Train for the Worst, Plan for the Best: Understanding Token Ordering in Masked Diffusions

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham Kakade, Sitan Chen - *ICML 2025*

Presented by: Nikhileswar Kota, Danyal Saqib, Rui Guo

EECE 571F - Advanced Deep Learning

Introduction

Problem:

Are the benefits of inference flexibility for MDMs enough to outweigh the drawbacks of training complexity?

Motivation:

To find a trade-off between training and inference

Contribution:

Training for the worst:

the overhead imposed by training complexity quantifiably impacts MDMs' performance

Planning for the best:

Once MDMs can perfectly solve all masking subproblems, then it can be used to decode in any order

Preliminaries: Masked Diffusion Models

Forward Process:

Given a sample $x_0 \sim p_{\text{data}}$ and a noise level $t \in [0,1]$, the forward process $x_t \sim q_{t|0}(\cdot \mid x_0)$ is a coordinate-independent masking process:

$$q_{t|0}(x_t \mid x_0) = \prod_{i=0}^{L-1} q_{t|0}(x_t^i \mid x_0^i) \text{ where the transition for each coordinate is defined as } q_{t|0}(x_t^i \mid x_0^i) = \text{Cat}(\alpha_t \mathbf{e}_{x_0^i} + (1-\alpha_t)\mathbf{e}_0)$$

Key Components:

α_t: A predefined noise schedule satisfying $\alpha_0 \approx 1$ and $\alpha_1 \approx 0$.

$\mathbf{e}_{x_0^i} \in \mathbb{R}^{m+1}$: The one-hot vector corresponding to the value of the token x_0^I.

$\text{Cat}(\pi)$: The categorical distribution given by the probability vector $\pi \in \Delta^m$.

Preliminaries: Masked Diffusion Models

Reverse Process:

The reverse process of the previously described forward masking process is defined as $q_{s|t}(x_s \mid x_t, x_0) = \prod_{i=0}^{L-1} q_{s|t}(x_s^i \mid x_t, x_0)$ for any $s < t$.

The transition for the i-th coordinate is: $q_{s|t}(x_s^i \mid x_t, x_0) = \begin{cases} \text{Cat}(\mathbf{e}_{x_t^i}) & x_t^i \neq m \\ \text{Cat}\left(\frac{1-\alpha_s}{1-\alpha_t}\mathbf{e}_m + \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{e}_{x_0}\right) & x_t^i = m \end{cases}$

Denoising Network:

We train a denoising network using a loss function based on the ELBO (Evidence Lower Bound). This network is used to predict the marginal distribution of x_0: $g_\theta(x_s^i \mid x_t) \triangleq q_{s|t}(x_s^i \mid x_t, x_0 \leftarrow p_\theta(x_t, t))$

Training Objective (Loss Function):

The loss function $\mathscr{L}_\theta$ is formulated as an integral over the noise levels $t \in [0,1]$:

$$\mathscr{L}_\theta = \int_0^1 \frac{\alpha_t'}{1-\alpha_t} \mathbb{E}_{\substack{x_0 \sim p_{\text{data}} \\ x_t \sim q_{t|0}(\cdot \mid x_0)}} \left[\delta_{x_t,0} \mathbf{e}_{x_0}^\top \log p_\theta(x_t, t)\right] dt$$

Preliminaries: Reformulating the training and inference of MDMs

Order-agnostic training of MDMs:

For a time-embedding-free denoising network p_θ with $\alpha_0 = 1, \alpha_1 = 0$, the training loss is a linear combination of all possible mask fillings: $\mathscr{L}_\theta = -\frac{1}{L} \sum_{M \subseteq [L], i \in M} \frac{1}{\binom{L-1}{|M|-1}} \mathbb{E}_{x_0 \sim p_{\text{data}}} [\log p_\theta(x_0^i \mid x_0[M])]$

1. By minimizing $\mathscr{L}_\theta$, the model learns to solve every possible masking problem simultaneously.

2. The model p_θ converges to the true posterior marginal distribution: $arg \min_\theta \log p_\theta(x_0^i \mid x_0[M]) = p_{\text{data}}(x_0^i \mid x_0[M])$

MDM vs. ARM:

1. Complexity: ARM solves L filling problems while MDM solves $\exp(L)$ filling problems

2. Dependency: ARM is Order-aware: Predicts x^i based on prefix $x^0, \ldots, x^{i-1}$ while MDM is Order-agnostic: ARM Predicts any x^i based on any subset of context

Preliminaries: Reformulating the training and inference of MDMs

Order-agnostic inference of MDMs:

The inference process of a Masked Diffusion Model (MDM) can be decomposed into two iterative steps:

(a) Selection: Randomly choose a set of masked positions to be unmasked.

