

When i.i.d. Factorization Breaks Language: The Structural Failure of Discrete Diffusion Language Models

Hyukhun Koh¹

Seongho Joo²

Haesung Pyun³

Kyomin Jung^{1,2}

¹IPAI, Seoul National University ²Dept. of ECE, Seoul National University

³Graduate School of Data Science, Seoul National University

{hyukhunkoh-ai, seonghojoo, haesung.pyun, kjung}@snu.ac.kr

Abstract

Discrete diffusion language models (DDLMs) offer potential advantages for language generation, such as parallel decoding and iterative refinement, but they often underperform autoregressive (AR) models and exhibit higher training instability. Recent approaches incorporate AR training signals to stabilize training, yet this introduces strong sequential inductive biases that limit the benefits of non-autoregressive generation. In this work, we analyze the structural causes of degeneration in non-AR DDLMs. We show that standard diffusion formulations rely on an i.i.d. corruption process and factorized sampling structure that implicitly assume token exchangeability, which conflicts with the non-exchangeable dependency structure of natural language. This mismatch leads to degeneration under semantic sparsity, where contextual signals are insufficient. To guarantee stable language generation, we propose the *Dependency-Coupling Principle*, a structural constraint that preserves necessary contextual dependencies during diffusion generation and provides insight into the performance gap between AR and DDLMs.

1 Introduction

Recently, diffusion-based generative models have emerged as a promising alternative paradigm for language modeling. Discrete Diffusion Language Models (DDLMs) offer potential advantages such as parallel generation and iterative refinement. However, despite substantial progress (Austin et al., 2023; Han et al., 2023; Lou et al., 2024; Sahoo et al., 2024), diffusion language models frequently underperform AutoRegressive (AR) models across a range of language generation tasks (Ye et al., 2025; Bie et al., 2025; Koh et al., 2025). Moreover, diffusion training often exhibits substantially higher training variance than AR models, which can lead to significant performance degradation after fine-tuning (Jia et al., 2026).

To address these issues, recent approaches (Liu et al., 2025; Bie et al., 2025) incorporate causal AR training signals. While these observations may appear to favor AR modeling as a backbone, it introduces strong sequential inductive biases (Goel et al., 2026) that may limit the ability of diffusion models to fully exploit Non-AutoRegressive (NAR) generation. This naturally raises a more fundamental question:

"Is there an intrinsic reason why previous DDLMs underperform AR models?"

Therefore, in this work, we investigate the conditions under which traditional NAR DDLMs become unstable or degenerate. In particular, we focus on the structural assumptions underlying diffusion models, including the use of an *i.i.d. corruption process* and the resulting *factorized training and sampling structure*. Standard diffusion formulations implicitly treat token positions as exchangeable variables. However, natural language tokens are fundamentally non-exchangeable: their distributions are governed by structured dependency relations rather than independent marginal statistics. As a result, the independence assumption underlying diffusion corruption conflicts with the dependency structure of language.

Recent work attempts to alleviate this mismatch through improved scheduling or uncertainty-based masking strategies, such as entropy-based block schedules (Koh et al., 2025) or confidence-based dependency estimation (Li et al., 2026). While these methods may improve local prediction quality, they do not resolve the underlying dependency mismatch. In particular, language dependency cannot be fully captured by token-level uncertainty measures such as entropy or perplexity.

Our analysis further shows that the observed degeneration cannot be resolved simply by increasing the number of diffusion steps in Section 5.2, nor is the observed degeneration not caused by insuffi-

cient model expressivity in Section 5.3. In addition, the factorized sampling structure cannot eliminate the underlying mismatch in Section 5.4. Consequently, under *semantic sparsity*—when contextual signals are limited—the factorized update process collapses toward marginal token distributions, leading to repetitive and incoherent outputs.

These observations suggest that the independence assumptions underlying standard diffusion training should be reconsidered for language modeling. **The goal of this paper is to characterize the structural conditions required for stable parallel language generation.** Motivated by this perspective, we formulate the *Dependency Coupling Principle* as a **necessary condition** for stable language generation. Under this view, AR models can be understood as a **maximally constrained special case**, where each prediction conditions on a strictly ordered prefix and generation proceeds sequentially. This construction trivially avoids dependency violations, explaining their robustness under data-sparse conditions (Kasa et al., 2025). However, this comes at the cost of enforcing a fixed directional factorization, which limits the ability to capture more general dependency structures and motivates efforts to overcome the intrinsic sequential bias of AR modeling (Sun et al., 2025b; Kumar et al., 2025; Binkowski et al., 2026). This perspective suggests that the key challenge is whether parallel generation can be made compatible with dependency-aware modeling.

2 Related Work

BERT-Generation Traditional language models have been dominated by autoregressive (AR) architectures, while encoder-only models such as BERT (Devlin et al., 2019; Liu et al., 2019; Warner et al., 2024) excel at classification and representation learning. Several approaches (Ghazvininejad et al., 2019; Chen et al., 2022; Goyal et al., 2022) have attempted to develop generative mechanisms, but their performance underperforms AR models. Our work builds upon recent discrete diffusion language model (DDLm) frameworks to enable effective generation with BERT.

Discrete Diffusion Language Models (DDLms) Recently, diffusion-based methods have largely explored to overcome the problem of AR inductive biases. Prior studies show that applying this supervision at every diffusion step rather than only at the first step (Li et al., 2022; Gong et al.,

2023; Sun et al., 2023) leads to improved performance (Lou et al., 2024; Sahoo et al., 2024). Therefore, DDLms such as D3PM (Austin et al., 2023) formulate text generation as iterative refinement, gradually reconstructing text from noise. Subsequent works (He et al., 2022; Lou et al., 2024; Sahoo et al., 2024) has explored efficiency improvements through BERT-based schedules. Diffusion-EAGS (Koh et al., 2025) introduces entropy-adaptive sampling to dynamically control denoising using pre-trained BERT, and outperform AR models in the toxicity mitigation task. The success of DDLms in conditional generation has led to scaled-up versions of these models and follow-up studies on likelihood-based partial masking strategies. However, they still tend to underperform compared to AR models on many generation tasks (Nie et al., 2025; Ye et al., 2025; Liu et al., 2025; Li et al., 2026). Recent analysis examines the limitations of mask diffusion, showing that DDLms exhibit strong locality bias (Sun et al., 2025a) and that mask tokens themselves act as distractors that impair contextual understanding (Piskorz et al., 2025). Therefore, recent approaches adopt AR models as backbones and train them with diffusion objectives (Liu et al., 2025; Bie et al., 2025). Most recently, block-wise semi-autoregressive diffusion models (e.g., SDAR (Cheng et al., 2026)) and AR-backbone hybrids (Liu et al., 2025; Bie et al., 2025; Arriola et al., 2025) narrow this gap by reintroducing partial sequential dependency; we analyze why such structural coupling helps through the lens of the Dependency Coupling Principle (Appendix J). This trend raises a fundamental question: *why do AR models remain so effective for language generation?* In this work, we investigate the structural conditions that enable stable language generation and analyze the conditions that diffusion models must satisfy in order to close the performance gap.

3 Language Modeling as Dependency Graph Generation

Before investigating the underlying problem of traditional DDLms, we first formalize the language modeling task. Natural language is not generated from exchangeable token marginals, but from structured dependency relations between tokens. The standard AR formulation captures this structure through a fixed linear factorization:

$$p(x) = \prod_{i=1}^L p(x_i \mid x_{<i}),$$

where $x = (x_1, \dots, x_L)$ is a token sequence of length L . This factorization imposes a total order over token generation. However, such a linear order is not intrinsic to language itself; rather, it is one particular serialization of a more general dependency structure.

To formalize this observation, we define language generation as a graph-based factorization problem. Let $G = (V, E)$ denote a directed acyclic graph (DAG) over token positions, where $V = \{1, \dots, L\}$ and E represents dependency relations between tokens. For each node $i \in V$, let $\text{pa}(i)$ denote the parent set of node i in G . Then the joint distribution of a sequence can be factorized as

$$p(x) = \prod_{i \in \text{topo}(G)} p(x_i \mid x_{\text{pa}(i)}), \quad (1)$$

where $\text{topo}(G)$ denotes any valid topological ordering of G . Equation (1) generalizes AR language modeling: the AR model is the special case where G is a chain graph. Under this view, language modeling is not fundamentally tied to a linear order, but to a *dependency-respecting generation process*.

A key consequence of Eq. (1) is that generation can in principle be partially parallelized. If multiple nodes have all their parents already generated, then those nodes may be updated simultaneously without violating the dependency structure. Hence, the natural objective for NAR generation is not unrestricted parallel generation, but *parallel generation conditioned on satisfied dependencies*.

Flaw of standard diffusion language models. Standard DDLMs *do not optimize such graph-based objective* in Eq. (1). Instead, under the i.i.d. corruption process in DDLMs, the resulting training objective induces a factorized denoising distribution that breaks the language dependency structure in Eq. (1). Consequently, the model learns factorized token recovery objectives of the form

$$p_\theta(x \mid x_t) = \prod_{i=1}^L p_\theta(x_i \mid x_t, t, i), \quad (2)$$

where x_t denotes a corrupted sequence at diffusion step t . Under severe corruption, especially in the fully masked regime,

$$x_T = [\text{[MASK]}, \dots, \text{[MASK]}], \quad (3)$$

Eq. (2) no longer guarantees conditioning on the dependency structure required by Eq. (1). As a result, the model is trained to recover tokens from

semantically sparse inputs without ensuring that the parent information necessary for coherent generation is available. We will demonstrate that the objective of diffusion-training under such semantically sparse conditions leads to marginal collapse in Section 4 and 5.

