Simple and Effective Masked Diffusion Language Models

Subham Sahoo

Marianne Arriola

Yair Schiff

Aaron Gokaslan

Edgar Marroquin

Justin Chiu

Alexander Rush

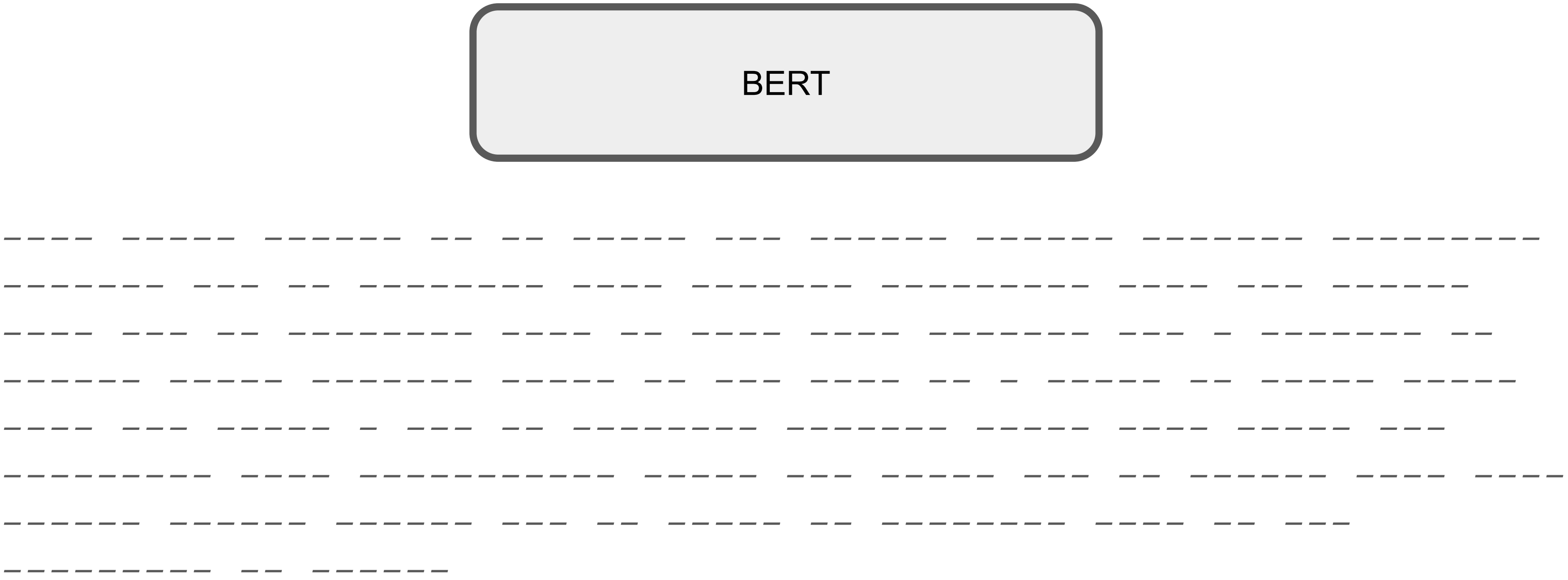
Volodymyr Kuleshov

Goal: Parallel Sampling from a Language Model

$$x \sim p_\theta(x)$$

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.

Sampling from a Masked Language Model



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.

$$x \sim p_\theta(x)$$

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.

