Remasking Discrete Diffusion Models with Inference-Time Scaling

Guanghan Wang, Yair Schiff

April 16th, 2025

ASAP Seminar

said

Effective discrete diffusion models (MDLM)

Improved sampling methods (ReMDM)

said

Effective discrete diffusion models (MDLM)

Improved sampling methods (ReMDM)

$$p_\theta(x)$$

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point.

Language Models

Language Models

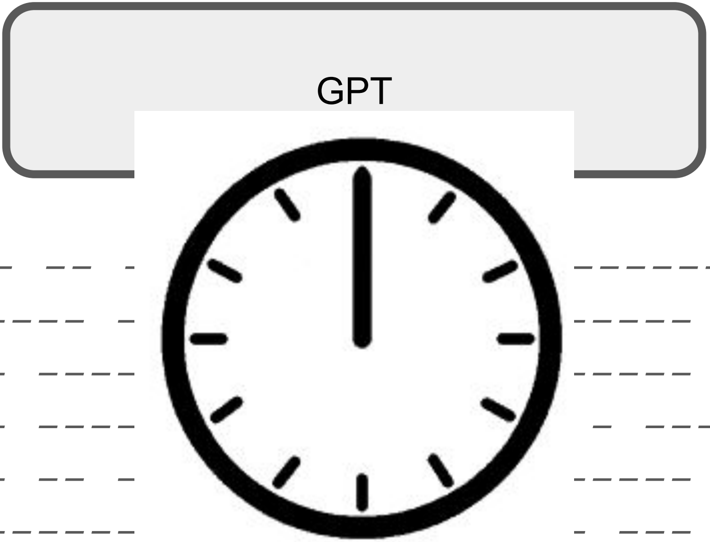
Many ----- ------ -- -- ----- --- ------ ------ ------- --------- ------- --- -- -------- ---- ------- --------- ---- --- ------ ---- --- -- -------- ---- -- ---- ---- ------- --- - ------- -- ------ ----- ------- ----- -- --- ---- -- - ----- -- ----- ----- ---- --- ----- - --- -- -------- ------- ----- ---- ----- --- --------- ---- ----------- ----- --- ----- --- -- ------ ---- ---- ------ ------ ------ --- -- ----- -- -------- ---- -- --- --------- -- ------

Language Models

GPT

```
Many years ------ -- -- ----- --- ------ ------ ------- ---------
------- --- -- -------- ---- ------- --------- ---- --- ------
---- --- -- -------- ---- -- ---- ---- ------- --- - ------- --
------ ----- ------- ----- -- --- ---- -- - ----- -- ----- -----
---- --- ----- - --- -- -------- ------- ----- ---- ----- ---
--------- ---- ----------- ----- --- ----- --- -- ------ ---- ----
------ ------ ------ --- -- ----- -- -------- ---- -- ---
--------- -- ------
```

Language Models

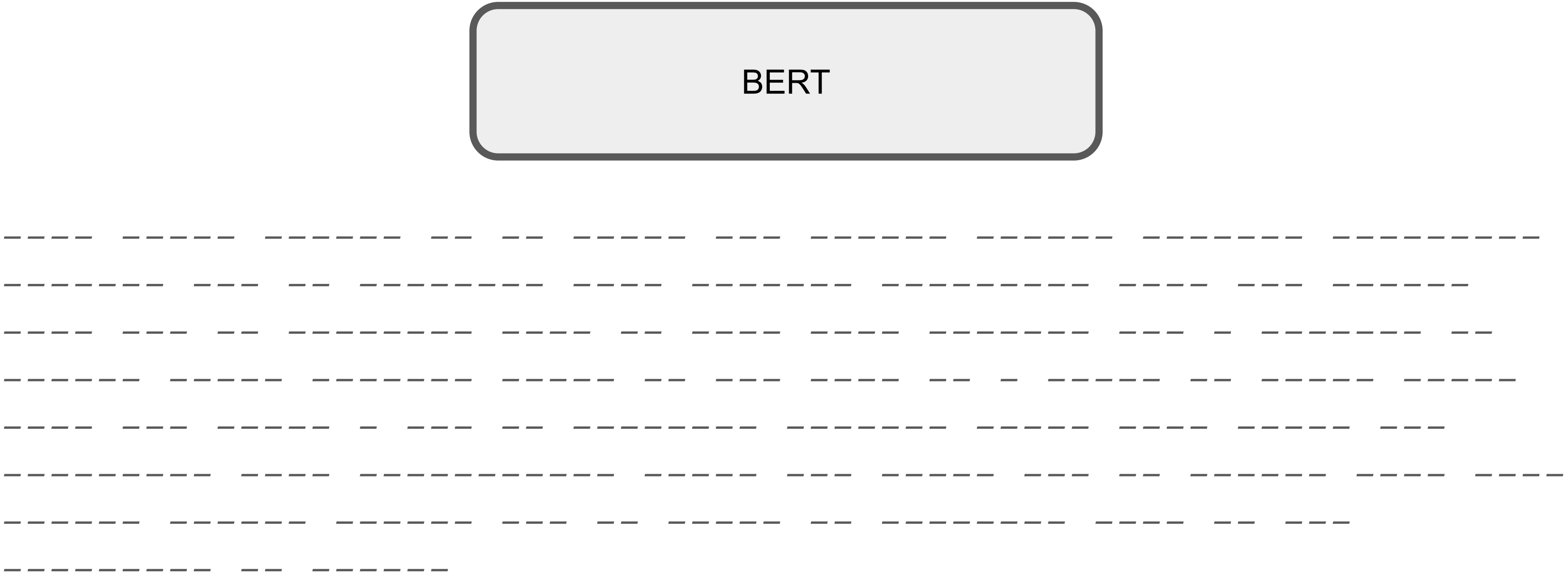


Many years later

Language Models

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point.

Masked Language Models



BERT

---- ----- ------ -- -- ----- --- ------ *squad,* ------- --------- ------- --- *to remember* ---- ------- --------- *when* --- ------ ---- --- *to* -------- ---- -- ---- ---- ------- --- - ------- -- *twenty adobe* ------- ----- -- --- ---- -- - ----- *of* ----- *water* ---- --- *along* - --- -- *polished* ------- *which* ---- ----- --- --------- ---- *prehistoric* ----- --- ----- --- -- ------ ---- *many* ------ ------ ------ --- *in* ----- -- -------- ---- -- --- --------- -- ------

BERT

Many years later, as he faced --- ------ squad, ------- --------- Buendía was to remember that distant --------- when his ------ took --- to discover ---- At that ---- ------- was a village of twenty adobe houses, built -- the bank of - river of clear water that ran along - --- of polished stones, which were ----- and --------- like prehistoric eggs. --- ----- --- so recent ---- many things lacked names, and in ----- -- -------- them it was --------- to ------

BERT

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant --------- when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of - river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world --- so recent that many things lacked names, and in order -- indicate them it was --------- to point.

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point.

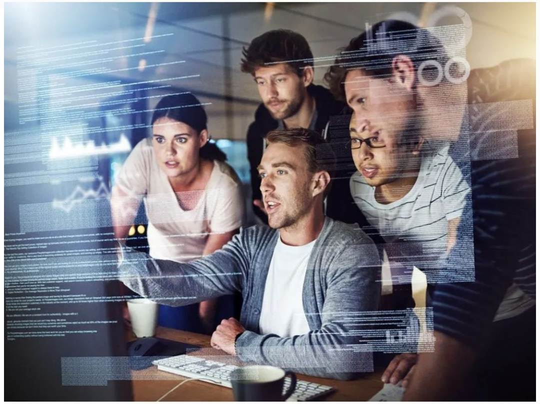
Forbes

INNOVATION > AI

Generative AI Gets Shaken Up By Newly Announced Text-Producing Diffusion LLMs

By Lance Eliot , Contributor. Dr. Lance B. Eliot is a world-renowned AI scientist... Follow Author

Mar 07, 2025, 10:19pm EST

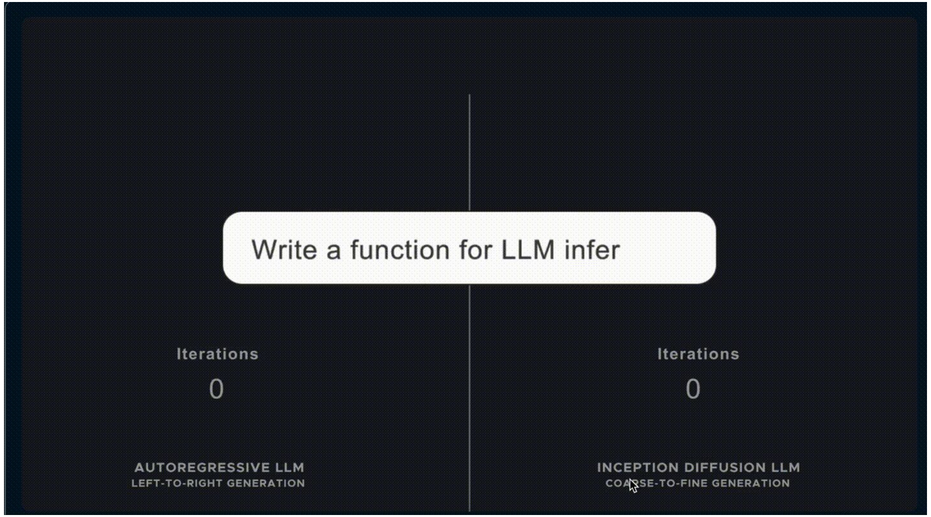


Diffusion LLMs are an exciting innovation that could shake up conventional generative AI and cause ... [+] GETTY

In today's column, I explore the exciting news that an alternative method to generative AI and large language models (LLMs) appears to be gaining interest and potentially provides some distinct advantages to conventional approaches. Here's the deal in a nutshell. The usual path to devising generative AI consists of what is known as autoregressive LLMs, while the promising new avenue is referred to as diffusion LLMs (dLLMs).