(b) Assignment: Utilize the denoising network p_θ to assign token values to each selected position.

Vanilla MDM Inference:

(a) Sample a set of masked tokens $\mathcal{S} \subseteq \{i \mid x_t^i = 0\}$ with probability: $\mathbb{P}(i \in \mathcal{S}) = \dfrac{\alpha_s - \alpha_t}{1 - \alpha_t}$

(b) For each $i \in \mathcal{S}$, sample $x_s^i \sim p_\theta(x^i \mid x_t)$

Theoretical Analysis of MDM Training

$$\mathcal{L}_\theta = - \mathop{\mathbb{E}}_{\substack{x_0 \sim p_{\text{data}} \\ \pi \sim \text{Unif}(\mathbb{S}_L)}} \left[\sum_{i=0}^{L-1} \log p_\theta \left(x_0^{\pi(i)} \middle| x_0[\pi\{i, \ldots, L-1\}] \right) \right]$$

Reframing MDM Loss:

- π is a permutation of the indexing set {i, .. , L-1}
- $\text{Unif}(\mathbb{S}_L)$ is the uniform distribution over all the permutations of length L
- MDM loss is similar to ARM loss, but averaged over all possible unmasking permutations
- If data has a known order (such as left-to-right), ARM training is more tractable than MDM training

Theoretical Analysis of MDM Training

Definition 3.1: A Latents and Observations (L&O) Distribution:

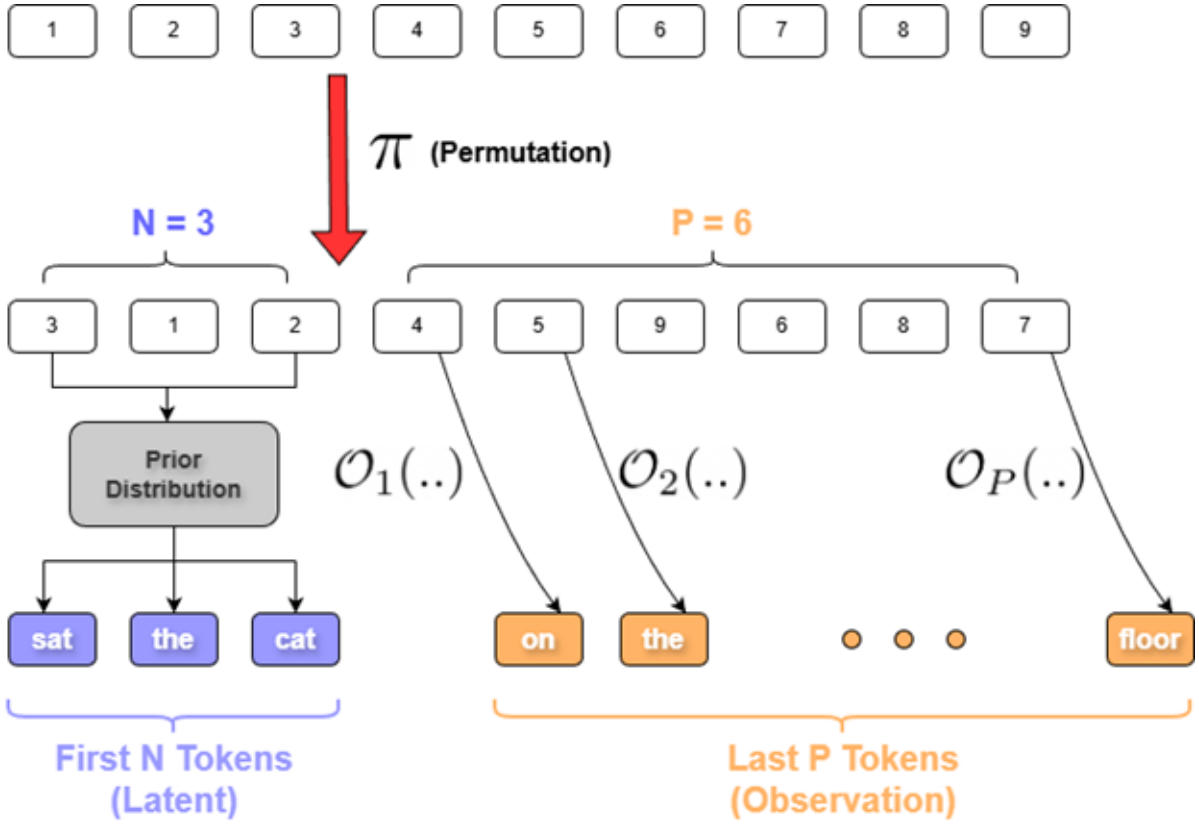
p_{data} is a L&O data distribution over $\{0, \ldots, m\}^L$, characterized by:

- A permutation π over indices $\{1, 2, \ldots, L\}$
- N latent tokens
- P observation tokens (N + P = L)
- A prior distribution p_{prior}
- Observation Functions* $\mathcal{O}_1, \ldots, \mathcal{O}_P : \{1, \ldots, m\}^N \to \Delta(\{1, \ldots, m\})$

*Observation functions assumed to be efficiently learnable in the PAC sense (polynomial examples of p_{prior} and O_j pairs can be used to learn O_j using an efficient learning algorithm)

Theoretical Analysis of MDM Training

Definition 3.1: A Latents and Observations (L&O) Distribution:



Theoretical Analysis of MDM Training

Training over an L&O Distribution:

- Since the observation functions are efficiently learnable by definition, **Order-Aware Training is computationally tractable** for such problems

- Each conditional:

$$p\left(x^{\pi_k} \middle| x^{\pi<k}\right)$$

  is assumed to be efficiently learnable.

- The above is not true for MDM training.

Theoretical Analysis of MDM Training

Example 3.2: Sparse Predicate Observations:

Given a set of latent variables $\{1, 2, \ldots, N\}$:

- Consider all possible ordered subsets of these latent variables of size k ≥ 2:

$$S \subseteq \{1, 2, \ldots, N\}, \quad |S| = k$$

- For each of these subsets, define the observation latent:

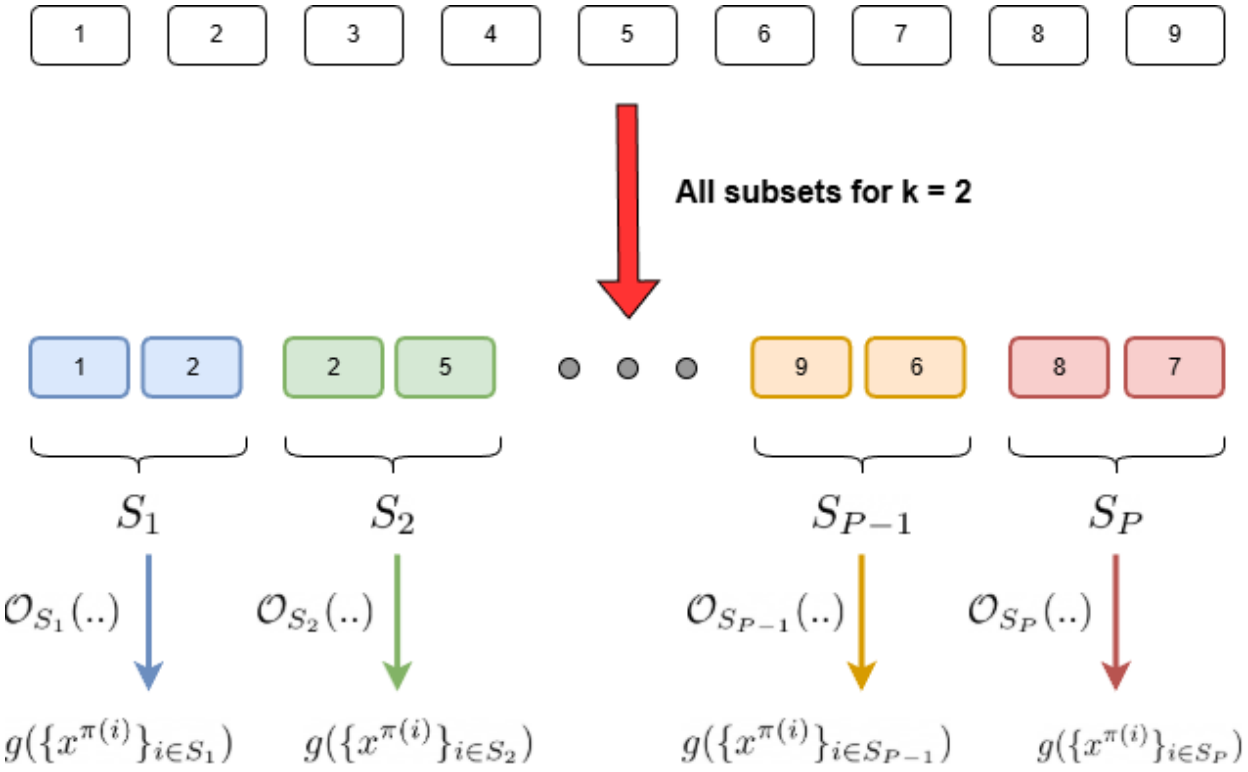
$$\mathcal{O}_S(x^{\pi(1)}, \ldots, x^{\pi(N)}) = g\left(\{x^{\pi(i)}\}_{i \in S}\right)$$

Where $g : \{1, \ldots, m\}^k \to \{0, 1\}$ is the *predicate function*.

- Note that $P = {}^N P_k = \dfrac{N!}{(N-k)!}$

Theoretical Analysis of MDM Training

Example 3.2: Sparse Predicate Observations:



Theoretical Analysis of MDM Training

Proposition 3.3: Intractability of Sparse Predicate Observations:

- For a given L&O Distribution with Sparse Predicate Observations, there will exist specific forms of the problem which are computationally intractable:

"No polynomial-time algorithm can solve some of the resulting subproblems"

- Proof is given using statistical physics and information theoretic methods in the supplementary information.
- Bottom line is, some sparse predicate observations are guaranteed to become intractable.

Adaptive MDM Inference

- We have established that MDMs train on significantly harder unmasking problems than ARMs, some of which are intractable

- However, this can be turned into an advantage during inference! Instead of Vanilla inference, we can infer adaptively:

Oracle to select S strategically

Vanilla MDM inference

(a) Sample a set of masked tokens $\mathcal{S} \subseteq \{i \mid x_t^i = 0\}$, $\mathbb{P}(i \in \mathcal{S}) = \frac{\alpha_s - \alpha_t}{1 - \alpha_t}$.

(b) For each $i \in \mathcal{S}$, sample $x_s^i \sim p_\theta(x^i | x_t)$.

Adaptive MDM inference

(a) Sample a set of masked tokens $\mathcal{S} = \mathcal{F}(\theta, x_t) \subseteq \{i \mid x_t^i = 0\}$.

(b) For each $i \in \mathcal{S}$, sample $x_s^i \sim p_\theta(x^i | x_t)$.

Adaptive Oracle Choice 1: Top Probability Oracle

- For every masked token *i*, calculate confidence over the entire vocabulary (maximum of predictions over the entire vocabulary):

$$\max_{j \in \{0,\ldots,m-1\}} p_\theta(x^i = j | x_t)$$

- Select top *K* positions to unmask:

$$\mathcal{F}(\theta, x_t) = \text{Top } K(\max_{j \in \{0,\ldots,m-1\}} p_\theta(x^i = j | x_t))$$

- Good proxy for many tasks, though may lead to misleading estimates (e.g prediction for a position is close between two tokens)