Challenges

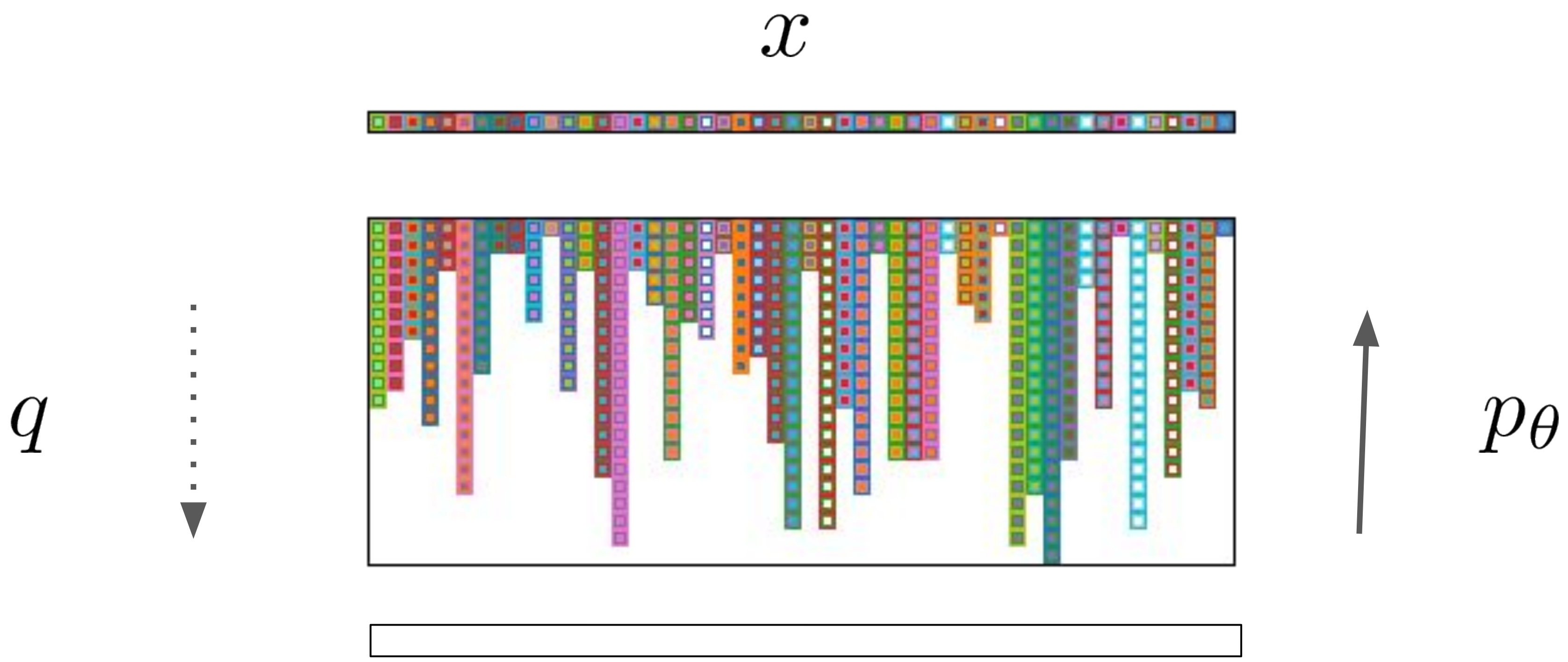
- What should the *noising process* look like for discrete sequence models?
- How should we *train a model* for parallel sampling?
- Can this process be made competitive with autoregressive models?

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.

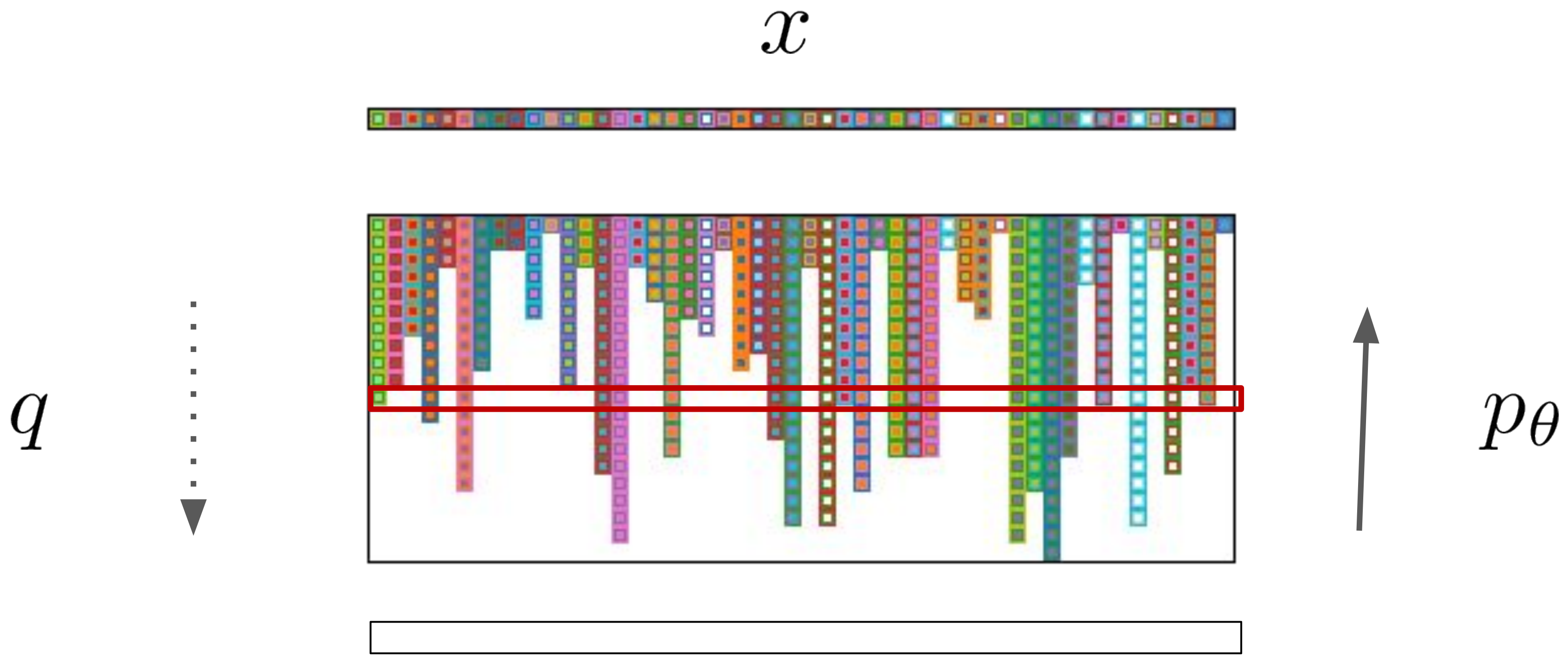
Many years later, -- he faced --- firing squad, ------- Aureliano Buendía was to remember that ------- afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the ---- of a river of clear water that ran ----- a bed of polished stones, which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

