Simple and Effective Masked Diffusion Language Models

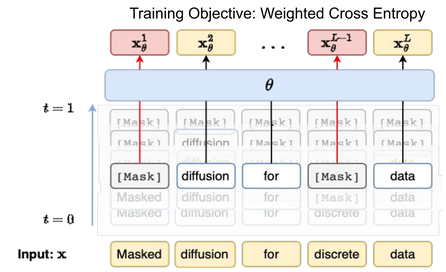
Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, Volodymyr Kuleshov
Cornell Tech, NYC, USA

Video Tutorial, Code, and more …

MOTIVATION

Unlike Autoregressive models, Diffusion Models
- ✓ Support **parallel generation**.
- ✓ Have potential to improve **long-term planning, controllable generation, and sampling speed**.

However, until recently Diffusion Models
- ✗ Exhibit a **performance gap** relative to AR models.



Simplified Masked Diffusion LM

- **Simple** interpolating masking noise processes
- Masking rate is **random**, not fixed
- Objective is a **variational lower bound**
- Admits fast **ancestral sampling**
- Novel substitution based parameterization
- Diffusion Objective is a **simple average** of MLM losses
- **Improved implementation** relative to previous masked diffusion

NOTATION

$\mathcal{V} \in \{\mathbf{x} \in \{0,1\}^K : \sum_{i=1}^K \mathbf{x}_i = 1\}$
Δ^K : Probability simplex
ℓ : Index in a sequence
L : Sequence length
$\mathbf{x} \in \mathcal{V}^L$: input sequence of length L
$\mathbf{x}^\ell \in \mathcal{V}$: Input token at index ℓ
$\bar{\mathbf{m}} \in \mathcal{V}$: mask token
$\mathbf{m}$: Sequence of L mask tokens
$\alpha_t \in [0,1]$: Signal level
$t \in [0,1]$: Diffusion time step
T : Total number of diffusion steps
$\mathbf{x}_\theta(\mathbf{z}_t) : \mathcal{V}^L \to (\Delta^K)^L$: Denoising model
$\mathbf{z}_t \in \mathcal{V}^L$: Masked input
$\langle \mathbf{a}, \mathbf{b} \rangle$: Dot product between 2 vectors $\mathbf{a}$ and $\mathbf{b}$

x

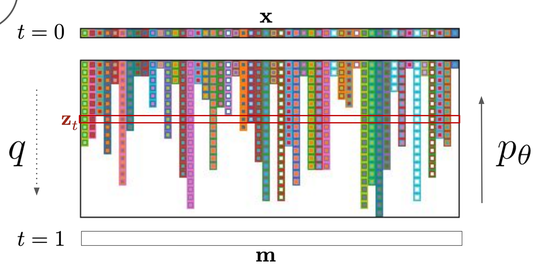
Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant afternoon when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of a river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs.

$\mathbf{z}_t$

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.

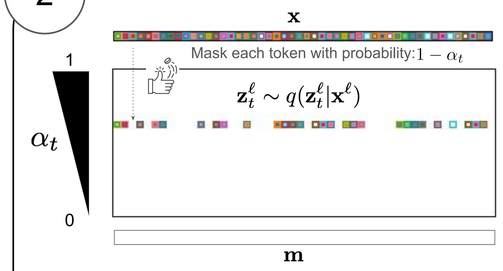
MASKED DIFFUSION LANGUAGE MODEL (MDLM)

1 Masked Diffusion



We consider **simplified diffusion processes** that interpolate between the clean data, $\mathbf{x}$, and the prior $\mathbf{m}$

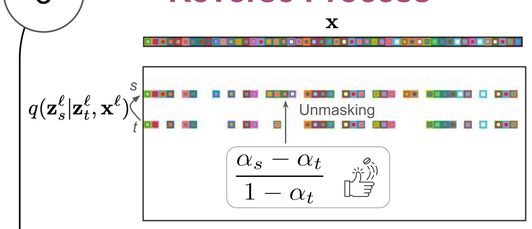
2 Forward Process



A diffusion process defines latents $\mathbf{z}_t^\ell$ for each token $\mathbf{x}^\ell$ for $t \in \{0, 1/T, \dots, 1\}$ using markov forward process inducing the marginals:

$$q(\mathbf{z}_t^\ell|\mathbf{x}^\ell) = \text{Cat}(\mathbf{z}_t^\ell; \alpha_t \mathbf{x}^\ell + (1-\alpha_t)\bar{\mathbf{m}})$$

3 Reverse Process



- The true reverse posterior for a token for a timestep $\mathbf{s} < \mathbf{t}$ is given as:

"Unmasked" tokens are never re-masked!

$$q(\mathbf{z}_s^\ell|\mathbf{z}_t^\ell, \mathbf{x}^\ell) = \begin{cases} \text{Cat}(\mathbf{z}_s^\ell; \mathbf{z}_t^\ell) & \mathbf{z}_t^\ell \neq \bar{\mathbf{m}} \\ \text{Cat}\left(\mathbf{z}_s^\ell; \frac{(1-\alpha_s)\bar{\mathbf{m}} + (\alpha_s - \alpha_t)\mathbf{x}^\ell}{1-\alpha_t}\right) & \mathbf{z}_t^\ell = \bar{\mathbf{m}} \end{cases}$$

- The approximate posterior is given as:

$$p_\theta(\mathbf{z}_s^\ell|\mathbf{z}_t) = q(\mathbf{z}_s^\ell|\mathbf{z}_t^\ell, \mathbf{x}^\ell = \mathbf{x}_\theta^\ell(\mathbf{z}_t))$$

where $\mathbf{x}_\theta(\mathbf{z}_t)$ is a BERT style denoising model.