Yes, indeed, dLLMs just might be a winner-winner chicken dinner. I will share with you how

inception



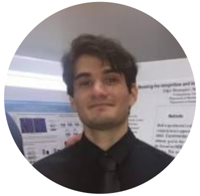
Simple and Effective Masked Diffusion Language Models

Subham Sahoo

Marianne Arriola

Yair Schiff

Aaron Gokaslan



Edgar Marroquin

Justin Chiu

Alexander Rush

Volodymyr Kuleshov

Diffusion Background

Notation:

Signal / “Clean” data $\mathbf{x}$

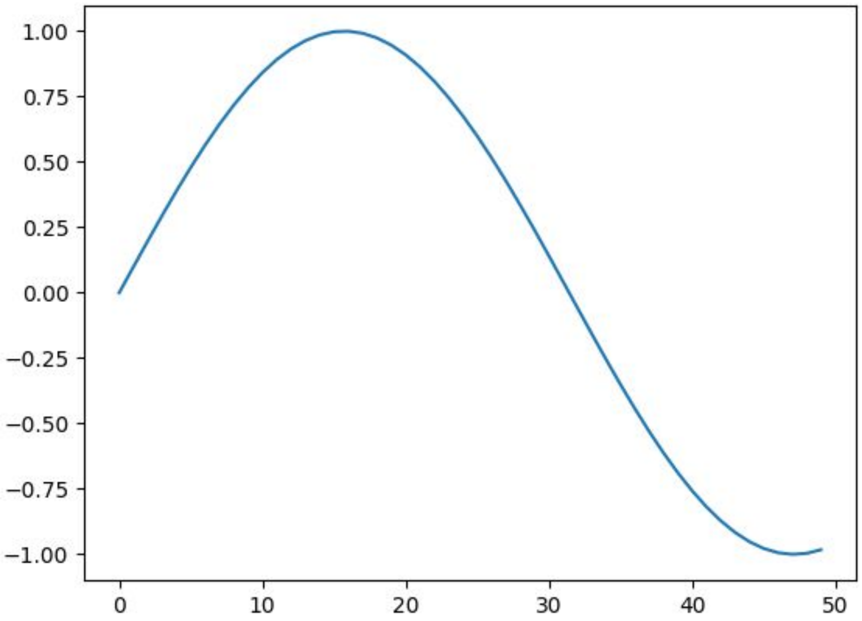
Latent variables / Noisy data $\mathbf{z}_t$