Masking Diffusion Language Model (MDLM)

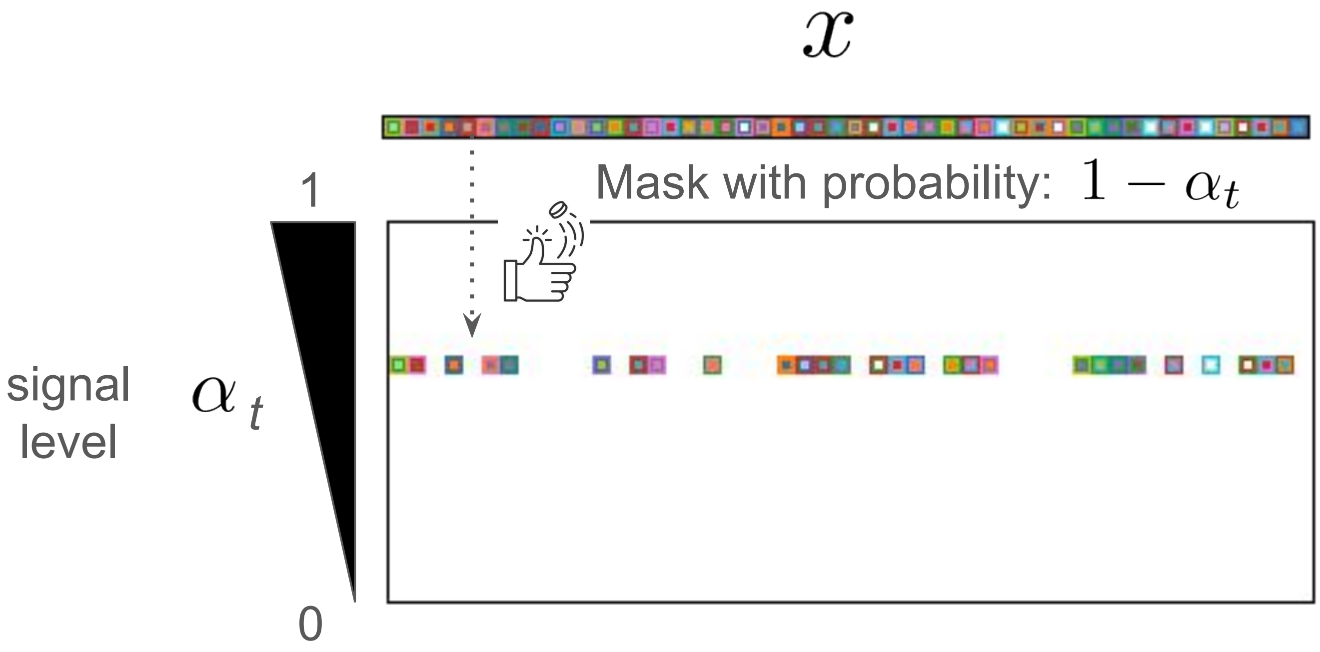
Our Goal: Discrete Masking Diffusion



Our Goal: Discrete Masking Diffusion



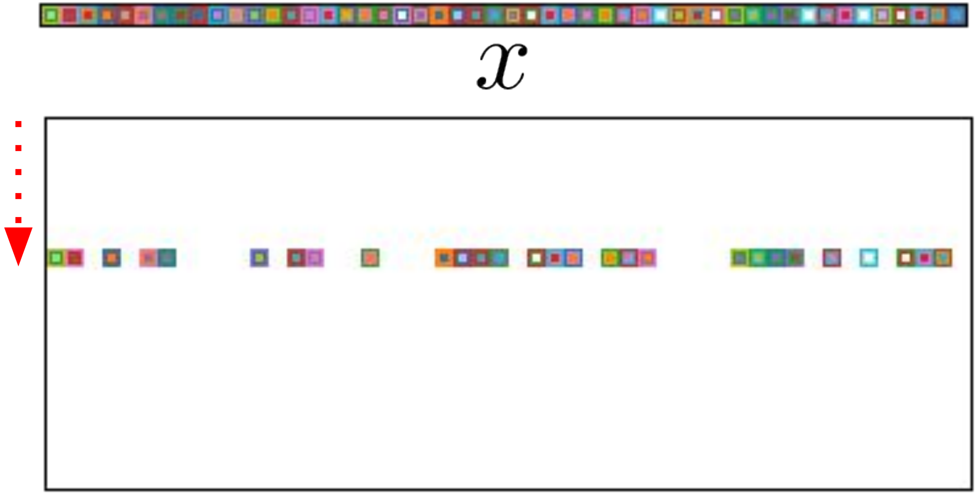
Masking Noise



Learning To Reverse

$$\underbrace{\mathbb{E}}_{t, z_t \sim q} \frac{\alpha'_t}{1 - \alpha_t} \log p_\theta(x|z_t)$$

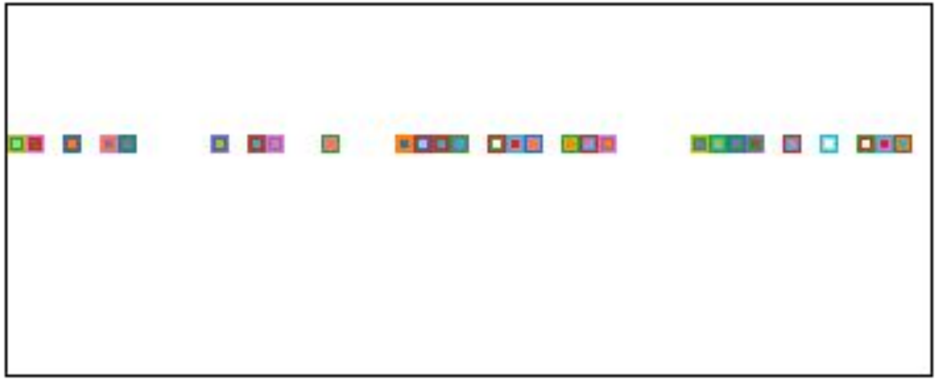
Contribution



1) Sample

Learning To Reverse

$$\mathbb{E}_{t,z_t\sim q}\underbrace{\frac{\alpha'_t}{1-\alpha_t}}\log p_\theta(x|z_t)$$



