how many tokens to unmask based on prediction confidence, have shown strong empirical performance. Despite this success, our theoretical understanding of confidence-based decoding remains limited. In this work, we develop the first theoretical analysis framework for confidence-based decoding in DLMs. We focus on an entropy sum-based strategy that continues unmasking tokens in a uniformly random order within each iteration until the cumulative entropy exceeds a threshold, and show that it achieves ε-accurate sampling in KL divergence with an expected number of iterations $\widetilde{O}(H(X_0)/\varepsilon)$, where $H(X_0)$ denotes the entropy of the target data distribution and $\widetilde{O}(\cdot)$ denotes an asymptotic upper bound up to constant and logarithmic factors. Notably, this strategy yields substantial sampling acceleration when the data distribution has low entropy relative to the sequence length, while automatically adapting to the intrinsic complexity of data without requiring prior knowledge or hyperparameter tuning. Overall, our results provide a theoretical foundation for confidence-based decoding and may inform the design of more efficient decoding strategies for DLMs.

1 Introduction

Diffusion language models (DLMs) (Sahoo et al., 2024; Shi et al., 2024) have recently emerged as a compelling alternative to the autoregressive (AR) paradigm for language modeling. Unlike AR models, DLMs use a non-causal generation mechanism that can produce multiple tokens in parallel at each iteration and relax the rigid left-to-right constraint. The parallel generation capability offers strong potential for sampling acceleration, while the non-causal structure also makes DLMs well suited to tasks that benefit from bidirectional reasoning or the enforcement of global constraints (Nie et al., 2025; Prabhudesai et al., 2025; Li et al., 2026b). At scale, DLMs have been shown to achieve performance competitive with AR models (Ye et al., 2025; Song et al., 2025; You et al., 2025; Zhu et al., 2025; Bie et al., 2025; Labs et al., 2025). Beyond their role as standalone generators, DLMs have also been used as speculative drafters to accelerate AR inference through parallel token generation (Chen et al., 2026; Cheng et al., 2026; Zhao and Cai, 2026b).

State-of-the-art DLMs are built upon *masked diffusion models* (MDMs) (Austin et al., 2021; Campbell et al., 2022; Lou et al., 2023), a class of discrete diffusion models that augment the vocabulary with an additional absorbing state, referred to as *mask*. An MDM is defined through two complementary

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

processes. In the forward (masking) process, tokens in a sequence drawn from the target data distribution are progressively replaced by the mask until the sequence becomes fully masked. The reverse (unmasking) process then seeks to invert this corruption by iteratively recovering masked tokens, thereby transforming a fully masked sequence into a fresh sample from the data distribution.

Central to the reverse process is a *mask predictor*, which learns the conditional distributions of masked tokens given the currently unmasked context at various masking levels (Lou et al., 2023; Ou et al., 2024; Zheng et al., 2024). The reverse dynamics are implemented by repeatedly sampling token values from these conditionals to impute masked positions, thereby moving from a fully masked sequence toward a data-like sample. In this sense, the mask predictor in MDMs serves as a discrete analog of the score estimator in continuous diffusion models (Song et al., 2020). As the joint distribution of multiple masked tokens is generally intractable in high-dimensional settings, practical mask predictors are typically trained to learn the conditional marginals of each masked token, and use the product of per-token conditional marginals as an approximation (Nie et al., 2025; Wu et al., 2025). In particular, a neural network-based mask predictor can output, in one forward pass, conditional marginal distributions for all currently masked positions.

While this factorized approximation enables parallel generation of DLMs, it also introduces an inevitable bias in the sampling stage—simultaneously unmasking multiple tokens in one iteration ignores their statistical dependence conditioned on the currently revealed context. At one extreme, unmasking all tokens in a single iteration incurs an error comparable to the discrepancy between the target distribution and the product of its marginals. At the other extreme, generating tokens one-by-one recovers the standard AR approach, which removes the factorization bias but sacrifices the parallel generation advantage of DLMs.

Therefore, the sampling speed and generation quality of DLMs depend critically on the decoding (or unmasking) strategy used at the inference stage, which determines both which positions to unmask and how many tokens to reveal at each iteration.

Parallel decoding strategies. Decoding strategies used in practice can be broadly categorized into two families.

- **Uniform decoding strategies.** This class of decoding strategies prescribes a fixed total iteration count T and a step-size schedule $s_1, s_2 \ldots, s_T$ with $s_t \geq 1$ and $\sum_{t=1}^{T} s_t = L$ (e.g., cosine schedules (Shi et al., 2024) or log-linear schedules (Sahoo et al., 2024; Lou et al., 2023)). At each iteration t, one samples a subset of size s_t uniformly at random from the remaining masked positions and then unmasks the tokens in the sampled subset. Note that this family of schemes treats all masked positions uniformly and thus belongs to the broader class of non-adaptive decoding strategies.
- **Confidence-based decoding strategies.** This collection of decoding strategies leverages the fact that a mask predictor provides not only a distribution but also a natural confidence signal for each token prediction. It is thus natural to use prediction confidence to guide unmasking decisions at each iteration, with newly revealed tokens providing additional context in later iterations (Chang et al., 2022; Zheng et al., 2023; Gong et al., 2024). In practice, confidence is often quantified by the entropy of the predicted distribution (Ben-Hamu et al., 2025), the mass of the most likely token (Nie et al., 2025; Yu et al., 2025; Wu et al., 2025), or the likelihood gap between the two most likely tokens (Kim et al., 2025), leading to a range of effective heuristics. Compared to uniform decoding strategies, these schemes determine the decoding order and set size in an adaptive manner.

In contrast to the rapid development of empirical decoding strategies, our theoretical understanding of their sampling efficiency remains limited. For uniform random schedules, substantial recent progress has been made in establishing theoretical guarantees for sampling performance (Li and Cai, 2025; Chen et al., 2025b; Lavenant and Zanella, 2025; Dmitriev et al., 2026; Zhao and Cai, 2026a). We defer a detailed discussion to Section 1.2. By comparison, despite their strong empirical performance, confidence-based decoding remains far less understood theoretically. The fundamental difficulty lies in their inherently adaptive decoding dynamics: the set of tokens unmasked at each iteration depends on the *realization* of the tokens generated so far, through the corresponding confidence levels produced by the mask predictor. This adaptivity raises several fundamental but unanswered questions. For instance, if the total number of iterations is not predetermined, how many iterations are needed to

unmask all tokens? What is the resulting sampling error? Can these schemes provably adapt to the (unknown) intrinsic complexity or dependence structure of the data distribution?

Such a theory-practice gap motivates the central question of our study:

Are confidence-based decoding strategies in DLMs provably efficient?

1.1 Our contributions

In this work, we take an initial step toward answering this question by developing sampling convergence guarantees for confidence-based decoding in DLMs, with the goal of demystifying their practical effectiveness.

We focus on a canonical confidence measure—entropy—and investigate entropy-based decoding that selects unmasking positions and set sizes based on the entropy of the predicted distribution at each masked position. Concretely, we study a representative *entropy sum-based strategy* that, at each iteration, unmasks tokens from the remaining masked positions in a uniformly random order until the running entropy sum of the newly unmasked tokens exceeds a prescribed threshold, thereby revealing as many tokens as possible per iteration while keeping the cumulative prediction uncertainty controlled.

We demonstrate that ε-accuracy in Kullback-Leibler (KL) divergence can be attained with an expected number of iterations that scales as

$$\widetilde{O}\bigg(\frac{H(X_0)}{\varepsilon}\bigg),$$

where $H(X_0)$ is the entropy of the data distribution and $\widetilde{O}(\cdot)$ hides a logarithmic factor in the sequence length L. This result implies a sublinear iteration complexity in the sequence length L when the data distribution has low entropy $H(X_0) \ll L$, thereby achieving provable sampling speedup compared to the $\Theta(L)$ iteration complexity of the AR decoding or naive uniform parallel decoding. Here, $\Theta(\cdot)$ denotes matching asymptotic upper and lower bounds up to constant factors. Notably, this scheme does not require any prior knowledge of the data distribution such as its entropy, and thus automatically adapts to the intrinsic low-complexity data structure to achieve sampling acceleration.

To establish these results, we introduce a novel analysis framework for adaptive decoding dynamics in DLMs that may be of independent interest. Overall, our results provide the first theoretical justification for the effectiveness of confidence-based unmasking strategies, and offer a principled explanation of when and why entropy-based decoding can achieve accelerated sampling without sacrificing generation fidelity. Our theoretical insights may inform the design of more efficient inference schemes in practice.

1.2 Other related work

Convergence theory for uniform decoding strategies. Existing sampling convergence analysis for DLMs largely focuses on uniform decoding strategies and two main viewpoints have been pursued. The first line of work directly studies the sampling error accumulated in the iterative unmasking procedure under the factorized approximation. For balanced step-size schedules (i.e., $s_t \asymp L/T$ for all t), Li and Cai (2025) established an $O(1/T)$ convergence rate in KL divergence, with leading factor equal to the sum of mutual information $\sum_{i=1}^{T} I(X^i; X^{\setminus i})$ between the i-th token X^i and the remaining tokens $X^{\setminus i} := (X^j)_{j\neq i}$ under the data distribution. They also proved matching lower bounds (up to constant factors), showing that the $O(1/T)$ decay is unimprovable for balanced step-size schedules in the worst case. Subsequently, Chen et al. (2025b) refined the convergence rate using the total correlation and dual total correlation of the data distribution, and developed optimal step-size schedules. These schedules can outperform balanced step-size schedules, albeit requiring either a priori estimates of these correlation measures or hyperparameter tuning. Recently, Zhao and Cai (2026a) showed that randomized step-size schedules can automatically adapt to intrinsic dependence structure underlying data, leading to improved iteration complexity for structured data distributions. A parallel viewpoint analyzes discrete diffusion models from the perspective of CTMC discretization. However, most results in this direction yield iteration complexity that scales at least linearly in the sequence length (Chen and Ying, 2024; Liang et al., 2025b; Feng et al., 2025; Ren et al., 2024, 2025; Liang et al., 2026). Consequently, they fail to capture the parallel-generation

regime in DLM decoding. A notable exception is Dmitriev et al. (2026), which developed sharp and adaptive guarantees for MDMs, with convergence rate governed by low-complexity measures of data. In addition, sample complexity guarantees have recently been established for DLMs (Wu and Cai, 2026b; Zhang et al., 2026; Cho and Wu, 2026).

Empirical confidence-based strategies. Early confidence-based approaches typically use a fixed-size strategy, which prescribes an unmasking size K and unmasks the top-K most confident tokens at each iteration (Chang et al., 2022). This idea has been instantiated with several ranking criteria, including top-K probability (Nie et al., 2025), top-K margin (Kim et al., 2025), and top-K entropy (Ben-Hamu et al., 2025). More recent work considers adaptive strategies, where the number of newly unmasked tokens is determined on the fly. Examples include threshold-based rules that reveal tokens whose confidence exceeds a prescribed level (Yu et al., 2025), and budget-based strategies that continue unmasking while the accumulated uncertainty remains within a budget (Ben-Hamu et al., 2025; Wu et al., 2025). The methods analyzed in this paper fall into this category. Beyond those one-step confidence rules, several works combine confidence information with more structured decoding procedures. For instance, Wei et al. (2025) proposes a two-stage method that alternates between exploratory and accelerated decoding strategies depending on the confidence region. In addition, another line of work seeks to improve decoding by explicitly planning ahead for future iterations rather than making purely myopic decisions, including lookahead-based exploration methods (Fu et al., 2025; Lee et al., 2025; Peng et al., 2025) and trial-and-verification procedures (Wu and Zhang, 2025; Li et al., 2026a). Finally, to mitigate the limitation of standard MDMs that fix tokens once they are unmasked, recent work has also studied confidence-based remasking mechanisms that allow low-confidence tokens to be revised in later iterations (Wang et al., 2025; Shi et al., 2024).

Theory for continuous diffusion models. A rich body of work has developed sampling convergence guarantees for continuous diffusion models in $\mathbb{R}^d$ (Lee et al., 2022; Chen et al., 2022; Lee et al., 2023; Chen et al., 2023a; Benton et al., 2023; Li and Yan, 2024b; Li et al., 2024b), with state-of-the-art iteration complexity of $\widetilde{O}(\sqrt{d/\varepsilon})$ in KL divergence and $\widetilde{O}(\sqrt{d}/\varepsilon)$ in total variation distance (Jiao et al., 2025b; Zhang et al., 2025). A variety of provably accelerated samplers have been further developed (Li and Cai, 2024; Li and Jiao, 2024; Gatmiry et al., 2026; Huang et al., 2024a; Li et al., 2024a, 2025b; Jiao et al., 2025a). A theme relevant to our work is the adaptivity to (unknown) low-complexity structure. Several recent works (Li and Yan, 2024a; Liang et al., 2025a; Huang et al., 2024b; Potaptchik et al., 2024; Li et al., 2025a) show that when the data distribution concentrates near a k-dimensional structure with $k \ll d$, the dependence on the ambient dimension d can be replaced by the intrinsic dimension k, yielding substantial sampling acceleration. Beyond sampling convergence, end-to-end statistical guarantees have been developed to demonstrate the optimality of diffusion-based distribution learning and sample generation (Oko et al., 2023; Wibisono et al., 2024; Zhang et al., 2024; Cai and Li, 2025; Gatmiry et al., 2024; Chen et al., 2024). Further results tailored to structured data have also been established, including distributions supported on low-dimensional linear subspaces or manifolds (Chen et al., 2023b; Tang and Yang, 2024; Azangulov et al., 2024; Wang et al., 2024; Yakovlev and Puchkin, 2025; Wu and Cai, 2026a), distributions with low-complexity dependence (Fan et al., 2025), and factor models (Chen et al., 2025a).

1.3 Notation

For any integer $n > 0$, let $[n] \coloneqq \{1, 2, \ldots, n\}$. For any set $\mathcal{S}$, we use the upper-case letter S to denote its cardinality $|\mathcal{S}|$. We denote by $\mathbb{X}$ the (discrete) vocabulary of tokens. We write M for the mask and extend the vocabulary by adding M to define $\overline{\mathbb{X}} = \mathbb{X} \cup \{\mathsf{M}\}$. Given a sequence $x \in \mathbb{X}^L$, we denote its i-th coordinate by x^i for $i \in [L]$. For any index set $\mathcal{S} \subseteq [L]$, let $x^{\mathcal{S}} = (x_i)_{i \in \mathcal{S}}$ be the subsequence in $\mathbb{X}^S$ formed by the entries of x with indices in $\mathcal{S}$. Define the projection operator $\mathcal{P}_{\mathcal{S}} : \mathbb{X}^L \to \overline{\mathbb{X}}^L$ by $[\mathcal{P}_{\mathcal{S}}(x)]^i = x^i$ if $i \in \mathcal{S}$ and $[\mathcal{P}_{\mathcal{S}}(x)]^i = \mathsf{M}$ otherwise.

For a random variable X, we use p_X to denote its distribution, and (by a slight abuse of notation) also its probability mass/density function. The entropy of X is defined by $H(X) = H(p_X) \coloneqq \mathbb{E}_{x \sim p_X}[-\log p_X(x)]$. For a random vector $(X, Y) \sim p_{X,Y}$, the Kullback-Leibler divergence between p_X and p_Y is $\mathsf{KL}(p_X \,\|\, p_Y) \coloneqq \mathbb{E}_{x \sim p_X}[\log \frac{p_X(x)}{p_Y(x)}]$. The conditional entropy of X given Y is defined as $H(X \mid Y) \coloneqq \mathbb{E}_{(x,y) \sim p_{X,Y}}[-\log p_{X|Y}(x \mid y)]$. The mutual information between X and Y is defined

as $I(X;Y) \coloneqq \mathsf{KL}(p_{X,Y} \,\|\, p_X p_Y)$. For a random vector $(X,Y,Z) \sim p_{X,Y,Z}$, the conditional mutual information between X and Y given Z is defined as $I(X;Y \mid Z) \coloneqq \mathsf{KL}(p_{X,Y,Z} \,\|\, p_{X|Z} p_{Y|Z} p_Z)$.

In addition to these standard information-theoretic measures, we also define *pointwise* conditional entropy and conditional mutual information as follows:

$$\mathtt{H}(X \mid Z) \coloneqq f(Z) \quad \text{with} \quad f(z) \coloneqq H(p_{X|Z=z}), \tag{1}$$

$$\mathtt{I}(X,Y \mid Z) \coloneqq g(Z) \quad \text{with} \quad g(z) \coloneqq \mathsf{KL}(p_{X,Y|Z=z} \,\|\, p_{X|Z=z} p_{Y|Z=z}). \tag{2}$$

We note that $\mathtt{H}(X \mid Z)$ and $\mathtt{I}(X,Y \mid Z)$ are random variables measurable with respect to Z. Moreover, taking expectation with respect to $Z \sim p_Z$ recovers the standard quantities:

$$\mathbb{E}_Z[\mathtt{H}(X \mid Z)] = H(X \mid Z) \quad \text{and} \quad \mathbb{E}_Z[\mathtt{I}(X,Y \mid Z)] = I(X;Y \mid Z).$$

2 Preliminaries

In this section, we provide an overview of the fundamental concepts underlying DLMs.

