Google DeepMind

Discrete Generative Modeling with Masked Diffusions

Jiaxin Shi
Google DeepMind
2025/2/28 @TTIC

jiaxins.io

Why Diffusion Models for Discrete Data

- Generating discrete data with parallel sampling

Mayor � � said � � � � � �
� that � new plan � � � � � � �

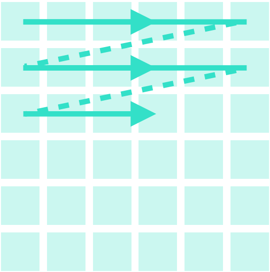
Mayor � Bowser said � meetings � Commissioner � on
Thursday that � new plan will be � board in � �

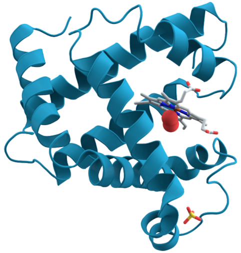
Mayor Muriel Bowser said after meetings � Commissioner � on
Thursday that � new plan will be � board in December �

Mayor Muriel Bowser said after meetings with Commissioner Busby on
Thursday that the new plan will be on board in December.

Why Diffusion Models for Discrete Data

- Generating discrete data with parallel sampling
- AR models require imposing an ordering which may be unnatural for many data types





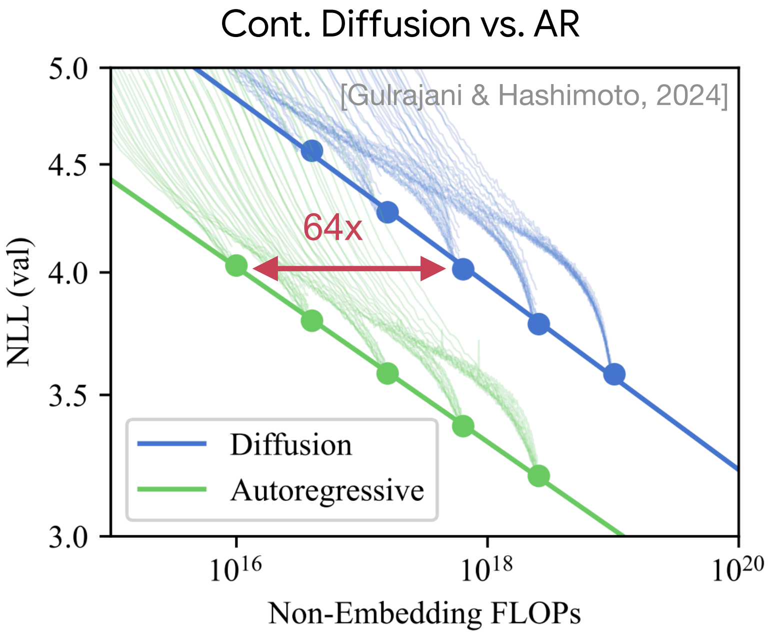
```
HGFGTLEHPIYKVAKQWSMVHDTTVYFSCGLHVAAHPATYVSM
TMLYHINMESFVNLEFCNFQTDDKYLEDPWARHEKYPIRKAIK
VSMDPNHGPVYCAKWDTILYMGKDGKERRTSAYMFTGVDEQHC
GRLFRITKSCWWGCCTLDNMKPDKAKACAEDMRRCRNIPVVQN
RNSKCRAIEWEIFQYWINCSTVVKTFAPCMFGFQFRFHYGYNY
DRETPVHAVNIINIWSAYKMTRYWCRIQCDSYWLWSGMTWRWC
CWEGSYKLMFCGWWRHFISKSMVTLGGHKKDDGRRWMLQSTHH
```