Diffusion timesteps $s, t \in [0, 1], s < t$

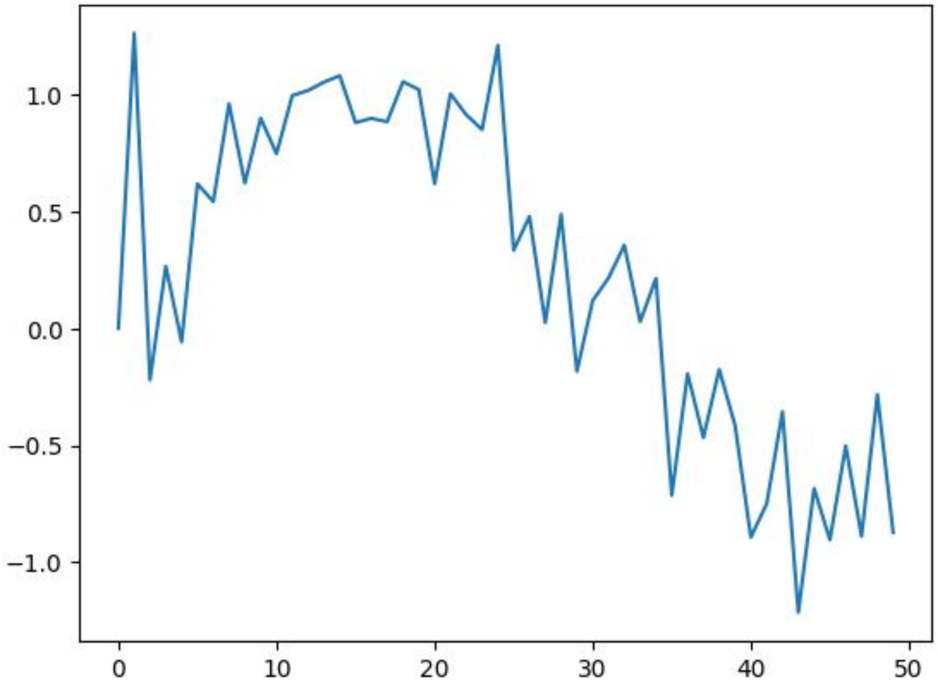
Forward / Noising process (fixed) q

Reverse / Denoising process (learned) p_θ

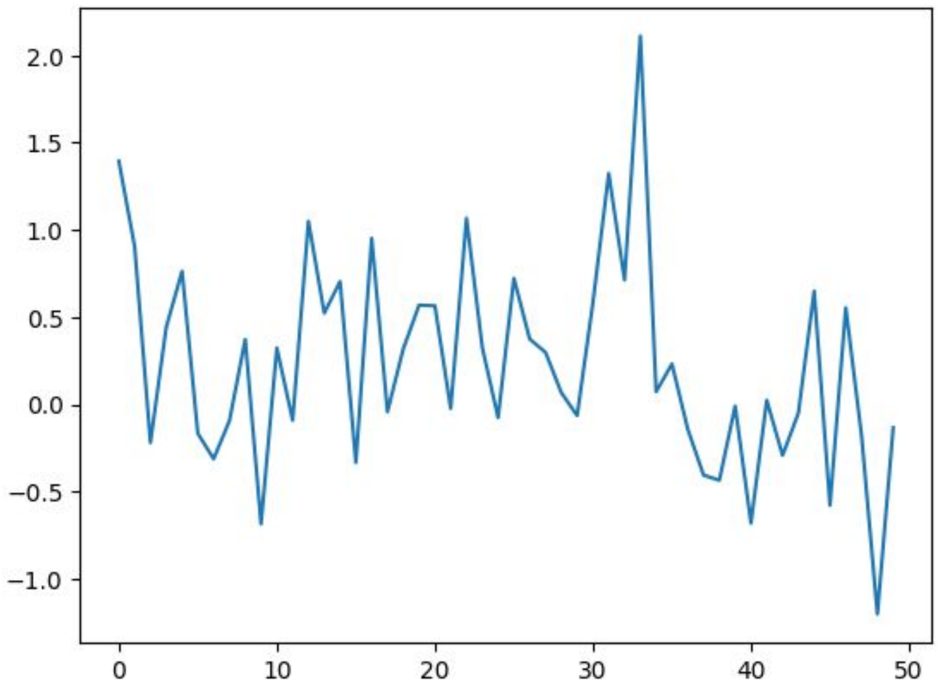
X



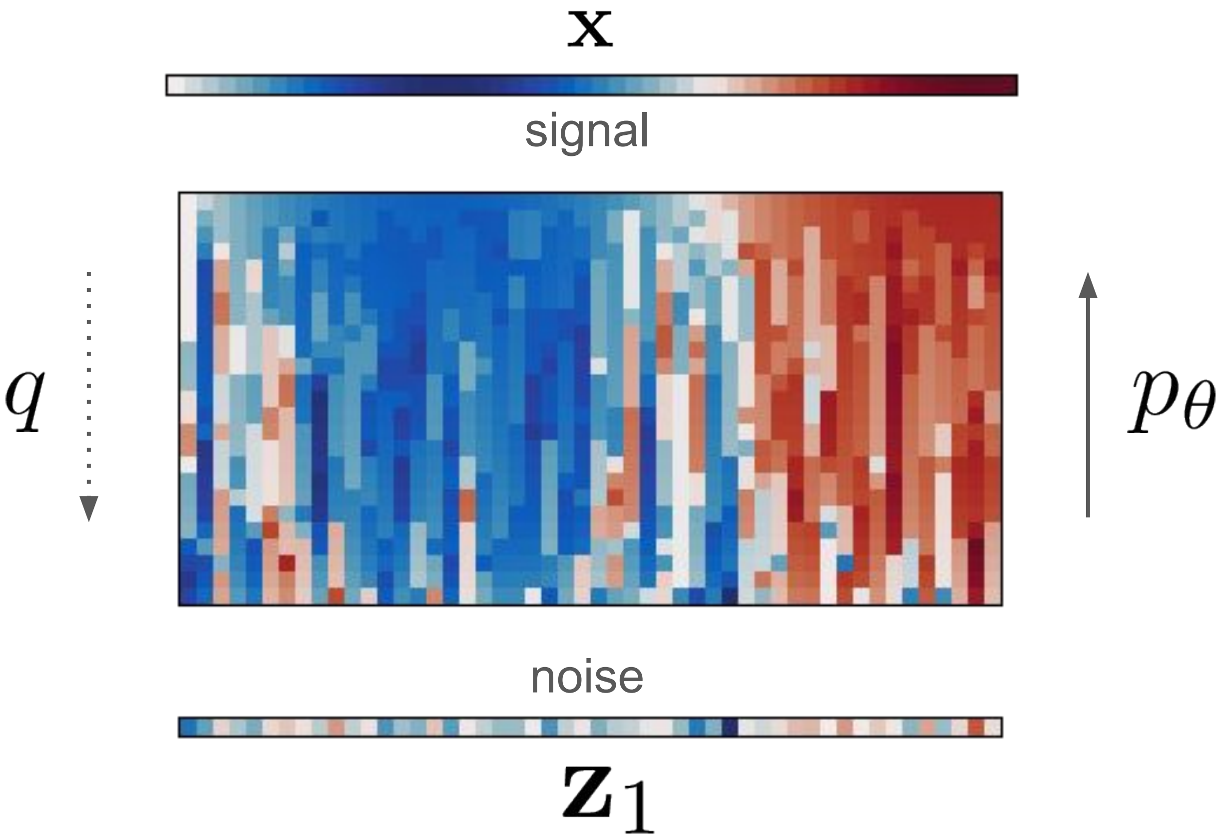
$$\mathbf{z}_t$$



$$\mathbf{z}_1$$



Continuous Diffusion



Use variational lower bound

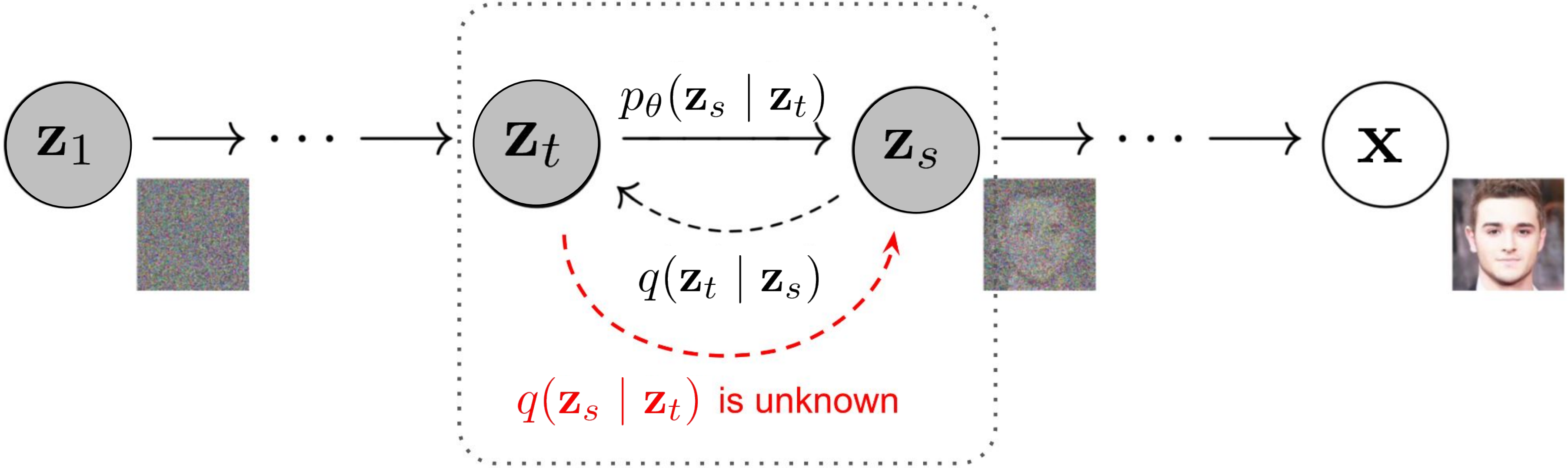
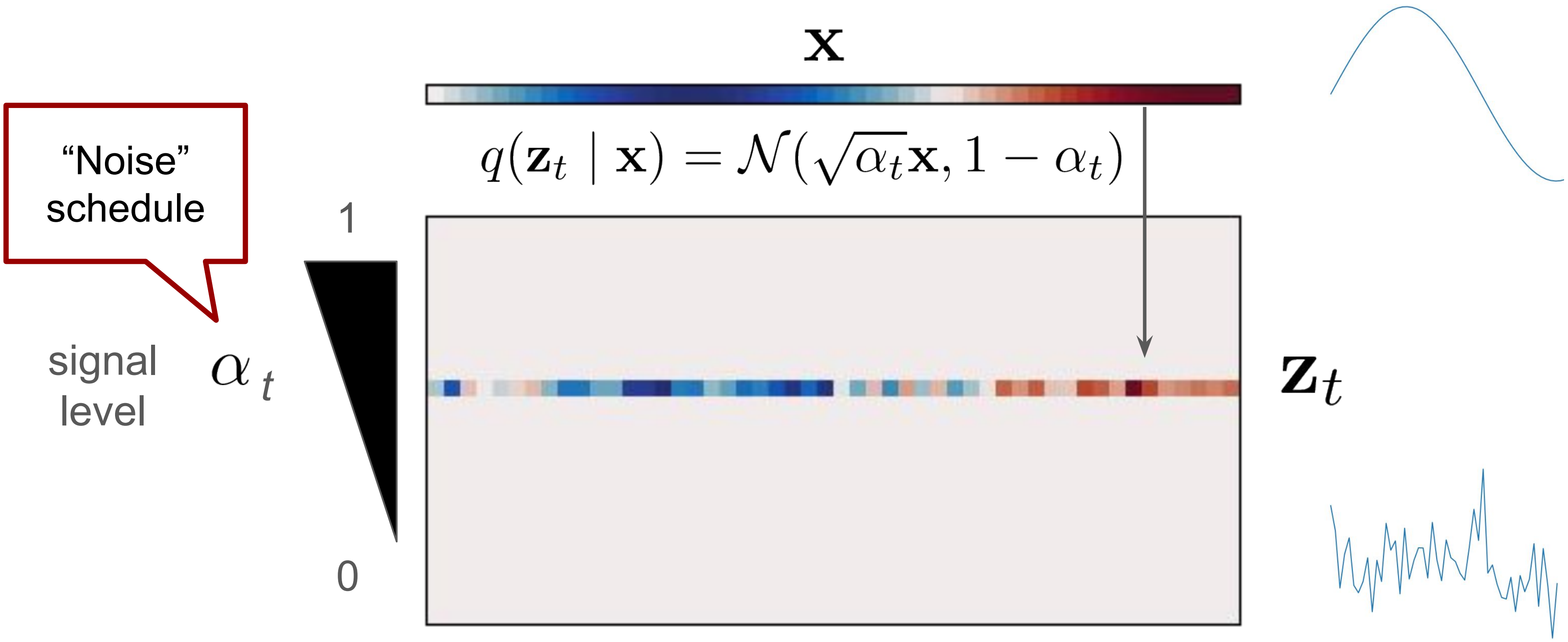


Image credit: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/, https://arxiv.org/abs/2006.11239

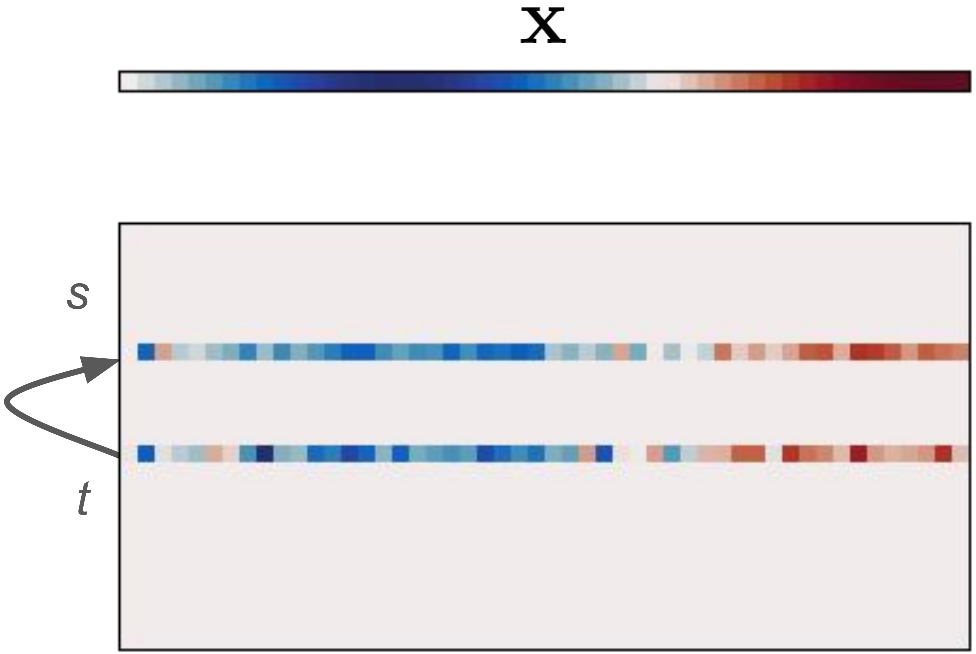
Diffusion Variational Objective

$$\mathbb{E}_q\left[\underbrace{-\log p_\theta(\mathbf{x}|\mathbf{z}_{t(0)})}_{\mathcal{L}_{\text{recons}}} + \underbrace{\sum_{i=1}^{T} \mathrm{KL}[q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x}) \| p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}))]}_{\mathcal{L}_{\text{diffusion}}}\right] + \underbrace{\mathrm{KL}[q(\mathbf{z}_{t(T)}|\mathbf{x}) \| p_\theta(\mathbf{z}_{t(T)}))]}_{\mathcal{L}_{\text{prior}}}$$

Gaussian Forward Process



Gaussian Forward Implies Gaussian Reverse



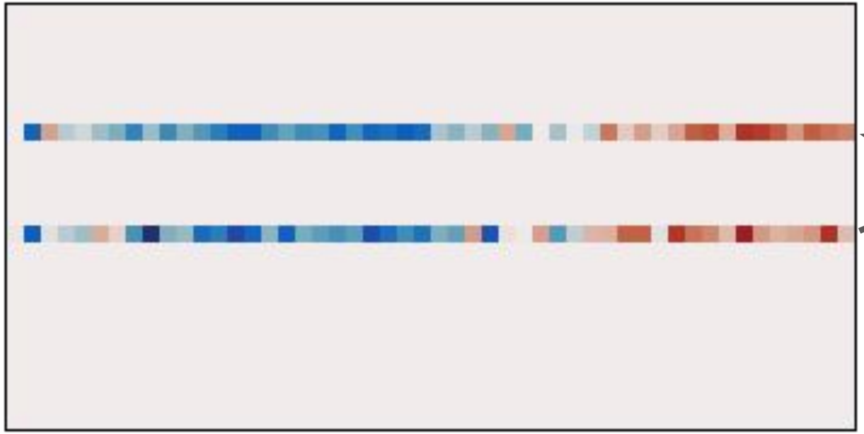
$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x})$$

Gaussian

Reverse Prediction Problem

$\mathbf{x}$

???



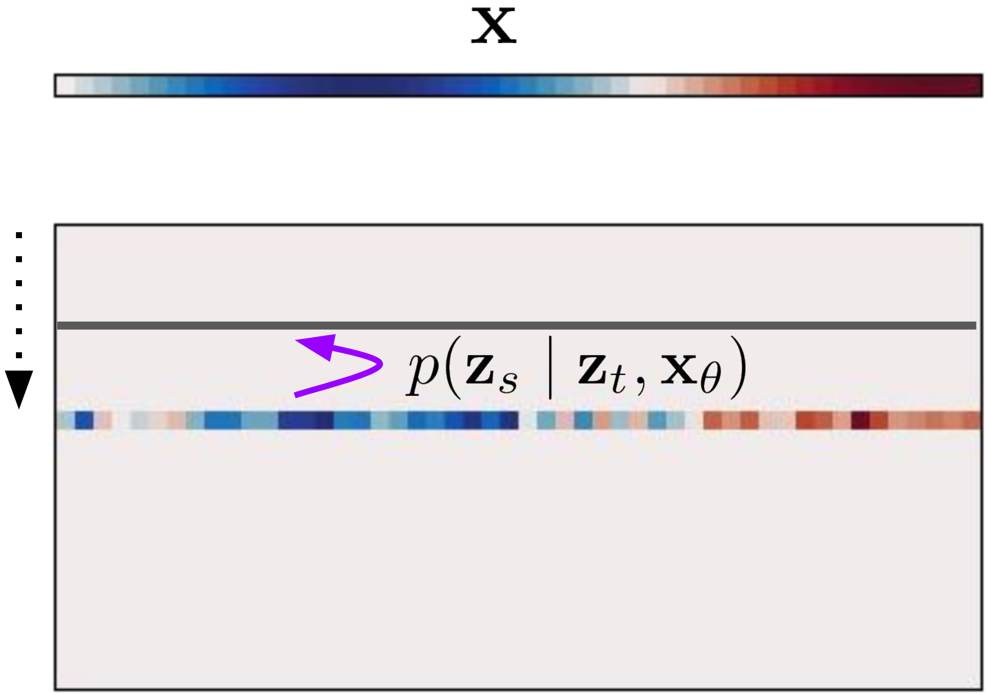
$p(\mathbf{z}_s \mid \mathbf{z}_t, \boxed{\mathbf{x}_\theta})$

Gaussian

Learning to Denoise

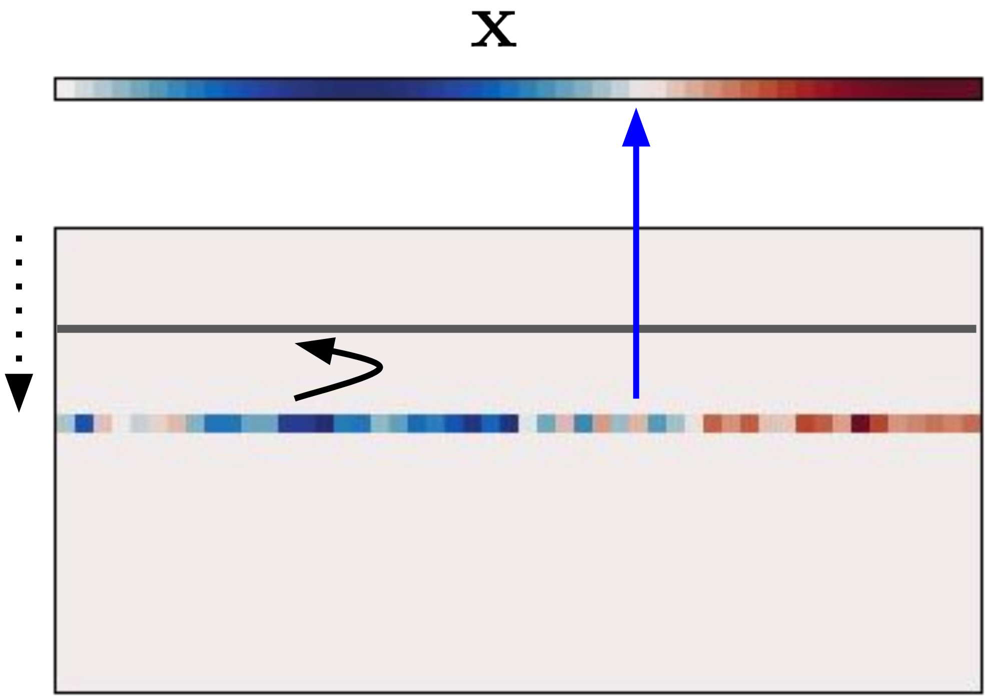
1) Sample t & Noise acc. to $q(\mathbf{z}_t \mid \mathbf{x})$