Forward (masking) process. Consider a token sequence $X_0 = (X_0^1, X_0^2, \ldots, X_0^L) \in \mathbb{X}^L$ sampled from the data distribution p_{data}. The forward process progressively corrupts X_0 by replacing its tokens with the mask symbol M until the entire sequence is masked. Formally, let X_t denote the partially masked sequence obtained after t masking steps and let $\mathcal{M}_t \coloneqq \{i \in [L] \colon X_t^i = \mathsf{M}\}$ denote the index set of masked tokens. At each step $t \geq 1$, we select a masking set $\mathcal{S}_t \subseteq [L] \setminus \mathcal{M}_{t-1}$ and update X_{t-1} to X_t by replacing tokens at positions in the set $\mathcal{S}_t$ with M:

$$X_t^i = \begin{cases} \mathsf{M}, & \text{if } i \in \mathcal{S}_t, \\ X_{t-1}^i, & \text{if } i \notin \mathcal{S}_t. \end{cases}$$

The masking process continues until all tokens have been masked.

Mask predictor training. DLMs are trained to reverse the above forward masking process by learning the conditional distribution of masked tokens given the unmasked ones. A notable feature of mask predictor training is that this posterior distribution is time-independent—it depends only on the unmasked context and not on the masking step t (Ou et al., 2024; Zheng et al., 2024). More precisely, for each token index $i \in [L]$ and any set of unmasked positions $\mathcal{S} \subseteq [L] \setminus \{i\}$, we define the true conditional marginal distribution of X_0^i given $X_0^{\mathcal{S}}$ under the data distribution p_{data} as

$$p^i(x \mid z^{\mathcal{S}}) \coloneqq \mathbb{P}\big\{X_0^i = x \mid X_0^{\mathcal{S}} = z^{\mathcal{S}}\big\} \qquad \forall\, x \in \mathbb{X},\, z^{\mathcal{S}} \in \mathbb{X}^{|\mathcal{S}|}. \tag{3}$$

The objective is to learn estimators $p_\theta^i(\cdot \mid \cdot)$ for each conditional marginal $p^i(\cdot \mid \cdot)$, thereby forming a mask predictor $p_\theta = \prod_{i=1}^L p_\theta^i$.

The mask predictor is trained by maximizing the log-likelihood of masked tokens conditioned on unmasked ones:

$$\max_{p_\theta = \prod_{i=1}^L p_\theta^i} \mathbb{E}_{X_0,\mathcal{S}} \left[\frac{1}{|\mathcal{S}^c|} \sum_{i \in \mathcal{S}^c} \log p_\theta^i\big(X_0^i \mid X_0^{\mathcal{S}}\big) \right], \tag{4}$$

where $X_0 \sim p_{\mathsf{data}}$ and $\mathcal{S}$ follows a prescribed masking-set distribution. This objective is equivalent to minimizing the KL divergence between the data distribution and the distribution induced by the product of learned conditional marginals over partially masked sequences. In practice, the expectation is approximated by averaging over a finite number of training samples and random masking patterns.

We remark that the product of the true conditional marginals $\prod_{i=1}^L p^i$ defined in (3) is the maximizer of the training objective (4).

Reverse (unmasking) process. Given a trained mask predictor p_θ, DLMs produce a token sequence through an iterative unmasking procedure. Starting from a fully masked sequence $Y_0 = (\mathsf{M}, \mathsf{M}, \ldots, \mathsf{M}) \in \overline{\mathbb{X}}^L$ with cumulative unmasking set $\mathcal{W}_0 = \varnothing$, we repeat the following two steps for iteration $t = 1, 2, \ldots$ until all tokens have been unmasked.

Algorithm 1 Entropy sum-based decoding strategy

Input: Mask predictor p_θ and entropy threshold η
Initialize $k \leftarrow 0$, $t \leftarrow 0$, token sequence $Y_0 \leftarrow (\mathsf{M}, \ldots, \mathsf{M})$
Draw a permutation $\Pi = (\Pi_1, \ldots, \Pi_L)$ of $[L]$ uniformly at random
while $k < L$ **do** ▷ Repeat until all tokens are generated
  $t \leftarrow t + 1$ ▷ Start iteration t
  $Y_t \leftarrow Y_{t-1}$, $\mathcal{D}_t \leftarrow \varnothing$, $S \leftarrow 0$ ▷ Initialize current iterate, unmasking set, and entropy sum
  repeat
    $k \leftarrow k + 1$
    sample $Y_t^{\Pi_k} \sim p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})$ ▷ Unmask token with index Π_k
    $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{\Pi_k\}$ ▷ Add index Π_k to unmasking set $\mathcal{D}_t$
    $S \leftarrow S + \mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})\big)$ ▷ Update entropy sum
  until $S > \eta$ or $k = L$ ▷ Iteration termination criteria
$T \leftarrow t$ ▷ Record total number of iterations
Output: Generated token sequence Y_T

1. **Select unmasking set.** Choose a subset $\mathcal{D}_t$ of the currently masked positions $[L] \setminus \mathcal{W}_{t-1}$ according to a *decoding strategy*, and update the cumulative unmasking set $\mathcal{W}_t = \mathcal{W}_{t-1} \cup \mathcal{D}_t$.
2. **Sample tokens.** Draw the tokens for all positions in $\mathcal{D}_t$ *independently* and *simultaneously* from the learned mask predictor $p_\theta = \prod_{i=1}^{L} p_\theta^i$, namely,

$$\widehat{Y}^i \sim p_\theta^i\big(\cdot \mid Y_{t-1}^{\mathcal{W}_{t-1}}\big), \quad \forall\, i \in \mathcal{D}_t. \tag{5}$$

   The sequence is then updated by inserting the newly sampled tokens:

$$Y_t^i = \begin{cases} \widehat{Y}^i, & \text{if } i \in \mathcal{D}_t, \\ Y_{t-1}^i, & \text{if } i \notin \mathcal{D}_t. \end{cases} \tag{6}$$

The output is the fully unmasked sequence Y_T, where T is the (random) number of iterations.

It is worth noting that given a fixed mask predictor p_θ, the sampling process of DLMs and the sampled distribution p_{Y_T} are completely determined by the choice of decoding strategy, which specifies the unmasking set $\mathcal{D}_t$ at each iteration t.

3 Main results

In this section, we present two entropy-based decoding strategies for DLMs and establish their theoretical guarantees. We begin with the entropy sum-based strategy, which is the main algorithmic result of the paper, and then study a maximum entropy-based variant to illustrate how the choice of entropy criterion affects sampling efficiency.

3.1 Entropy sum-based decoding strategy

We first introduce the entropy sum-based decoding strategy. Its key feature is that each decoding iteration is determined by the total uncertainty, measured by entropy, accumulated over the current set of unmasked tokens. This leads to an adaptive rule that enlarges the unmasking set automatically in low-uncertainty regions and shrinks it when the mask predictor is uncertain.

Decoding procedure. At a high level, the entropy sum-based strategy scans the remaining masked positions in a random order and continues unmasking tokens until the cumulative entropy of the current set exceeds a threshold.

Concretely, we first draw a permutation Π of $[L]$ uniformly at random, where Π_k represents the index of the k-th token to be unmasked. We then repeat the following decoding procedure until all tokens have been generated. At each iteration $t \geq 1$, we initialize the current unmasking set $\mathcal{D}_t = \varnothing$ and the running entropy sum $S = 0$. Recall that for any set $\mathcal{W}$, we use the upper-case letter $W = |\mathcal{W}|$ to denote its size. Given the partially unmasked sequence Y_{t-1} from the previous iteration, in which the

tokens indexed by the cumulative unmasking set

$$\mathcal{W}_{t-1} = \bigcup_{i=1}^{t-1} \mathcal{D}_i = \{\Pi_1, \ldots, \Pi_{W_{t-1}}\}$$

have been generated, we continue along the permutation and examine the positions Π_k for

$$k = W_{t-1} + 1, W_{t-1} + 2, \ldots$$

For each such position Π_k, we compute the entropy of the corresponding conditional marginal $p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})$, generate the token $Y_t^{\Pi_k}$ from this marginal, add the index Π_k to the current unmasking set $\mathcal{D}_t$, and update the running entropy sum S by $\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})\big)$[1].

The iteration terminates once the running entropy sum S exceeds a prescribed threshold η or all L tokens have been generated. Equivalently, the index set of newly unmasked tokens at iteration t has the form

$$\mathcal{D}_t = \{\Pi_k \colon W_{t-1} < k \le k_t\},$$

where k_t is defined by

$$k_t \coloneqq \min\left\{ i > W_{t-1} : \sum_{k=W_{t-1}+1}^{i} \mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})\big) > \eta \right\} \wedge L.$$

In other words, k_t is the first permutation index beyond W_{t-1} such that the cumulative entropy over permutation indices $k = W_{t-1} + 1, \ldots, k_t$ exceeds η, or L if no such index appears before the sequence is fully unmasked.

The complete decoding procedure is summarized in Algorithm 1.

Remark 1. The random permutation Π is introduced solely to simplify the presentation and analysis. Equivalently, the algorithm may be viewed as selecting the next token from the remaining masked positions uniformly at random.

An important observation is that, for each iteration t, conditioned on the permutation Π, the cumulative unmasking set $\mathcal{W}_{t-1}$, and the partially unmasked sequence Y_{t-1}, the unmasking set $\mathcal{D}_t$ (and thus the cumulative unmasking set $\mathcal{W}_t = \mathcal{W}_{t-1} \cup \mathcal{D}_t$) is fully determined by the conditional entropies:

$$\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})\big), \quad k = W_{t-1} + 1, W_{t-1} + 2, \ldots, L.$$

Once $\mathcal{D}_t$ is determined, we obtain the iterate Y_t by generating the tokens in $\mathcal{D}_t$ independently according to the corresponding conditional distributions

$$Y_t^i \sim p_\theta^i(\cdot \mid Y_{t-1}), \quad i \in \mathcal{D}_t,$$

or equivalently,

$$Y_t^{\Pi_k} \sim p_\theta^{\Pi_k}(\cdot \mid Y_{t-1}), \quad W_{t-1} + 1 \le k \le W_t.$$

Therefore, the overall sampling process can be summarized by the following Markov chain:

$$\begin{aligned} \Pi \to \big\{\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_0)\big)\big\}_{k=1}^{L} \to (\mathcal{D}_1, \mathcal{W}_1) \to Y_1 \to \big\{\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_1)\big)\big\}_{k=W_1+1}^{L} \to (\mathcal{D}_2, \mathcal{W}_2) \to Y_2 \\ \to \cdots \to \big\{\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{T-1})\big)\big\}_{k=W_{T-1}+1}^{L} \to (\mathcal{D}_T, \mathcal{W}_T) \to Y_T, \end{aligned} \tag{7}$$

where T denotes the (random) total number of iterations.

3.2 Theoretical guarantee

We now state the performance guarantee for the entropy sum-based decoding scheme (Algorithm 1) in Theorem 1. The proof sketch can be found in Section A.2.

[1]Here, we use the notation $\mathtt{H}$ defined in (1) to denote the pointwise conditional entropy. It is a random variable depending on Y_{t-1}, as opposed to the standard conditional entropy H that averages over the randomness in Y_{t-1}.

Theorem 1. *Assume that the mask predictor p_θ is optimal, namely, $p_\theta = \prod_{i=1}^{L} p^i$ is the product of true conditional marginals. For any $\varepsilon > 0$, set the threshold to $\eta = \varepsilon/\big(4(\log_2 L + 1)\big)$. Then the entropy sum-based strategy in Algorithm 1 achieves ε-accurate sampling in KL divergence, i.e., $\mathbb{E}_\Pi\big[\mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi})\big] \le \varepsilon$. Moreover, the expected number of iterations satisfies*

$$\mathbb{E}[T] \le 4\left(\frac{H(X_0)}{\varepsilon} + 1\right)(\log_2 L + 1) + 1, \tag{8}$$

where $H(X_0)$ denotes the entropy of the target data distribution.

To the best of our knowledge, this result provides the first theoretical guarantee for confidence-based decoding strategies in DLMs. We discuss several important implications below.

- *Iteration complexity.* The expected number of iterations required to achieve ε-accuracy in KL divergence is at most

$$\widetilde{O}\left(\frac{H(X_0)}{\varepsilon}\right). \tag{9}$$

  Up to a logarithmic dependence on the sequence length L, this iteration complexity depends only on the entropy of the target data distribution, which can be viewed as a measure of its intrinsic complexity. Consequently, when the target distribution has low entropy, namely $H(X_0) \ll L$, the entropy sum-based decoding scheme yields a substantial sampling speedup.

- *Adaptivity to intrinsic structure.* The above sampling acceleration is achieved without requiring any prior knowledge of the data distribution or hyperparameter tuning. This highlights the distribution-agnostic adaptivity of the entropy-based decoding—it automatically exploits low-complexity structure in the target distribution to accelerate generation.

3.3 Connection to empirical entropy sum-based strategies

Algorithm 1 is closely related to the entropy-based decoding strategy recently proposed in Ben-Hamu et al. (2025). Their method sorts the remaining tokens by conditional entropy and selects the largest subset for which the gap between the entropy sum and the maximum entropy remains below a prescribed threshold. Numerical results in Ben-Hamu et al. (2025) demonstrate that this strategy achieves substantial generation speedups while matching or surpassing the generation quality of uniform unmasking schemes. The main difference is that Algorithm 1 scans token positions in a uniformly random order instead of a deterministic sorted order.

Despite its empirical success, entropy sum-based decoding with sorted ordering can incur large KL sampling error even for a low-entropy target distribution. The following proposition formalizes this limitation; its proof can be found in Appendix C.

Proposition 1. *For any entropy budget H and threshold η satisfying $H = O(\eta L)$, there exists a distribution over $\{0,1\}^L$ with entropy at most H such that the entropy-sum decoding strategy with sorted ordering requires $\Omega(H/\eta)$ iterations and incurs KL sampling error $\Omega(H)$.*

Therefore, we adopt a uniformly random ordering to make the adaptive decoding process analytically tractable and enables us to establish nontrivial KL sampling error guarantees. Our main objective is to provide a theoretical justification for a central component of confidence-based decoding: adaptive set-size selection where the entropy budget controls how many tokens are decoded at each iteration. Determining under what conditions similar guarantees can be established for deterministic sorted ordering remains an interesting direction for future work.

3.4 Effect of entropy criterion

A natural question is how the choice of entropy criterion affects decoding efficiency. To investigate this, we present a second entropy-based decoding strategy—the *maximum entropy-based strategy*—which differs from the entropy sum-based strategy in a single design choice. Instead of controlling the cumulative entropy of the entire set, it terminates an iteration as soon as any single token's conditional entropy exceeds a prescribed threshold, subject to an explicit cap on the set size.

Algorithm 2 Maximum entropy-based decoding strategy

Input: Mask predictor p_θ, entropy threshold η, and maximum unmasking size $S_{\max}$
Initialize $k \leftarrow 0$, $t \leftarrow 0$, and token sequence $Y_0 \leftarrow (\mathsf{M}, \ldots, \mathsf{M})$
Draw a permutation $\Pi = (\Pi_1, \ldots, \Pi_L)$ of $[L]$ uniformly at random
while $k < L$ **do** ▷ Repeat until all tokens are generated
  $t \leftarrow t + 1$ ▷ Start iteration t
  $Y_t \leftarrow Y_{t-1}$, $\mathcal{D}_t \leftarrow \varnothing$ ▷ Initialize current iterate and unmasking set
  repeat
    $k \leftarrow k + 1$
    $Y_t^{\Pi_k} \sim p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})$ ▷ Unmask token with index Π_k
    $\mathcal{D}_t \leftarrow \mathcal{D}_t \cup \{\Pi_k\}$ ▷ Add index Π_k to unmasking set $\mathcal{D}_t$
  until $\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{t-1})\big) > \eta$, or $|\mathcal{D}_t| = S_{\max}$, or $k = L$ ▷ Iteration termination criteria
$T \leftarrow t$ ▷ Record total number of iterations
Output: Generated token sequence Y_T

As we will see shortly, this strategy still achieves an iteration complexity sublinear in the sequence length. However, its sampling guarantee is weaker in two respects: the iteration complexity bound is looser, and the parameter choice requires knowledge of the target distribution. This contrast highlights the importance of how entropy is incorporated into the design of decoding strategies.

Maximum entropy-based decoding strategy. As before, we begin by drawing a random permutation Π of $[L]$. At iteration $t \geq 1$, given the previous iterate Y_{t-1} with cumulative unmasking set $\mathcal{W}_{t-1}$, we scan the remaining positions in the permutation order and keep adding them to the current set $\mathcal{D}_t$. The iteration ends as soon as one of the following occurs:

(i) the entropy of the most recently unmasked token exceeds the threshold η (this token is included in the set $\mathcal{D}_t$);

(ii) the current set reaches the cap size $S_{\max}$;

(iii) all L tokens have been generated.

Formally, the index set $\mathcal{D}_t$ of newly unmasked tokens at iteration t is given by

$$\mathcal{D}_t = \big\{\Pi_k \colon W_{t-1} < k \leq k_t\big\},$$

where k_t is the first permutation index after W_{t-1} for which the entropy exceeds η, or $W_{t-1} + S_{\max}$ if no such index appears within the next $S_{\max}$ positions, or L if no such index appears before all tokens are generated. In other words, k_t is defined by

$$k_t \coloneqq \min\big\{i > W_{t-1} : \mathtt{H}\big(p_\theta^{\Pi_i}(\cdot \mid Y_{t-1})\big) > \eta\big\} \wedge (W_{t-1} + S_{\max}) \wedge L.$$

A complete description of the unmasking scheme is given in Algorithm 2.