1) Sample

2) Weight by step change

Learning To Reverse

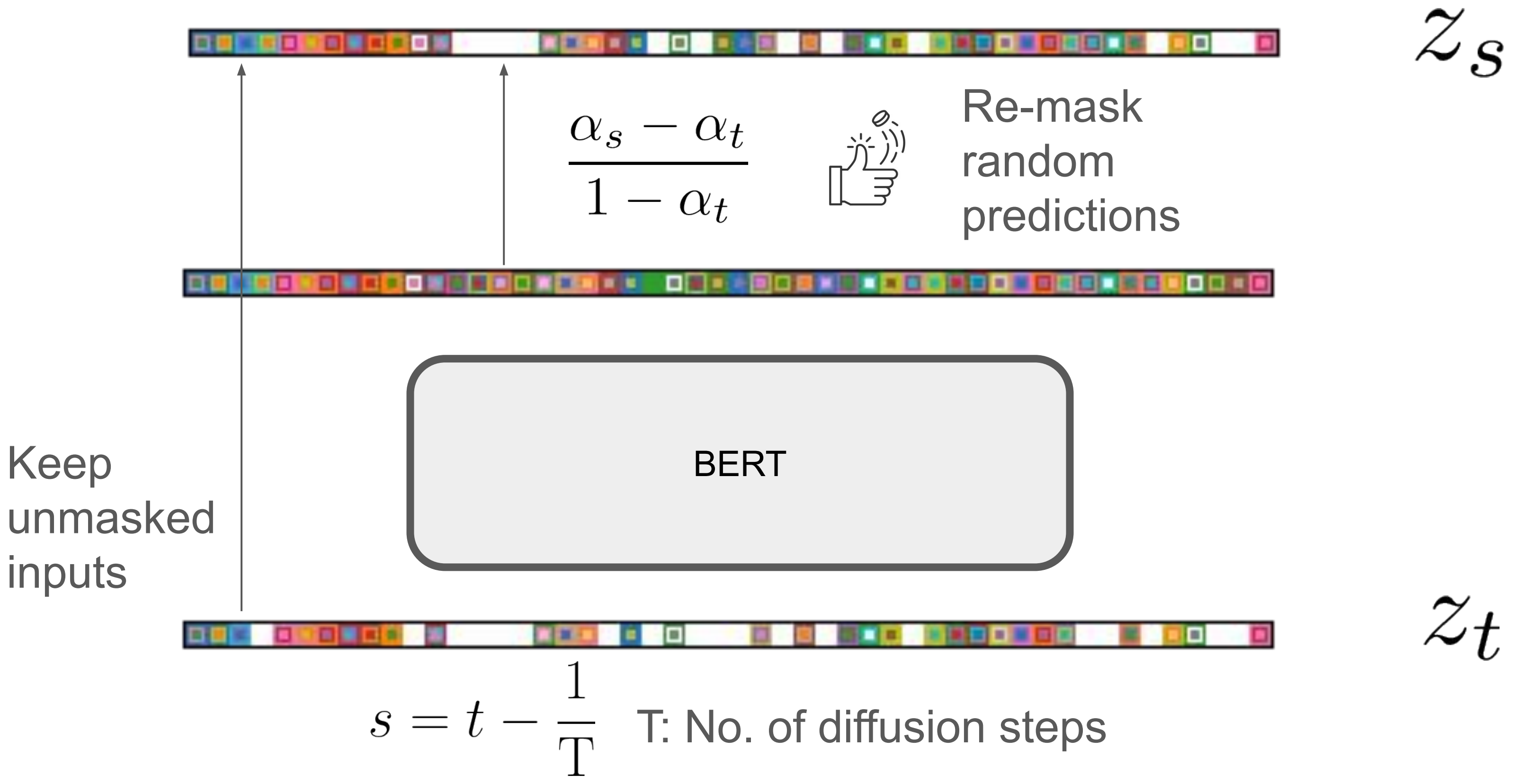
$$\mathbb{E}_{t, z_t \sim q} \frac{\alpha'_t}{1 - \alpha_t} \underbrace{\log p_\theta(x|z_t)}$$

1) Sample