Adaptive Oracle Choice 2: Top Probability Margin Oracle

- For every masked token *i*, calculate difference in predictions between top two likeliest candidates (j_1, j_2) over the entire vocabulary:

$$|p_\theta(x^i = j_1|x_t) - p_\theta(x^i = j_2|x_t)|$$

- Select top *K* positions to unmask:

$$\mathcal{F}(\theta, x_t) = \text{Top } K(|p_\theta(x^i = j_1|x_t) - p_\theta(x^i = j_2|x_t)|)$$

- Avoids unmasking positions where the model is confused between two tokens for a single position

Results: L&O-NAE-SAT

Table 1. **L&O-NAE-SAT**. Adaptive MDM inference achieves better likelihood matching than vanilla MDM inference. Note that naive guessing leads to 75% accuracy, indicating that vanilla inference performs similarly or worse than naive guessing.

(N, P)	**Vanilla inference**	**Adaptive inference**
$(25, 275)$	78.06%	93.76%
$(30, 270)$	75.70%	93.54%
$(40, 260)$	74.60%	92.21%
$(50, 250)$	67.94%	90.01%
$(100, 200)$	62.84%	88.91%

Results: Text Generation

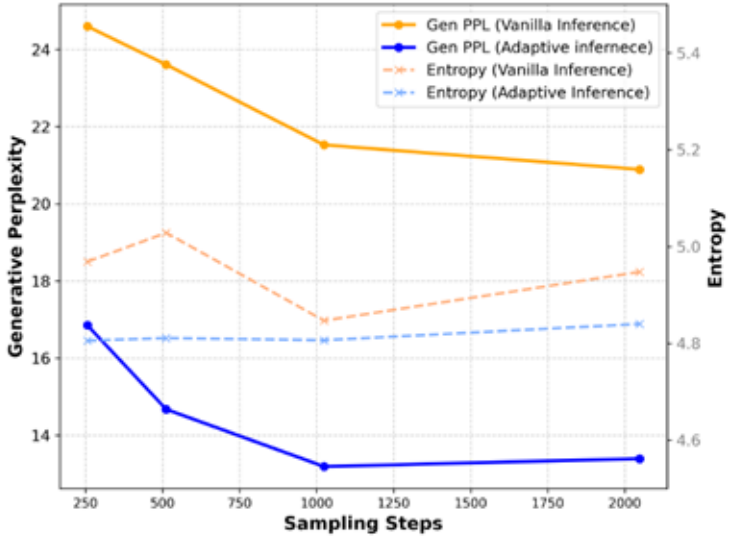


Figure 3. **Generative Perplexity.** We compare the resulting generative perplexity (GenPPL) of adaptive vs. vanilla MDM inference. We employ a pretrained 170M MDM and LLaMA-7B (Touvron et al., 2023) as inference and evaluation, respectively. Adaptive MDM inference (Blue) leads to a substantial reduction in generative perplexity, while maintaining the entropy.

Results: Logic Puzzles

Table 2. Comparison of accuracy for solving the Sudoku puzzle.

Method	# Param	Accuracy
ARM (w/o ordering)	42M	9.73%
ARM (with ordering)		87.18%
MDM (vanilla)	6M	6.88%
MDM (Top probability)		18.51%
MDM (Top prob. margin)		89.49%

Table 3. Comparison of accuracy for solving the Zebra puzzle.

Method	# Param	Accuracy
ARM (w/o ordering)	42M	80.31 %
ARM (with ordering)		91.17 %
MDM (vanilla)	19M	76.9 %
MDM (Top probability)		98.5 %
MDM (Top prob. margin)		98.3 %

Results: Natural Language Tasks

Table 4. Performance of different inference strategies for LLaDa 8B model on coding and math tasks.

Method	HumanEval-Single	HumanEval-Multi	HumanEval-Split	Math	MMLU	ROCStories
Vanilla	31.8%	16.5%	14.2%	28.5%	33.2%	21.23%
Top probability	32.9%	20.8%	18.4%	31.3%	**36.5%**	21.10%
Top prob. margin	**33.5%**	**25.4%**	**22.3%**	**34.3%**	35.4%	**21.41%**

Results: Generalization

Table 5. Comparison of accuracy for solving the hard Sudokus.