Compared with the entropy sum-based decoding scheme, this strategy differs only in the stopping criterion: (1) it reacts to the entropy of a single token rather than the cumulative entropy of the whole set, and (2) it explicitly imposes a maximum unmasking size $S_{\max}$. Despite this difference, these two entropy-based strategies share the same Markovian structure: the unmasking sets $\mathcal{D}_t$ and $\mathcal{W}_t$ at iteration t are fully determined by the previous iterate Y_{t-1} and cumulative unmasking set $\mathcal{W}_{t-1}$.

Theoretical guarantee. We now present the performance guarantee of the maximum entropy-based inference scheme (Algorithm 2) in Theorem 2 below. The proof can be found in Appendix D.

Theorem 2. *Assume that the mask predictor p_θ is optimal, namely, $p_\theta = \prod_{i=1}^L p^i$ is the product of true conditional marginals. For any $\varepsilon > 0$, set the entropy threshold to $\eta = \sqrt{\varepsilon H(X_0)/L}$ and the maximum unmasking size to $S_{\max} = \lfloor \sqrt{\varepsilon L / H(X_0)} \rfloor \vee 1$. Then the maximum entropy-based unmasking strategy in Algorithm 2 achieves ε-accurate sampling in KL divergence, namely, $\mathbb{E}_\Pi\big[\mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi})\big] \leq \varepsilon$. Moreover, the expected number of iterations satisfies*

$$\mathbb{E}[T] \leq 2\sqrt{\frac{H(X_0)L}{\varepsilon}} + 1. \tag{10}$$

In words, the maximum entropy-based decoding strategy attains ε-accuracy in KL divergence with an expected number of iterations scaling as $O\big(\sqrt{H(X_0)L/\varepsilon}\big)$. As a consequence, when the entropy of the target data distribution satisfies $H(X_0) \ll L$, it achieves a sublinear iteration complexity in the sequence length L, thereby implying a provable speedup over fully sequential generation.

Nevertheless, compared with the entropy sum-based strategy, this maximum entropy-based strategy has two important shortcomings. First, in the parallel sampling regime, where the iteration complexity is smaller than the sequence length L, the bound in (10) is worse than that in (9) by a factor of $\sqrt{\varepsilon L/H(X_0)} \gg 1$. Second, the parameter choices η and $S_{\max}$ needed to achieve a sublinear iteration complexity require explicit knowledge of the entropy $H(X_0)$ of the target data distribution. This stands in stark contrast to the entropy sum-based strategy, which is fully distribution-adaptive.

Since the theoretical guarantees for these two entropy-based decoding strategies are established under the same analysis framework, comparing them suggests that the choice of entropy criterion plays a critical role in efficient entropy-based decoding. An important future direction is thus to develop a deeper understanding of how different entropy criteria affect the sampling performance of DLMs.

4 Discussion

In this work, we have shown that entropy-based unmasking strategies can provably accelerate sampling, achieving an iteration complexity that is sublinear in the sequence length when the data distribution has low entropy. Notably, they reveal that entropy-based strategies can adapt to the unknown intrinsic complexity of the data distribution without requiring any prior knowledge or hyperparameter tuning.

Moving forward, it would be of interest to understand the performance of the entropy-based decoding schemes based on a deterministic order (e.g., a sorted order) instead of the random order considered in this work. Also, it would be valuable to extend our analysis to other confidence measures, such as the mass of the most likely token and the mass gap between the two most likely tokens. Second, it would be important to understand the impact of training error in the mask predictor on the sampling performance under confidence-based decoding, which may yield insights into the design of more effective training objectives for DLMs. Third, another important direction is to establish lower bounds on the iteration complexity of confidence-based decoding schemes, which would shed light on the fundamental limits of sampling acceleration when using this class of strategies.

Acknowledgments and Disclosure of Funding

C. Cai is supported in part by the NSF CAREER Award CCF-2541600 and the NSF grant DMS-2515333. G. Li is supported in part by the Chinese University of Hong Kong Direct Grant for Research and the Hong Kong Research Grants Council ECS 24305724 and GRF 14307525.

References

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. (2021). Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993.

Azangulov, I., Deligiannidis, G., and Rousseau, J. (2024). Convergence of diffusion models under the manifold hypothesis in high-dimensions. *arXiv preprint arXiv:2409.18804*.

Ben-Hamu, H., Gat, I., Severo, D., Nolte, N., and Karrer, B. (2025). Accelerated sampling from masked diffusion models via entropy bounded unmasking. *arXiv preprint arXiv:2505.24857*.

Benton, J., De Bortoli, V., Doucet, A., and Deligiannidis, G. (2023). Linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*.

Bie, T., Cao, M., Chen, K., Du, L., Gong, M., Gong, Z., Gu, Y., Hu, J., Huang, Z., Lan, Z., et al. (2025). Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*.

Cai, C. and Li, G. (2025). Minimax optimality of the probability flow ode for diffusion models. *arXiv preprint arXiv:2503.09583*.

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. (2022). A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279.

Chang, H., Zhang, H., Jiang, L., Liu, C., and Freeman, W. T. (2022). Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325.

Chen, H., Lee, H., and Lu, J. (2023a). Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR.

Chen, H. and Ying, L. (2024). Convergence analysis of discrete diffusion model: Exact implementation through uniformization. *arXiv preprint arXiv:2402.08095*.

Chen, J., Liang, Y., and Liu, Z. (2026). Dflash: Block diffusion for flash speculative decoding. *arXiv preprint arXiv:2602.06036*.

Chen, M., Huang, K., Zhao, T., and Wang, M. (2023b). Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *International Conference on Machine Learning*, pages 4672–4712. PMLR.

Chen, M., Xu, R., Xu, Y., and Zhang, R. (2025a). Diffusion factor models: Generating high-dimensional returns with factor structure. *arXiv preprint arXiv:2504.06566*.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. R. (2022). Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*.

Chen, S., Cong, K., and Li, J. (2025b). Optimal inference schedules for masked diffusion models. *arXiv preprint arXiv:2511.04647*.

Chen, S., Kontonis, V., and Shah, K. (2024). Learning general gaussian mixtures with efficient score matching. *arXiv preprint arXiv:2404.18893*.

Cheng, X., Yu, X., Shao, C., Li, J., Xiong, Y., Qian, Y., Zhu, J., Ma, S., Zhang, X., Ye, J., et al. (2026). Dspark: Confidence-scheduled speculative decoding with semi-autoregressive generation. *arXiv preprint arXiv:2607.05147*.

Cho, C. and Wu, Y. (2026). Minimax optimality of score-entropy discrete diffusion. *arXiv preprint arXiv:2608.20635*.

Dmitriev, D., Huang, Z., and Wei, Y. (2026). Efficient sampling with discrete diffusion models: Sharp and adaptive guarantees. *arXiv preprint arXiv:2602.15008*.

Fan, J., Gu, Y., and Li, X. (2025). Optimal estimation of a factorizable density using diffusion models with relu neural networks. *arXiv preprint arXiv:2510.03994*.

Feng, G., Geng, Y., Guan, J., Wu, W., Wang, L., and He, D. (2025). Theoretical benefit and limitation of diffusion language model. *arXiv preprint arXiv:2502.09622*.

Fu, H., Huang, B., Adams, V., Wang, C., Srinivasan, V., and Jiao, J. (2025). From bits to rounds: Parallel decoding with exploration for diffusion language models. *arXiv preprint arXiv:2511.21103*.

Gatmiry, K., Chen, S., and Salim, A. (2026). High-accuracy and dimension-free sampling with diffusions. *arXiv preprint arXiv:2601.10708*.

Gatmiry, K., Kelner, J., and Lee, H. (2024). Learning mixtures of gaussians using diffusion models. *arXiv preprint arXiv:2404.18869*.

Gong, S., Agarwal, S., Zhang, Y., Ye, J., Zheng, L., Li, M., An, C., Zhao, P., Bi, W., Han, J., et al. (2024). Scaling diffusion language models via adaptation from autoregressive models. *arXiv preprint arXiv:2410.17891*.

Huang, D. Z., Huang, J., and Lin, Z. (2024a). Convergence analysis of probability flow ODE for score-based generative models. *arXiv preprint arXiv:2404.09730*.

Huang, Z., Wei, Y., and Chen, Y. (2024b). Denoising diffusion probabilistic models are optimally adaptive to unknown low dimensionality. *arXiv preprint arXiv:2410.18784*.

Jiao, Y., Li, N., Cai, C., and Li, G. (2025a). Are first-order diffusion samplers really slower? a fast forward-value approach. *arXiv preprint arXiv:2512.24927*.

Jiao, Y., Zhou, Y., and Li, G. (2025b). Optimal convergence analysis of ddpm for general distributions. *arXiv preprint arXiv:2510.27562*.

Kim, J., Shah, K., Kontonis, V., Kakade, S., and Chen, S. (2025). Train for the worst, plan for the best: Understanding token ordering in masked diffusions. *arXiv preprint arXiv:2502.06768*.

Labs, I., Khanna, S., Kharbanda, S., Li, S., Varma, H., Wang, E., Birnbaum, S., Luo, Z., Miraoui, Y., Palrecha, A., et al. (2025). Mercury: Ultra-fast language models based on diffusion. *arXiv preprint arXiv:2506.17298*.

Lavenant, H. and Zanella, G. (2025). Error bounds and optimal schedules for masked diffusions with factorized approximations. *arXiv preprint arXiv:2510.25544*.

Lee, H., Lu, J., and Tan, Y. (2022). Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35:22870–22882.

Lee, H., Lu, J., and Tan, Y. (2023). Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR.

Lee, S., Kim, S., Park, J., and Park, D. (2025). Lookahead unmasking elicits accurate decoding in diffusion language models. *arXiv preprint arXiv:2511.05563*.

Li, G. and Cai, C. (2024). Provable acceleration for diffusion models under minimal assumptions. *arXiv preprint arXiv:2410.23285*.

Li, G. and Cai, C. (2025). Breaking ar's sampling bottleneck: Provable acceleration via diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Li, G., Cai, C., and Wei, Y. (2025a). Dimension-free convergence of diffusion models for approximate gaussian mixtures. *arXiv preprint arXiv:2504.05300*.

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. (2024a). Accelerating convergence of score-based diffusion models, provably. *arXiv preprint arXiv:2403.03852*.

Li, G. and Jiao, Y. (2024). Improved convergence rate for diffusion probabilistic models. *arXiv preprint arXiv:2410.13738*.

Li, G., Wei, Y., Chi, Y., and Chen, Y. (2024b). A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*.

Li, G. and Yan, Y. (2024a). Adapting to unknown low-dimensional structures in score-based diffusion models. *arXiv preprint arXiv:2405.14861*.

Li, G. and Yan, Y. (2024b). $O(d/T)$ convergence theory for diffusion probabilistic models under minimal assumptions. *arXiv preprint arXiv:2409.18959*.

Li, G., Zhou, Y., Wei, Y., and Chen, Y. (2025b). Faster diffusion models via higher-order approximation. *arXiv preprint arXiv:2506.24042*.

Li, M., Jiang, H., Cheng, S., Fu, H., Cai, Y., Huang, B., Ye, T., Chen, X., and Van Hentenryck, P. (2026a). Plan, verify and fill: A structured parallel decoding approach for diffusion language models. *arXiv preprint arXiv:2601.12247*.

Li, N., Jiao, Y., Cai, C., and Li, G. (2026b). Convergeflow: Language flow with provable convergence to token embeddings. *arXiv preprint arXiv:2608.23551*.

Liang, J., Huang, Z., and Chen, Y. (2025a). Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*.

Liang, Y., Huang, R., Lai, L., Shroff, N., and Liang, Y. (2025b). Absorb and converge: Provable convergence guarantee for absorbing discrete diffusion models. *arXiv preprint arXiv:2506.02318*.

Liang, Y., Tan, Z., Shroff, N., and Liang, Y. (2026). Sharp convergence rates for masked diffusion models. *arXiv preprint arXiv:2602.22505*.

Lou, A., Meng, C., and Ermon, S. (2023). Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. (2025). Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Oko, K., Akiyama, S., and Suzuki, T. (2023). Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR.

Ou, J., Nie, S., Xue, K., Zhu, F., Sun, J., Li, Z., and Li, C. (2024). Your absorbing discrete diffusion secretly models the conditional distributions of clean data. *arXiv preprint arXiv:2406.03736*.

Peng, F. Z., Bezemek, Z., Patel, S., Rector-Brooks, J., Yao, S., Bose, A. J., Tong, A., and Chatterjee, P. (2025). Path planning for masked diffusion model sampling. *arXiv preprint arXiv:2502.03540*.

Potaptchik, P., Azangulov, I., and Deligiannidis, G. (2024). Linear convergence of diffusion models under the manifold hypothesis. *arXiv preprint arXiv:2410.09046*.

Prabhudesai, M., Wu, M., Zadeh, A., Fragkiadaki, K., and Pathak, D. (2025). Diffusion beats autoregressive in data-constrained settings. *arXiv preprint arXiv:2507.15857*.

Ren, Y., Chen, H., Rotskoff, G. M., and Ying, L. (2024). How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework. *arXiv preprint arXiv:2410.03601*.

Ren, Y., Chen, H., Zhu, Y., Guo, W., Chen, Y., Rotskoff, G. M., Tao, M., and Ying, L. (2025). Fast solvers for discrete diffusion models: Theory and applications of high-order algorithms. *arXiv preprint arXiv:2502.00234*.

Sahoo, S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J., Rush, A., and Kuleshov, V. (2024). Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. (2024). Simplified and generalized masked diffusion for discrete data. *Advances in neural information processing systems*, 37:103131–103167.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. (2020). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Song, Y., Zhang, Z., Luo, C., Gao, P., Xia, F., Luo, H., Li, Z., Yang, Y., Yu, H., Qu, X., et al. (2025). Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*.

Tang, R. and Yang, Y. (2024). Adaptivity of diffusion models to manifold structures. In *International Conference on Artificial Intelligence and Statistics*, pages 1648–1656. PMLR.

Wang, G., Schiff, Y., Sahoo, S. S., and Kuleshov, V. (2025). Remasking discrete diffusion models with inference-time scaling. *arXiv preprint arXiv:2503.00307*.

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. (2024). Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*.

Wei, Q., Zhang, Y., Liu, Z., Liu, D., and Zhang, L. (2025). Accelerating diffusion large language models with slowfast sampling: The three golden principles. *arXiv preprint arXiv:2506.10848*.

Wibisono, A., Wu, Y., and Yang, K. Y. (2024). Optimal score estimation via empirical bayes smoothing. *arXiv preprint arXiv:2402.07747*.

Wu, C., Zhang, H., Xue, S., Liu, Z., Diao, S., Zhu, L., Luo, P., Han, S., and Xie, E. (2025). Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.

Wu, J. and Cai, C. (2026a). Diffusion models are statistically optimal for learning low-dimensional multi-modal distributions. *arXiv preprint arXiv:2605.30153*.

Wu, J. and Cai, C. (2026b). Transformer-based masked diffusion models provably learn distributions with latent structure. *arXiv preprint arXiv:2610.12544*.

Wu, S. and Zhang, J. (2025). Free draft-and-verification: Toward lossless parallel decoding for diffusion large language models. *arXiv preprint arXiv:2510.00294*.

Yakovlev, K. and Puchkin, N. (2025). Generalization error bound for denoising score matching under relaxed manifold assumption. *arXiv preprint arXiv:2502.13662*.

Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., and Kong, L. (2025). Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

You, Z., Nie, S., Zhang, X., Hu, J., Zhou, J., Lu, Z., Wen, J.-R., and Li, C. (2025). Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.

Yu, R., Ma, X., and Wang, X. (2025). Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*.

Zhang, K., Yin, H., Liang, F., and Liu, J. (2024). Minimax optimality of score-based diffusion models: Beyond the density lower bound assumptions. *arXiv preprint arXiv:2402.15602*.

Zhang, M. S., Huan, S., Huang, J., Boffi, N. M., Chen, S., and Chewi, S. (2025). Sublinear iterations can suffice even for ddpms. *arXiv preprint arXiv:2511.04844*.

Zhang, Z., Fu, H., Yang, Z., Wang, M., Zhao, T., and Chen, M. (2026). Generalization bounds for discrete diffusion: Statistical advantage of masking. In *Forty-third International Conference on Machine Learning*.

Zhao, Y. and Cai, C. (2026a). Adaptation to intrinsic dependence in diffusion language models. *arXiv preprint arXiv:2602.20126*.

Zhao, Y. and Cai, C. (2026b). Training parallel speculative draft models by directly minimizing expected decoding rounds. *arXiv preprint arXiv:2610.10411*.

Zheng, K., Chen, Y., Mao, H., Liu, M.-Y., Zhu, J., and Zhang, Q. (2024). Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling. *arXiv preprint arXiv:2409.02908*.

Zheng, L., Yuan, J., Yu, L., and Kong, L. (2023). A reparameterized discrete diffusion model for text generation. *arXiv preprint arXiv:2302.05737*.

Zhu, F., Wang, R., Nie, S., Zhang, X., Wu, C., Hu, J., Zhou, J., Chen, J., Lin, Y., Wen, J.-R., et al. (2025). Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *arXiv preprint arXiv:2505.19223*.

A Analysis

In this section, we outline the proof strategy for the main theory (Theorem 1) for the entropy sum-based decoding strategy.

A.1 Preliminary facts

Before presenting the proof sketch, we collect several basic facts and notational conventions that will be used throughout the analysis.