2) Weight by step change
3) Reconstruct

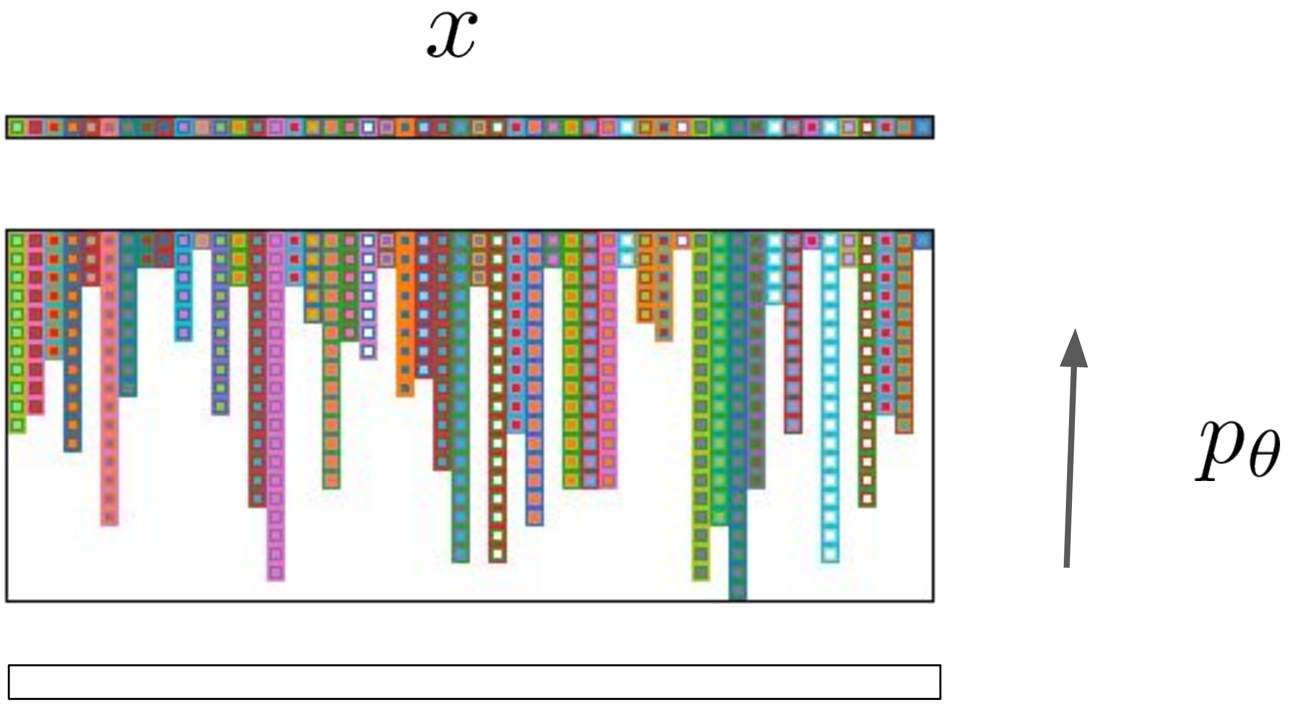
$$\log p_\theta(\mathbf{x}|\mathbf{z}_t) = \sum_{l}^{L} \log \langle \mathbf{x}_\theta^\ell(\mathbf{z}_t), \mathbf{x}^\ell \rangle$$