Video Tutorial, Code, and more …

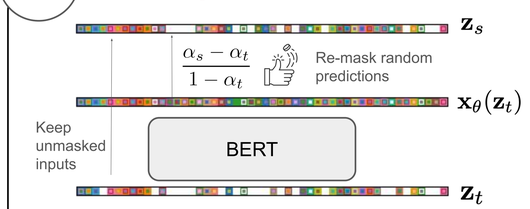
4 Training Objective

Denoising model $\mathbf{x}_\theta(\mathbf{z}_t)$ uses **SUBS**titution based parameterization:
- **Zero Masking Probabilities**: $\langle \mathbf{x}_\theta^\ell(\mathbf{z}_t), \bar{\mathbf{m}} \rangle = 0$
- **Carry Over Unmasking**: $\mathbf{x}_\theta^\ell(\mathbf{z}_t) = \mathbf{z}_t^\ell \text{ if } \mathbf{z}_t^\ell \neq \bar{\mathbf{m}}$

$T \to \infty$ + **SUBS** parameterization yields the following simplified Negative Evidence Lower Bound (N-ELBO):

$$\mathbb{E}_q \frac{\alpha_t'}{1-\alpha_t} \sum_\ell^L \log \langle \mathbf{x}_\theta^\ell(\mathbf{z}_t), \mathbf{x}^\ell \rangle$$

5 Sample Generation



Given a noisy sample $\mathbf{z}_t$, we construct the clean input $\mathbf{x}_\theta(\mathbf{z}_t)$. We then re-mask each token independently with a specified probability, ensuring that tokens unmasked in $\mathbf{z}_t$ are not re-masked

RESULTS

MDLM Perplexity Approaches AR

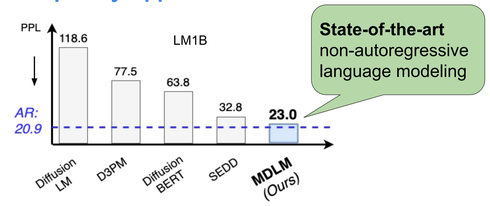


Table 3: Zero-shot validation perplexities (↓) of models trained for 524B tokens on OWT. All perplexities for diffusion models are upper bounds.

	PTB	Wikitext	LM1B	Lambada	AG News	Pubmed	Arxiv
AR (Retrained)	**82.05**	**25.75**	**51.25**	51.28	**52.09**	49.01	41.73
SEDD (Retrained)	100.09	34.28	68.20	49.86	62.09	44.53	38.48
MDLM (Ours)	95.26	32.83	67.01	**47.52**	61.15	**41.89**	**37.37**

Semi-AR Text Generation: Generating Sequences of Arbitrary Length

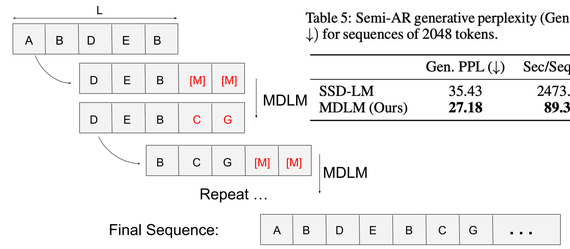
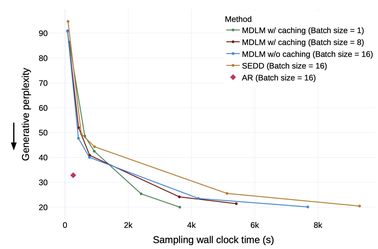


Table 5: Semi-AR generative perplexity (Gen. PPL; ↓) for sequences of 2048 tokens.

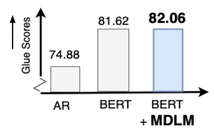
	Gen. PPL (↓)	Sec/Seq (↓)
SSD-LM	35.43	2473.9
MDLM (Ours)	**27.18**	**89.3**

Faster Sampler



Our denoising model is not conditioned on the diffusion time step, hence, the predictions of the models can be cached saving function calls to the denoising model. [TODO]

MDLM Preserves Representation Learning Capabilities



Trained BERT on C4
Fine-tuned with MDLM
Evaluated it on GLUE

REFERENCES

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. Advances in Neural Information Processing Systems, 34:17981–17993, 2021.
[2] Zhengfu He, Tianxiang Sun, Kuanning Wang, Xuanjing Huang, and Xipeng Qiu. Diffusion-bert: Improving generative masked language models with diffusion models. arXiv preprint arXiv:2211.15029, 2022.
[3] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. Advances in Neural Information Processing Systems, 35:4328–4343, 2022.
[4] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion language modeling by estimating the ratios of the data distribution. arXiv preprint arXiv:2310.16834, 2023.