Iterate representation. Starting from the fully masked initialization $Y_0 = (\mathsf{M}, \ldots, \mathsf{M})$, the iterate Y_t at step t is completely characterized by the cumulative unmasking set $\mathcal{W}_t$ together with the revealed token values on that set, namely $Y_t^{\mathcal{W}_t}$, since $Y_t^i = \mathsf{M}$ for all $i \in \mathcal{W}_t^{\mathrm{c}}$. Thus, the entropy-based sampler can be equivalently viewed as iteratively filling in the coordinates indexed by $\mathcal{W}_t$. To simplify notation, we suppress the time index t in the iterate Y_t whenever it is clear from context, and we also write $X = X_0$.

Permutation-induced order. Given a fixed permutation Π of $[L]$, the sampler scans token indices in the order $\Pi_1, \Pi_2, \ldots, \Pi_L$, where Π_k denotes the index of the k-th token to be unmasked. Let Π^{-1} denote the inverse permutation so that $\Pi_i^{-1} = k$ gives the unmasking order k of the token with index i. Moreover, for each token index $i \in [L]$, there exists a unique iteration t_i at which token i is unmasked, that is, the iteration satisfying $i \in \mathcal{D}_{t_i} = \mathcal{W}_{t_i} \setminus \mathcal{W}_{t_i - 1}$.

Unmasking sets. The entropy-based sampling procedure produces an increasing sequence of cumulative unmasking sets $\varnothing = \mathcal{W}_0 \subseteq \mathcal{W}_1 \subseteq \cdots \subseteq \mathcal{W}_T = [L]$ where T is the random total number of iterations. For convenience of analysis, we extend this sequence by defining $\mathcal{W}_t \coloneqq [L]$ for $t > T$ so that the sequence $(\mathcal{W}_t)_{t=0}^L$ is defined for all $0 \le t \le L$. The incremental unmasking sets $(\mathcal{D}_t)_{t=1}^L$ are extended in an analogous manner by setting $\mathcal{D}_t \coloneqq \varnothing$ for $t > T$.

In addition, since tokens are revealed according to the permutation order, the cumulative unmasking set at iteration t can be expressed as

$$\mathcal{W}_t = \{\Pi_1, \ldots, \Pi_{W_t}\}, \tag{11}$$

where we recall the notation $W_t \coloneqq |\mathcal{W}_t|$. Equivalently, the unmasking set $\mathcal{W}_t$ consists of the first W_t indices in the permutation Π. Similarly, the set of newly unmasked tokens at iteration t can be written as

$$\mathcal{D}_t = \mathcal{W}_t \setminus \mathcal{W}_{t-1} = \{\Pi_{W_{t-1}+1}, \ldots, \Pi_{W_t}\}. \tag{12}$$

Dependence structure of sampling process. Finally, we record two critical consequences of the Markovian structure of the entropy sum-based sampling procedure as shown in (7).

- First, for a fixed permutation Π, the unmasking set $\mathcal{D}_t$ at iteration t, and hence the cumulative unmasking set $\mathcal{W}_t$, are completely determined by the cumulative set $\mathcal{W}_{t-1}$ and sampled token values $Y^{\mathcal{W}_{t-1}} = (Y^{\Pi_1}, \ldots, Y^{\Pi_{|\mathcal{W}_{t-1}|}})$ from the previous iteration.
- Second, once the permutation Π and the fully sampled token sequence $Y = (Y^1, \ldots, Y^L) \in \mathbb{X}^L$ are given, the entire sampling trajectory is uniquely determined. In particular, one can recover the total number of iterations T, the cumulative unmasking sets $(\mathcal{W}_t)_{t=0}^L$, and the incremental unmasking sets $(\mathcal{D}_t)_{t=1}^L$. In other words, the total number of iterations T and sets $(\mathcal{W}_{t-1}, \mathcal{D}_t)_{t=1}^L$ are deterministic functions of the sampled token sequence Y and permutation Π, i.e.,

$$T = T(Y, \Pi) \qquad \text{and} \qquad \mathcal{W}_{t-1} = \mathcal{W}_{t-1}(Y, \Pi), \quad \mathcal{D}_t = \mathcal{D}_t(Y, \Pi), \quad t \in [L]. \tag{13}$$

A.2 Proof sketch of Theorem 1

Having introduced the necessary preliminaries, we now outline the proof of Theorem 1 for the entropy sum-based decoding strategy.

Step 1: Express KL error along unmasking trajectory. We begin by expressing both the sampled distribution and the data distribution along the adaptive unmasking trajectory. The key insight is that, by the characterization in (13), once the permutation π and the sampled sequence x are given, the realized total iteration count T and the realized sets $(w_{t-1}, d_t)_{t=1}^L$ are fully determined.

Concretely, when the mask predictor is optimal, conditioned on a fixed permutation $\Pi = \pi$, the entropy-based sampler generates the tokens in each unmasking set independently according to the

true conditional marginals $p_\theta^i = p^i$ defined in (3). Hence the density of the sampled distribution factorizes within each set as a product of conditional marginals:

$$\log p_{Y_T|\Pi=\pi}(x) = \sum_{t=1}^{T} \sum_{i \in d_t} \log p(x^i \mid x^{w_{t-1}}) = \sum_{t=1}^{L} \mathbb{1}\{t \le T\} \sum_{i \in d_t} \log p(x^i \mid x^{w_{t-1}}), \quad x \in \mathbb{X}^L, \tag{14a}$$

where the total number of iterations $T = T(x, \pi)$ and the sets $d_t = d_t(x, \pi)$ and $w_{t-1} = w_{t-1}(x, \pi)$ for $1 \le t \le L$ are uniquely determined by x and π.

By contrast, under the data distribution, the tokens revealed at each iteration follow their true conditional joint distribution. Grouping the density according to the same unmasking sets $(w_{t-1}, d_t)_{t=1}^L$ gives

$$\log p_{X_0}(x) = \sum_{t=1}^{T} \log p(x^{d_t} \mid x^{w_{t-1}}) = \sum_{t=1}^{L} \mathbb{1}\{t \le T\} \log p(x^{d_t} \mid x^{w_{t-1}}). \tag{14b}$$

Using the two factorizations above, let us calculate the KL divergence between the data and sampled distributions. Fixing a permutation $\Pi = \pi$, we can combine (14a) and (14b) to derive

$$\begin{aligned}
&\mathsf{KL}\big(p_{X_0} \,\|\, p_{Y_T|\Pi=\pi}\big) \\
&\quad = \mathbb{E}_{X \sim p_{X_0}} \left[\sum_{t=1}^{L} \mathbb{1}\{t \le T\} \log \frac{p(X^{\mathcal{D}_t} \mid X^{\mathcal{W}_{t-1}})}{\prod_{i \in \mathcal{D}_t} p(X^i \mid X^{\mathcal{W}_{t-1}})} \,\Big|\, \Pi = \pi \right] \\
&\quad = \mathbb{E}\left[\sum_{t=1}^{L} \mathbb{1}\{t \le T\} \mathbb{E}\left[\log \frac{p(X^{\mathcal{D}_t} \mid X^{\mathcal{W}_{t-1}})}{\prod_{i \in \mathcal{D}_t} p(X^i \mid X^{\mathcal{W}_{t-1}})} \,\Big|\, \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi = \pi \right] \Big|\, \Pi = \pi \right],
\end{aligned} \tag{15}$$

where the last line holds due to the tower property, linearity of expectation, and the fact that $\{t \le T\} = \{\mathcal{W}_{t-1} \ne [L]\}$ can be determined by $\mathcal{W}_{t-1}$. Importantly, the only randomness here (with $\Pi = \pi$ fixed) comes from $X \sim p_{X_0}$. Once X is fixed, the iteration count T, the sets $(\mathcal{W}_{t-1}, \mathcal{D}_t)_{t=1}^L$ and the subsequences $(X^{\mathcal{W}_{t-1}}, X^{\mathcal{D}_t})_{t=1}^L$ are all determined.

We next study the inner conditional expectation, which corresponds to the contribution of iteration t to the overall KL sampling error. Fix an iteration $t \in [L]$, and condition on $\Pi = \pi$ and $(\mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) = (w_{t-1}, x^{w_{t-1}})$. Under this conditioning, the set $\mathcal{D}_t = d_t$ is deterministic, and we can derive

$$\begin{aligned}
&\mathbb{E}\left[\log \frac{p(X^{d_t} \mid X^{w_{t-1}})}{\prod_{i \in d_t} p(X^i \mid X^{w_{t-1}})} \,\Big|\, (\mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) = (w_{t-1}, x^{w_{t-1}}), \Pi = \pi \right] \\
&\quad \overset{\text{(i)}}{=} \mathbb{E}_{X^{d_t}} \left[\log \frac{p(X^{d_t} \mid x^{w_{t-1}})}{\prod_{i \in d_t} p(X^i \mid x^{w_{t-1}})} \,\Big|\, (\mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) = (w_{t-1}, x^{w_{t-1}}), \Pi = \pi \right] \\
&\quad \overset{\text{(ii)}}{=} \mathsf{KL}\Big(p^{d_t}(\cdot \mid x^{w_{t-1}}) \,\Big\|\, \prod_{i \in d_t} p^i(\cdot \mid x^{w_{t-1}}) \Big),
\end{aligned} \tag{16}$$

where (i) is true because the unmasking set $\mathcal{D}_t = d_t$ becomes deterministic conditioned on $\Pi = \pi$ and $(\mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) = (w_{t-1}, x^{w_{t-1}})$; (ii) holds because the expectation is taken with respect to X^{d_t} under the true conditional distribution $p(\cdot \mid x^{w_{t-1}})$, which is independent of Π.

Substituting (16) into (15) and then averaging over both $X \sim p_{X_0}$ and Π yields

$$\mathbb{E}_\Pi\big[\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi})\big] = \mathbb{E}_{\Pi, X \sim p_{X_0}} \left[\sum_{t=1}^{T} \mathsf{KL}\Big(p^{\mathcal{D}_t}(\cdot \mid X^{\mathcal{W}_{t-1}}) \,\Big\|\, \prod_{i \in \mathcal{D}_t} p^i(\cdot \mid X^{\mathcal{W}_{t-1}}) \Big) \right]. \tag{17}$$

Thus, the expected KL sampling error decomposes as a sum over iterations t, where each term measures the discrepancy between the true conditional joint distribution of the tokens unmasked at iteration t and the product of their conditional marginals.

Step 2: Decompose KL error into sum of mutual information. In light of the expression for the total KL error in (17), it suffices to analyze the contribution from each iteration. This reduces to understanding the KL divergence between a joint distribution and the product of its marginals.

To this end, Lemma 1 below states that this KL divergence can be decomposed into a sum of mutual information terms, each involving one token and the tokens with larger indices. The proof can be found in Appendix B.1.

Lemma 1. *Consider a random sequence $(Z^1, \ldots, Z^d)$ with joint distribution $p_{Z^1,\ldots,Z^d}$ and marginal distributions p_{Z^i} for $i \in [d]$. The KL divergence between the joint distribution and the product of marginals is given by*

$$\mathsf{KL}\Big(p_{Z^1,\ldots,Z^d} \,\Big\|\, \prod_{i\in[d]} p_{Z^i}\Big) = \sum_{i=1}^{d-1} I(Z^i; Z^{i+1}, Z^{i+2}, \ldots, Z^d). \tag{18}$$

Moreover, for any $j \in [d]$, the KL divergence can be upper bounded as

$$\mathsf{KL}\Big(p_{Z^1,\ldots,Z^d} \,\Big\|\, \prod_{i=1}^{d} p_{Z^i}\Big) \le \sum_{i\in[d]\setminus\{j\}} I(Z^i; Z^1, \ldots, Z^{i-1}, Z^{i+1} \ldots, Z^d). \tag{19}$$

Remark 2. Both expressions involve a sum of $d-1$ terms, but they differ in structure. In the identity (18), the last index $i = d$ is excluded, and each term measures the mutual information between Z^i and the variables with larger indices $Z^{i+1}, Z^{i+2}, \ldots, Z^d$. In the upper bound (19), one instead excludes an arbitrary index $j \in [d]$, and each remaining term measures the mutual information between Z^i and all other variables Z^k with $k \neq i$.

We shall apply the upper bound (19) in Lemma 1 to control the per-iteration KL sampling error. The key point is that, by construction, each iteration stops as soon as the running entropy sum of the currently unmasked tokens exceeds the threshold η. Consequently, within each iteration, there can be at most one token whose entropy is larger than η. Motivated by this observation, define

$$j_t \coloneqq \arg\max_{i\in\mathcal{D}_t} \mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}), \tag{20}$$

with ties broken arbitrarily. Then every index $i \in \mathcal{D}_t \setminus \{j_t\}$ satisfies

$$\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) \le \eta.$$

Here, we remind the reader that $\mathtt{H}(X \mid Z)$ denotes the entropy of X conditioned on a realization of Z (without taking the expectation of Z), so it is a function of Z. Similarly, $\mathtt{I}(X; Y_T \mid Z)$ denotes the mutual information between X and Y given a realization of Z.

Now applying (19) in Lemma 1 with $p_{Z^1,\ldots,Z^d} = p^{D_t}(\cdot \mid x^{w_{t-1}})$ and $j = j_t$, we obtain

$$\begin{aligned}
&\mathsf{KL}\Big(p^{\mathcal{D}_t}(\cdot \mid X^{\mathcal{W}_{t-1}}) \,\Big\|\, \prod_{i\in\mathcal{D}_t} p^i(\cdot \mid X^{\mathcal{W}_{t-1}})\Big) \\
&\le \sum_{i\in\mathcal{D}_t} \mathtt{I}(X^i; X^{\mathcal{D}_t\setminus\{i\}} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi)\, \mathbb{1}\{i \neq j_t\} \\
&\le \sum_{i\in\mathcal{D}_t} \mathtt{I}(X^i; X^{\mathcal{D}_t\setminus\{i\}} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi)\, \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) \le \eta\big\}.
\end{aligned} \tag{21}$$

Remark 3. Conditioned on $\mathcal{W}_{t-1}$ and $X^{\mathcal{W}_{t-1}}$, the conditional entropy $\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}})$ is independent of the permutation Π, i.e.,

$$\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) = \mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi).$$

This is in contrast to the mutual information terms, which depend explicitly on the unmasking set $\mathcal{D}_t$, and hence on Π.

Thus, the KL sampling error contributed by iteration t is controlled by a sum of mutual information, where each term quantifies the dependence between one unmasked token and the rest of the tokens unmasked in iteration t, conditioned on the tokens revealed previously. The indicator function in (21)

highlights an important feature of the entropy-sum stopping rule: within each iteration, only tokens with conditional entropy at most η contribute to this KL error bound. As will be clear momentarily, this indicator function will be essential in the subsequent analysis.

Substituting the bound (21) into the KL error expression in (17), we find

$$\begin{aligned}
&\mathbb{E}_{\Pi}\Big[\mathsf{KL}\big(p_{X_0} \,\|\, p_{Y_T|\Pi}\big)\Big] \\
&\quad\le \mathbb{E}_{X,\Pi}\left[\sum_{t=1}^{T}\sum_{i\in\mathcal{D}_t} \mathtt{I}(X^i; X^{\mathcal{D}_t\setminus\{i\}} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi)\,\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) \le \eta\big\}\right] \\
&\quad= \sum_{i=1}^{L}\mathbb{E}_{X,\Pi}\Big[\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}\setminus\{i\}} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\,\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \le \eta\big\}\Big], \quad (22)
\end{aligned}$$

where the last step rearranges the summation order from the iteration index t to the token index i, where we recall that t_i denotes the index of the iteration at which token i is unmasked.

In summary, the expected KL sampling error is bounded by a sum of mutual information terms over token index $i \in [L]$. Each term measures the dependence between token i and the other unmasked tokens in the same iteration, conditioned on the already unmasked tokens and the permutation.

Step 3: Average over random unmasking order. Equipped with the KL error bound in (22), it remains to control each token-wise contribution separately.

To this end, fix a token index i and consider the corresponding summand:

$$\mathbb{E}_{X,\Pi_{\setminus i},\Pi_i^{-1}}\Big[\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}\setminus\{i\}} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \le \eta\big\}\Big]. \quad (23)$$

To make the dependence on the permutation explicit, we decompose the randomness of Π into the position Π_i^{-1} of token i in the permutation Π and the relative ordering of all other tokens, denoted by $\Pi_{\setminus i}$. Concretely, we define

$$\Pi_{\setminus i} \coloneqq \big(\Pi_1, \ldots, \Pi_{\Pi_i^{-1}-1}, \Pi_{\Pi_i^{-1}+1}, \ldots, \Pi_L\big). \quad (24)$$

That is, the permutation $\Pi_{\setminus i}$ is obtained from Π by removing the index i while preserving the relative order of all other tokens.

Our goal is to show, after averaging over the randomness of the position Π_i^{-1}, the contribution of token i to the KL sampling error admits a clean and tractable upper bound. However, the major technical difficulty arises from the complicated dependence between the set $\mathcal{D}_{t_i}\setminus\{i\}$—the tokens revealed in the same iteration as token i—and the full permutation Π. In particular, the iteration index t_i and the set $\mathcal{D}_{t_i}$ both vary with Π_i^{-1}, making it challenging to directly analyze (23).

To address this, recall that $D_t \coloneqq |\mathcal{D}_t|$ denotes the size of the unmasking set $\mathcal{D}_t$ at iteration $t \in [L]$. We then introduce a sequence of dyadic size envelopes $(\overline{D}_t)_{t=1}^{L}$ in $\mathbb{R}$, defined recursively by $\overline{D}_1 = 0$ and, for $2 \le t \le L$,

$$\overline{D}_t \coloneqq \begin{cases} \overline{D}_{t-1}, & \text{if } D_{t-1} < \overline{D}_{t-1}, \\ 2D_{t-1}, & \text{if } D_{t-1} \ge \overline{D}_{t-1}. \end{cases} \quad (25)$$

In essence, the envelope remains unchanged as long as the previous set size stays below it; once the previous set size reaches the envelope, the envelope is doubled.