ℓ : token index L: Sequence length

One Step of Generation

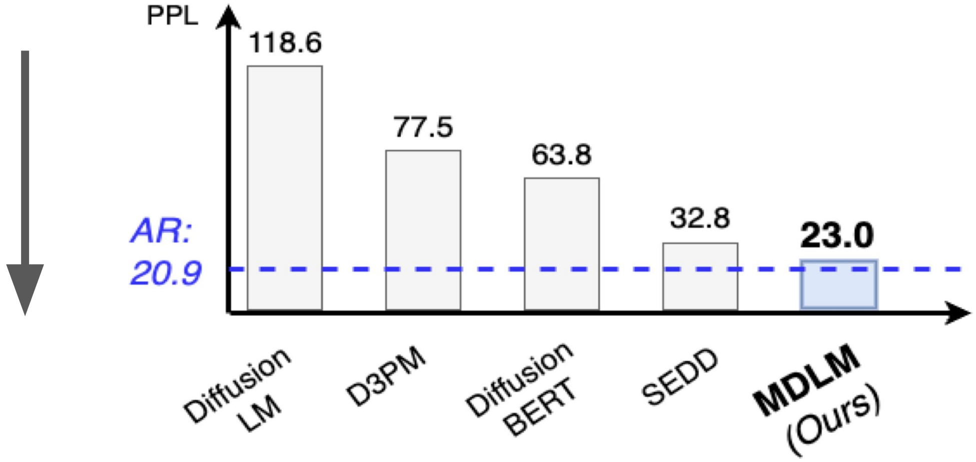


Generation

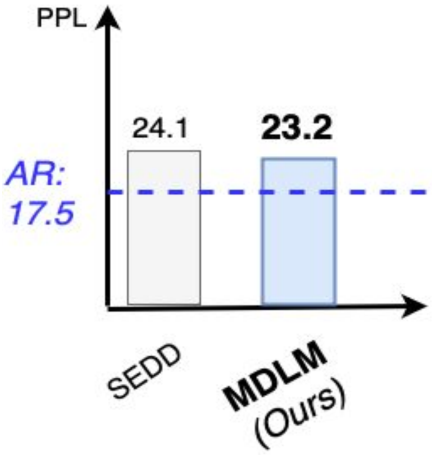


Experiments

Likelihood Evaluation

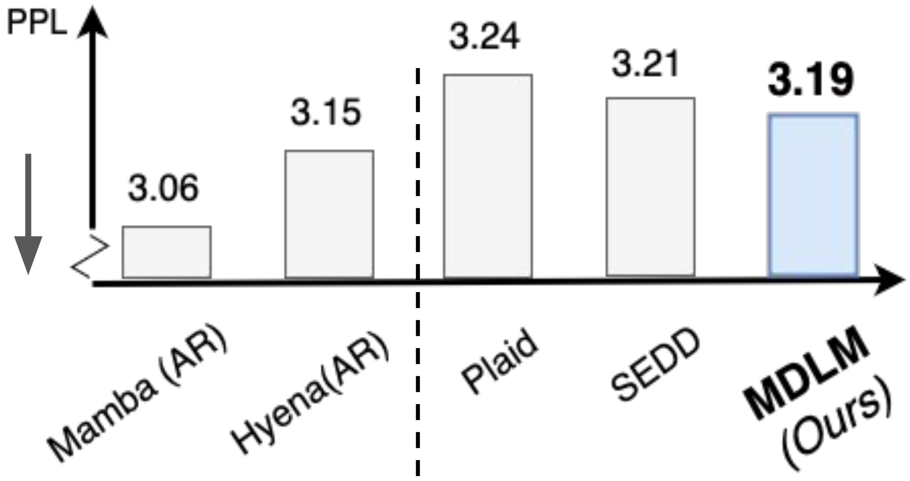


LM1B

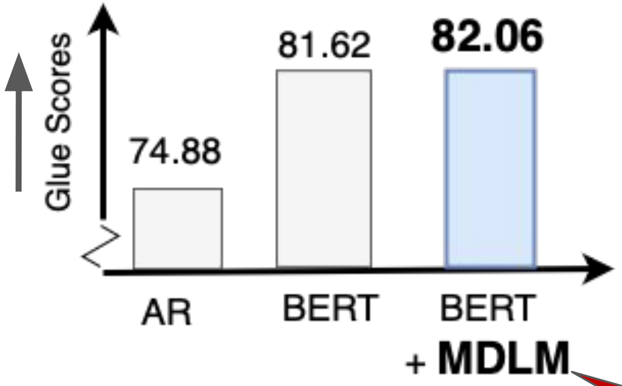


OpenWebText

Applying MDLM to Genomics



Representation learning + Generative modeling



Trained BERT on C4
Finetuned with MDLM
Evaluated it on GLUE

Generative BERT

Other contributions

- Derivation of the Rao Blackwellized Objective

$$\mathbb{E}_{q,t}\left[-\log p_\theta(\mathbf{x}|\mathbf{z}_{t(0)})+T\left[\frac{\alpha_s-\alpha_t}{1-\alpha_t}\log\frac{\alpha_t\langle\mathbf{x}_\