4 Problem of One-step Corruption with I.I.D.

4.1 I.I.D. Corruption in DDLM

Existing DDLMs adapt DDPM principles to text by replacing Gaussian corruption with masking, uniform noise, or token replacement. Critically, like continuous diffusion, they adopt *i.i.d. assumption*: noise is sampled independently at each position. This manifests in two ways: (1) independent noise addition during the forward process, and (2) factorized decoding during the reverse process.

4.2 Why I.I.D. Leads to Collapse in Language

4.2.1 The Mathematical Insight

In this section, we mathematically prove that the i.i.d. assumption, when applied to a fully masked sequence, forces the model to learn a specific type of collapsed distribution. We show that under the i.i.d. corruption assumption, the optimal denoising model inevitably degenerates into a position-wise marginal predictor, erasing all joint dependencies in the sequence.

The Setup: Full-MASK Input Consider the initial state of generation where the sequence provides no semantic information, the full-MASK state. Let $\mathcal{V}$ be vocabulary and $Y = (Y_1, \dots, Y_L)$ be the target sequence of length L . Our objective is to analyze the behavior of the learned probability distribution $p_\theta(y \mid x_T)$ and demonstrate that it inherently collapses to the empirical position-wise marginal distribution q_i :

$$q_i(v) := \Pr_{Y \sim p_{\text{data}}} (Y_i = v \mid |Y| = L), \quad v \in \mathcal{V}.$$

The Constraint: Independent Factorization under Full-MASK

Following the existing DDLM’s formulation (Sahoo et al., 2024; Lou et al., 2024), the one-step denoising distribution for a Non-AutoRegressive (NAR) model factorizes over positions. This factorization arises due to *both* (1) independent noise addition at each position, and (2) independently predicting the target. Since x_T contains no semantic information, the model predicts each token conditioned only on its position

index i :

$$p_\theta(y \mid x_T) = \prod_{i=1}^L p_\theta(y_i \mid x_T, i).$$

The Cross-Entropy loss of one-step denoising decomposes linearly due to this factorization:

$$\mathcal{L}(\theta) = \mathbb{E}_{Y \sim p_{\text{data}}} \left[\sum_{i=1}^L \text{CE}(p_\theta(\cdot \mid x_T, i), \mathbf{1}_{Y_i}) \right], \quad (4)$$

where $\mathbf{1}_{Y_i}$ denotes one-hot encoding of random label Y_i , with $Y_i \sim q_i$.

The Result: Convergence to Marginals The global minimizer of the objective in Eq. (4) can be derived via convex optimization. We formally state that under the i.i.d. assumption, the model p_θ inevitably converges to the empirical position-wise marginal distribution, denoted as $q_i(v)$.

Lemma 1 (Marginal Collapse of the Optimal Solution). *The unique optimal solution p_θ^* for the loss in Eq. (4) satisfies:*

$$p_\theta^*(v \mid x_T, i) = q_i(v), \quad \forall i \in \{1, \dots, L\}.$$

The proof Lemma 1 is in Appendix B, and we refer to this phenomenon as marginal collapse.

4.2.2 Implication

Loss of Joint Dependency The proof above demonstrates that under the full-MASK sequence x_T , the model discards all joint dependencies (e.g., subject-verb agreement or semantic correlation) and learns only *which word is likely to appear at position i* . In other words, the learned joint distribution factorizes as a product of independent position-wise marginals:

$$p_\theta(y \mid x_T) \approx \prod_{i=1}^L q_i(y_i).$$

By minimizing the timestep objective independently per position, the model fundamentally fails to capture the syntactic and semantic correlations inherent in language.

4.3 Why Collapse Appears As Repetition

4.3.1 The Linguistic Insight

Supported by formal semantics (Stalnaker, 1978; Heim and Kratzer, 1998) and empirical BERT studies (Devlin et al., 2019; Warner et al., 2024) (Figure 1), we posit that absolute position carries negligible semantic information compared to contextual

signals, particularly under semantically sparse conditions such as the full-MASK state in the form (3). Importantly, the position refers to the absolute index (e.g., 1, 2, 3, ...) rather than structural or dependency-based positions. Prior studies show that token prediction in neural language models depends strongly on surrounding context and syntactic dependencies (Linzen et al., 2016; Khandelwal et al., 2018; Giulianelli et al., 2018). Furthermore, natural language corpora consist of sentences with varying lengths, which further weakens the predictive power of absolute position indices. We experimentally verify that these observations remain valid in the diffusion setting by comparing the predictive information provided by contextual tokens and absolute position (Figure 5). Based on **these linguistic studies of formal semantics and the standard experimental setup of engineering studies**, we adopt the following assumption.

Assumption 1 (Semantic Dominance and Negligible Positional Info). *Let Y denote a token-level random variable drawn from the empirical data distribution $\mathcal{D}$, I its position index, and let C denote the set of contextual, unmasked tokens excluding the target token. We assume that linguistic meaning is dominated by context, rendering the mutual information provided by absolute position negligible in the form (3):*

$$MI(Y; C) \gg MI(Y; I),$$

where $MI(\cdot; \cdot)$ denotes mutual information, and $MI(Y; I)$ is sufficiently small.

Lemma 2 (Near-homogeneity of Position Marginals). *Under Assumption 1, the empirical position-wise marginals q_i remain close to the global unigram distribution q_{global} .*

The proof of Lemma 2 is in Appendix D. As the position-wise marginals collapse toward the global marginal and become nearly position-independent, the resulting predictions manifest as repetition.

4.3.2 Implication

Breaking Language Dependency Combined with the factorized decoding structure, the resulting joint model collapses toward

$$p_\theta(y \mid x_T) \approx \prod_{i=1}^L q_{\text{global}}(y_i),$$

which corresponds to independent sampling from the global unigram distribution. This degener-

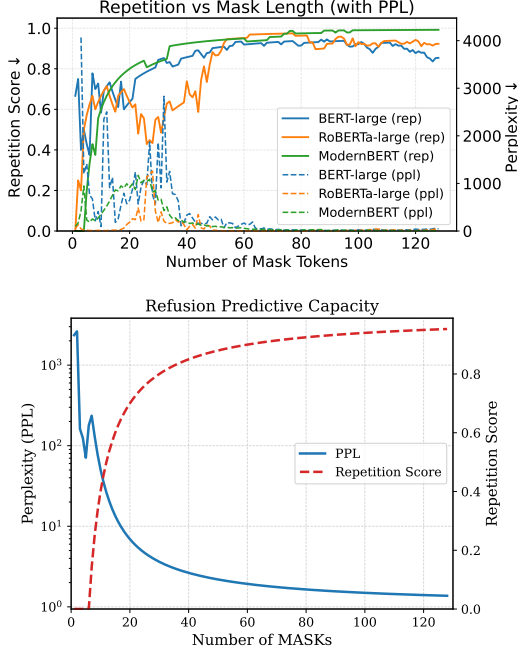


Figure 1: The repetition score $\downarrow$ is computed as $1 - \frac{|\mathcal{U}|}{|\mathcal{T}|}$, where $\mathcal{U}$ is the set of unique tokens in a generated sample and $\mathcal{T}$ is the number of all generated tokens. Perplexity $\downarrow$ is computed using GPT-2 XL. We use pre-trained models including large BERT*, RoBERTa*, ModernBERT*, Refusion-7B*. We construct fully masked sequences and denoise them in a single step.

ation removes the dependency structure that characterizes natural language generation. This factorization is particularly detrimental because natural language tokens are fundamentally non-exchangeable. Unlike image pixels, which exhibit partial invariance to permutation, linguistic tokens are strictly ordered and governed by structural dependencies. Even a single permutation or independent resampling can disrupt syntactic hierarchy (e.g., dependency parse trees) and semantic coherence. Therefore, the i.i.d. assumption, which treats tokens as isolated and exchangeable units, violates the structural principles of language and inevitably leads to generation degeneration. In other words, diffusion-based token recovery primarily aligns marginal token statistics rather than the dependency structure that defines natural language generation.

4.3.3 Empirical Evidence

In Figure 1, as the number of masked tokens increases, pre-trained BERT variants and a recent DDLM (Refusion (Li et al., 2026)) degenerate into generating identical tokens repeatedly (rep-

*BERT-large, RoBERTa-large, ModernBERT-large, and Refusion-7B.

etition score $\rightarrow 1.0$). This behavior indicates that the generated sequence becomes dominated by tokens with high unconditional probability (e.g., the most frequent tokens in the corpus), effectively repeating the mode of the global marginal distribution ($v^* = \operatorname{argmax} q_{\text{global}}$). In other words, when contextual information is largely absent, the optimal predictor of iterative NAR models collapses to marginal token statistics. This regime directly corresponds to the high-noise states in DDLMs, where most tokens remain masked and contextual signals are limited. Thus, the observed degeneration reflects a structural limitation of factorized prediction under sparse context, which persists into early denoising steps and explains marginal collapse in multi-step diffusion.

5 Problem of Multi-step Corruption with I.I.D.

In the previous section, we identified a fundamental flaw in the diffusion objective at the single-step level. We now show that this issue persists in multi-step diffusion. Although additional denoising steps increase the available context, they do not resolve the factorization issue rooted in the training objective, even in recent large DDLMs (Li et al., 2026). In particular, during inference, early denoising stages remain highly information-sparse, where the optimal predictor reduces to position-wise marginals. We analyze three mechanisms underlying this persistence: (1) increasing the number of diffusion steps amplifies repetition, (2) uncertainty-based estimates fail to capture linguistic dependency, and (3) dependency structure remains unenforced during intra-step decoding due to the i.i.d. assumption.