Using the size-envelopes constructed above, the following lemma shows that, after averaging over the random position of token i, the mutual-information term in (23) can be bounded by a much simpler quantity involving only the conditional entropy of X^i and the event that the corresponding set size crosses the envelope. The proof is deferred to Appendix B.2.

Lemma 2. *For any token index $i \in [L]$, we have*

$$\begin{aligned}
&\mathbb{E}_{X,\Pi_{\setminus i},\Pi_i^{-1}}\Big[\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}\setminus\{i\}} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \le \eta\big\}\Big] \\
&\quad\le 2\,\mathbb{E}_{X,\Pi_{\setminus i}}\Big[\big(\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \wedge \eta\big)\mathbb{1}\big\{D_{t_i} \ge \overline{D}_{t_i}\big\}\Big]. \quad (26)
\end{aligned}$$

In words, after averaging over the random unmasking order of token i, its contribution to the KL sampling error is controlled by its truncated conditional entropy, and contributes only on the event that the size of its unmasking set is at least as large as the corresponding envelope. Thus, the intricate dependence structure in (23) is reduced to a much more tractable indicator of whether a size-doubling event occurs.

Step 4: Control KL error by entropy-based stopping rule. Summing the bound (26) from Lemma 2 over all tokens and combining it with the KL error expression in (22), we obtain

$$\mathbb{E}_{\Pi}\big[\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi})\big] \le 2\,\mathbb{E}\left[\sum_{i=1}^{L}\big(\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi_{\setminus i}) \wedge \eta\big)\mathbb{1}\big\{D_{t_i} \ge \overline{D}_{t_i}\big\}\right]. \quad (27)$$

We now reindex the sum by iteration rather than token. Since $\mathcal{D}_t$ is exactly the set of tokens revealed at iteration t, the right-hand side of (27) becomes

$$\begin{aligned}&\mathbb{E}\left[\sum_{i=1}^{L}\big(\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi_{\setminus i}) \wedge \eta\big)\mathbb{1}\big\{D_{t_i} \ge \overline{D}_{t_i}\big\}\right]\\ &\quad= \mathbb{E}\left[\sum_{t=1}^{L}\mathbb{1}\{D_t \ge \overline{D}_t\}\sum_{i\in\mathcal{D}_t}\big(\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi_{\setminus i}) \wedge \eta\big)\right]\end{aligned}$$

Next, we bound the inner sum using the stopping rule. By construction, within each iteration the cumulative entropy remains below η until the final token is added. Therefore, the total contribution of all but the last token is at most η, while the final token contributes at most η after truncation. Consequently,

$$\sum_{i\in\mathcal{D}_t}\big(\mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi_{\setminus i}) \wedge \eta\big) \le 2\eta,$$

and hence

$$\mathbb{E}\left[\sum_{i=1}^{L}\big(\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi_{\setminus i}) \wedge \eta\big)\mathbb{1}\big\{D_{t_i} \ge \overline{D}_{t_i}\big\}\right] \le 2\eta\cdot\mathbb{E}\left[\sum_{t=1}^{L}\mathbb{1}\{D_t \ge \overline{D}_t\}\right]. \quad (28)$$

It remains to bound the number of iterations for which the event $D_t \ge \overline{D}_t$ occurs. Let $s_1 < s_2 < \dots$ be the iterations at which this event holds. By the definition of the size-envelope sequence in (25), whenever $D_{s_j} \ge \overline{D}_{s_j}$ occurs, the envelope at the next such iteration must have doubled. Thus, we have

$$D_{s_{j+1}} \ge \overline{D}_{s_{j+1}} = 2D_{s_j},$$

and by induction,

$$D_{s_j} \ge 2^{j-1}D_{s_1} = 2^{j-1}.$$

Since $D_{s_j} \le L$ for all j, this implies that the number of such iterations satisfies

$$\sum_{t=1}^{L}\mathbb{1}\{D_t \ge \overline{D}_t\} \le \log_2 L + 1. \quad (29)$$

Finally, combining (27)–(29) gives the final bound on the expected KL sampling error:

$$\mathbb{E}_{\Pi}\big[\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi})\big] \le 4\eta(\log_2 L + 1). \quad (30)$$

Setting $\eta = \varepsilon/(4(\log_2 L + 1))$ immediately yields the desired result $\mathbb{E}_{\Pi}\big[\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi})\big] \le \varepsilon$.

Step 5: Bound expected number of iterations. Finally, let us study the expected number of iterations $\mathbb{E}[T]$. To this end, we use the following entropy identity, whose proof is deferred to Appendix B.3.

Lemma 3. *For any fixed permutation Π, the entropy of the data distribution admits the following decomposition:*

$$H(X_0) = \mathbb{E}_{X\sim p_{X_0}}\left[\sum_{t=1}^{T} \mathtt{H}(X^{\mathcal{D}_t} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi)\right] \tag{31}$$

$$= \mathbb{E}_{X\sim p_{X_0}}\left[\sum_{t=1}^{T}\sum_{i\in\mathcal{D}_t} \mathtt{H}(X^{i} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi)\right] - \mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi}). \tag{32}$$

We now combine this identity with the stopping rule. By construction, every iteration except the last one accumulates conditional entropy at least η, that is,

$$\sum_{i\in\mathcal{D}_t} \mathtt{H}(X^{i} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi) \geq \eta,$$

for all $t < T$. Therefore, substituting this into the entropy identity (32) yields

$$H(X_0) \geq \mathbb{E}[T-1]\eta - \mathbb{E}_{\Pi}\big[\mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi})\big].$$

Using the KL bound from (30), we further obtain

$$H(X_0) \geq \big(\mathbb{E}[T] - 1 - 4(\log_2 L + 1)\big)\eta.$$

Therefore, rearranging the above inequality and substituting $\eta = \varepsilon/(4(\log_2 L + 1))$ into it yields

$$\mathbb{E}[T] \leq \frac{H(X_0)}{\eta} + 4(\log_2 L + 1) + 1 = 4\left(\frac{H(X_0)}{\varepsilon} + 1\right)(\log_2 L + 1) + 1.$$

This completes the proof of Theorem 1.

B Proof of lemmas for Theorem 1

B.1 Proof of Lemma 1

To simplify the notation, we write $Z^{>i} = (Z^{i+1}, \ldots, Z^d)$ for any $1 \leq i < d$ and define $Z^{<i}$, $Z^{\geq i}$, and $Z^{\leq i}$ analogously. In addition, we denote $Z^{\setminus i} \coloneqq (Z^1, \ldots, Z^{i-1}, Z^{i+1}, \ldots, Z^d)$ for any $1 \leq i \leq d$. Using the chain rule, one can derive

$$\begin{aligned}
\mathsf{KL}\Big(p_{Z^1,\ldots,Z^d} \,\Big\|\, \prod_{i=1}^{d} p_{Z^i}\Big) &= \int p_{Z^1,\ldots,Z^d}(z^1,\ldots,z^d) \log \frac{\prod_{i=1}^{d} p_{Z^i|Z^{>i}}(z^i \mid z^{>i})}{\prod_{i=1}^{d} p_{Z^i}(z^i)} \,\mathrm{d}z^1 \ldots \mathrm{d}z^d \\
&= \sum_{i=1}^{d-1} \int p_{Z^{<i}|Z^{\geq i}}(z^{<i} \mid z^{\geq i}) \,\mathrm{d}z^{<i} \int p_{Z^{\geq i}}(z^{\geq i}) \log \frac{p_{Z^i|Z^{>i}}(z^i \mid z^{>i})}{p_{Z^i}(z^i)} \,\mathrm{d}z^{\geq i} \\
&= \sum_{i=1}^{d-1} \int p_{Z^i,Z^{>i}}(z^i, z^{>i}) \log \frac{p_{Z^i|Z^{>i}}(z^i \mid z^{>i})}{p_{Z^i}(z^i)} \,\mathrm{d}z^i \,\mathrm{d}z^{>i} \\
&= \sum_{i=1}^{d-1} I(Z^i; Z^{>i}),
\end{aligned}$$

where the second step holds because $p_{Z^d|Z^{>d}}(z^d \mid z^{>d}) = p_{Z^d}(z^d)$. This proves (18).

Turning to (19), let us fix an arbitrary $j \in [d]$. We shall apply the same identity as in (18) but choose an ordering of the random variables such that Z^j placed last. Define

$$(Y^1, \ldots, Y^d) \coloneqq (Z^1, \ldots, Z^{j-1}, Z^{j+1}, \ldots, Z^d, Z^j).$$

Recognizing that the KL divergence is invariant to the ordering of the random variables, one has

$$\mathsf{KL}(p_{Z^1,\ldots,Z^d} \,\|\, \prod_{i=1}^{d} p_{Z^i}) = \mathsf{KL}(p_{Y^1,\ldots,Y^d} \,\|\, \prod_{i=1}^{d} p_{Y^i}) = \sum_{k=1}^{d-1} I(Y^k; Y^{>k}),$$

where the last step applies (18) to the random sequence $(Y^1, \ldots, Y^d)$.

Note that for each $1 \le k < d$, Y^k is one of the random variables Z^i with $i \neq j$ and $Y^{>k} \subseteq Z^{\setminus i}$. Therefore, by the monotonicity of mutual information, one has

$$I(Y^k; Y^{>k}) \le I(Z^i; Z^{\setminus i}).$$

Summing over all $i \neq j$ gives the desired bound in (19):

$$\mathsf{KL}(p_{Z^1,\ldots,Z^d} \,\|\, \prod_{i=1}^{d} p_{Z^i}) \le \sum_{i \neq j} I(Z^i; Z^{\setminus i}).$$

This completes the proof of Lemma 1.

B.2 Proof of Lemma 2

Preparation: auxiliary sequence. Before presenting the details, let us first make several preparations to address the statistical coupling issue mentioned in Section A.

Consider the auxiliary sampling process where we want to unmask the token sequence except the i-th token, i.e.,

$$X^{\setminus i} = (X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^L),$$

and run the entropy sum-based sampling procedure (Algorithm 1) according to the order induced by the auxiliary permutation in (24):

$$\Pi_{\setminus i} \coloneqq \big(\Pi_1, \ldots, \Pi_{\Pi_i^{-1}-1}, \Pi_{\Pi_i^{-1}+1}, \ldots, \Pi_L\big).$$

For each $t \in [L]$, let $\mathcal{W}_t^{\setminus i}$ and $\mathcal{D}_t^{\setminus i}$ denote the cumulative and incremental unmasking sets at iteration t in this auxiliary sampling process, respectively. In addition, we denote their sizes by $W_t^{\setminus i} \coloneqq |\mathcal{W}_t^{\setminus i}|$ and $D_t^{\setminus i} \coloneqq |\mathcal{D}_t^{\setminus i}|$.

Moreover, similar to the size envelopes $(\overline{D}_t)_{t=1}^L$ in (25), we define the size envelopes associated with the auxiliary process $(\overline{D}_t^{\setminus i})_{t=1}^L$ recursively by $\overline{D}_1^{\setminus i} \coloneqq 0$, and for $t \ge 2$,

$$\overline{D}_t^{\setminus i} \coloneqq \begin{cases} \overline{D}_{t-1}^{\setminus i}, & \text{if } D_{t-1}^{\setminus i} < \overline{D}_{t-1}^{\setminus i}, \\ 2D_{t-1}^{\setminus i}, & \text{if } D_{t-1}^{\setminus i} \ge \overline{D}_{t-1}^{\setminus i}. \end{cases} \tag{33}$$

Below we make several important observations about the auxiliary sampling process.

1. By construction, the auxiliary unmasking sets $(\mathcal{D}_t^{\setminus i})_{t \ge 1}$ are produced by running Algorithm 1 using the auxiliary permutation $\Pi_{\setminus i}$. Hence, for any iteration $t \ge 1$, the set $\mathcal{D}_t^{\setminus i}$ is independent of the unmasking order Π_i^{-1} of token i in the original process.
2. For any iteration $t \ge 1$, conditioned on $\Pi_{\setminus i}$ and $(\mathcal{W}_{t-1}^{\setminus i}, X^{\mathcal{W}_{t-1}^{\setminus i}})$, the unmasking set $\mathcal{D}_t^{\setminus i}$ at iteration t in the auxiliary process is deterministic.
3. Recall that t_i denotes the iteration at which the token with index i is unmasked in the original sampling process. Note that before this iteration, the original sampling process never uses the information contained in the i-th token to determine unmasking sets and sample tokens. This implies that the unmasking sets and sampled tokens before iteration t_i coincide between the original and auxiliary processes. In particular, $\mathcal{W}_t^{\setminus i} = \mathcal{W}_t$ and $\mathcal{D}_t^{\setminus i} = \mathcal{D}_t$ for all $t < t_i$ in both processes.

To proceed, we make explicit some statistical dependence.

- After conditioning on $\Pi_{\setminus i}$ and $X^{\setminus i}$, the only remaining randomness in the original sampling process that affects the iteration at which token i is unmasked is the insertion position Π_i^{-1} of token i in the permutation Π. To highlight this, we write

  $$t_i = t_i(\Pi_i^{-1})$$

  for the iteration at which token i is unmasked when its insertion position is Π_i^{-1}.

- More generally, in the original sampling process, the unmasking set $\mathcal{D}_{t_i}$ containing token i depends on Π_i^{-1} through both the iteration index and the order of token i within that iteration. Hence, we write
$$\mathcal{D}(\Pi_i^{-1}) \coloneqq \mathcal{D}_{t_i(\Pi_i^{-1})}(\Pi_i^{-1}) \tag{34}$$
to emphasize this dependence on the insertion position Π_i^{-1}.
- Now fix an iteration $s \in [L]$ and condition on $(\mathcal{W}_{s-1}^{\diagdown i}, X^{\mathcal{W}_{s-1}^{\diagdown i}})$ and $\Pi_{\diagdown i}$. Under this conditioning, the auxiliary unmasking set $\mathcal{D}_s^{\diagdown i}$ and its size $D_s^{\diagdown i}$ are deterministic. In the original sampling process, the set of insertion positions p of token i in the permutation Π that lead to $t_i(p) = s$ is precisely the unmasking set $\mathcal{D}_s^{\diagdown i}$ in the auxiliary process, i.e.,
$$\{p \in [L] : t_i(p) = s\} = \mathcal{D}_s^{\diagdown i}.$$
To see this, note that each index except the last one in $\mathcal{D}_s^{\diagdown i}$ has the property that the sum up to the following token is still below the threshold. Therefore, token i can be inserted at any such position without changing the iteration at which it is unmasked. On the other hand, the last index in $\mathcal{D}_s^{\diagdown i}$ is the one at which the cumulative sum including the following token exceeds the threshold. Therefore, for token i to be unmasked at iteration s, this is the last position at which it can be inserted.
- However, if token i is inserted at position $p \in \mathcal{D}_s^{\diagdown i}$ so that it is unmasked at iteration s in the original sampling process, the resulting unmasking set $\mathcal{D}_s$ may still vary with p, because a high entropy of token i can lead to early termination of iteration s. Accordingly, for positions $p \in \mathcal{D}_s^{\diagdown i}$, we write
$$D_s(p) = |\mathcal{D}_s(p)|$$
to denote the size of the original unmasking set at iteration s when token i is inserted at position p in the permutation Π.
More precisely, the set of all other unmasked tokens in the same iteration s as token i in the original process, is a subset of the s-th unmasking set in the auxiliary process, i.e.,
$$\mathcal{D}(p) \setminus \{i\} = \mathcal{D}_s(p) \setminus \{i\} \subseteq \mathcal{D}_s^{\diagdown i}, \quad \forall\, p \in \mathcal{D}_s^{\diagdown i}. \tag{35}$$
This is because inserting token i into position $p \in \mathcal{D}_s^{\diagdown i}$ in the permutation Π can only make iteration s terminate earlier in the original process than in the auxiliary process since the entropy of token i can be large.
- To organize these insertion positions, it is convenient to index them relative to the start of the s-th auxiliary set $\mathcal{D}_s^{\diagdown i}$. Recall $W_s^{\diagdown i} \coloneqq |\mathcal{W}_s^{\diagdown i}|$ and $D_s^{\diagdown i} \coloneqq |\mathcal{D}_s^{\diagdown i}|$. For each absolute insertion position $q \in [L]$, let
$$r \coloneqq \big((q-1) \bmod D_s^{\diagdown i}\big) + 1,$$
be the relative position of q within the set $\mathcal{D}_s^{\diagdown i}$, and define
$$D_s^{\mathsf{rel}}(q) \coloneqq D_s(r + W_{s-1}^{\diagdown i}). \tag{36}$$
In other words, $D_s^{\mathsf{rel}}(q)$ represents the size of the s-th original unmasking set $\mathcal{D}_s$ obtained when token i is inserted at relative position r within the s-th auxiliary unmasking set $\mathcal{D}_s^{\diagdown i}$. In particular, we introduce the module operation in (36) to handle the case when q is larger than the auxiliary set size $D_s^{\diagdown i}$.
- Finally, note that conditioned on $t_i(p) = s$, the envelope $\overline{D}_s$ is determined by the original process up to iteration $s-1$, and therefore does not depend on the relative index q of the insertion position p within the s-th auxiliary set $\mathcal{D}_s^{\diagdown i}$.

Proof of Lemma 2. Having the above preparations, we are ready to prove Lemma 2.