Learning to Denoise



1) Sample & noise
2) Denoise and weigh by contribution of $\mathbf{x}_\theta$ to posterior

Learning to Denoise



1) Sample & noise
2) Denoise and weigh
3) Score our reconstruction

Simple Discrete Masking Diffusion

Notation:

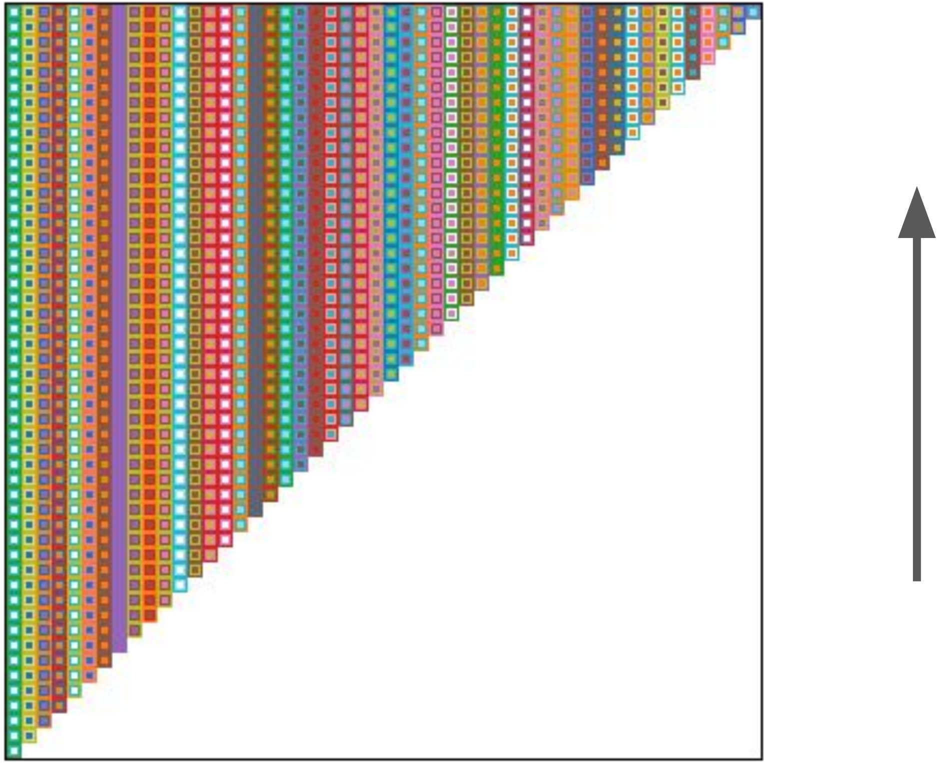
Vocabulary $\mathcal{V}$

One-hot representations $\mathbf{x}, \mathbf{z}_t \in \{0,1\}^{|\mathcal{V}|} \subset \Delta^{|\mathcal{V}|}$

Special "[MASK]" one-hot $\mathbf{m}$

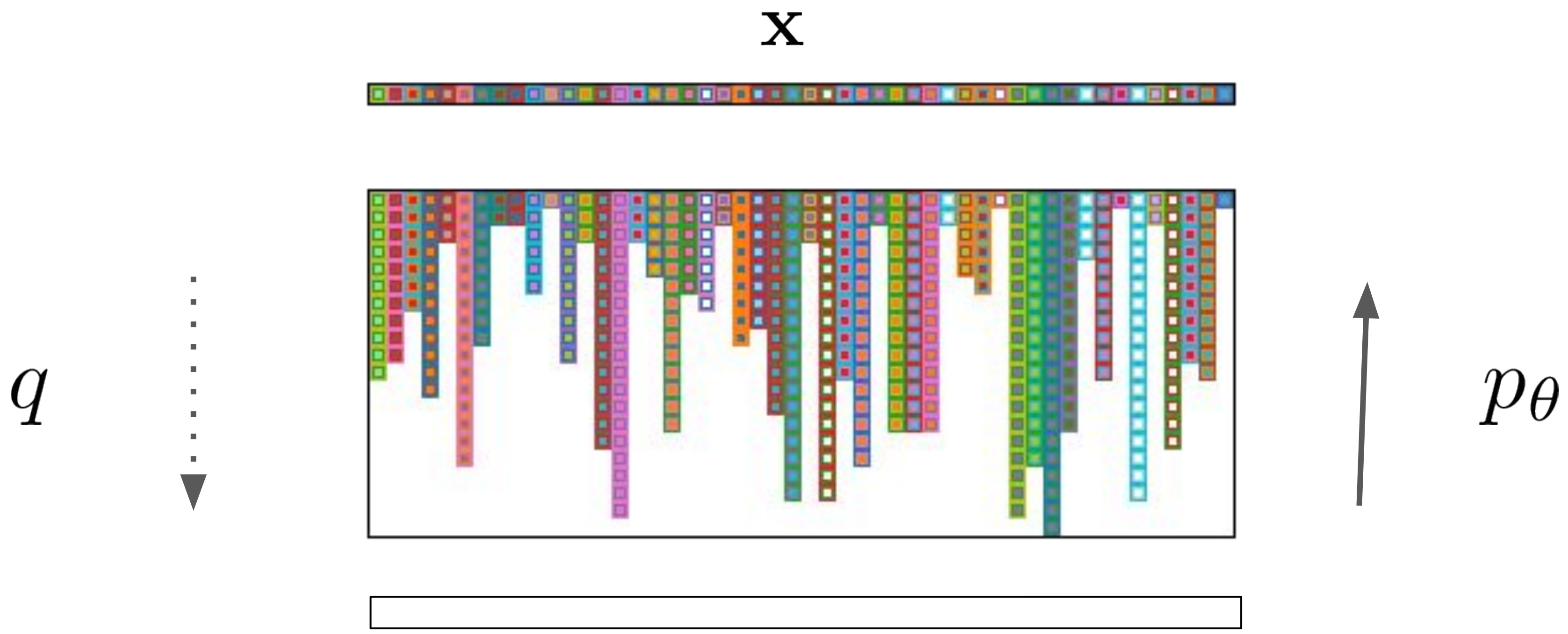
Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point.

Standard Model: Autoregressive Unmasking



p_θ

Our Goal: Discrete Masking Diffusion



Our Goal: Discrete Masking Diffusion

$$q(\mathbf{z}_t \mid \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_t; \alpha_t \mathbf{x} + (1 - \alpha_t)\mathbf{m})$$

Our Goal: Discrete Masking Diffusion

$$q(\mathbf{z}_t \mid \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_t; \alpha_t \mathbf{x} + (1 - \alpha_t)\mathbf{m})$$

Marginal over latent variable at time t

Our Goal: Discrete Masking Diffusion

$$q(\mathbf{z}_t \mid \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_t; \alpha_t \mathbf{x} + (1 - \alpha_t)\mathbf{m})$$

With probability α_t token remains unchanged and with probability $1 - \alpha_t$ it transitions to mask

Our Goal: Discrete Masking Diffusion

$$q(\mathbf{z}_t \mid \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_t; \alpha_t \mathbf{x} + (1 - \alpha_t)\mathbf{m})$$

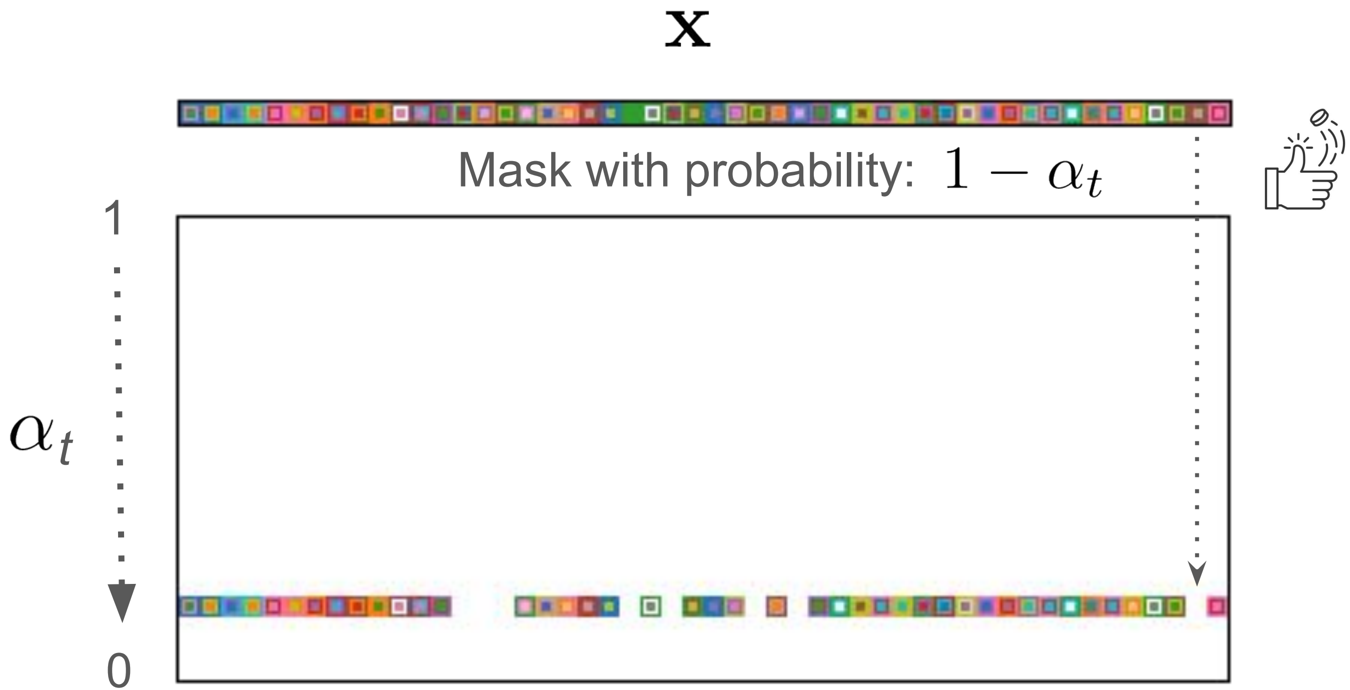
α_t is monotonically decreasing from 1 to 0