theta(\mathbf{z}_t,t),\mathbf{m}\rangle+(1-\alpha_t)}{(1-\alpha_t)\langle\mathbf{x}_\theta(\mathbf{z}_t,t),\mathbf{x}\rangle}\right.\right.$$

$$\left.\left.+\frac{1-\alpha_s}{1-\alpha_t}\log\frac{(1-\alpha_s)(\alpha_t\langle\mathbf{x}_\theta(\mathbf{z}_t,t),\mathbf{m}\rangle+(1-\alpha_t))}{(1-\alpha_t)(\alpha_s\langle\mathbf{x}_\theta(\mathbf{z}_t,t),\mathbf{m}\rangle+(1-\alpha_s))}\right]\langle\mathbf{z}_t,\mathbf{m}\rangle\right]$$

D3PM

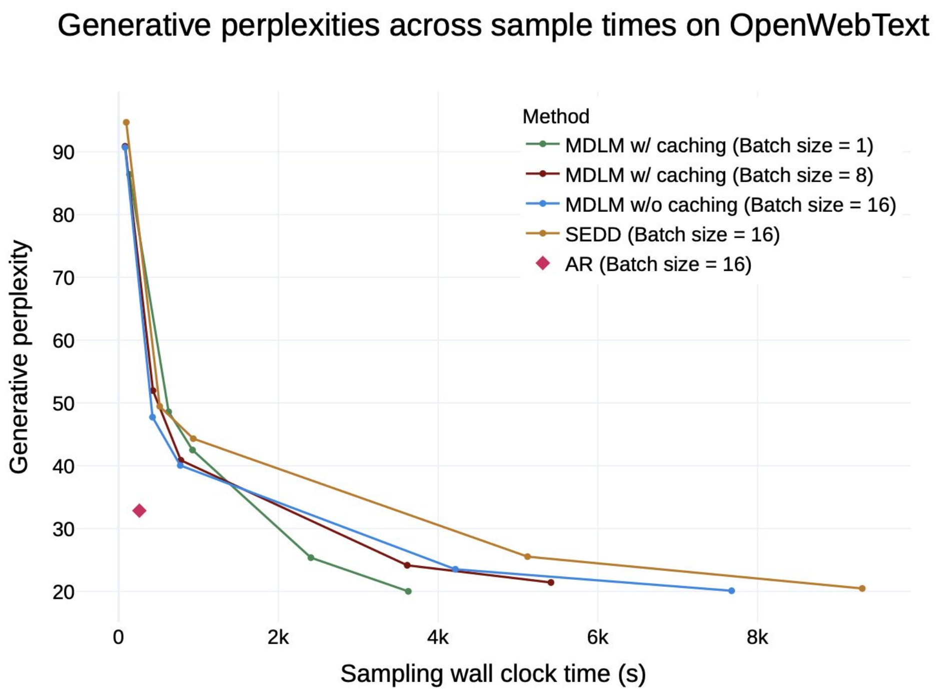
Strictly better

$$\mathbb{E}_{t\sim\mathcal{U}[0,1],q(\mathbf{z}_t|\mathbf{x})}\left[\frac{\alpha'_t}{1-\alpha_t}\log\langle\mathbf{x}_\theta(\mathbf{z}_t,t),\mathbf{x}\rangle\right]$$