Our goal is to analyze the contribution of token i to the expected KL sampling error in (23) by taking expectation with respect to the randomness of Π_i^{-1}, namely, the random insertion position Π_i^{-1} of token i in the permutation Π. Hence, for each $p \in [L]$, define
$$F(p) \coloneqq \mathtt{I}(X^i; X^{\mathcal{D}(p)\setminus\{i\}} \mid \mathcal{W}_{t_i(p)-1}^{\diagdown i}, X^{\mathcal{W}_{t_i(p)-1}^{\diagdown i}}, \Pi_{\diagdown i}) \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i(p)-1}^{\diagdown i}, X^{\mathcal{W}_{t_i(p)-1}^{\diagdown i}}) \le \eta\big\}. \tag{37}$$

We can then express (23) as

$$\mathbb{E}_{X,\Pi}\Big[\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}\setminus\{i\}} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \le \eta\big\}\Big]$$
$$= \mathbb{E}\big[F(\Pi_i^{-1})\big] = \mathbb{E}_{X^{\setminus i},\Pi_{\setminus i}}\Big[\mathbb{E}_{\Pi_i^{-1}}\big[F(\Pi_i^{-1}) \mid X^{\setminus i}, \Pi_{\setminus i}\big]\Big]. \tag{38}$$

In what follows, we focus on the inner expectation with respect to Π_i^{-1} in (38). As Π_i^{-1} is uniformly distributed over $[L]$ conditioned on $X^{\setminus i}$ and $\Pi_{\setminus i}$, we have

$$\begin{aligned}
&\mathbb{E}_{\Pi_i^{-1}}\big[F(\Pi_i^{-1}) \mid X^{\setminus i}, \Pi_{\setminus i}\big] \\
&\overset{\text{(i)}}{=} \frac{1}{L}\sum_{p=1}^{L} \mathtt{I}(X^i; X^{\mathcal{D}(p)\setminus\{i\}} \mid \mathcal{W}^{\setminus i}_{t_i(p)-1}, X^{\mathcal{W}^{\setminus i}_{t_i(p)-1}}, \Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}^{\setminus i}_{t_i(p)-1}, X^{\mathcal{W}^{\setminus i}_{t_i(p)-1}}) \le \eta\big\} \\
&\overset{\text{(ii)}}{=} \frac{1}{L}\sum_{s=1}^{L}\sum_{p=1}^{L} \mathbb{1}\{t_i(p)=s\}\mathtt{I}(X^i; X^{\mathcal{D}(p)\setminus\{i\}} \mid \mathcal{W}^{\setminus i}_{t_i(p)-1}, X^{\mathcal{W}^{\setminus i}_{t_i(p)-1}}, \Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}^{\setminus i}_{t_i(p)-1}, X^{\mathcal{W}^{\setminus i}_{t_i(p)-1}}) \le \eta\big\} \\
&\overset{\text{(iii)}}{=} \frac{1}{L}\sum_{s=1}^{L}\sum_{p\in\mathcal{D}_s^{\setminus i}} \mathtt{I}(X^i; X^{\mathcal{D}(p)\setminus\{i\}} \mid \mathcal{W}^{\setminus i}_{s-1}, X^{\mathcal{W}^{\setminus i}_{s-1}}, \Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}^{\setminus i}_{s-1}, X^{\mathcal{W}^{\setminus i}_{s-1}}) \le \eta\big\} \\
&\overset{\text{(iv)}}{=} \frac{1}{L}\sum_{s=1}^{L}\sum_{q=1}^{D_s^{\setminus i}} \mathtt{I}(X^i; X^{\mathcal{D}(q+W^{\setminus i}_{s-1})\setminus\{i\}} \mid \mathcal{W}^{\setminus i}_{s-1}, X^{\mathcal{W}^{\setminus i}_{s-1}}, \Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}^{\setminus i}_{s-1}, X^{\mathcal{W}^{\setminus i}_{s-1}}) \le \eta\big\}
\end{aligned} \tag{39}$$

Here, (i) follows from the fact that Π_i^{-1} is uniformly distributed over $[L]$ independent of $X^{\setminus i}$ and $\Pi_{\setminus i}$ and that conditioned on $X^{\setminus i}$ and $\Pi_{\setminus i}$, t_i is determined given $\Pi_i^{-1} = p$; (ii) rearranges the summation by grouping the insertion positions with the same iteration index t_i; (iii) arises from the fact $\{p \in [L] : t_i(p) = s\} = \mathcal{D}_s^{\setminus i}$; (iv) rewrites the position p in terms of the relative position q within the auxiliary unmasking set $\mathcal{D}_s^{\setminus i}$.

Next, let us swap the order of double summation in (39). The technical challenge is that the size of the second summation, or equivalently the size of the auxiliary unmasking set $\mathcal{D}_s^{\setminus i}$, varies across different iterations s. To ensure that the swapped summation is well-organized, we can leverage the dyadic size envelopes $(\overline{D}_s)_{s=1}^L$ to obtain an upper bound. This is established by the following lemma, whose proof is deferred to Appendix B.4.

Lemma 4. *Recall the definition of $\overline{D}_s$ and $D_s^{\mathsf{rel}}(q)$ in* (25) *and* (36)*, respectively. The sum in* (39) *satisfies*

$$\begin{aligned}
&\sum_{s=1}^{L}\sum_{q=1}^{D_s^{\setminus i}} \mathtt{I}(X^i; X^{\mathcal{D}(q+W^{\setminus i}_{s-1})\setminus\{i\}} \mid \mathcal{W}^{\setminus i}_{s-1}, X^{\mathcal{W}^{\setminus i}_{s-1}}, \Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}^{\setminus i}_{s-1}, X^{\mathcal{W}^{\setminus i}_{s-1}}) \le \eta\big\} \\
&\quad\le \sum_{s=1}^{L}\sum_{k=s}^{L}\sum_{q=1}^{2D_s^{\setminus i}\wedge D_k^{\setminus i}} \mathtt{I}(X^i; X^{\mathcal{D}(q+W^{\setminus i}_{k-1})\setminus\{i\}} \mid \mathcal{W}^{\setminus i}_{k-1}, X^{\mathcal{W}^{\setminus i}_{k-1}}, \Pi_{\setminus i}) \\
&\qquad\qquad \cdot \mathbb{1}\{D_s^{\mathsf{rel}}(q) \ge \overline{D}_s\} \cdot \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}^{\setminus i}_{k-1}, X^{\mathcal{W}^{\setminus i}_{k-1}}) \le \eta\big\}.
\end{aligned} \tag{40}$$

Equipped with Lemma 4, we can continue the derivation. Recall (35) that

$$\mathcal{D}(p) \setminus \{i\} \subseteq \mathcal{D}_k^{\setminus i}, \quad \forall\, p \in \mathcal{D}_k^{\setminus i},$$

or equivalently, rewriting the absolute position p in terms of the relative position q within the auxiliary unmasking set $\mathcal{D}_k^{\setminus i}$, we have

$$\mathcal{D}(q + W_{k-1}^{\setminus i}) \setminus \{i\} \subseteq \mathcal{D}_k^{\setminus i}, \quad \forall\, 1 \le q \le D_k^{\setminus i}.$$

Using the monotonicity of mutual information, we can upper bound the right-hand side of (40) as

$$
\begin{aligned}
&\sum_{s=1}^{L}\sum_{k=s}^{L}\sum_{q=1}^{2D_s^{\diagdown i}\wedge D_k^{\diagdown i}} \mathbb{1}\{D_s^{\mathsf{rel}}(q)\geq \overline{D}_s\}\mathtt{I}(X^i; X^{\mathcal{D}(q+W_{k-1}^{\diagdown i})\setminus\{i\}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}})\leq \eta\big\} \\
&\quad\leq \sum_{s=1}^{L}\sum_{k=s}^{L}\sum_{q=1}^{2D_s^{\diagdown i}\wedge D_k^{\diagdown i}} \mathbb{1}\{D_s^{\mathsf{rel}}(q)\geq \overline{D}_s\}\mathtt{I}(X^i; X^{\mathcal{D}_k^{\diagdown i}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}})\leq \eta\big\} \\
&\quad\leq \sum_{s=1}^{L}\sum_{q=1}^{2D_s^{\diagdown i}} \mathbb{1}\{D_s^{\mathsf{rel}}(q)\geq \overline{D}_s\}\sum_{k=s}^{L}\mathtt{I}(X^i; X^{\mathcal{D}_k^{\diagdown i}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}})\leq \eta\big\} \\
&\quad= 2\sum_{s=1}^{L}\sum_{q=1}^{D_s^{\diagdown i}} \mathbb{1}\{D_s^{\mathsf{rel}}(q)\geq \overline{D}_s\}\sum_{k=s}^{L}\mathtt{I}(X^i; X^{\mathcal{D}_k^{\diagdown i}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}})\leq \eta\big\} \\
&\quad= 2\sum_{s=1}^{L}\sum_{p=1}^{L} \mathbb{1}\{t_i(p)=s\}\mathbb{1}\{D_s(p)\geq \overline{D}_s\}\sum_{k=s}^{L}\mathtt{I}(X^i; X^{\mathcal{D}_k^{\diagdown i}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}})\leq \eta\big\},
\end{aligned}
\tag{41}
$$

where the penultimate step is due to the definition that $D_s^{\mathsf{rel}}(q)$ is a function of q modulo $D_s^{\diagdown i}$, and the last step is true because for $q \in [D_s^{\diagdown i}]$, it corresponds to an insertion position p such that $t_i(p) = s$ with $D_s(p) = D_s^{\mathsf{rel}}(q)$.

Putting (40)–(41) together and substituting it into (39), we obtain that conditioned on $X^{\diagdown i}$ and $\Pi_{\diagdown i}$,

$$
\begin{aligned}
&\mathbb{E}_{\Pi_i^{-1}}\big[F(\Pi_i^{-1}) \mid X^{\diagdown i}, \Pi_{\diagdown i}\big] \\
&\quad\leq \frac{2}{L}\sum_{s=1}^{L}\sum_{p=1}^{L}\mathbb{1}\{t_i(p)=s\}\mathbb{1}\{D_s(p)\geq \overline{D}_s\} \\
&\qquad\qquad \cdot \sum_{k=s}^{L}\mathtt{I}(X^i; X^{\mathcal{D}_k^{\diagdown i}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}\big(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i}\big)\leq \eta\big\}.
\end{aligned}
\tag{42}
$$

To control the sum of the mutual information in the above display, the key insight is that the auxiliary unmasking sets satisfy $\mathcal{W}_k^{\diagdown i} = \mathcal{W}_{k-1}^{\diagdown i} \cup \mathcal{D}_k^{\diagdown i}$, which allows us to leverage the chain rule of mutual information to control the sum. This is accomplished in the following lemma, whose proof can be found in Appendix B.5.

Lemma 5. *For any token index $i \in [L]$ and iteration index $s \in [L]$, one has*

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{k=s}^{L}\mathtt{I}(X^i; X^{\mathcal{D}_k^{\diagdown i}} \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}}, \Pi_{\diagdown i})\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\diagdown i}, X^{\mathcal{W}_{k-1}^{\diagdown i}})\leq \eta\big\} \,\Big|\, \mathcal{W}_{s-1}^{\diagdown i}, X^{\mathcal{W}_{s-1}^{\diagdown i}}, \Pi_{\diagdown i}\right] \\
&\quad\leq \mathtt{H}(X^i \mid \mathcal{W}_{s-1}^{\diagdown i}, X^{\mathcal{W}_{s-1}^{\diagdown i}}) \wedge \eta.
\end{aligned}
\tag{43}
$$

With the help of Lemma 5, let us compute the expectation over $X^{\diagdown i}$ and $\Pi_{\diagdown i}$ on both sides of (42). For each $s \in [L]$, the expectation of the s-th term on the right-hand-side of (42) can be upper bounded

as

$$\mathbb{E}\Bigg[\mathbb{E}\Bigg[\sum_{p=1}^{L}\mathbb{1}\{t_i(p)=s\}\mathbb{1}\{D_s(p)\geq\overline{D}_s\}$$
$$\cdot\sum_{k=s}^{L}\mathtt{I}(X^i;X^{\mathcal{D}_k^{\setminus i}}\mid\mathcal{W}_{k-1}^{\setminus i},X^{\mathcal{W}_{k-1}^{\setminus i}},\Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}\big(X^i\mid\mathcal{W}_{k-1}^{\setminus i},X^{\mathcal{W}_{k-1}^{\setminus i}}\big)\leq\eta\big\}\,\Big|\,\mathcal{W}_{s-1}^{\setminus i},X^{\mathcal{W}_{s-1}^{\setminus i}},\Pi_{\setminus i}\Bigg]\Bigg]$$
$$\overset{\text{(i)}}{\leq}\mathbb{E}\Bigg[\sum_{p=1}^{L}\mathbb{1}\{t_i(p)=s\}\mathbb{1}\{D_s(p)\geq\overline{D}_s\}$$
$$\cdot\,\mathbb{E}\Bigg[\sum_{k=s}^{L}\mathtt{I}(X^i;X^{\mathcal{D}_k^{\setminus i}}\mid\mathcal{W}_{k-1}^{\setminus i},X^{\mathcal{W}_{k-1}^{\setminus i}},\Pi_{\setminus i})\mathbb{1}\big\{\mathtt{H}(X^i\mid\mathcal{W}_{k-1}^{\setminus i},X^{\mathcal{W}_{k-1}^{\setminus i}})\leq\eta\big\}\mid\mathcal{W}_{s-1}^{\setminus i},X^{\mathcal{W}_{s-1}^{\setminus i}},\Pi_{\setminus i}\Bigg]\Bigg]$$
$$\overset{\text{(ii)}}{\leq}\mathbb{E}\Bigg[\sum_{p=1}^{L}\mathbb{1}\{t_i(p)=s\}\mathbb{1}\big\{D_s(p)\geq\overline{D}_s\big\}\big(\mathtt{H}(X^i\mid\mathcal{W}_{s-1}^{\setminus i},X^{\mathcal{W}_{s-1}^{\setminus i}})\wedge\eta\big)\Bigg],$$

where (i) holds because $\{t_i(p)=s\}$ and $\{D_s(p)\geq\overline{D}_s\}$ are determined by $(\mathcal{W}_{s-1}^{\setminus i},X^{\mathcal{W}_{s-1}^{\setminus i}})$ and $\Pi_{\setminus i}$; (ii) applies (43) in Lemma 5. Summing the above bound over $s\in[L]$ yields

$$\begin{aligned}\mathbb{E}\big[F(\Pi_i^{-1})\big]&\leq\frac{2}{L}\mathbb{E}\Bigg[\sum_{s=1}^{L}\sum_{p=1}^{L}\mathbb{1}\{t_i(p)=s\}\mathbb{1}\big\{D_s(p)\geq\overline{D}_s\big\}\big(\mathtt{H}(X^i\mid\mathcal{W}_{s-1}^{\setminus i},X^{\mathcal{W}_{s-1}^{\setminus i}})\wedge\eta\big)\Bigg]\\&\overset{\text{(i)}}{=}\frac{2}{L}\mathbb{E}\Bigg[\sum_{p=1}^{L}\mathbb{1}\big\{D_{t_i(p)}\geq\overline{D}_{t_i(p)}\big\}\big(\mathtt{H}(X^i\mid\mathcal{W}_{t_i(p)-1},X^{\mathcal{W}_{t_i(p)-1}})\wedge\eta\big)\Bigg]\\&\overset{\text{(ii)}}{=}2\,\mathbb{E}_{X^{\setminus i},\Pi_{\setminus i}}\Big[\mathbb{E}_{\Pi_i^{-1}}\big[\mathbb{1}\{D_{t_i}\geq\overline{D}_{t_i}\}\big(\mathtt{H}(X^i\mid\mathcal{W}_{t_i-1},X^{\mathcal{W}_{t_i-1}})\wedge\eta\big)\mid X^{\setminus i},\Pi_{\setminus i}\big]\Big],\\&=2\,\mathbb{E}\Big[\mathbb{1}\{D_{t_i}\geq\overline{D}_{t_i}\}\big(\mathtt{H}(X^i\mid\mathcal{W}_{t_i-1},X^{\mathcal{W}_{t_i-1}})\wedge\eta\big)\Big],\end{aligned}$$

where (i) swaps the summation order and uses the fact that $\mathcal{W}_{t_i-1}^{\setminus i}=\mathcal{W}_{t_i-1}$; (ii) is true because Π_i^{-1} is uniformly random over $[L]$ and independent of $X^{\setminus i}$ and $\Pi_{\setminus i}$.

Combining the above bound with (38), we arrive at the desired claim in (26). This completes the proof of Lemma 2.