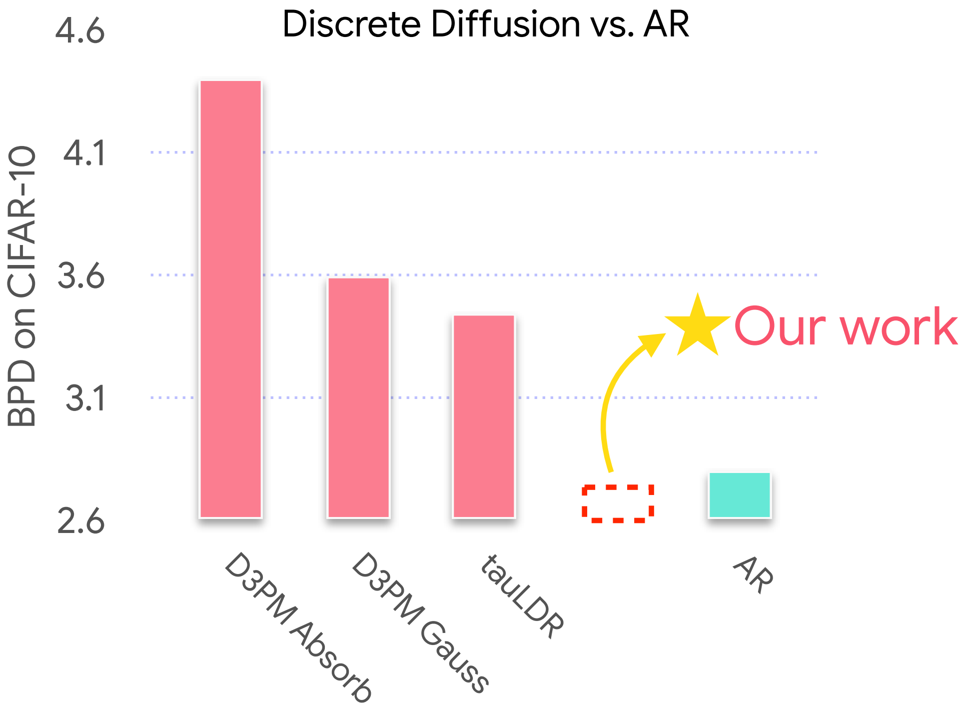
Duration (min)	IMDB Rating	Genre	Award
150	6.5	Action	Nominated
95	8.3	Romantic	Won
120	5.2	Horror	None

https://www.minecraft.net/en-us/article/build-your-very-own-custom-mobs
Shi, Juntong, et (2025). TabDiff: a mixed-type diffusion model for tabular data generation

Challenge

Diffusion yet to match AR performance on discrete data

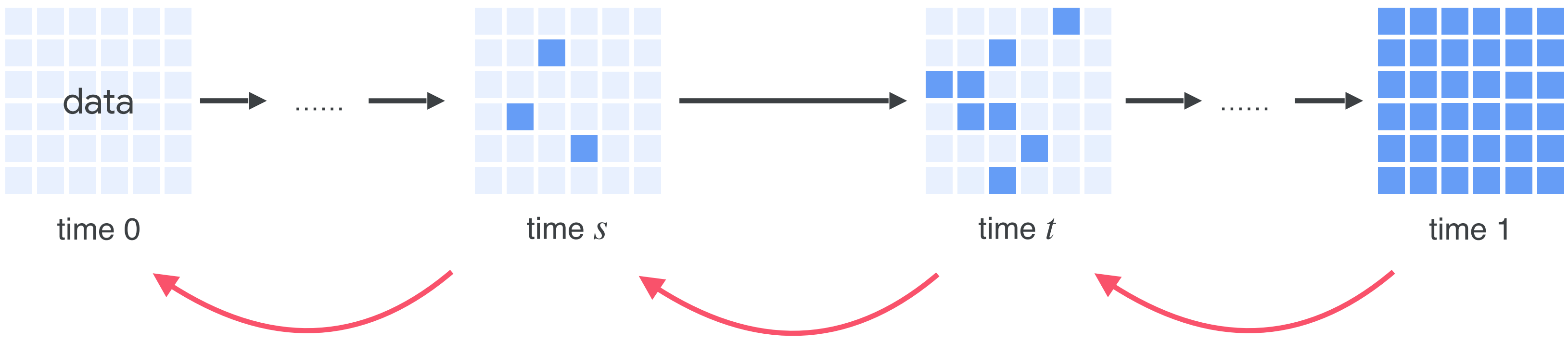


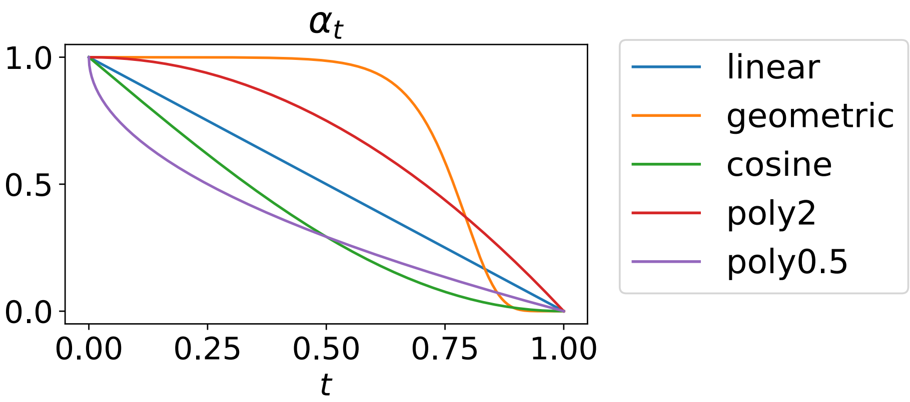


Gulrajani & Hashimoto (2024). Likelihood-based diffusion language models.