5.1 More Denoising Steps Amplify Repetition

One might argue that the marginal collapse could be mitigated by revealing tokens more gradually. However, as we show next, the problem persists even under fine-grained multi-step diffusion schedules by how training objective manifests during inference, particularly when the model operates under information-sparse conditions in early denoising steps. Theoretically, the number of diffusion steps increases, the transition probability of revealing informative tokens becomes smaller during early timesteps. Consequently, the model spends more iterations operating in semantically sparse states, reinforcing marginal token statistics

and leading to increasingly repetitive outputs.

5.1.1 Empirical Evidence

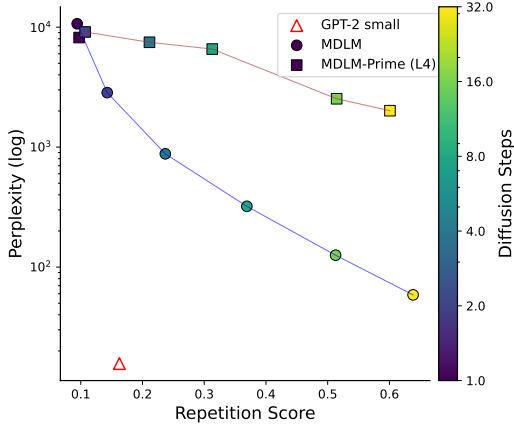


Figure 2: Trade-off between repetition score ($\downarrow$) and perplexity ($\downarrow$). Color indicates the number of diffusion steps. Lines show inference trajectories as steps decrease.

To empirically examine this phenomenon, we evaluate representative diffusion language models including MDLM (Sahoo et al., 2024) and MDLM-Prime (Chao et al., 2025a) trained on the OpenWebText dataset[‡] (Gokaslan et al., 2019). We sample 100 sequences with the same decoding temperature and sampling strategy fixed across different diffusion steps to isolate the effect of the diffusion schedule, following the official repository. Figure 2 reveals a trade-off between likelihood and generation quality. As the number of diffusion steps increases, the perplexity decreases, indicating improved marginal token prediction. However, the repetition score simultaneously increases, suggesting degeneration in token diversity. This behavior supports our analysis: additional denoising steps prolong the model’s exposure to semantically sparse states, where the training objective reinforces marginal statistics rather than dependency-conditioned predictions.

5.2 The Problem of Early Steps In DDLMs

Then why does increasing the number of denoising steps not alleviate this collapse? In the high-noise regime ($t \approx T$) of MDLMs, standard noise schedules (e.g., cosine or linear) induce only a marginal decrease in noise levels between consecutive steps. Consequently, the probability of unmasking a token remains negligible across a long prefix of the trajectory. This results in idle diffusion steps (Chao et al., 2025b), where the sequence state x_t changes little from x_{t+1} and remains predominantly composed

of [MASK] tokens.

This stagnation forces the model to perform denoising repeatedly on nearly identical, information-poor inputs. Because x_t is dominated by masks, the model effectively remains trapped in the full-MASK regime described in Section 4.2 for a substantial portion of training. During these idle steps, the absence of contextual information causes the training signal to be dominated by the position-wise marginal objective (q_i). Instead of learning to recover joint dependencies, the model expends capacity reinforcing collapsed unigram predictions. Repeated exposure to such inputs thus reinforces the marginal statistics of the dataset while providing insufficient signal for learning joint linguistic dependencies.

5.2.1 Empirical Evidence

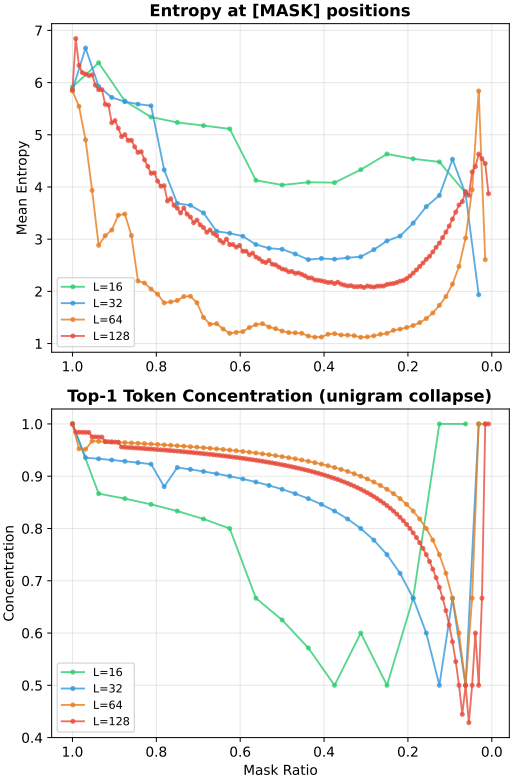


Figure 3: We evaluate sequences of varying lengths (16, 32, 64, and 128). During denoising, we reveal one token per step and measure the per-position prediction entropy at each step (top). In addition, we obtain the top-1 (via argmax) at each position and measure the degree to which these top tokens are identical across positions (bottom).

We investigate the Refusion-7B* to see whether scale-up resolves the marginal collapse by scrutinizing both the entropy and top-1 confidence of each position at each denoising step. As shown in Figure 3, we observe two consistent patterns. First, the model predictions exhibit high entropy

Step t	Mask Ratio	Avg. Top Prob.	Dominant Token
0	1.00	0.066	space
1	0.97	0.053	space
2	0.94	0.078	space
4	0.88	0.11	space
8	0.75	0.34	space
12	0.62	0.47	space
16	0.50	0.51	space
20	0.38	0.58	space
24	0.25	0.61	space
28	0.12	0.35	digits
31	0.03	0.45	symbol

Table 1: Prediction statistics during denoising showing the stagnation regime.

in early steps. At the same time, the **top-1 predictions across positions are highly homogeneous**, often collapsing to a small set of high-frequency tokens regardless of sequence length. This indicates that the model relies primarily on marginal token statistics rather than contextual dependencies, as the predicted distribution remains largely unchanged during the early diffusion steps (Figure 6). To further examine this effect, we conduct a linear denoising schedule and track the proportion of masked positions throughout the process. As shown in Table 1, denoising steps remain dominated by space tokens, confirming that the model spends a substantial portion of the diffusion trajectory in this mask-trapped regime. Interestingly, the average top-1 probability increases while predictions are dominated by high-frequency tokens (e.g., space). As more tokens become revealed, contextual information sufficiently emerges and the dominant token changes only in final steps.

5.3 Adaptive Scheduling Does Not Fix The Structural Issue

Traditional unconditional DDLMs often rely on heuristic scheduling strategies that prioritize either average confidence statistics across samples (Koh et al., 2025) or sample-specific uncertainty (Li et al., 2026). These approaches typically select tokens to reveal based on prediction confidence or entropy, implicitly assuming that low-uncertainty tokens provide useful signals for generation. Such strategies have shown promising performance in *strongly conditioned generation tasks*, where sufficient semantic context is available. In these settings, diffusion-based models can even outperform AR models (Koh et al., 2025). However, as generation becomes more open-ended and the amount of available context decreases (i.e., semantic sparsity increases), their performance deteriorates and often falls below that of AR models (Liu et al.,

2025; Bie et al., 2025). This phenomenon persists even in recent large-scale diffusion language models, despite incorporating dependency-aware heuristics (Li et al., 2026). This suggests that increasing model scale alone does not resolve the structural limitations of diffusion generation.

These results highlight a key limitation of uncertainty-based scheduling. **Confidence or entropy measures struggle to capture mutual information** between tokens and the underlying dependency structure, particularly under the semantically sparse states of early diffusion steps. Consequently, uncertainty-based scheduling may prioritize tokens with low local uncertainty without necessarily revealing tokens that are informative for recovering the underlying dependency structure.

5.3.1 Empirical Evidence

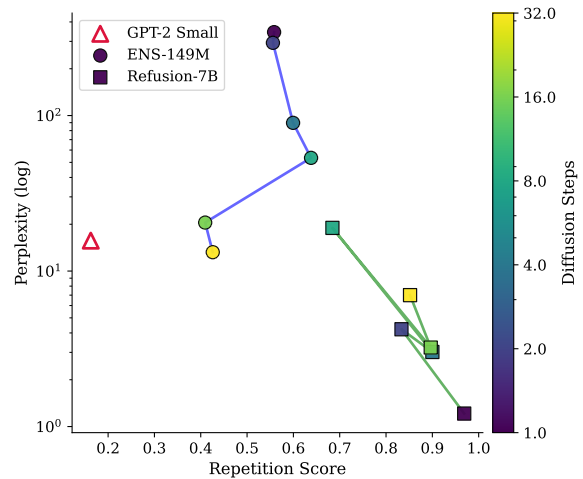


Figure 4: The repetition score does not decrease below 0.4 for both uncertainty-based DDLMs.

We conduct an experiment under the same experimental setting described in Section 5.1.1. In this experiment, we follow the prior works in which entropy-based (ENS-149M) (Koh et al., 2025) and confidence-based (Refusion-7B) (Li et al., 2026) estimation is used as a proxy for language dependency. However, as shown in Figure 4, diffusion models still exhibit persistent repetition. Even when the number of denoising steps increases, the repetition score does not fall below 0.4. Qualitative inspection of generated samples further confirms that the models frequently produce repetitive outputs (See Attached Materials). Notably, this **de-generation persists even when the size of model (Refusion) is 7B**, indicating that confidence-based scheduling alone is insufficient to resolve the problem. Examples are in Appendix G. Additional discussion is in Appendix G.5.