MDLM

- Derivation of the Rao Blackwellized ELBO

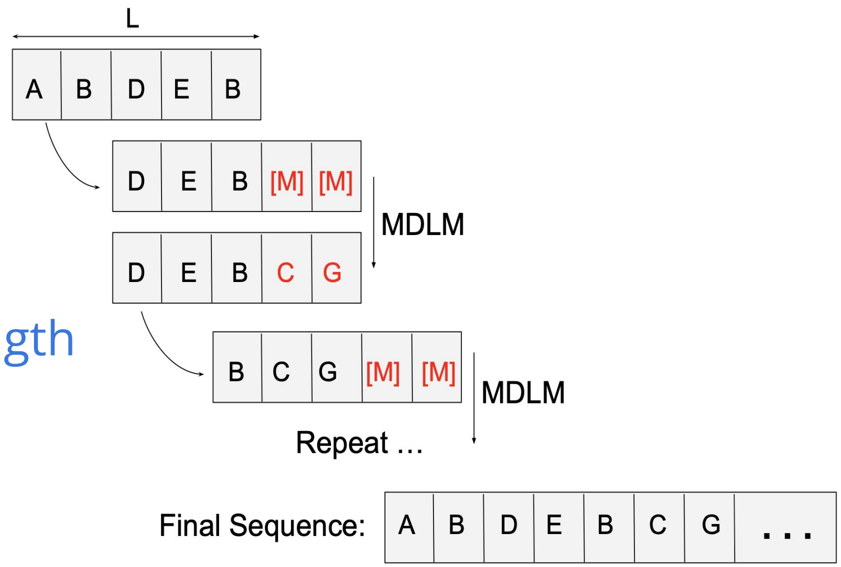
- Faster Sampler



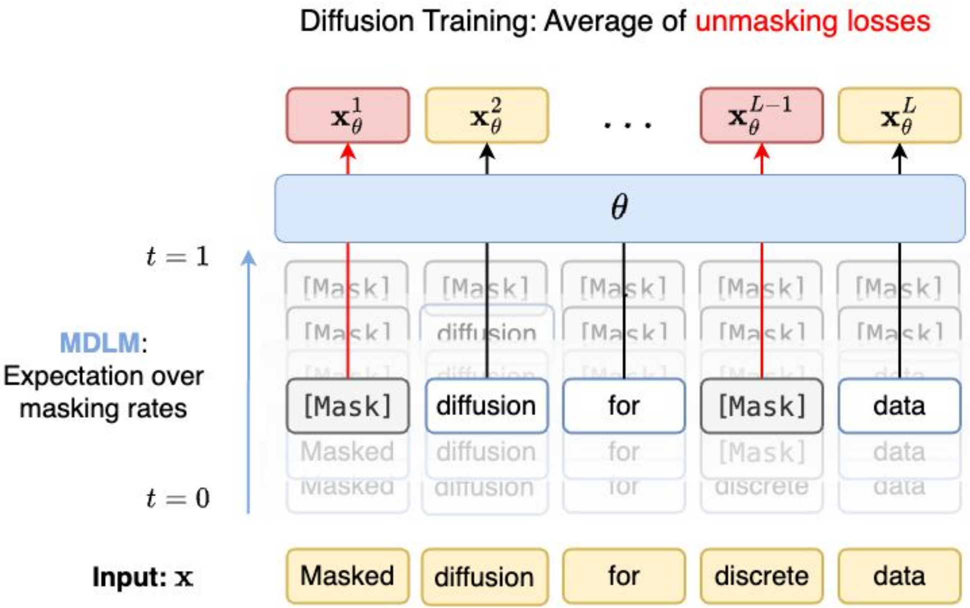
- Derivation of the Rao Blackwellized ELBO

- Faster Sampler

- Generating Sequences of Arbitrary Length



Conclusion



Diffusion Training: Average of unmasking losses
$\mathbf{x}_\theta^1$
$\mathbf{x}_\theta^2$
. . .
$\mathbf{x}_\theta^{L-1}$
$\mathbf{x}_\theta^L$
θ
$t = 1$
MDLM:
Expectation over masking rates
$t = 0$
[Mask] diffusion for [Mask] data
Masked diffusion for [Mask] data
Input: $\mathbf{x}$
Masked diffusion for discrete data