B.3 Proof of Lemma 3

We begin with the proof of (31). Recall from (14b) that for a fixed permutation $\Pi=\pi$, the data density can be expressed as

$$\log p_{X_0}(x)=\sum_{t=1}^{\tau}\log p(x^{d_t}\mid x^{w_{t-1}}),$$

where the realized number of iterations τ and unmasking sets $(d_t,w_{t-1})_{t\in[\tau]}$ are determined by $x=(x^1,\ldots,x^L)$ and π. One can then apply the same argument as in (15). Specifically, taking expectation over $X\sim p_{X_0}$ yields that for any fixed $\Pi=\pi$,

$$\begin{aligned}H(X_0)=\mathbb{E}_{X\sim p_{X_0}}\big[-\log p_{X_0}(X)\big]&=\mathbb{E}_X\Bigg[-\sum_{t=1}^{T}\log p(X^{d_t}\mid X^{w_{t-1}})\,\Big|\,\Pi=\pi\Bigg]\\&=\mathbb{E}_X\Bigg[\sum_{t=1}^{T}\mathtt{H}(X^{d_t}\mid w_{t-1},X^{w_{t-1}},\pi)\Bigg].\end{aligned}$$

Next, let us turn to (32). Recall the density of the sampled distribution in (14a):

$$\log p_{Y|\Pi=\pi}(x)=\sum_{t=1}^{\tau}\sum_{i\in d_t}\log p(x^i\mid x^{w_{t-1}}).$$

We can then express the entropy of X_0 as

$$\begin{aligned} H(X_0) &= \mathbb{E}_{X\sim p_{X_0}}\left[-\log p_{Y_T|\Pi=\pi}(X) - \log \frac{p_{X_0}(X)}{p_{Y_T|\Pi=\pi}(X)}\right] \\ &= \mathbb{E}_X\left[-\log p_{Y_T|\Pi=\pi}(X)\right] - \mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi=\pi}) \\ &= \mathbb{E}_X\left[\sum_{t=1}^{T}\sum_{i\in\mathcal{D}_t} \mathtt{H}(X^i \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \pi)\right] - \mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi=\pi}). \end{aligned}$$

B.4 Proof of Lemma 4

We prove Lemma 4 by showing that every term on the left-hand-side of (40) is covered by a term on the right-hand-side.

Fix an arbitrary $1 \le k \le L$ and $1 \le q \le D_k^{\searrow i}$. Recall the definitions of $\overline{D}_s$ and $D_s^{\mathsf{rel}}(q)$ in (25) and (36), respectively. Define

$$s = s(k,q) \coloneqq \max\left\{u \le k \,:\, \overline{D}_u \le q\right\}. \tag{44}$$

This set is non-empty because $\overline{D}_1 = 0$, so s is well-defined.

In what follows, let us show that the right-hand-side term with the index (s,k,q) matches the term indexed by (k,q) on the left-hand-side.

To begin with, let us verify that the indicator $\mathbb{1}\{D_s^{\mathsf{rel}}(q) \ge \overline{D}_s\}$ is 1. We first claim that the relative size $D_s^{\mathsf{rel}}(q)$ is at least q, i.e.,

$$D_s^{\mathsf{rel}}(q) \ge q. \tag{45}$$

To see this, recall that any position $1 \le q \le D_s^{\searrow i}$ guarantees that token i is unmasked at iteration s. As the sampling process must first unmask the first $q-1$ tokens in the set and then unmask token i, the unmasking set size must satisfy $D_s^{\mathsf{rel}}(q) = |\mathcal{D}_s(q)| \ge q$. This establishes (45).

We can then combine (45) with the definition of s in (44) to find

$$D_s^{\mathsf{rel}}(q) \ge q \ge \overline{D}_s.$$

Thus the indicator $\mathbb{1}\{D_s^{\mathsf{rel}}(q) \ge \overline{D}_s\}$ is 1.

As we choose $q \le D_k^{\searrow i}$, it remains to prove that

$$q \le 2D_s^{\searrow i}.$$

If $s = k$, then the above inequality holds trivially because $q \le D_k^{\searrow i} \le 2D_s^{\searrow i}$.

If $s < k$, by the definition of s in (44), we must have $\overline{D}_{s+1} > q \ge \overline{D}_s$. On the other hand, by the recursive definition of $\overline{D}_{s+1}$ in (25), $\overline{D}_{s+1}$ can only take a value in $\{\overline{D}_s, 2D_s\}$. This implies that $\overline{D}_{s+1} = 2D_s$. Moreover, because $q \le D_k^{\searrow i}$ guarantees that token i is unmasked at iteration k with $k > s$, we know that $D_s = D_s^{\searrow i}$ by by Observation B3. Therefore, this leads to

$$\overline{D}_{s+1} = 2D_s = 2D_s^{\searrow i} > q.$$

Hence, in either case, we have

$$q \le 2D_s^{\searrow i} \wedge D_k^{\searrow i}.$$

This proves that each term indexed by (k,q) on the left-hand-side corresponds to a term indexed by (s,k,q) on the right-hand-side, and the proof of Lemma 4 is complete.

B.5 Proof of Lemma 5

Fix a token index i and an iteration index s. Before proceeding, we first summarize some notation and observations. For convenience of notation, for each $s \le k \le L$, define the filtration

$$\mathcal{F}_{k-1} := \sigma\big(\mathcal{W}_{k-1}^{\searrow i}, X^{\mathcal{W}_{k-1}^{\searrow i}}, \Pi_{\searrow i}\big).$$

In addition, denote the pointwise conditional entropy and mutual information by

$$h_{k-1} := \mathtt{H}(X^i \mid \mathcal{F}_{k-1}) = \mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}) = \mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}),$$
$$\Delta_k := \mathtt{I}(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{F}_{k-1}) = \mathtt{I}(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}).$$

By the chain rule for mutual information, we can express the conditional mutual information Δ_k as

$$\begin{aligned}\Delta_k &= \mathtt{H}(X^i \mid \mathcal{F}_{k-1}) - \mathbb{E}_{X^{\mathcal{D}_k^{\setminus i}}}\Big[\mathtt{H}\big(X^i \mid X^{\mathcal{D}_k^{\setminus i}}, \mathcal{F}_{k-1}\big) \mid \mathcal{F}_{k-1}\Big] \\ &= h_{k-1} - \mathbb{E}\big[h_k \mid \mathcal{F}_{k-1}\big], \end{aligned} \tag{46}$$

where the second equality follows from the fact that $\mathcal{F}_k = \sigma(\mathcal{W}_k^{\setminus i}, X^{\mathcal{W}_k^{\setminus i}}, \Pi_{\setminus i}) = \sigma(\mathcal{F}_{k-1}, X^{\mathcal{D}_k^{\setminus i}})$ because given $\mathcal{W}_k^{\setminus i}$, $X^{\mathcal{W}_k^{\setminus i}}$, and $\Pi_{\setminus i}$, the unmasking set $\mathcal{D}_k^{\setminus i}$ is determined. In particular, $\Delta_k \geq 0$ due to the non-negativity of mutual information, and $(h_k)_k$ is a nonnegative supermartingale.

Finally, define the first entrance time into the low-entropy region by

$$\tau := \min\{s \leq k \leq L : h_{k-1} \leq \eta\},$$

with the convention that $\tau = \infty$ if the set is empty.

With the above definitions, the identity in (43) can be rewritten as

$$\mathbb{E}\left[\sum_{k=s}^{L} \Delta_k \mathbb{1}\{h_{k-1} \leq \eta\} \mid \mathcal{F}_{s-1}\right] \leq h_{s-1} \wedge \eta, \tag{47}$$

and the remainder of the proof focuses on establishing this inequality.

By the definition of τ, every term selected by the indicator $\mathbb{1}\{h_{k-1} \leq \eta\}$ occurs at or after τ. Combining this with the fact that $\Delta_k \geq 0$, we have

$$\mathbb{E}\left[\sum_{k=s}^{L} \Delta_k \mathbb{1}\{h_{k-1} \leq \eta\} \;\Big|\; \mathcal{F}_{s-1}\right] \leq \mathbb{E}\left[\mathbb{1}\{\tau \leq L\} \sum_{k=\tau}^{L} \Delta_k \;\Big|\; \mathcal{F}_{s-1}\right]. \tag{48}$$

Since $h_{k-1} = \mathtt{H}(X^i \mid \mathcal{F}_{k-1})$ is $\mathcal{F}_{k-1}$-measurable, we have $\{\tau = k\} \in \mathcal{F}_{k-1}$ for every $s \leq k \leq L$. Decomposing the right-hand side of (48) over the value of τ and using the tower property with $\mathcal{F}_{s-1} \subseteq \mathcal{F}_{k-1}$ yields

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}\{\tau \leq L\} \sum_{k=\tau}^{L} \Delta_k \;\Big|\; \mathcal{F}_{s-1}\right] &= \sum_{k=s}^{L} \mathbb{E}\left[\mathbb{1}\{\tau = k\}\, \mathbb{E}\left[\sum_{j=k}^{L} \Delta_j \;\Big|\; \mathcal{F}_{k-1}\right] \;\Big|\; \mathcal{F}_{s-1}\right] \\ &= \sum_{k=s}^{L} \mathbb{E}\Big[\mathbb{1}\{\tau = k\}\big(h_{k-1} - \mathbb{E}[h_L \mid \mathcal{F}_{k-1}]\big) \;\Big|\; \mathcal{F}_{s-1}\Big] \\ &\leq \sum_{k=s}^{L} \mathbb{E}\big[\mathbb{1}\{\tau = k\}\, h_{k-1} \;\big|\; \mathcal{F}_{s-1}\big] \\ &\leq \eta\, \mathbb{P}\{\tau \leq L \mid \mathcal{F}_{s-1}\} \leq \eta, \end{aligned}$$

where the second equality uses the telescoping identity (46) from k to L, the first inequality uses $h_L \geq 0$, and the second inequality holds because $h_{k-1} \leq \eta$ when $\tau = k$.

On the other hand, dropping the indicator and applying the same telescoping identity from iteration s gives

$$\mathbb{E}\left[\sum_{k=s}^{L} \Delta_k \mathbb{1}\{h_{k-1} \leq \eta\} \;\Big|\; \mathcal{F}_{s-1}\right] \leq \mathbb{E}\left[\sum_{k=s}^{L} \Delta_k \;\Big|\; \mathcal{F}_{s-1}\right] = h_{s-1} - \mathbb{E}\big[h_L \mid \mathcal{F}_{s-1}\big] \leq h_{s-1}.$$

Combining the two bounds, we obtain (47), which completes the proof of Lemma 5.

C Proof of Proposition 1

Let $m = \lfloor H/\eta \rfloor$ and define

$$\delta_k = \frac{\delta\eta}{\log(k+1)}, \qquad 1 \le k \le m,$$

where $\delta > 0$ is sufficiently small. For each $1 \le k \le m$, let $X_0^{2k-1} = X_0^{2k} = Z_k$, where the random variables $Z_1, \dots, Z_m$ are independent and $Z_k \sim \mathrm{Bern}(p_k)$, with p_k chosen so that

$$-p_k \log p_k - (1-p_k)\log(1-p_k) = \eta - \delta_k.$$

All remaining coordinates are deterministic: $X_0^j = 0, j > 2m$.

For $1 \le k \le m$, we have

$$H(X_0^{2k-1}) = H(X_0^{2k}) = H(X_0^{2k-1}, X_0^{2k}) = \eta - \delta_k,$$

and hence

$$I(X_0^{2k-1}; X_0^{2k}) = \eta - \delta_k.$$

Therefore,

$$H(X_0) = \sum_{k=1}^{m} H(Z_k) = \sum_{k=1}^{m} (\eta - \delta_k) < m\eta \le H.$$

Under sorted ordering, all deterministic coordinates, whose conditional entropy is zero, are selected first. The first correlated pair is then added to the same iteration. Since

$$H(X_0^{2k-1}) = \eta - \delta_k < \eta$$

but

$$H(X_0^{2k-1}) + H(X_0^{2k}) = 2(\eta - \delta_k) > \eta$$

for sufficiently small δ, each subsequent iteration contains exactly one correlated pair:

$$D_1 = \{j > 2m\} \cup \{1, 2\}, \qquad D_t = \{2t-1, 2t\}, \quad 2 \le t \le m.$$

Consequently, the number of iterations is

$$T = m = \left\lfloor \frac{H}{\eta} \right\rfloor.$$

Moreover, each pair is decoded independently within the same iteration, so its contribution to the KL error is its mutual information. Hence,

$$\mathsf{KL}(p_{X_0} \| p_Y) = \sum_{t=1}^{m} I(X_0^{2t-1}; X_0^{2t}) = \sum_{t=1}^{m} (\eta - \delta_t) = \Theta(H).$$

This construction shows that selecting tokens solely according to low marginal entropy does not guarantee small sampling error.

D Proof of Theorem 2

D.1 Proof sketch of Theorem 2

The high-level proof strategy for Theorem 2 is similar to that for Theorem 1. Hence, we will mainly focus on the key differences in the analysis, and refer readers to Section A.2 for the common technical details.

Step 1: Decompose KL error into sum of mutual information. To begin with, notice that the sampling process of the maximum entropy-based strategy in Algorithm 2 also admits a Markovian structure:

$$\begin{aligned}\Pi \to \big\{\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_0)\big)\big\}_{k=1}^{L} \to (\mathcal{D}_1, \mathcal{W}_1) \to Y_1 \to \big\{\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_1)\big)\big\}_{k=W_1+1}^{L} \to (\mathcal{D}_2, \mathcal{W}_2) \to Y_2 \\ \to \cdots \to \big\{\mathtt{H}\big(p_\theta^{\Pi_k}(\cdot \mid Y_{T-1})\big)\big\}_{k=W_{T-1}+1}^{L} \to (\mathcal{D}_T, \mathcal{W}_T) \to Y_T. \end{aligned} \tag{49}$$

Hence, the KL sampling error takes the same form as (17) in Section A.2, i.e.,

$$\mathbb{E}_\Pi\big[\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi})\big] = \mathbb{E}_{\Pi, X\sim p_{X_0}}\left[\sum_{t=1}^{T} \mathsf{KL}\Big(p^{\mathcal{D}_t}(\cdot \mid X^{\mathcal{W}_{t-1}}) \,\Big\|\, \prod_{i\in\mathcal{D}_t} p^i(\cdot \mid X^{\mathcal{W}_{t-1}})\Big)\right].$$

It thus suffices to control the per-iteration KL error term in (16) for the maximum entropy-based strategy.

Recall that for the entropy sum-based strategy (Algorithm 1), we handle this term via the KL divergence bound in (19) from Lemma 1. Here, we instead apply the KL divergence identity (18) from the same lemma. Specifically, combining (18) with the representation of the unmasking set $\mathcal{D}_t$ in (12):

$$\mathcal{D}_t = \mathcal{W}_t \setminus \mathcal{W}_{t-1} = \{\Pi_{|\mathcal{W}_{t-1}|+1}, \ldots, \Pi_{|\mathcal{W}_t|}\},$$

we can express the KL error from iteration t in (16) as

$$\begin{aligned}&\mathsf{KL}\Big(p^{d_t}(\cdot \mid x^{w_{t-1}}) \,\Big\|\, \prod_{i\in d_t} p^i(\cdot \mid x^{w_{t-1}})\Big) \\ &\quad= \sum_{k=1}^{|d_t|-1} I\big(X^{\pi_{|w_{t-1}|+k}}; X^{\pi_{|w_{t-1}|+k+1}}, \ldots, X^{\pi_{|w_t|}} \,\big|\, (\mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) = (w_{t-1}, x^{w_{t-1}}), \Pi = \pi\big).\end{aligned} \tag{50}$$

That is, the contribution of iteration t is given by a sum of mutual information terms, where the k-th term measures the mutual information between the k-th unmasked token and the tokens that are unmasked later within the same iteration. In particular, the last unmasked token in the iteration $X^{\pi_{|w_t|}}$ does not contribute to the KL error.

Substituting (50) back into the expression of the KL sampling error (17) yields

$$\begin{aligned}&\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi}) \\ &\quad= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{|\mathcal{D}_t|-1} \mathtt{I}(X^{\Pi_{|\mathcal{W}_{t-1}|+k}}; X^{\Pi_{|\mathcal{W}_{t-1}|+k+1}}, \ldots, X^{\Pi_{|\mathcal{W}_t|}} \mid \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}, \Pi)\right] \\ &\quad= \mathbb{E}\left[\sum_{k=1}^{L} \mathtt{I}(X^{\Pi_k}; X^{\Pi_{k+1}}, \ldots, X^{\Pi_{|\mathcal{W}_{t_{\Pi_k}}|}} \mid \mathcal{W}_{t_{\Pi_k}-1}, X^{\mathcal{W}_{t_{\Pi_k}-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^{\Pi_k} \mid \mathcal{W}_{t_{\Pi_k}-1}, X^{\mathcal{W}_{t_{\Pi_k}-1}}) < \eta\big\}\right],\end{aligned} \tag{51}$$

where the second line reindexes by the absolute unmasking order $k \in [L]$ instead of the relative order within each iteration. Here, we recall that t_i denotes the iteration at which token i is unmasked and thus t_{Π_k} stands for the iteration at which the k-th unmasked token Π_k is revealed. The indicator function arises because only non iteration-ending tokens contribute to the KL error and they satisfy the low-entropy condition by the maximum entropy-based stopping rule.

We next regroup the sum by token index $i \in [L]$ instead of unmasking order $k \in [L]$. To this end, let $\mathcal{L}^i$ be the index set of the tokens that are unmasked in the same iteration as token i but are revealed after it, namely,

$$\mathcal{L}^i \coloneqq \big\{\Pi_{\Pi_i^{-1}+1}, \Pi_{\Pi_i^{-1}+2}, \ldots, \Pi_{|\mathcal{W}_{t_i}|}\big\}. \tag{52}$$

Note that $\mathcal{L}^i$ is a deterministic function of the permutation Π and the previous iterate $(\mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}})$. Rewriting the sum in (51) accordingly yields

$$\mathbb{E}_\Pi\big[\mathsf{KL}(p_{X_0} \,\|\, p_{Y_T|\Pi})\big] = \sum_{i=1}^{L} \mathbb{E}\Big[\mathtt{I}(X^i; X^{\mathcal{L}^i} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi) < \eta\big\}\Big]. \tag{53}$$

In a nutshell, this demonstrates that the expected KL sampling error can be decomposed into a sum of mutual information terms over token index $i \in [L]$, where each term measures the correlation between token i and the tokens that are unmasked later within the same iteration, conditioned on the tokens revealed in prior iterations and the permutation.