5.4 Intra-Step I.I.D Factorization Persists

Even perfect scheduling is insufficient because intra-step decoding remains independent. Even if dependency relations could somehow be properly estimated during the scheduling stage, the reverse diffusion update *within each step* still remains independently factorized. Specifically, the reverse diffusion transition assumes position independence:

$$p_{\theta}(x_{t-1} | x_t) = \prod_{i=1}^L p_{\theta}(x_{t-1,i} | x_t, t, i).$$

This assumption corresponds to the standard formulation of parallel diffusion decoding, where token updates are sampled independently conditioned on the corrupted state. However, such *naive parallel sampling* ignores dependency relations between simultaneously updated positions. **As a result, even if dependency structures are estimated during scheduling, they are not enforced during sampling.** Preserving linguistic dependencies therefore requires *dependency-aware sampling* that conditions on the current token reveal set, rather than sampling all target positions independently, as empirically illustrated in Appendix G.3.

6 Minimal Condition for DDLMs

In Sections 4 and 5, we show that the i.i.d. assumption underlying DDLMs is fundamentally incompatible with the language dependency structure required in Eq. (1). To resolve this structural conflict, the following two conditions must hold. We refer to these conditions collectively as *Dependency Coupling Principle*.

Dependency Coupling Principle

Let $V_t(i)$ be the visible tokens when predicting the token at masked position i . Stable language generation requires:

C1. Information availability:

$$MI(Y_i; Y_{V_t(i)}) > 0$$

for each masked token to be predicted.

C2. Conditional Dependency Decoupling:

$$MI(Y_i; Y_j | X_t) = 0$$

for all tokens updated in parallel.

Mutual Information Availability This condition is not automatically satisfied by the visible set $V_t(i)$ (a.k.s. sufficient context). In semantically sparse diffusion states, the currently visible tokens may carry little or no predictive information about the target token, i.e.,

$$MI(Y_i; Y_{V_t(i)}) \approx 0.$$

In that case, the marginal predictor may remain optimal under cross-entropy loss, leading to position-wise collapse. As discussed in Section 5.3, conditional entropy is generally a poor proxy under semantically sparse diffusion states; a more appropriate measure is therefore required. By C1, the optimal predictor for the cross-entropy objective is no longer the marginal predictor (See Details in Appendix I).

Conditional Dependency Decoupling Information availability alone is insufficient for parallel revealing, as the predicted tokens themselves may still have strong interdependencies. Even when the visible context is sufficient to predict each token individually, parallel revealing should respect the dependency structure among the predicted tokens themselves. Tokens that are tightly coupled should therefore not be revealed simultaneously.

As a simple proof-of-concept approximation of C1 and C2 in Appendix I.1, we can first use a NAR model to predict all masked tokens in parallel from the current visible context X_t . This stage provides candidate token hypotheses based on available context, loosely approximating C1. Next, a dependency parser is applied to these hypotheses, and parallel updates are restricted to tokens that are not directly connected, thereby introducing a simple structural constraint that approximates C2. However, to achieve a fully scalable approach that does not rely on external dependency parsers, it would be necessary to develop a model-internal estimator that can infer such dependency structures directly from data. We leave this direction for future work.

7 Discussion & Conclusion

Diagnosing block-causal DDLMs through the DCP. Block-causal methods such as LLaDA (Nie et al., 2025; Bie et al., 2025), Dream (Ye et al., 2025), and Block Diffusion (Arriola et al., 2025) generate text block-by-block with causal attention between blocks. The DCP exposes two structural limitations. From the C1 (information availability)

perspective, the number of tokens decoded in parallel should adapt to the available context: when context is rich (low mask ratio) larger blocks are safe, but when context is sparse (e.g., the first block with no prompt) blocks should shrink. Fixed block sizes violate C1, and our mask-ratio sweep quantifies the boundary—coherent below ≈ 0.5 , collapse above ≈ 0.7 (Appendix J.2). Calibrating block size to the information state at each step—generating more when the model knows more, less when it knows less—remains an important open challenge. From the C2 (conditional dependency decoupling) perspective, positions decoded in parallel should be conditionally independent given the visible context, yet block methods assign *contiguous* spans, and linguistically independent positions do not necessarily form contiguous spans. A more principled scheme would identify positions whose predictions do not influence each other given the current context and decode those in parallel, enabling safe parallelism without the i.i.d. assumption.

Why confidence-based scheduling is not enough.

Confidence- and entropy-based schedulers (Li et al., 2026; Koh et al., 2025) succeed only when C1 is already satisfied by the existing context: they prioritize positions where the model is already informative, which in turn supplies context for subsequent positions. However, as block size or mask ratio grows, the confidence signal itself becomes unreliable—when all positions face semantic sparsity, every confidence estimate is uniformly uninformative (Section 5.3, Figure 4). Confidence scheduling therefore is insufficient to guarantee unless sufficient context are provided. We empirically ground both observations on mainstream DDLMs (up to 16B parameters) in Appendix J: the collapse persists across training objectives, decoding parallelism, and scale, and is reversed only when dependency information is supplied—either sequentially (Appendix J.3) or through explicit dependency coupling (Appendix J.5).

Taken together, our analysis suggests that the key challenge in DDLMs is not merely improving denoising accuracy, but fundamentally reconciling parallel generation with dependency-aware modeling. Our analysis indicates that this requires moving beyond fully factorized objectives toward mechanisms that preserve conditional dependencies without reverting to strictly sequential generation. This perspective points toward a new class of NAR language models that retain the efficiency

of parallel decoding while capturing globally consistent dependency structures, overcoming the prefix bias inherent in traditional AR models (Sun et al., 2025b; Kumar et al., 2025; Binkowski et al., 2026).

Limitations

Our goal is not to modify the training objective, but to identify and characterize a fundamental structural limitation of existing diffusion language models. By isolating this effect at inference time, we show that the observed degeneration arises from the objective itself rather than specific architectural or implementation choices. Addressing this limitation would require fundamentally altering the training objective to incorporate dependency-aware coupling, which constitutes a separate research direction beyond the scope of this work.

To support our analysis, we provide a simple proof-of-concept algorithm that partially instantiates the proposed principles. This is not intended as a complete solution, but as a minimal demonstration that incorporating dependency-aware constraints can mitigate degeneration in practice.

Importantly, we do not attempt to explicitly estimate mutual information or fully decouple conditional dependencies at the representation level. These would require reliable semantic inference under information-sparse conditions, which is itself a challenging and unresolved problem, and thus lies beyond the scope of this work.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No. RS-2023-00229780, Development of Artificial Intelligence Technology for Process-focused Evaluation (Student’s Learning Diagnosis)]. This work was also supported by LYWAY & computational resources from the 2026 High-Performance Computing Support Project (GPU), managed by the National IT Industry Promotion Agency (NIPA) [No. RQT-26-040786], and partially supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) [RS-2025-02263628] and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211343, Artificial Intelligence Grad-

uate School Program (Seoul National University)]. K. Jung is with ASRI, Seoul National University, Korea.

References

- Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Cheung. 2022. [Why exposure bias matters: An imitation learning perspective of error accumulation in language generation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 700–710, Dublin, Ireland. Association for Computational Linguistics.
- Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. 2025. [Block diffusion: Interpolating between autoregressive and diffusion language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2023. [Structured denoising diffusion models in discrete state-spaces](#).
- Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Ling Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. 2025. [Llada2.0: Scaling up diffusion language models to 100b](#).
- Jakub Binkowski, Kamil Adamczewski, and Tomasz Jan Kajdanowicz. 2026. [Attention sinks as internal signals for hallucination detection in large language models](#). In *Workshop on Scientific Methods for Understanding Deep Learning*.
- Chen-Hao Chao, Wei-Fang Sun, Hanwen Liang, Chun-Yi Lee, and Rahul G. Krishnan. 2025a. [Beyond masked and unmasked: Discrete diffusion models via partial masking](#).
- Chen-Hao Chao, Wei-Fang Sun, Hanwen Liang, Chun-Yi Lee, and Rahul G. Krishnan. 2025b. [Beyond Masked and Unmasked: Discrete Diffusion Models via Partial Masking](#). In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.
- Cheng Chen, Yichun Yin, Lifeng Shang, Xin Jiang, Yujia Qin, Fengyu Wang, Zhi Wang, Xiao Chen, Zhiyuan Liu, and Qun Liu. 2022. [bert2BERT: Towards reusable pretrained language models](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2134–2148, Dublin, Ireland. Association for Computational Linguistics.
- Shuang Cheng, Yihan Bian, Dawei Liu, Yuhua Jiang, Yihao Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, and Bowen Zhou. 2026. [SDAR: A synergistic diffusion-AutoRegression paradigm for scalable sequence generation](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 22058–22075, San Diego, California, United States. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. [Mask-predict: Parallel decoding of conditional masked language models](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6112–6121, Hong Kong, China. Association for Computational Linguistics.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. [Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.
- Raghavv Goel, Risheek Garrepalli, Sudhanshu Agrawal, Chris Lott, Mingu Lee, and Fatih Porikli. 2026. [Skip to the good part: Representation structure and inference-time layer skipping in diffusion vs. autoregressive llms](#).
- Aaron Gokaslan, Vanya Cohen, Ellie Pavlick, and Stefanie Tellex. 2019. [Openwebtext corpus](#). <http://Skylion007.github.io/OpenWebTextCorpus>.
- Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. 2023. [Diffuseq: Sequence to sequence text generation with diffusion models](#). In *The Eleventh International Conference on Learning Representations*.
- Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. [Exposing the implicit energy networks behind masked language models via metropolis-hastings](#).
- Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. 2023. [Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control](#).
- Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. 2022. [Diffusionbert: Improving generative masked language models with diffusion models](#).
- Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell, Malden, MA.