Our Goal: Discrete Masking Diffusion

$$q(\mathbf{z}_t \mid \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_t; \alpha_t \mathbf{x} + (1 - \alpha_t)\mathbf{m})$$

$$q(\mathbf{z}_t \mid \mathbf{x}) = \mathcal{N}(\mathbf{z}_t; \sqrt{\alpha_t}\mathbf{x}, (1 - \alpha_t)\mathbf{I})$$

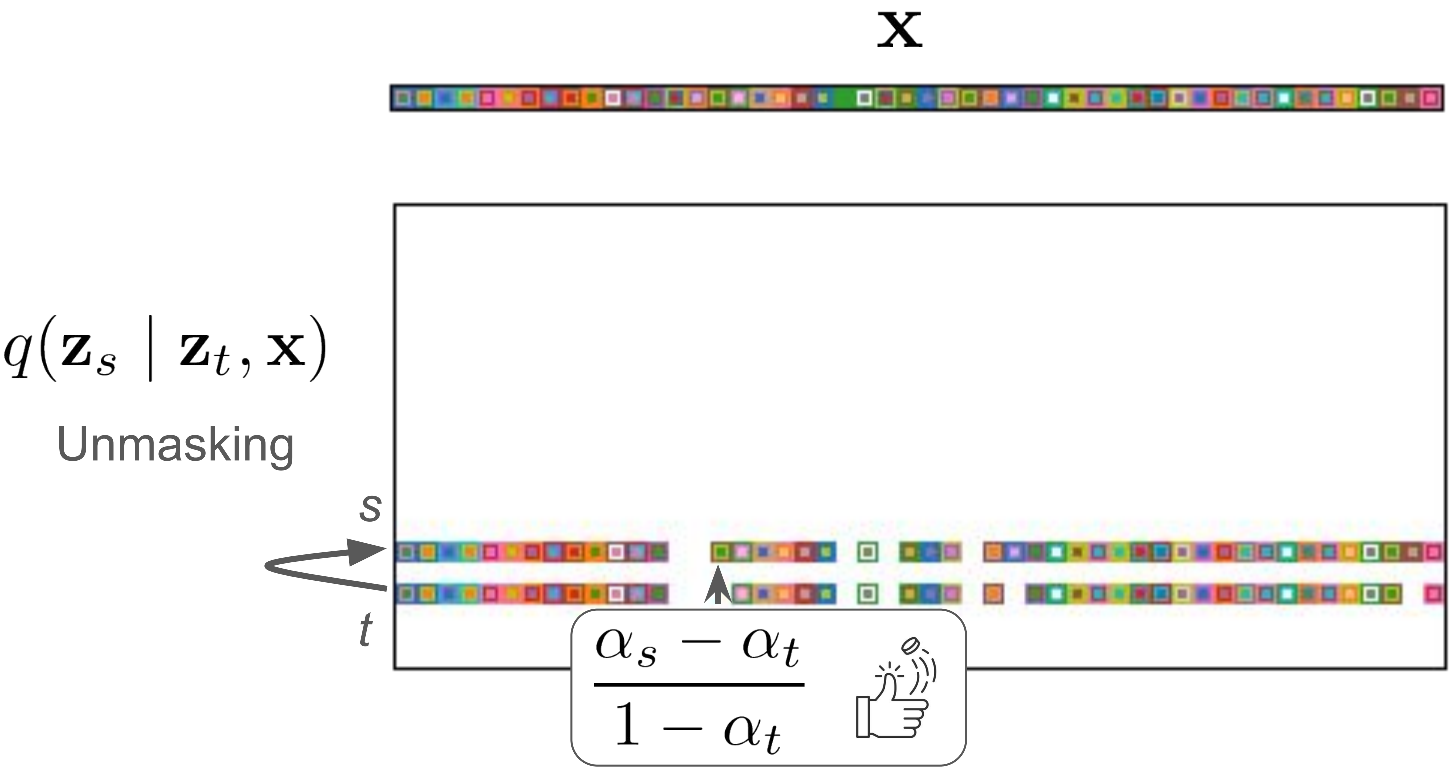
Masking Noise



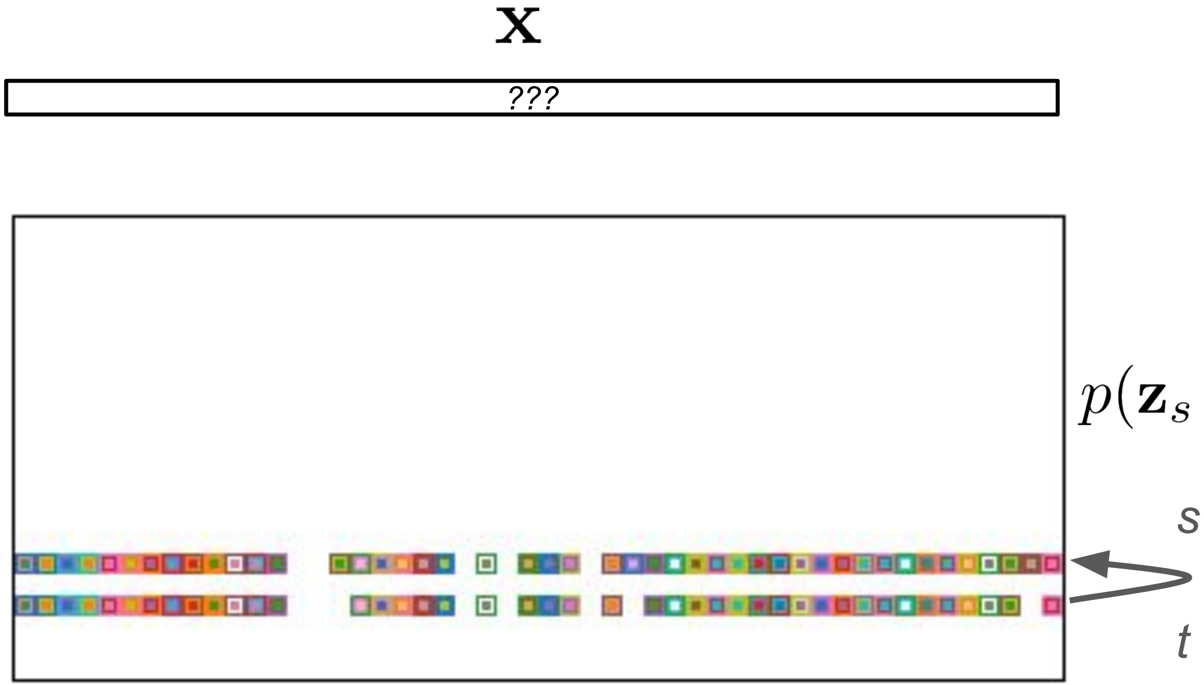
Masking Forward Implies Unmasking Reverse (posterior)

$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}(\mathbf{z}_s; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}), & \mathbf{z_t} = \mathbf{m} \end{cases}$$

Masking Forward Implies Unmasking Reverse



Reverse Prediction Problem



$$p(\mathbf{z}_s \mid \mathbf{z}_t, \boxed{\mathbf{x}_\theta})$$

Learned model should “respect” the diffusion process

$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}_s; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z_t} = \mathbf{m} \end{cases}$$

$$\mathbf{x}_\theta \quad \text{s.t.} \quad \begin{cases} \mathbf{x}_\theta^\top \mathbf{z}_t = 1, & \mathbf{z}_t \neq \mathbf{m} \\ \mathbf{x}_\theta^\top \mathbf{z}_t = 0, & \mathbf{z}_t = \mathbf{m} \end{cases}$$

Masked Diffusion Variational Objective

$$\mathbb{E}_q \left[\underbrace{-\log p_\theta(\mathbf{x}|\mathbf{z}_{t(0)})}_{\mathcal{L}_{\text{recons}}} + \underbrace{\sum_{i=1}^{T} \text{KL}[q(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)}, \mathbf{x}) \| p_\theta(\mathbf{z}_{s(i)}|\mathbf{z}_{t(i)})]}_{\mathcal{L}_{\text{diffusion}}} \right] + \underbrace{\text{KL}[q(\mathbf{z}_{t(T)}|\mathbf{x}) \| p_\theta(\mathbf{z}_{t(T)})]}_{\mathcal{L}_{\text{prior}}}$$

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{\alpha_{t(i)} - \alpha_{s(i)}}{1 - \alpha_{t(i)}} \log \langle \mathbf{x}_\theta(\mathbf{z}_{t(i)}), \mathbf{x} \rangle \right]$$

Masked Diffusion Variational Objective

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{\alpha_{t(i)} - \alpha_{s(i)}}{1 - \alpha_{t(i)}} \log \langle \mathbf{x}_\theta(\mathbf{z}_{t(i)}), \mathbf{x} \rangle \right]$$

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{-(\alpha_{s(i)} - \alpha_{t(i)})}{1 - \alpha_{t(i)}} \log \langle \mathbf{x}_\theta(\mathbf{z}_{t(i)}), \mathbf{x} \rangle \right]$$

Masked Diffusion Variational Objective

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{\boxed{-}(\alpha_{s(i)} - \alpha_{t(i)})}{1 - \alpha_{t(i)}} \boxed{\log \langle \mathbf{x}_\theta(\mathbf{z}_{t(i)}), \mathbf{x} \rangle} \right]$$

The 'familiar' cross-entropy loss

Masked Diffusion Variational Objective

$$\mathbf{x}_\theta^\top \mathbf{z}_t = 1, \quad \mathbf{z}_t \neq \mathbf{m}$$

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{-(\alpha_{s(i)} - \alpha_{t(i)})}{1 - \alpha_{t(i)}} \boxed{\log \langle \mathbf{x}_\theta(\mathbf{z}_{t(i)}), \mathbf{x} \rangle} \right]$$

Only masked
tokens contribute!

Masked Diffusion Variational Objective

$$\mathbf{x}_\theta^\top \mathbf{z}_t = 1, \quad \mathbf{z}_t \neq \mathbf{m}$$

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{-(\alpha_{s(i)} - \alpha_{t(i)})}{1 - \alpha_{t(i)}} \boxed{\log \langle \mathbf{x}_\theta(\mathbf{z}_{t(i)}), \mathbf{x} \rangle} \right]$$

Only masked tokens contribute!