We note that in contrast to the analogous expression (22) for the entropy sum-based scheme, the mutual information term here involves only the tokens unmasked after token i within the iteration, i.e., $X^{\mathcal{L}^i}$, rather than all other tokens in the same set, i.e., $X^{\mathcal{D}_{t_1} \setminus \{i\}}$.

Step 2: Bound KL error by entropy-based stopping rule. Equipped with the KL error decomposition in (53), it suffices to fix a token index $i \in [L]$ and focus on the i-th term:

$$\mathbb{E}_{\Pi,X}\Big[\mathtt{I}(X^i; X^{\mathcal{L}^i} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi) < \eta\big\}\Big]. \tag{54}$$

This is established in the following lemma; with proof deferred to Appendix D.2.

Lemma 6. *For any token index $i \in [L]$, we have*

$$\mathbb{E}_{X,\Pi_{\setminus i},\Pi_i^{-1}}\Big[\mathtt{I}(X^i; X^{\mathcal{L}^i} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) < \eta\big\}\Big] \le \frac{S_{\max}}{L}\eta. \tag{55}$$

With Lemma 6 in hand, plugging (55) into (54) and summing over all token indices $i \in [L]$, we find that the expected KL sampling error in (53) can be bounded as

$$\mathbb{E}_{\Pi}\big[\mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi})\big] \le S_{\max}\eta.$$

Setting the parameters as

$$S_{\max} = \sqrt{\frac{\varepsilon L}{H(X_0)}} \qquad \text{and} \qquad \eta = \frac{H(X_0)S_{\max}}{L} = \sqrt{\frac{\varepsilon H(X_0)}{L}}$$

yields the desired KL error bound $\mathbb{E}_{\Pi}\big[\mathsf{KL}(p_{X_0} \parallel p_{Y_T|\Pi})\big] \le \varepsilon$.

Step 3: Control expected number of iterations. As for the expected total number of iterations $\mathbb{E}[T]$, it is controlled by the following lemma; its proof is deferred to Appendix D.3.

Lemma 7. *The expected number of iterations $\mathbb{E}[T]$ is upper bounded as*

$$\mathbb{E}[T] \le \frac{2L}{S_{\max}} + 1. \tag{56}$$

In other words, the expected number of iterations is of the same order as if each iteration unmasks exactly $S_{\max}$ tokens.

The claim in (10) then follows by plugging in the choice of $S_{\max}$ into the above bound. We now complete the proof of Theorem 2.

D.2 Proof of Lemma 6

Our strategy is first to take the expectation over Π_i^{-1}, the insertion position of token i in the permutation Π. As in Lemma 2, the main technical challenge lies in the statistical dependence between $\mathcal{L}^i$ (the set of tokens unmasked after token i in iteration t) and Π_i^{-1}, making it difficult to take the expectation over the randomness of Π in (54).

To this end, similar to the proof for Lemma 2 in Appendix B.2, we consider the auxiliary sampling process in which we unmask the entire token sequence except for the token with index i. We work with $X^{\setminus i} = (X^1, \ldots, X^{i-1}, X^{i+1}, \ldots, X^L)$ and run the maximum entropy-based sampling procedure (Algorithm 2) in the order induced by the permutation

$$\Pi_{\setminus i} \coloneqq \big(\Pi_1, \ldots, \Pi_{\Pi_i^{-1}-1}, \Pi_{\Pi_i^{-1}+1}, \ldots, \Pi_L\big).$$

This auxiliary sampling process produces a sequence of cumulative unmasking sets $(\mathcal{W}_t^{\setminus i})_{t=0}^L$ and a sequence of unmasking sets $(\mathcal{D}_t^{\setminus i})_{t=1}^L$. Here, we slightly abuse notation and continue using $\mathcal{W}^{\setminus i}$

and $\mathcal{D}^{\setminus i}$ to denote the cumulative and incremental unmasking sets in the auxiliary process as in Appendix B.2, but now they are defined based on the maximum entropy-based sampling procedure instead of the entropy sum-based sampling procedure.

As before, we make several key observations.

1. For any iteration $t \geq 1$, conditioned on $(\mathcal{W}_{t-1}^{\setminus i}, X^{\mathcal{W}_{t-1}^{\setminus i}})$ and $\Pi_{\setminus i}$, the auxiliary sets $(\mathcal{D}_t^{\setminus i}, \mathcal{W}_t^{\setminus i})$ are fully determined. In addition, they are independent of the insertion position Π_i^{-1} for token i.
2. By definition, $\mathcal{D}_{t_i}^{\setminus i}$ stands for the index set of tokens unmasked at the iteration where token i is unmasked in the auxiliary sampling process. Because the auxiliary process never considers the token with index i, the set $\mathcal{D}_{t_i}^{\setminus i}$ consists of either all low-entropy indices up to (and including) the first high-entropy index among indices not equal to i, or low-entropy indices with size at most $s_{\max}$. Consequently, the index set of tokens that are unmasked in the same iteration as token i but after it in the original process must be a subset of the auxiliary unmasking set, i.e.,
$$\mathcal{L}^i = \{\Pi_{\Pi_i^{-1}+1}, \ldots, \Pi_{|\mathcal{W}_{t_i}|}\} \subseteq \mathcal{D}_{t_i}^{\setminus i}.$$
3. By construction, we have $\mathcal{W}_k = \mathcal{W}_k^{\setminus i}$ and $\mathcal{D}_k = \mathcal{D}_k^{\setminus i}$ for all $1 \leq k < t_i$, and hence
$$\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}^{\setminus i}} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi_{\setminus i}, \Pi_i^{-1}) = \mathtt{I}(X^i; X^{\mathcal{D}_{t_i}^{\setminus i}} \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}, \Pi_{\setminus i}, \Pi_i^{-1})$$
$$\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) = \mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}).$$

In light of these observations, we can bound the target in (54) as

$$\begin{aligned}
&\mathtt{I}(X^i; X^{\mathcal{L}^i} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \leq \eta\} \\
&\quad \leq \mathtt{I}(X^i; X^{\mathcal{D}_{t_i}^{\setminus i}} \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}, \Pi)\mathbb{1}\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}, X^{\mathcal{W}_{t_i-1}}) \leq \eta\}. \\
&\quad = \mathtt{I}(X^i; X^{\mathcal{D}_{t_i}^{\setminus i}} \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}, \Pi_{\setminus i}, \Pi_i^{-1})\mathbb{1}\{\mathtt{H}\big(X^i \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}\big) \leq \eta\} \qquad (57)
\end{aligned}$$

where the second line arises from Observation D2 that $\mathcal{L}^i \subseteq \mathcal{D}_{t_i}^{\setminus i}$ and the monotonicity of mutual information; the last line holds due to Observation D3.

Therefore, it suffices to control the right-hand side of (57). Towards this, we define the random variable

$$\begin{aligned}
G_k &\coloneqq \mathtt{I}\big(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}, \Pi_i^{-1}\big)\mathbb{1}\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}) \leq \eta\} \\
&= \mathtt{I}\big(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}\big)\mathbb{1}\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}) \leq \eta\}, \qquad (58)
\end{aligned}$$

for each $k \in [L]$. Here, the second line arises from Observation D1, namely, conditioned on $\Pi_{\setminus i}$ and $(\mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}})$, $\mathcal{D}_k^{\setminus i}$ is independent of Π_i^{-1}. We can then write the target quantity to be controlled on the right-hand side of (57) as

$$\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}^{\setminus i}} \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}, \Pi)\mathbb{1}\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}) \leq \eta\} = G_{t_i} = \sum_{k=1}^{L} \mathbb{1}\{t_i = k\} G_k, \qquad (59)$$

To proceed, writing

$$\mathbb{E}\big[G_{t_i}\big] = \mathbb{E}_{X^{\setminus i}, \Pi_{\setminus i}}\Big[\mathbb{E}_{\Pi_i^{-1}}\big[G_{t_i} \mid X^{\setminus i}, \Pi_{\setminus i}\big]\Big],$$

we focus on the inner expectation $\mathbb{E}_{\Pi_i^{-1}}\big[G_{t_i} \mid X^{\setminus i}, \Pi_{\setminus i}\big]$, where the expectation is taken over the randomness of Π_i^{-1}, the insertion position of token i in the permutation Π, conditioned on $X^{\setminus i}$ and $\Pi_{\setminus i}$.

By Observation D1 and the definition in (58), G_k is deterministic conditioned on $\Pi_{\setminus i}$ and $X^{\setminus i}$ for each $k \in [L]$. Therefore, the inner expectation can be computed as

$$\mathbb{E}_{\Pi_i^{-1}}\big[G_{t_i} \mid X^{\setminus i}, \Pi_{\setminus i}\big] = \mathbb{E}_{\Pi_i^{-1}}\left[\sum_{k=1}^{L} G_k \mathbb{1}\{t_i = k\} \mid X^{\setminus i}, \Pi_{\setminus i}\right] = \sum_{k=1}^{L} G_k \mathbb{P}\big\{t_i = k \mid X^{\setminus i}, \Pi_{\setminus i}\big\}. \tag{60}$$

This suggests that we need to study the random iteration index t_i at which token i is unmasked in the original sampling process, which is determined by the insertion position Π_i^{-1} of token i in the permutation Π. This is accomplished by the following lemma, whose proof is deferred to Appendix D.4.

Lemma 8. *Conditioned on $\Pi_{\setminus i}$ and $X^{\setminus i}$, the probability that token i is unmasked at iteration k in the original sampling process is upper bounded by*

$$\mathbb{P}_{\Pi_i^{-1}}\big\{t_i = k \mid X^{\setminus i}, \Pi_{\setminus i}\big\} \le \frac{s_{\max}}{L}, \quad \forall\, k \in [L].$$

With Lemma 8 in hand, we can continue the derivation in (60) as

$$\begin{aligned}
\mathbb{E}_{\Pi_i^{-1}}\big[G_{t_i} \mid X^{\setminus i}, \Pi_{\setminus i}\big] &= \sum_{k=1}^{L} G_k \mathbb{P}\big\{t_i = k \mid X^{\setminus i}, \Pi_{\setminus i}\big\} \\
&\le \frac{s_{\max}}{L} \sum_{k=1}^{L} G_k \\
&= \frac{s_{\max}}{L} \sum_{k=1}^{L} \mathtt{I}(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}) \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}) \le \eta\big\},
\end{aligned}$$

where the last line follows from the definition of G_k in (58). Taking the expectation over $\Pi_{\setminus i}$ and $X^{\setminus i}$ on both sides, we obtain

$$\mathbb{E}[G_{t_i}] \le \frac{s_{\max}}{L} \mathbb{E}\left[\sum_{k=1}^{L} \mathtt{I}(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}) \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}) \le \eta\big\}\right]. \tag{61}$$

Now applying Lemma 5 with $s = 1$, the summation on the right-hand side of (61) satisfies

$$\mathbb{E}\left[\sum_{k=1}^{L} \mathtt{I}(X^i; X^{\mathcal{D}_k^{\setminus i}} \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}, \Pi_{\setminus i}) \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{k-1}^{\setminus i}, X^{\mathcal{W}_{k-1}^{\setminus i}}) \le \eta\big\}\right] \le \mathtt{H}(X^i) \wedge \eta \le \eta.$$

Plugging this into (61) and combining it with (59), we can bound the expectation of the right-hand side of (57) by

$$\mathbb{E}\big[\mathtt{I}(X^i; X^{\mathcal{D}_{t_i}^{\setminus i}} \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}, \Pi) \mathbb{1}\big\{\mathtt{H}(X^i \mid \mathcal{W}_{t_i-1}^{\setminus i}, X^{\mathcal{W}_{t_i-1}^{\setminus i}}) \le \eta\big\}\big] \le \frac{s_{\max}}{L}\eta.$$

Finally, putting this together with (57), we finish the proof of Lemma 6.

D.3 Proof of Lemma 7

By the termination condition of the maximum entropy-based sampling procedure, we can split iterations into three types:

- cap iterations: end because $D_t = |\mathcal{D}_t| = s_{\max}$;
- high-entropy iterations: end because the stopping token has entropy exceeding η;
- the final iteration, which contributes at most 1 to the total number of iterations T.

Let T_{cap} and T_{high} denote the number of the first two types, respectively. We have $T \le T_{\mathsf{cap}} + T_{\mathsf{high}} + 1$.

For T_{cap}, since each cap iteration unmasks $s_{\max}$ tokens, we have

$$T_{\mathsf{cap}} \le L/s_{\max}. \tag{62}$$

Therefore, the remainder of the proof focuses on controlling $\mathbb{E}[T_{\mathsf{high}}]$.

Fix an arbitrary $t \in [T]$. Suppose that the t-th iteration is a high-entropy iteration. Let $i_t \in \mathcal{D}_t$ be the index of the last unmasked token at iteration t, i.e., $i_t = \Pi_{|\mathcal{W}_t|}$. By the stopping rule, we know that

$$\eta < \mathtt{H}(X^{i_t} \mid \Pi, \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}) \le \mathtt{H}(X^{\mathcal{D}_t} \mid \Pi, \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}),$$

where the last inequality follows from the monotonicity of entropy. This implies that

$$\mathbb{1}\{\text{iteration } t \text{ is high-entropy}\} \le \frac{1}{\eta}\mathtt{H}(X^{\mathcal{D}_t} \mid \Pi, \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}}).$$

Summing over $t \in [T]$ and taking expectation, we obtain

$$\mathbb{E}[T_{\mathsf{high}}] \le \frac{1}{\eta}\mathbb{E}\left[\sum_{t=1}^{T} \mathtt{H}(X^{\mathcal{D}_t} \mid \Pi, \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}})\right]. \tag{63}$$

As shown in (31) from Lemma 3, we know that the sum on the right-hand side of the above display satisfies

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathtt{H}(X^{\mathcal{D}_t} \mid \Pi, \mathcal{W}_{t-1}, X^{\mathcal{W}_{t-1}})\right] = H(X_0),$$

Consequently, we obtain

$$\mathbb{E}[T_{\mathsf{high}}] \le \frac{H(X_0)}{\eta} = \frac{L}{s_{\max}}, \tag{64}$$

where the last step is due to our choice $\eta = H(X_0)s_{\max}/L$.

Combining (62) and (64), we conclude that

$$\mathbb{E}[T] \le \mathbb{E}[T_{\mathsf{cap}}] + \mathbb{E}[T_{\mathsf{high}}] + 1 \le \frac{2L}{s_{\max}} + 1.$$

D.4 Proof of Lemma 8

Fix an arbitrary token index $i \in [L]$ and condition on an arbitrary realization of $\Pi_{\setminus i}$ and $X^{\setminus i}$. Recall the auxiliary sampling process defined by removing token i. Conditioned on $\Pi_{\setminus i}$ and $X^{\setminus i}$, the auxiliary unmasking sets $(\mathcal{D}_t^{\setminus i}, \mathcal{W}_t^{\setminus i})_{t\in[L]}$ are deterministic.

Because Π is uniformly random over all permutations of $[L]$, after conditioning on the relative order of all indices except i, the insertion position Π_i^{-1} of token i is uniformly random over $[L]$:

$$\mathbb{P}\{\Pi_i^{-1} = r \mid \Pi_{\setminus i}, X^{\setminus i}\} = \frac{1}{L}, \quad r = 1, \dots, L.$$

Now fix an arbitrary $k \in [L]$ and consider the event $\{t_i = k\}$, i.e., token i is unmasked at iteration k in the original sampling process. By Observation B3, the original and auxiliary processes coincide before iteration k. In particular, both processes have unmasked the same set of tokens $\mathcal{W}_{k-1}^{\setminus i}$. Therefore, token i must appear after the first $W_{k-1}^{\setminus i} = |\mathcal{W}_{k-1}^{\setminus i}|$ tokens in the full permutation Π, implying that the insertion position of token i satisfies $\Pi_i^{-1} > W_{k-1}^{\setminus i}$.

Meanwhile, during iteration k, before token i is encountered, the original process sees exactly the same previously revealed tokens and the same sequence of candidate indices as the auxiliary process. Hence, before reaching token i, it cannot proceed past the stopping point of iteration k in the auxiliary process. In particular, if i were inserted after the block $\mathcal{D}_k^{\setminus i}$, then the original process would stop iteration k before unmasking token i. Thus, token i must be inserted within the block

of permutation positions corresponding to $\mathcal{D}_k^{\diagdown i}$, which means that the insertion position of token i satisfies $\Pi_i^{-1} \leq W_k^{\diagdown i} = |\mathcal{W}_k^{\diagdown i}|$.

Combining the above two observations, we find

$$\mathbb{P}\big\{t_i = k \mid \Pi_{\diagdown i}, X^{\diagdown i}\big\} \leq \mathbb{P}\big\{W_{k-1}^{\diagdown i} < \Pi_i^{-1} \leq W_k^{\diagdown i} \mid \Pi_{\diagdown i}, X^{\diagdown i}\big\} = \frac{D_k^{\diagdown i}}{L} \leq \frac{s_{\max}}{L},$$

where the last step holds because by construction of the maximum entropy-based scheme, each iteration unmasks at most $s_{\max}$ tokens.

As the above bound holds for any $k \in [L]$, we finish the proof of Lemma 8.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The paper discusses the limitations of the work performed by the authors.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For each theoretical result, the paper provides the full set of assumptions and a complete (and correct) proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: This paper does not include experiments requiring code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: This paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.