- Fei Huang, Tianhua Tao, Hao Zhou, Lei Li, and Minlie Huang. 2022. [On the learning of non-autoregressive transformers](#).
- Mengni Jia, Mengyu Zhou, Yihao Liu, Xiaoxi Jiang, and GuanJunJiang. 2026. [Bringing stability to diffusion: Decomposing and reducing variance of training masked diffusion models](#). In *The Fourteenth International Conference on Learning Representations*.
- Siva Rajesh Kasa, Karan Gupta, Sumegh Roychowdhury, Ashutosh Kumar, Yaswanth Biruduraju, Santhosh Kumar Kasa, Pattisapu Nikhil Priyatam, Arindam Bhattacharya, Shailendra Agarwal, and Vijay Huddar. 2025. [Generative or discriminative? revisiting text classification in the era of transformers](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9615–9637, Suzhou, China. Association for Computational Linguistics.
- Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. 2018. [Sharp nearby, fuzzy far away: How neural language models use context](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 284–294, Melbourne, Australia. Association for Computational Linguistics.
- Hyukhun Koh, Minha Jhang, Dohyung Kim, Sangmook Lee, and Kyomin Jung. 2025. [Conditional \[MASK\] discrete diffusion language model](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8910–8934, Suzhou, China. Association for Computational Linguistics.
- Ashwin Kumar, Yuzi He, Aram H. Markosyan, Bobbie Chern, and Imanol Arrieta-Ibarra. 2025. [Detecting prefix bias in llm-based reward models](#).
- Jia-Nan Li, Jian Guan, Wei Wu, and Chongxuan Li. 2026. [Refusion: A diffusion large language model with parallel autoregressive decoding](#). In *The Fourteenth International Conference on Learning Representations*.
- Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. 2022. [Diffusion-LM improves controllable text generation](#). In *Advances in Neural Information Processing Systems*.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Jingyu Liu, Xin Dong, Zhifan Ye, Rishabh Mehta, Yonggan Fu, Vartika Singh, Jan Kautz, Ce Zhang, and Pavlo Molchanov. 2025. [Tidar: Think in diffusion, talk in autoregression](#).
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized bert pretraining approach](#).
- Aaron Lou, Chenlin Meng, and Stefano Ermon. 2024. [Discrete diffusion language modeling by estimating the ratios of the data distribution](#).
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. [Large language diffusion models](#).
- Iftitahu Nimah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. 2023. [NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada. Association for Computational Linguistics.
- Julianna Piskorz, Cristina Pinneri, Alvaro Correia, Moatasem Alfarra, Risheek Garrepalli, and Christos Louizos. 2025. [Masks can be distracting: On context comprehension in diffusion language models](#).
- Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. 2024. [Simple and effective masked diffusion language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Robert Stalnaker. 1978. Assertion. *Syntax and Semantics (New York Academic Press)*, 9:315–332.
- Haocheng Sun, Cynthia Xin Wen, and Edward Hong Wang. 2025a. [Why mask diffusion does not work](#).
- Haoran Sun, Lijun Yu, Bo Dai, Dale Schuurmans, and Hanjun Dai. 2023. [Score-based continuous-time discrete diffusion models](#). In *The Eleventh International Conference on Learning Representations*.
- Yiliu Sun, Zicheng Zhao, Yang Wei, Yanfang Zhang, and Chen Gong. 2025b. [Well begun, half done: Reinforcement learning with prefix optimization for llm reasoning](#).
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#).
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. [Dream 7b: Diffusion large language models](#).

A Necessity of Cross-Entropy on Vocabulary

However, in this subsection, we show that diffusion with MSE regression in hidden space is fundamentally insufficient for modeling discrete token distributions, which motivates the need for explicit vocabulary-level supervision such as cross-entropy.

Formally, let the vocabulary be $\mathcal{V} = \{1, \dots, V\}$, and for each token $v \in \mathcal{V}$, let its embedding be denoted by a vector $e_v \in \mathbb{R}^d$. Consider a model that, given context x , predicts a hidden vector $h_\theta(x) \in \mathbb{R}^d$, and is trained with the MSE objective

$$\mathcal{L}_{\text{MSE}}(\theta) = \mathbb{E}_{(x,y) \sim p_{\text{data}}} \|h_\theta(x) - e_y\|^2,$$

where y is the next token. Then, the Bayes-optimal predictor under this MSE loss is

$$h^*(x) = \mathbb{E}_{y \sim p(y|x)}[e_y] = \sum_{v \in \mathcal{V}} p_{\text{data}}(v | x) e_v.$$

Specifically, for fixed context x , minimizing $\mathbb{E}_y \|h - e_y\|^2$ in h is equivalent to minimizing $\|h\|^2 - 2h^\top \mathbb{E}[e_y]$, which attains its minimum at $h = \mathbb{E}[e_y]$. That is, the MSE-based objective between diffusion timesteps makes the model learn the *mean embedding* of the next-token distribution, not the full discrete distribution $p(\cdot | x)$. In particular: **(1)** If the embedding set $\{e_i \in \mathbb{R}^d\}_{i=0}^{|\mathcal{V}|}$ are not affinely independent (and in practical language applications $\mathcal{V} \gg d$, so they cannot be), then many distinct discrete distributions $p_{\text{data}} \neq q_{\text{data}}$ get mapped to the same mean embedding. **(2)** Thus, the regression objective is *insufficient* to recover the true next-token distribution and induces a many-to-one projection.

Since hidden-state regression alone does not fully capture the discrete distribution, incorporating cross-entropy-based vocabulary supervision is beneficial for diffusion loss.

B The Proof of Lemma 1

Proof. The expected loss in Eq. (4) decomposes over positions due to the factorization. By definition of the marginal distribution q_i , we can rewrite the expectation:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^L \mathbb{E}_{Y \sim p_{\text{data}}} [\text{CE}(p_\theta(\cdot | x_T, i), \mathbf{1}_{Y_i})] \\ &= \sum_{i=1}^L \mathbb{E}_{Y_i \sim q_i} [\text{CE}(p_\theta(\cdot | x_T, i), \mathbf{1}_{Y_i})]. \end{aligned}$$

Consequently, this reduces to finding optimal p_i for each position i minimizing the Cross-Entropy :

$$\begin{aligned} \min_{p_i \in \Delta^{|\mathcal{V}|-1}} \mathbb{E}_{Y_i \sim q_i} [\text{CE}(p_i, \mathbf{1}_{Y_i})] \\ = \min_{p_i} (H(q_i) + \text{KL}(q_i \| p_i)). \end{aligned}$$

Since $H(q_i)$ is constant with respect to p_i , minimizing the loss is equivalent to minimizing the Kullback-Leibler divergence $\text{KL}(q_i \| p_i)$. The unique minimizer of this term is $p_i = q_i$.

C Verification of Assumption 1

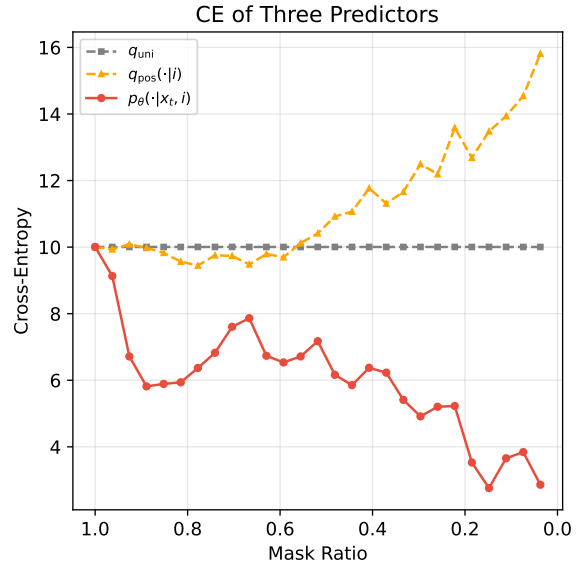


Figure 5: Lower cross-entropy indicates more informative predictors. As the mask ratio decreases, p_θ incorporates increasing amounts of contextual information, while q_{pos} remains limited to positional priors. The widening gap between q_{pos} and p_θ indicates the increasing informational value of context. Considering our findings that entropy-based measurements are unreliable in extremely low-information regimes, we evaluate the predictors after a sufficient amount of context has been revealed (mask ratio < 0.4). Under this condition, the position-only predictor does not provide useful information and in fact deteriorates (higher cross-entropy), whereas the context-based predictor becomes increasingly informative (lower cross-entropy).

Let y_i denote the ground-truth token at position i , and let x_t denote the partially revealed sequence at denoising step t , where unrevealed positions remain masked. We compare three predictors that differ in the amount of contextual information they exploit. To empirically validate our assumption, we conduct the following analysis using the Refusion model (Li et al., 2026).

Unigram predictor. We define a context-free baseline $q_{\text{uni}}(y)$ as the position-averaged predic-

tive distribution under the fully masked input x_T . Concretely, we first compute the model prediction at each position under the full-mask input and then average across positions:

$$q_{\text{uni}}(y) = \frac{1}{L} \sum_{i=1}^L p_{\theta}(y \mid x_T, i). \quad (5)$$

This predictor approximates a marginal unigram-like distribution that does not use contextual information.