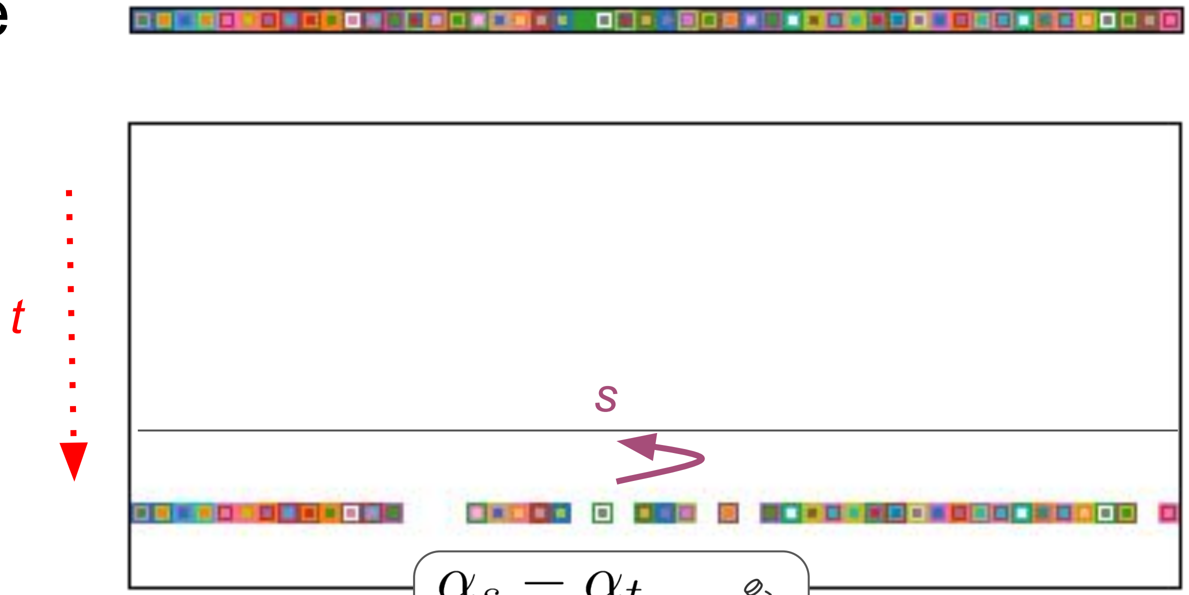
Learning To Reverse

$$\mathbb{E}_{t, \mathbf{z}_t \sim q(\mathbf{z}_t | \mathbf{x})}$$

1) Sample *t* and noise from *q*

Learning To Reverse

$$\mathbb{E}_{t,\mathbf{z}_t \sim q(\mathbf{z}_t|\mathbf{x})} \frac{\alpha_s - \alpha_t}{1 - \alpha_t}$$



$$\frac{\alpha_s - \alpha_t}{1 - \alpha_t}$$

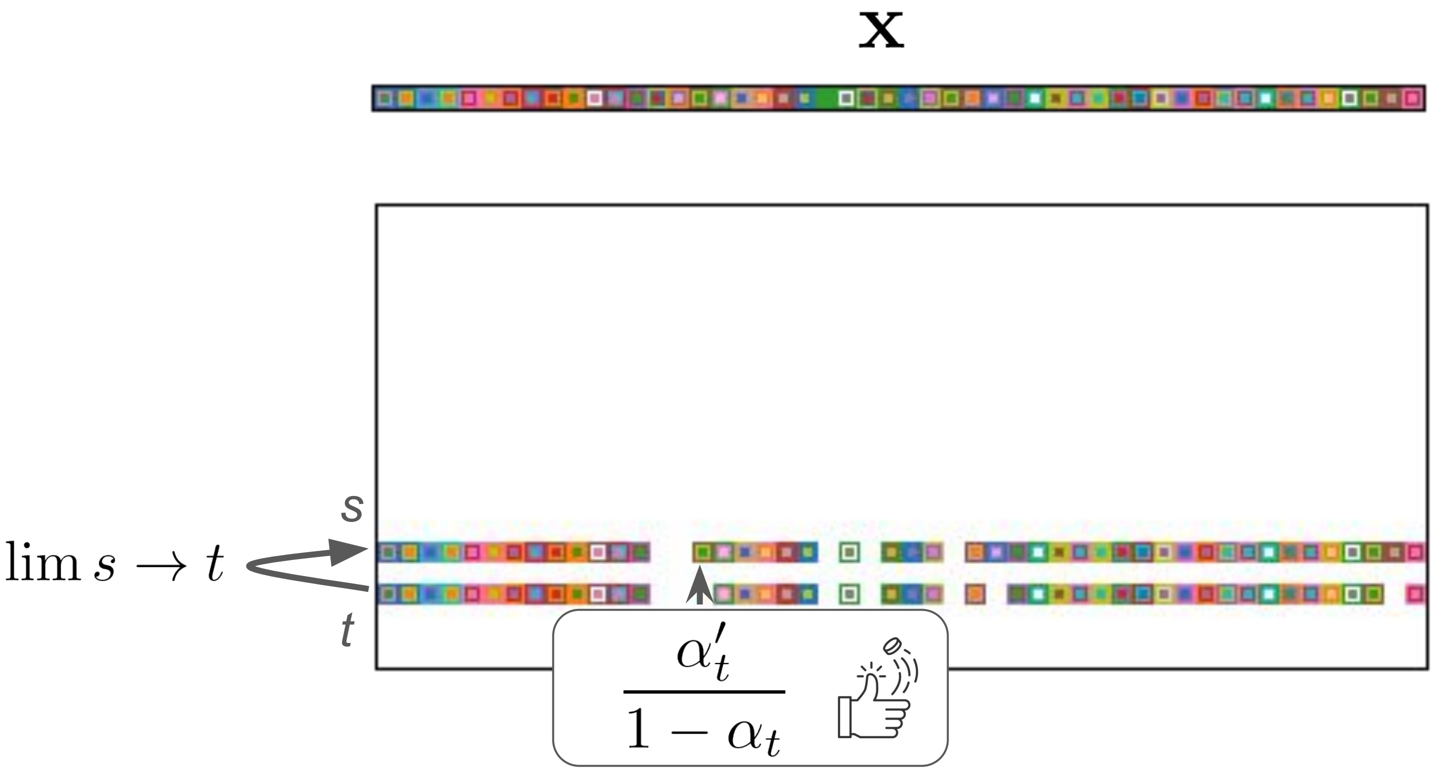
1) Sample and noise
2) Denoise and weigh by chance of unmasking

Learning To Reverse

$$\mathop{\mathbb{E}}_{t, \mathbf{z}_t \sim q(\mathbf{z}_t | \mathbf{x})} \frac{\alpha_s - \alpha_t}{1 - \alpha_t} \log \langle \mathbf{x}_\theta, \mathbf{x} \rangle$$

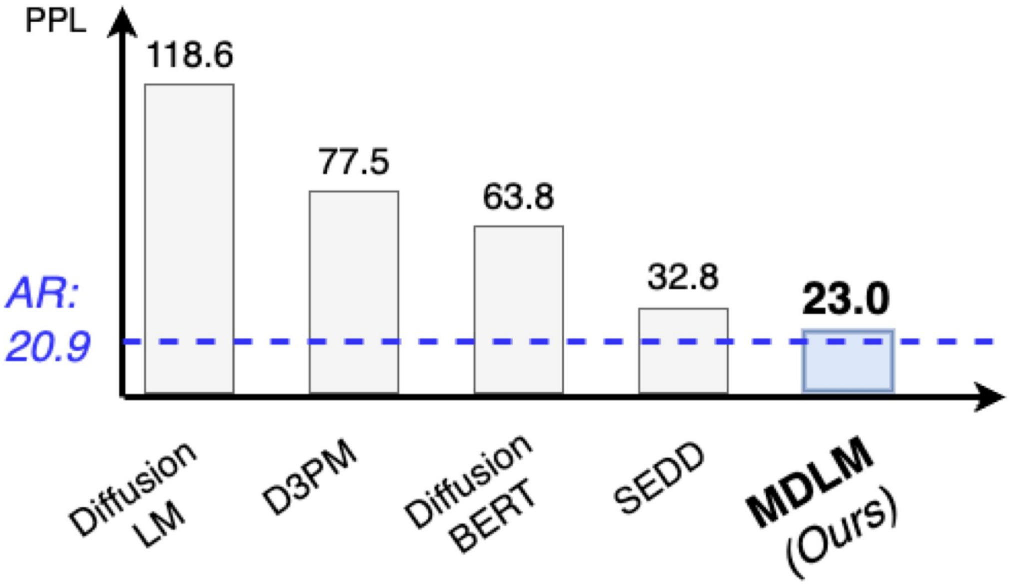
1) Sample and noise
2) Denoise and weigh
3) Score match

Continuous time Markov Chain



Experiments

Closing the gap to AR models



Results from LM1B dataset

Representation learning + Generative modeling

Table 4: GLUE evaluation results. Evaluation measures (↑) are F1 score for QQP and MRPC, Spearman correlations for STS-B, and accuracy for the rest. For MNLI, we report match/mismatch accuracies.

	MNLI (m/mm)	QQP	QNLI	SST-2	COLA	STS-B	MRPC	RTE	Avg
AR	80.94/80.78	86.98	86.16	90.14	33.43	84.32	83.88	47.29	74.88
BERT	84.43/85.35	88.41	**90.46**	**92.20**	54.81	**88.41**	89.16	61.37	81.62
+MDLM-FT	**84.76/85.07**	**88.49**	90.30	**92.20**	**57.69**	87.48	**90.53**	**62.09**	**82.06**

MDLM yields generative model <u>without</u> loss in representation learning capabilities

Effective discrete diffusion models (MDLM)

Improved sampling methods (ReMDM)

Recall:

$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}_s; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z_t} = \mathbf{m} \end{cases}$$

$$\Big\downarrow$$

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\theta + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z}_t = \mathbf{m} \end{cases}$$

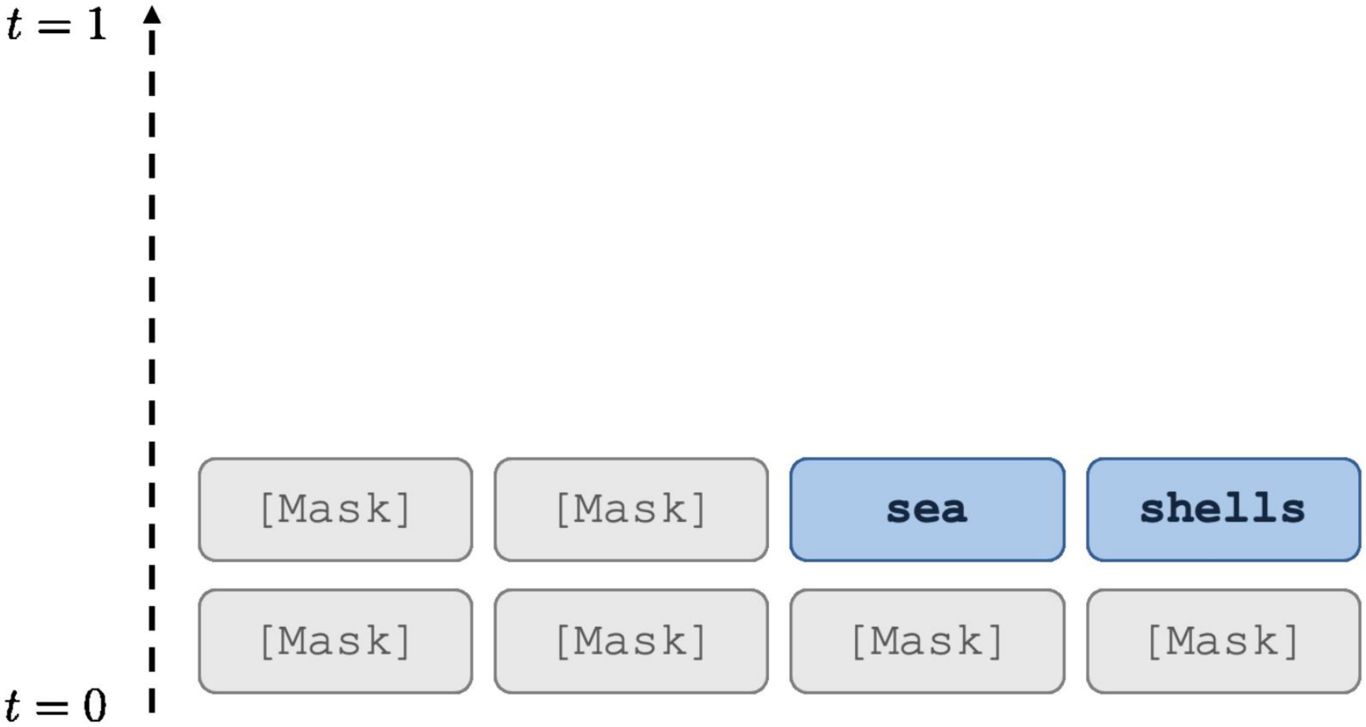
Recall:

$$q(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}_s; \frac{\alpha_s - \alpha_t}{1 - \alpha_t}\mathbf{x} + \frac{1 - \alpha_s}{1 - \alpha_t}\mathbf{m}\right), & \mathbf{z}_\mathbf{t} = \mathbf{m} \end{cases}$$

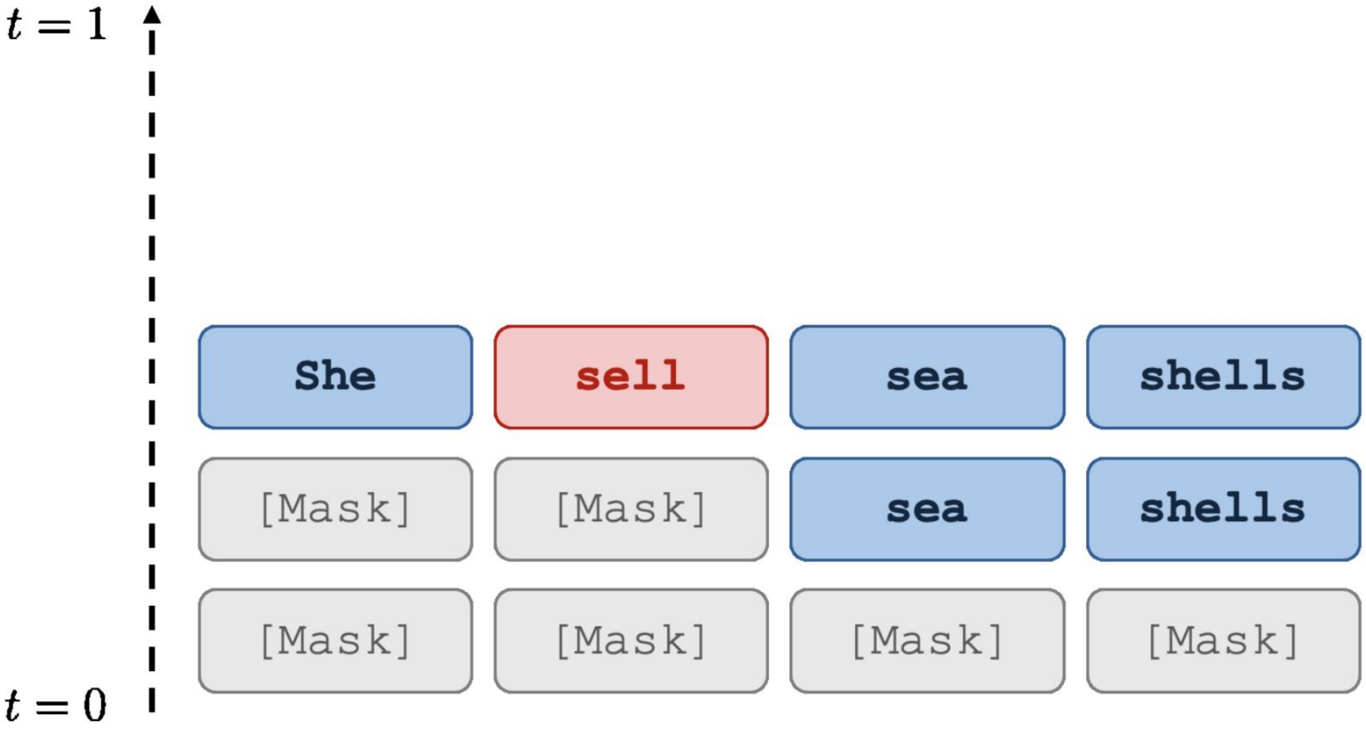
Unmasked tokens are ‘locked-in’ (even if incorrect!)

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}; \frac{\alpha_s - \alpha_t}{1 - \alpha_t}\mathbf{x}_\theta + \frac{1 - \alpha_s}{1 - \alpha_t}\mathbf{m}\right), & \mathbf{z}_t = \mathbf{m} \end{cases}$$

An example of Decoding Mistakes



An example of Decoding Mistakes



Summary:

Unmasked tokens are
'locked-in' (even if incorrect!)

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\theta + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z}_t = \mathbf{m} \end{cases}$$

- Bad quality
- Bad Inference-time scaling
- Bad controllability

Summary:

Unmasked tokens are 'locked-in' (even if incorrect!)

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\theta + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z}_t = \mathbf{m} \end{cases}$$

- Bad quality
- Bad Inference-time scaling
- Bad controllability

Summary:

Unmasked tokens are 'locked-in' (even if incorrect!)

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; \mathbf{z}_t), & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}\left(\mathbf{z}; \frac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbf{x}_\theta + \frac{1-\alpha_s}{1-\alpha_t}\mathbf{m}\right), & \mathbf{z}_t = \mathbf{m} \end{cases}$$

- Bad quality
- Bad Inference-time scaling
- Bad controllability

Remasking Discrete Diffusion Models with Inference-Time Scaling


Guanghan Wang*


Yair Schiff*


Subham Sahoo


Volodymyr Kuleshov

*Equal contribution

Goal: Enable remasking posterior

$$q(\mathbf{z}_s \mid \mathbf{z}_t = \mathbf{x}, \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_s; ?)$$

ReMasking Diffusion Models (ReMDM)

$$q(\mathbf{z}_s \mid \mathbf{z}_t = \mathbf{x}, \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_s; ?)$$

$$\downarrow$$

$$q_\sigma(\mathbf{z}_s \mid \mathbf{z}_t = \mathbf{x}, \mathbf{x}) = \mathrm{Cat}(\mathbf{z}_s; (1 - \sigma_t)\mathbf{x} + \sigma_t \mathbf{m})$$

ReMDM: (Re)masking posteriors

$$q_\sigma(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \text{Cat}(\mathbf{z}_s; (1-\sigma_t)\mathbf{x} + \sigma_t \mathbf{m}) & \mathbf{z}_t \neq \mathbf{m} \\ \text{Cat}(\mathbf{z}_s; \frac{\alpha_s - (1-\sigma_t)\alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s-\sigma_t\alpha_t}{1-\alpha_t}\mathbf{m}) & \mathbf{z}_t = \mathbf{m} \end{cases}$$

Theorem. *Given these posteriors q_σ, the marginal distributions do not change relative to the original masked diffusion language models.*

ReMDM: (Re)masking posteriors

$$q_\sigma(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; (1-\sigma_t)\mathbf{x} + \sigma_t \mathbf{m}) & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}(\mathbf{z}_s; \frac{\alpha_s - (1-\sigma_t)\alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s-\sigma_t\alpha_t}{1-\alpha_t}\mathbf{m}) & \mathbf{z}_t = \mathbf{m} \end{cases}$$

$$\sigma_t \uparrow \quad q(\mathbf{z}_s = \mathbf{x}; \mathbf{z}_t = \mathbf{m}, \mathbf{x}) \uparrow$$

ReMDM: (Re)masking posteriors

$$q_\sigma(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; (1-\sigma_t)\mathbf{x} + \sigma_t \mathbf{m}) & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}(\mathbf{z}_s; \frac{\alpha_s - (1-\sigma_t)\alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s-\sigma_t\alpha_t}{1-\alpha_t}\mathbf{m}) & \mathbf{z}_t = \mathbf{m} \end{cases}$$

$$\downarrow$$

$$\sigma_t \uparrow \qquad q(\mathbf{z}_s = \mathbf{x}; \mathbf{z}_t = \mathbf{m}, \mathbf{x}) \uparrow$$

$$\downarrow$$

encourage generate-then-refine sampling

Recall: Masked Diffusion Variational Objective

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \Big[\frac{\alpha_t - \alpha_s}{1 - \alpha_t} \log \langle \mathbf{x}_\theta(\mathbf{z}_t), \mathbf{x} \rangle \Big]$$

ReMDM objective is reweighted version of MDLM

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \Big[\frac{\alpha_t - \alpha_s}{1 - \alpha_t} \log \langle \mathbf{x}_\theta(\mathbf{z}_t), \mathbf{x} \rangle \Big]$$

$$\sigma_t \uparrow \quad \mathcal{L}^{\sigma}_{diffusion} \uparrow$$

$$\mathcal{L}^{\sigma}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_{q_\sigma} \Big[\frac{(1 - \sigma_t)\alpha_t - \alpha_s}{1 - \alpha_t} \log \langle \mathbf{x}_\theta(\mathbf{z}_t), \mathbf{x} \rangle \Big]$$

ReMDM objective is reweighted version of MDLM

$$\mathcal{L}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_q \left[\frac{\alpha_t - \alpha_s}{1 - \alpha_t} \log \langle \mathbf{x}_\theta(\mathbf{z}_t), \mathbf{x} \rangle \right]$$

Re-use pre-trained $\mathbf{x}_\theta$ from MDLM

$$\mathcal{L}^{\sigma}_{\text{diffusion}} = \sum_{i=1}^{T} \mathbb{E}_{q_\sigma} \left[\frac{(1 - \sigma_t)\alpha_t - \alpha_s}{1 - \alpha_t} \log \langle \mathbf{x}_\theta(\mathbf{z}_t), \mathbf{x} \rangle \right]$$

Recall: ReMDM posterior

$$q_\sigma(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x}) = \begin{cases} \mathrm{Cat}(\mathbf{z}_s; (1-\sigma_t)\mathbf{x} + \sigma_t \mathbf{m}) & \mathbf{z}_t \neq \mathbf{m} \\ \mathrm{Cat}(\mathbf{z}_s; \frac{\alpha_s - (1-\sigma_t)\alpha_t}{1-\alpha_t}\mathbf{x} + \frac{1-\alpha_s-\sigma_t\alpha_t}{1-\alpha_t}\mathbf{m}) & \mathbf{z}_t = \mathbf{m} \end{cases}$$

$$\Big\downarrow$$

$$0 \leq \sigma_t \leq \min\left\{1, \frac{1-\alpha_s}{\alpha_t}\right\} =: \sigma_t^{max}$$