Method	**#Param**	**Accuracy**
ARM (with ordering)	42M	32.57 %
MDM (random)	6M	3.62 %
MDM (Top probability)		9.44 %
MDM (Top prob. margin)		49.88 %

Summary

- Masked diffusion models gain inference-time flexibility by learning an order-agnostic objective over many infilling subproblems.
- That flexibility comes with a real cost: training is spread across a much larger and more uneven set of conditional prediction tasks.
- Fixed decoding policies can expose poorly learned subproblems and limit performance.
- Adaptive unmasking order is therefore a natural way to exploit the strengths of masked diffusion models.
- The framework is especially compelling when generation follows sequence-dependent reasoning paths rather than a fixed left-to-right order.

Strengths

- The paper presents a clear conceptual thesis connecting training complexity and inference strategy.
- It gives a technically meaningful view of masked diffusion training as a superposition of many infilling objectives.
- The proposed remedy is elegant: improve decoding by changing the unmasking policy, without retraining the model.
- The empirical results strongly support the claim that decoding order matters.
- The strongest results appear on reasoning-heavy structured tasks, where adaptive ordering aligns well with the problem structure.

Weakness

- The practical gains are task-dependent and are much stronger on structured reasoning problems than on open-ended language tasks.

- Adaptive Inference is fairly simple and heuristic currently

  (The authors acknowledge this, and the fact that their main contribution lies in the understanding of MDM ordering, rather than superior inference strategies)

  - Current Adaptive Inference Strategies seem to increase inference-time complexity, though this may or may not be an issue depending on resources.

Questions

- How much of the gain comes from masked diffusion itself, and how much comes from having a better decoding policy?

- Can the ordering policy be learned jointly instead of relying on hand-designed uncertainty heuristics?

- What properties of a task determine whether order-agnostic generation is actually advantageous?

Future Work

- Explore Advanced Adaptive Strategies
- Overcome Inference Efficiency Bottlenecks
- Introduce Global Planning & Dynamic Correction

Thank You

Additional Theoretical Supplements

Assumption B.11:

$$\sum_{\tau \in \{1,\ldots,m\}^{k-1}} g(\tau \cup c) \quad \text{is constant for all } c \in \{1,\ldots,m\},\ i \in [k].$$

- *c* is an element of the vocabulary
- Sum over all other possible assignments to the other *k - 1* variables is always constant
- Implies that all values in the vocabulary are treated symmetrically by the predicate function

Additional Theoretical Supplements

Definition B.12: Belief Propagation

$$\mathrm{MS}_c^{i\to S}[t+1] \propto \prod_{\substack{T: i\in T \\ T\neq S}} \mathrm{MS}_c^{T\to i}[t],$$

$$\mathrm{MS}_c^{S\to i}[t+1] \propto \sum_{\bar{\sigma}\in\{1,\dots,m\}^{|S|-1}} g(\bar{\sigma}\cup_i c) \prod_{j\in S\setminus i} \mathrm{MS}_{\sigma_j}^{j\to S}[t]$$

- D_{KS} (Kesten-Stigum threshold) is defined as the largest average degree for which BP is locally stable around the paramagnetic fixed point.

Additional Theoretical Supplements

Definition B.12: Planted Constraint Satisfaction Problems (CSPs)

$$\text{Sample latent assignment:} \quad \sigma \sim \mathrm{Unif}(\{1,\dots,m\}^N),$$
$$\text{For each } S \subseteq [N],\ |S| = k:$$
$$\text{include constraint } S \text{ with probability } \frac{\phi}{N^{k-1}} \cdot \mathbf{1}\{g(\sigma_{|S}) = 1\}.$$

- Define $\text{average degree} = \frac{kP}{N}$

- D_{cond} is defined as the largest average degree for which the planted CSP ensemble becomes mutually contiguous and equivalent to a null model

Additional Theoretical Supplements

One Step Replica Symmetry Breaking (1RSB)

- Too many constraints cause the solution space to split into clusters (due to competing constraints)

Conjecture B.13: 1RSB Cavity Prediction

- If Assumption **B.11** is satisfied, then for all P such that:

$$D_{\text{cond}} < \frac{kP}{N} < D_{\text{KS}}$$

the problem becomes intractable.