Position-only predictor. We define a position-only predictor as the model’s predictive distribution at position i under the fully masked input:

$$q_{\text{pos}}(y \mid i) = p_{\theta}(y \mid x_T, i). \quad (6)$$

Since all tokens are masked in x_T , this predictor can only exploit absolute positional bias.

Corrupted-context predictor. Finally, we consider the model’s predictive distribution at step t given the partially revealed sequence x_t :

$$p_{\theta}(y_i \mid x_t, i). \quad (7)$$

This predictor conditions on the visible tokens already revealed up to step t and therefore reflects the predictive value of the surviving contextual information.

To quantify the predictive benefit of visible context beyond positional bias, we define the *context advantage* as

$$\Delta_t = \mathbb{E}_{i \in M_t} [\text{CE}_{\text{pos}}^{(i)} - \text{CE}_{\text{ctx},t}^{(i)}],$$

where CE denotes as cross-entropy. A positive Δ_t indicates that the visible context in x_t improves prediction relative to the position-only baseline, while values near zero indicate that the remaining context provides little predictive signal beyond positional bias.

Cross-entropy evaluation. For each step t , we evaluate these predictors only on the positions that remain masked, ensuring that all predictions target unrevealed tokens under the same denoising state. For a masked position i , we compute the token-level cross-entropy against the ground-truth token y_i :

$$\begin{aligned} \text{CE}_{\text{uni}}^{(i)} &= -\log q_{\text{uni}}(y_i), \\ \text{CE}_{\text{pos}}^{(i)} &= -\log q_{\text{pos}}(y_i \mid i), \\ \text{CE}_{\text{ctx},t}^{(i)} &= -\log p_{\theta}(y_i \mid x_t, i). \end{aligned}$$

We then average these quantities over the set of masked positions M_t at step t .

To isolate contextual informativeness from error propagation, revealed tokens are replaced with the ground-truth tokens rather than the model’s own predictions during the denoising simulation. This oracle-reveal protocol ensures that the measured gain reflects the informational value of the available context itself, rather than artifacts introduced by accumulated generation errors.

D The Proof of Lemma 2

Proof. The mutual information $MI(Y; I)$ can be expressed as the average KL divergence between local marginals q_i and the global marginal q_{global} :

$$MI(Y; I) = \frac{1}{L} \sum_{i=1}^L \text{KL}(q_i \parallel q_{\text{global}}).$$

By Assumption 1, $MI(Y; I)$ is sufficiently small in the form (3). Applying Pinsker’s inequality,

$$\delta(q_i, q_{\text{global}})^2 \leq \frac{1}{2} \text{KL}(q_i \parallel q_{\text{global}}).$$

Averaging over i gives

$$\frac{1}{L} \sum_{i=1}^L \delta(q_i, q_{\text{global}})^2 \leq \frac{1}{2} MI(Y; I).$$

Therefore, when $MI(Y; I)$ is small, the position-wise marginals are close to the global unigram distribution on average. Moreover, by Jensen’s inequality,

$$\frac{1}{L} \sum_{i=1}^L \delta(q_i, q_{\text{global}}) \leq \sqrt{\frac{1}{2} MI(Y; I)}.$$

Thus, under semantically sparse conditions, the optimal predictor is dominated by the global unigram distribution.

D.1 Conditions for the Mutual Information Decomposition under Full-MASK

We justify the equality

$$MI(Y; I) = \frac{1}{L} \sum_{i=1}^L \text{KL}(q_i \parallel q),$$

where q is q_{global} , and verify that all required conditions are satisfied in the full-MASK setting.

D.2 Uniform Prior over Positions

Let $I \in \{1, \dots, L\}$ denote a uniformly sampled position index. By construction,

$$p(I = i) = \frac{1}{L},$$

which reflects that each token position is treated symmetrically under the i.i.d. corruption process and the full-MASK training objective.

D.3 Conditional Distributions

Let Y denote the token random variable at a randomly selected position I . We define the conditional distribution

$$q_i(y) \triangleq p(Y = y \mid I = i),$$

where Y follows empirical distribution. In the full-MASK regime, the model receives no semantic context and can condition only on the position index i . At optimum, the full-MASK cross-entropy objective yields

$$p_\theta(y \mid x_T, i) = q_i(y),$$

where q_i coincides with the empirical position-wise marginal distribution at position i .

D.4 Mixture (Marginal) Distribution

The marginal distribution of Y is obtained by marginalizing over I :

$$\begin{aligned} q(y) &= p(Y = y) = \sum_{i=1}^L p(I = i) p(Y = y \mid I = i) \\ &= \frac{1}{L} \sum_{i=1}^L q_i(y). \end{aligned}$$

Thus, q is the uniform mixture of the position-wise marginals $\{q_i\}_{i=1}^L$.

D.5 Absolute Continuity

For all i , we have

$$q_i \ll q.$$

This holds automatically since q is a convex combination of $\{q_i\}$. Indeed, if $q(y) = 0$, then

$$\frac{1}{L} \sum_{i=1}^L q_i(y) = 0 \quad \Rightarrow \quad q_i(y) = 0 \quad \forall i,$$

which excludes the possibility that $q_i(y) > 0$ while $q(y) = 0$.

D.6 Resulting Decomposition

Under the above conditions, the mutual information between the token identity Y and the position index I admits the decomposition by using the identities $p(Y \mid I = i) = q_i$ and $p(Y) = q$:

$$\begin{aligned} MI(Y; I) &= \sum_{i=1}^L p(I = i) \text{KL}(q_i \parallel q) \\ &= \frac{1}{L} \sum_{i=1}^L \text{KL}(q_i \parallel q), \end{aligned}$$

where KL denotes Kullback–Leibler divergence. This identity is purely information-theoretic and holds independently of the model parameterization. In the full-MASK setting, it characterizes the amount of information token identity retains about absolute position.

E Derivation

E.1 Details of Derivation

Recall that $Y := Y_i$ denotes a target token at a masked position and $V := Y_{V_i(i)}$ denote the visible context at reverse step t . Consider the log-loss (negative log-likelihood) of a conditional predictor $q(y \mid v)$:

$$\mathcal{L}(q) = \mathbb{E}_{(Y,V)} [-\log q(Y \mid V)].$$

Step 1: Write the difference of losses.

$$\begin{aligned} \mathcal{L}(q_{\text{marg}}) - \mathcal{L}(q^*) &= \mathbb{E}_{(Y,V)} [-\log p(Y) + \log p(Y \mid V)] \\ &= \mathbb{E}_{(Y,V)} \left[\log \frac{p(Y \mid V)}{p(Y)} \right]. \end{aligned}$$

Step 2: Apply the law of total expectation. Using $\mathbb{E}_{(Y,V)}[\cdot] = \mathbb{E}_V[\mathbb{E}_{Y|V}[\cdot]]$, we obtain

$$\mathcal{L}(q_{\text{marg}}) - \mathcal{L}(q^*) = \mathbb{E}_V \left[\mathbb{E}_{Y|V} \left[\log \frac{p(Y \mid V)}{p(Y)} \right] \right].$$

Step 3: Inner expectation as a KL divergence. Fix $V = v$. Then

$$\begin{aligned} \mathbb{E}_{Y|V=v} \left[\log \frac{p(Y \mid V = v)}{p(Y)} \right] &= \sum_y p(y \mid v) \log \frac{p(y \mid v)}{p(y)} \\ &= \text{KL}(p(Y \mid V = v) \parallel p(Y)). \end{aligned}$$

Substituting Step 2 into Step 3 yields

$$\mathcal{L}(q_{\text{marg}}) - \mathcal{L}(q^*) = \mathbb{E}_V [\text{KL}(p(Y \mid V) \parallel p(Y))].$$

Step 4 By definition of expectation and KL divergence,

$$\begin{aligned} \mathbb{E}_V[\text{KL}(p(Y | V) \| p(Y))] \\ &= \sum_v p(v) \text{KL}(p(Y | v) \| p(Y)) \\ &= \sum_v p(v) \sum_y p(y | v) \log \frac{p(y | v)}{p(y)}. \end{aligned}$$

Step 5: Joint distribution. Since $p(v)p(y | v) = p(y, v)$, we can rewrite Step 4 as

$$\begin{aligned} \sum_v p(v) \sum_y p(y | v) \log \frac{p(y | v)}{p(y)} \\ &= \sum_v \sum_y p(v)p(y | v) \log \frac{p(y | v)}{p(y)} \\ &= \sum_v \sum_y p(y, v) \log \frac{p(y | v)}{p(y)}. \end{aligned}$$

Step 6: A joint-to-product form. Using $p(y | v) = \frac{p(y, v)}{p(v)}$, we have

$$\log \frac{p(y | v)}{p(y)} = \log \frac{p(y, v)}{p(y)p(v)}.$$

Then Step 5 yields

$$\begin{aligned} \mathbb{E}_V[\text{KL}(p(Y | V) \| p(Y))] \\ &= \sum_v \sum_y p(y, v) \log \frac{p(y, v)}{p(y)p(v)}. \end{aligned}$$

Step 7: Mutual information. By definition, the mutual information between Y and V is

$$I(Y; V) = \sum_y \sum_v p(y, v) \log \frac{p(y, v)}{p(y)p(v)}.$$

So, we conclude

$$\mathbb{E}_V[\text{KL}(p(Y | V) \| p(Y))] = I(Y; V).$$

Final Step Finally, we obtain the desired relationship:

$$\mathcal{L}(q_{\text{marg}}) - \mathcal{L}(q^*) = I(Y; V).$$

When $I(Y; C) > 0$, the marginal predictor incurs strictly larger loss than the conditional predictor. Hence the optimal prediction must depend on the visible context.