ReMDM strategies

$$0 \leq \sigma_t \leq \min\left\{1, \frac{1-\alpha_s}{\alpha_t}\right\} =: \sigma_t^{max}$$

- ReMDM-cap $\quad \sigma_t = \min\{\eta_{cap}, \frac{1-\alpha_s}{\alpha_t}\} \quad \eta_{cap} \in [0, 1]$

ReMDM strategies

$$0 \leq \sigma_t \leq \min\left\{1, \frac{1-\alpha_s}{\alpha_t}\right\} =: \sigma_t^{max}$$

- ReMDM-cap $\quad \sigma_t = \min\{\eta_{cap}, \frac{1-\alpha_s}{\alpha_t}\} \quad \eta_{cap} \in [0,1]$

- ReMDM-rescale $\quad \sigma_t = \eta_{rescale} \cdot \sigma_{max} \quad \eta_{rescale} \in [0,1]$

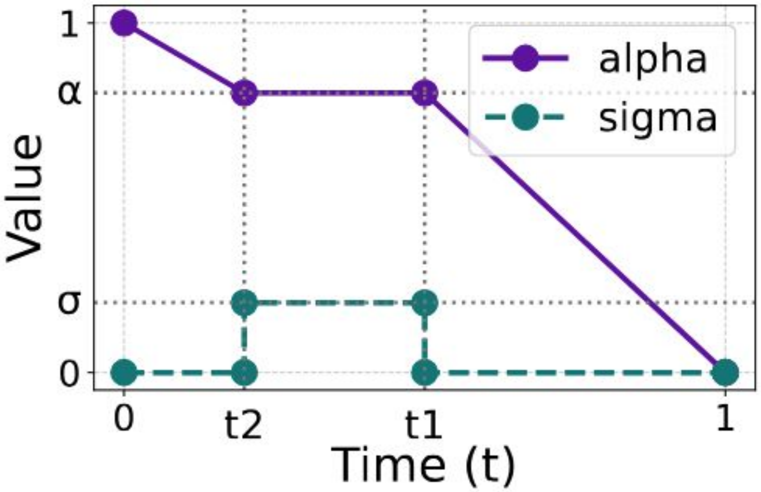
ReMDM strategies

$$0 \le \sigma_t \le \min\left\{1, \frac{1-\alpha_s}{\alpha_t}\right\} =: \sigma_t^{max}$$

- ReMDM-cap $\sigma_t = \min\{\eta_{cap}, \frac{1-\alpha_s}{\alpha_t}\} \quad \eta_{cap} \in [0, 1]$
- ReMDM-rescale $\sigma_t = \eta_{rescale} \cdot \sigma_{max} \quad \eta_{rescale} \in [0, 1]$
- ReMDM-conf $\sigma_t^{(\ell)} = \eta_{conf}^{(\ell)} \cdot \sigma_t$, where $\eta_{conf}^{(\ell)} = \frac{\exp(-\psi_t^{(\ell)})}{\sum_{l=1}^{L} \exp(-\psi_t^{(\ell')})}$

“Turning on” ReMDM: ReMDM-loop

In the beginning of generation, remasking brings little benefit and slows down generation



Algorithm 1 Sampling with ReMDM.

```
// Differences to standard MDLM
sampling noted in brown.
```

Input: pre-trained denoising network $\mathbf{x}_\theta$ (e.g., MDLM), number of timesteps T, noise schedule α_t, remasking schedule σ_t.

Initialize $\mathbf{z}_t = \boldsymbol{m}$.

for $i = T$ **to** 1 **do**

  $t = i/T, s = (i-1)/T$.

  Set α_t, α_s according to noise schedule.

  Set $\sigma_t \in [0, \sigma_t^{max}]$ according to remasking schedule.

  Compute approximate posterior:

$$p_\theta(\mathbf{z}_s \mid \mathbf{z}_t) = q_\sigma(\mathbf{z}_s \mid \mathbf{z}_t, \mathbf{x} = \mathbf{x}_\theta(\mathbf{z}_t))$$
$$= \begin{cases} \text{Cat}(\mathbf{z}_s; (1-\sigma_t)\mathbf{x}_\theta + \sigma_t \boldsymbol{m}), & \mathbf{z}_t \neq \boldsymbol{m} \\ \text{Cat}(\mathbf{z}_s; \frac{\alpha_s - (1-\sigma_t)\alpha_t}{1-\alpha_t}\mathbf{x}_\theta + \frac{1-\alpha_s-\sigma_t\alpha_t}{1-\alpha_t}\boldsymbol{m}), & \mathbf{z}_t = \boldsymbol{m} \end{cases}$$

  Sample $\mathbf{z}_s \sim p_\theta$.

  Set $\mathbf{z}_t = \mathbf{z}_s$.

end for

Output: $\mathbf{z}_t$.

Benefits of ReMDM vs. MDLM

✗ MDLM cannot correct mistakes

✓ ReMDM **can fix errors** via remasking

✗ MDLM can make at most L changes

✓ ReMDM can **benefit from increased test-time compute**

Experiments

Table 1. ReMDM improves sample quality in the case of inference-time scaling and faster sampling. ReMDM outperforms state-of-the-art masked diffusion models (SEDD; Lou et al. (2024), MDLM; Sahoo et al. (2024)) and masked diffusion models with corrector samplers such as Forward-Backward (FB; Campbell et al. (2022)) and Discrete Flow Matching (DFM; Gat et al. (2024)) corrector samplers. † indicates nucleus sampling. For each T, the best diffusion MAUVE score is **bolded**.

Method	MAUVE (↑)			Gen PPL. (↓)			Entropy (↑)		
Data	1.00			14.8			5.44		
AR *(T=1024)* †	0.760			12.1			5.22		
	T=1024	*T=2048*	*T=4096*	*T=1024*	*T=2048*	*T=4096*	*T=1024*	*T=2048*	*T=4096*
SEDD (absorb)	0.008	0.008	0.009	104.7	103.2	102.5	5.62	5.61	5.61
MDLM†	0.042	0.037	0.035	51.3	51.3	50.9	5.46	5.46	5.45
MDLM+FB†	0.133	0.197	0.243	33.8	28.6	22.8	5.35	5.28	5.18
MDLM+DFM†	0.254	0.294	0.269	21.7	21.0	20.7	5.20	5.19	5.17
ReMDM†	**0.403**	**0.610**	**0.656**	28.6	22.8	17.6	5.38	5.30	5.20
	T=128	*T=256*	*T=512*	*T=128*	*T=256*	*T=512*	*T=128*	*T=256*	*T=512*
SEDD (absorb)	0.007	0.007	0.008	119.2	110.1	107.2	5.65	5.63	5.62
MDLM†	0.015	0.023	0.031	61.5	55.8	53.0	5.52	5.49	5.48
MDLM+FB†	**0.064**	0.084	0.100	42.8	39.6	37.1	5.44	5.41	5.38
MDLM+DFM†	0.041	0.144	0.211	37.9	26.5	23.3	5.31	5.26	5.23
ReMDM†	0.057	**0.216**	**0.350**	42.5	30.5	21.1	5.43	5.34	5.21

Table 2. ReMDM produces the highest quality images. Values reflect FID / IS for varying T on discretized ImageNet conditional generation. For each metric and T, the best value is **bolded**.

Metric	Sampler	$T = 16$	$T = 32$	$T = 64$
FID (↓)	MaskGiT	**6.74**	**4.92**	4.85
	MDLM	7.88	5.37	4.69
	ReMDM	7.40	$\boxed{\textbf{4.92}}$	$\boxed{\textbf{4.45}}$
IS (↑)	MaskGiT	**155.32**	181.57	196.38
	MDLM	140.97	169.79	187.93
	ReMDM	145.27	$\boxed{\textbf{182.05}}$	$\boxed{\textbf{209.45}}$

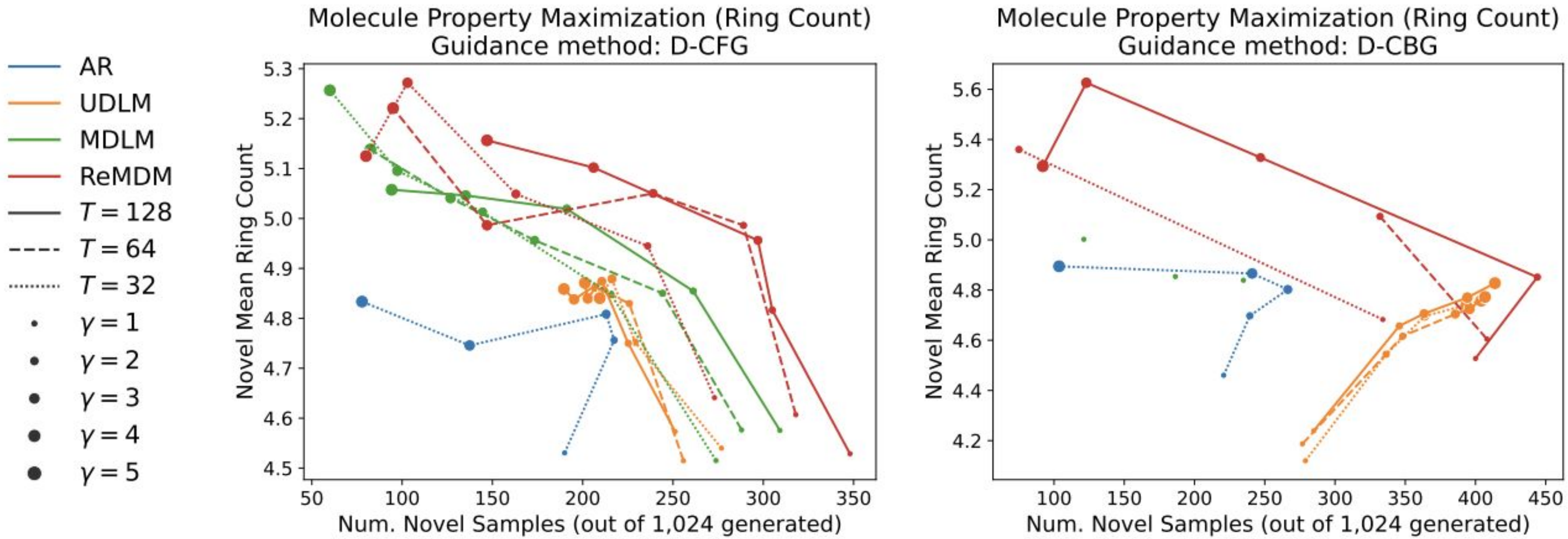


Figure 4. ReMDM improves steer-ability by extending the novelty-property maximization frontier. Controlled generation for ring count maximization on QM9 dataset with varying inference compute T and guidance strength γ. *(Left)* Discrete classifier-free guidance (D-CFG). *(Right)* Discrete classifier-based guidance (D-CBG) and FUDGE for AR.

Thank you!