Full-Mask Regime Retention: $D_t = \text{JS}(p_\theta(\cdot | x_t), p_\theta(\cdot | x_T))$

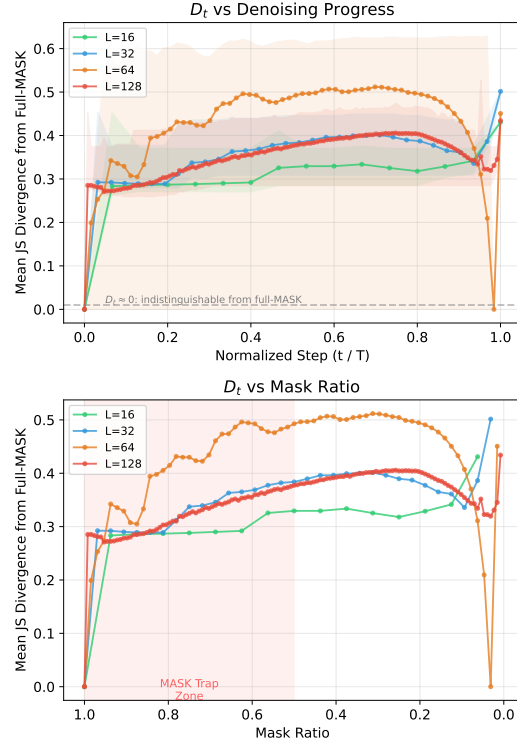


Figure 6: JS divergence between the prediction distribution at step t and the full-mask distribution. The divergence rises sharply at the first denoising step (≈ 0.3) and remains nearly constant thereafter, indicating that early collapse toward marginal token statistics persists throughout denoising.

F Retention Verification

F.1 Persistence of the Full-Mask Prediction Regime

To examine whether early denoising states differ meaningfully from the fully masked condition, we measure the Jensen–Shannon (JS) divergence between the token distributions predicted at each denoising step and those predicted under the fully masked input.

Let x_T denote the fully masked state and x_t the corrupted state at denoising step t . For each position i that remains masked at step t , we compute

$$D_t = \frac{1}{|M_t|} \sum_{i \in M_t} \text{JS}(p_\theta(y_i | x_t), p_\theta(y_i | x_T)),$$

where M_t denotes the set of positions that remain masked at step t .

This metric measures how far the model’s predictions at step t deviate from those produced under the fully masked input. If D_t remains small, the model’s predictions are effectively indistinguishable from the full-mask regime, indicating that the

We observe a rapid increase in divergence at the first denoising step, after which the value stabilizes around approximately 0.3 and changes little throughout the remaining steps.

The subsequent stability of D_t indicates that this degenerate prediction regime persists throughout the denoising trajectory. In other words, although tokens are gradually revealed, the model’s predictive distribution for the remaining masked positions does not substantially move away from the marginal-dominated regime established at the earliest steps.

G Examples

MDLM Repetition:

“real real real real real real real real
real real real real real real real real
real real real real real real real real
real real real real”

“1.5 million 0.0% 13.1% 13.0% 3.2% 4.”

“32 step: By: The New York Times, The New York Times, The New York Times, New York, NY, USA September 1, 2014 12:00”

“8 step: The first the of the the the of the
the the of the first the of the the the
of the the the of the the first.jpg”

"2 step: The first of of the the the the
the the the the the the the the the
the the the the the the the the.."

```

32 step: 100 # 1 1
16 step: 10 1
8 step: 1. 1. . 1. 1. .. 1. ... . .
4 step: 1. 1
2 step: #0. 00
1 step:

```

“of which was designed to protect people on both sides of the spectrum, as they’re perceived and noted in real life. A plurality contended owners who created their own”

“It’s an unusual day, but there is one more day you may never forget. We all love to imagine all the people who made it: the people who”

“On Friday, June 30, 2016, the official release of the first-ever video footage of the first episode of the Game of Us was announced.”

4432

G.3 Sampling Controlled Experiment

The following results compare parallel sampling with sequential intra-step prediction. L denotes the sequence length, and S denotes the number of denoising steps. Sequential generation means all tokens are generated in the AR manner within target blocks.

[$L=64, S=16$] Parallel

“→ has advanced,,,,, and,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
and,,,,,,,,,,,,,,,,,,,,,”

[$L=64, S=16$] Sequential

“→ has advanced the point,, what are the potential of a world where technology can...”

G.4 Predictive Capacity Examples

ModernBERT

The following of the the is\n\n\n\n\n..

Refusion

The 10th of the ...

G.5 Empirical Evidence: The Paradox of Low Perplexity

For uncertainty-based methods (Li et al., 2026), confidence-based scheduling optimizes conditional entropy, not mutual information. In practice, Figure 1 presents a paradox where extreme linguistic collapse coexists with high statistical likelihood. As the number of masked tokens increases, pre-trained BERT-variants degenerate into generating identical tokens repeatedly, but at the same time the perplexity (PPL), evaluated by a pre-trained GPT-2 XL model, remains extremely low or decreases. Language dependency fundamentally relies on mutual information between tokens, which cannot be reliably estimated under semantically sparse diffusion states. In fact, confidence-based method often underperforms compared to similar size of AR baselines in the previous study (Li et al., 2026). In addition, prior studies demonstrate that traditional likelihood-based measures are often poorly aligned with human judgments of language quality (Arora et al., 2022; Nimah et al., 2023; Huang et al., 2022).

H AI ASSISTANCE DISCLOSURE

This paper received limited assistance from AI tools (e.g., ChatGPT) for (i) language proofreading and copy-editing, and (ii) preliminary exploration

of prior work (e.g., keyword suggestions and citation candidates). All technical contributions, experimental designs, implementations, analyses, and final claims are by the authors. All cited references were verified by the authors; no citations were accepted solely on the basis of AI output. No private or sensitive data were shared with AI tools beyond what is explicitly described in the paper.

I Language Dependence Prevents Marginal Collapse

Lemma 3 (Language Dependence Prevents Marginal Collapse). *Fix a reverse step t and a masked position $i \notin V_t$. Assume that there exists language dependence between Y_i and the random token variable $Y_{V_t(i)}$, i.e.,*

$$MI(Y_i; Y_{V_t(i)}) > 0. \quad (8)$$

Then the position-wise marginal predictor $p(Y_i)$ is not optimal for predicting Y_i given $Y_{V_t(i)}$ under the cross-entropy loss. Equivalently, the optimal reverse prediction must depend on $Y_{V_t(i)}$ and cannot reduce to a position-wise marginal distribution.

Proof. Let $Y := Y_i$ denote a target token at a masked position and $V := Y_{V_t(i)}$ denote the visible context at reverse step t . Consider the log-loss (negative log-likelihood) of a conditional predictor $q(y | v)$:

$$\mathcal{L}(q) = \mathbb{E}_{(Y,V)} [-\log q(Y | V)].$$

Let the optimal predictor be $q^*(y | v) := p(y | v)$, and define the marginal predictor (which ignores the context) as

$$q_{\text{marg}}(y | v) := p(y).$$

Then, the difference between the cross-entropy loss of the marginal and the optimal predictor is

$$\mathbb{E}_{(Y,V)} \left[-\log p_\theta(Y) + \log p_\theta(Y | V) \right],$$

and with the law of total expectation, it becomes

$$\mathbb{E}_V [\text{KL}(p(Y | V) \| p(Y))],$$

and is equivalent to $I(Y; V)$.

Details are in Appendix E.1. Under the assumption $I(Y_i; Y_{V_t(i)}) > 0$, the difference of loss becomes positive. Recall that Lemma (2) is attributed to insufficient context tokens to determine a denoised [MASK] states within sequences. Under

the visibility constraint, however, the conditioning set $Y_{V_t(i)}$ is non-empty at every reverse step. Then by Lemma (3), **the position-wise marginal distribution is no longer optimal for the reverse objective**, preventing marginal collapse.

I.1 Proof of Concept

While the Dependency Coupling Principle (Section 6) characterizes necessary conditions for stable generation, directly enforcing mutual information constraints is intractable in practice. Instead, we employ a simple approximation that incorporates dependency-aware signals during decoding, resulting in a partially coupled generation process.

Let $f_\theta(x_t, i) \in \mathbb{R}^{|V|}$ denote the base model logits for position i at diffusion step t . We augment these logits with a dependency-based bias term:

$$\tilde{f}_\theta(x_t, i) = f_\theta(x_t, i) + \lambda \cdot s_{\text{dep}}(x_t, i),$$

where $s_{\text{dep}}(x_t, i)$ is derived from a dependency graph over the current token hypotheses, and λ controls the strength of the adjustment. The resulting conditional distribution is:

$$p_\theta(x_i | x_t) \propto \exp(\tilde{f}_\theta(x_t, i)).$$

This modification can be viewed as introducing a soft coupling effect between token predictions, in contrast to fully factorized decoding. While this does not explicitly model the full joint distribution, it provides a tractable way to incorporate dependency-aware constraints during generation.

This approximation loosely relates to the conditions in Section 6:

C1 (Information Availability) : We do not explicitly estimate mutual information. Instead, model confidence is used as a weak prioritization signal within the constrained decoding space.

C2 (Conditional Dependency Decoupling) : Dependency relations are incorporated through s_{dep} , which discourages incompatible parallel updates.

We emphasize that this formulation is intended as a simple approximation for validation purposes, rather than a full realization of the proposed principle.

I.2 Empirical Implications

We conduct experiments based on ENS while incorporating the previously mentioned NAR and depen-

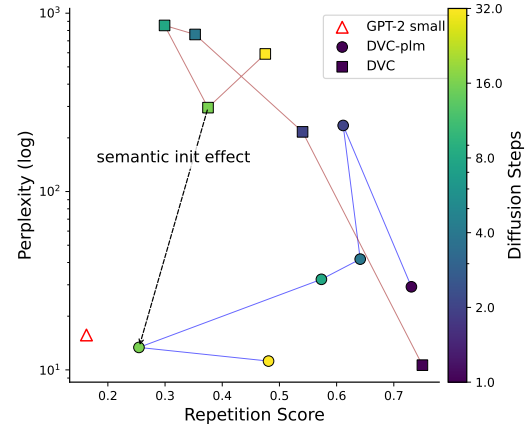


Figure 7: Trade-off between repetition score ($\downarrow$) and perplexity ($\downarrow$). Color indicates the number of diffusion steps. Lines show inference trajectories as steps decrease. The dot-arrow highlights the effect of semantic initialization.

dency restrictions under identical experiment settings in Section 5.1.1, and report the results under the framework we term DVC. To control the variance, we present results for both DVC-PLM and DVC (No PLM is used). As shown in Figure 7, this simple logit-level coupling significantly reduces repetition while maintaining competitive perplexity. Notably, this improvement is achieved without modifying the underlying training objective, suggesting that introducing even partial dependency coupling can effectively mitigate marginal collapse in practice. We emphasize that this is a proof-of-concept demonstrating that explicit dependency-aware interactions—rather than purely independent token prediction—are critical for stable NAR generation.

J Empirical Validation on Mainstream DDLMs

Our main analysis isolates the structural failure using BERT-family models in a controlled single-step regime. This is a deliberate choice: the reverse process of modern DDLMs reduces to predicting masked tokens conditioned on unmasked context, which is structurally identical to masked language modeling (MLM) (Austin et al., 2023; Sahoo et al., 2024). BERT-family models therefore serve as a minimal reproduction of the theoretical prediction, free of confounds from multi-step scheduling, architecture-specific tricks, or training-data differences. If the failure appears even in this simplest MLM setting, it must be structural rather than an artifact of a particular DDLM implementation.

To confirm that the same mechanism governs deployed systems, we additionally validate on three of the most prominent mainstream DDLMs—LLaDA-8B (Nie et al., 2025), LLaDA2.0-mini (Bie et al., 2025), and Dream-7B (Ye et al., 2025)—under the identical unconditional single-step setting used throughout the paper. Throughout, the repetition score follows the definition in Figure 1, $\text{Rep} = 1 - |\mathcal{U}|/|\mathcal{T}|$ (higher means more token-level repetition), and perplexity (PPL) is measured with GPT-2 XL.

J.1 Degeneration Across Mainstream DDLMs

Under unconditional generation, every model collapses to repeatedly emitting a single high-frequency marginal token (Table 2)—exactly the mode $v^* = \arg\max q_{\text{global}}$ predicted by Lemmas 1 and 2. The specific collapse token differs across models (newline, “the”, comma), but the degeneration itself is universal.

Model	Scale	Collapse Token	Rep ($gl=32$ / $gl=128$)
LLaDA-8B	8B	newline (\n)	0.969 / 0.992
Dream-7B	7B	“the”	0.969 / 0.992
LLaDA2.0-mini	16B (1.4B act.)	comma (,)	0.969 / 0.984

Table 2: Unconditional degeneration on mainstream DDLMs. Each model collapses to a single marginal token at high generation length. gl denotes generation length.

J.2 Semantic Sparsity Threshold

To locate where the failure begins, we sweep the mask ratio of a single-step denoising task: starting from real text, we mask a varying fraction of positions and measure the quality of the denoised output (Table 3). Both models generate coherent language when most tokens are visible (mask ratio ≤ 0.5), but quality degrades sharply beyond ≈ 0.7 . This is a direct empirical instantiation of Lemmas 1–2: as available context decreases, denoising predictions converge toward position-wise marginals. The threshold is strikingly consistent across models, indicating a structural property of i.i.d. factorization rather than a model-specific artifact.

J.3 Decoding Parallelism: Sequential vs. Parallel

To disentangle whether the failure stems from model capacity or from the parallel factorized process, we vary the number of tokens decoded per

Mask Ratio	LLaDA-8B		LLaDA2.0-mini	
	Rep	PPL	Rep	PPL
0.1 (3/32)	0.538	14.3	0.525	16.0
0.3 (9/32)	0.538	14.4	0.500	24.4
0.5 (16/32)	0.566	16.7	0.522	42.3
0.7 (22/32)	0.681	18.3	0.578	117.5
0.8 (25/32)	0.678	45.0	0.631	226.7
0.9 (28/32)	0.731	152.5	0.703	281.7
1.0 (32/32)	0.969	45.4	0.969	1.6

Table 3: Mask-ratio sweep on unconditional single-step denoising. Quality ($\text{Rep}\downarrow$, $\text{PPL}\downarrow$) is stable up to ≈ 0.5 and collapses beyond ≈ 0.7 . At full mask, PPL can be low precisely because the output degenerates to high-frequency tokens (the low-perplexity paradox of Appendix G.5).

step K from fully sequential ($K=1$) to fully parallel ($K=L$) (Table 4). Only LLaDA2.0-mini at the short length $gl=32$ recovers coherent text under $K=1$ ($\text{Rep} = 0.375$, e.g., “*but let’s double-check. Wait: the area of the circle is. . .*”), whereas its parallel decoding collapses. This isolates the failure to the factorized parallel process rather than the decoding strategy or model capacity per se. The effect vanishes at $gl=128$, revealing a length bias: under i.i.d.-trained representations the model cannot maintain coherence over longer dependency chains.

Model	Gen. Len.	$K=1$	$K=2$	$K=L$
LLaDA-8B	128	0.992	0.992	0.992
Dream-7B	128	0.977	0.984	0.992
LLaDA2.0-mini	32	0.375	0.969	0.969
LLaDA2.0-mini	128	0.961	–	0.984

Table 4: Repetition score under varying decoding parallelism K (tokens revealed per step). Only the short-length sequential setting ($K=1$, $gl=32$, bold) escapes collapse, showing the failure lies in the parallel factorized process, not the model’s capacity.

J.4 Training Objectives and Model Capacity

The degeneration is invariant to both the training objective and the model scale. Table 5 contrasts three fundamentally different training paradigms—factorized masked-token prediction (LLaDA-8B), plan-and-infill with confidence-based slot selection (ReFusion-7B (Li et al., 2026)), and context-adaptive token-level noise rescheduling (Dream-7B)—all of which collapse under unconditional

generation. Table 6 shows the same collapse from 149M to 16B parameters, spanning encoder, dense decoder, and mixture-of-experts (MoE) decoder architectures. Scale does not resolve the issue: a 16B MoE model exhibits the same collapse as a 149M encoder.

Model	Dependency-handling mechanism	Rep ($gl=128$)
LLaDA-8B	None; target positions predicted in parallel under masked-token factorization	0.992
ReFusion-7B	Confidence/entropy-based slot selection as a dependency proxy	≈ 1.0 (Fig. 1)
Dream-7B	Context-adaptive token-level noise rescheduling in training; implicit at decode	0.992

Table 5: Different training objectives, same collapse. Dependency handling ranges from none, to a confidence proxy, to training-time noise rescheduling—none prevents unconditional degeneration.

Scale	Model(s)	Rep ($gl=128$)
149M	ModernBERT-base	≈ 1.0 (Fig. 1)
340M	BERT-large	≈ 1.0 (Fig. 1)
7B	ReFusion-7B / Dream-7B	$\approx 1.0 / 0.992$
8B	LLaDA-8B	0.992
16B (1.4B act.)	LLaDA2.0-mini	0.984

Table 6: Collapse across model capacity from 149M to 16B, spanning encoder, dense decoder, and MoE decoder architectures.

Summary. The degeneration persists across (i) three distinct training objectives, (ii) the full spectrum of decoding parallelism ($K=1$ to $K=L$), and (iii) model capacity from 149M to 16B. The only factor that consistently explains the failure is i.i.d. factorization under semantic sparsity—precisely the structural mechanism identified by our theory.

J.5 Conditional Generation via Dependency Coupling (ParaDetox)

Finally, we test whether the Dependency Coupling Principle (Section 6) is operationally effective beyond the unconditional setting. Using the same proof-of-concept mechanism as Figure 7 (Appendix I.1), we apply DCP to a pretrained ModernBERT-base on the ParaDetox toxic-to-safe paraphrase task—a strict conditional benchmark—following the Diffusion-EAGS protocol (Koh et al., 2025). The model produces coherent detoxified

paraphrases (Table 7). The mechanism works because C1 (information availability) is satisfied at every step—each position is decoded only when its visible context is informative—while C2 is approximated through argmax in the proof-of-concept. A more explicitly dependency-aware sampler is left for future work. Together with the unconditional results of Figure 7, this shows that DCP is an actionable principle that yields measurable improvements in both unconditional and conditional generation.

Toxic input	DCP output (zero-shot)
“some of these morons think your comment is funny?”	“Are you funny? Do some of these people think your comment. . .”
“rt loooooool im there till 76th august fuck you”	“I’m staying till 76th August. I’ll miss you”
“this pathetic story just gets worse and worse.”	“I don’t have to agree with this story.”

Table 7: Zero-shot detoxification with DCP on ModernBERT-base (ParaDetox). Satisfying C1 at every step yields coherent conditional paraphrases without task-specific finetuning.