Masked Diffusion Models

Also known as absorbing diffusion, first proposed in Austin et al. (2021)



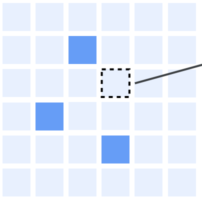


Masking schedule α_t: The expected proportion of unmasked tokens at t

Austin et al. (2021). Structured denoising diffusion models in discrete state-spaces.

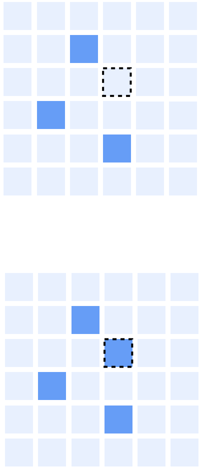
Masked Diffusion Models

Forward process $q(x_t \mid x_s) = \prod_{n=1}^{N} q(x_t^{(n)} \mid x_s^{(n)})$



time s

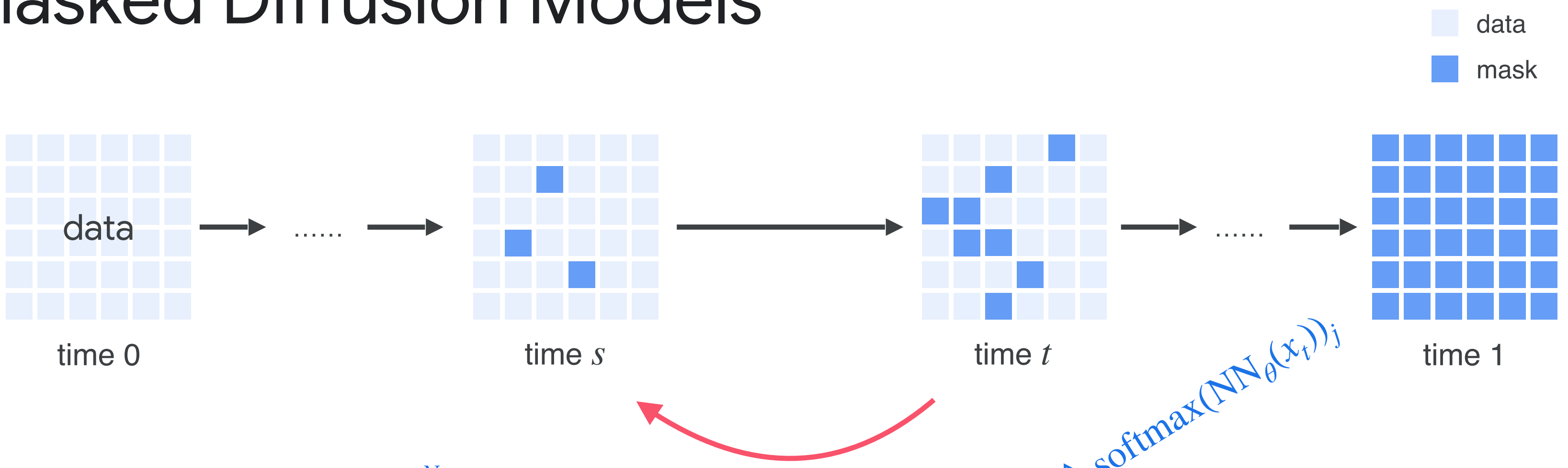
$$\begin{cases} \text{w/ prob. } \dfrac{\alpha_t}{\alpha_s}\text{, remains unmasked} \\ \\ \text{w/ prob. } 1 - \dfrac{\alpha_t}{\alpha_s}\text{, mask} \end{cases}$$



time t

Transition matrix: $\bar{Q}(s,t) = \frac{\alpha_t}{\alpha_s} I + \left(1 - \frac{\alpha_t}{\alpha_s}\right) \mathbf{1} e_m^\top$

Masked Diffusion Models



Reverse process $q(x_s \mid x_t) \approx \prod_{n=1}^{N} q(x_s^{(n)} \mid x_t)$ **as** $s \to t$

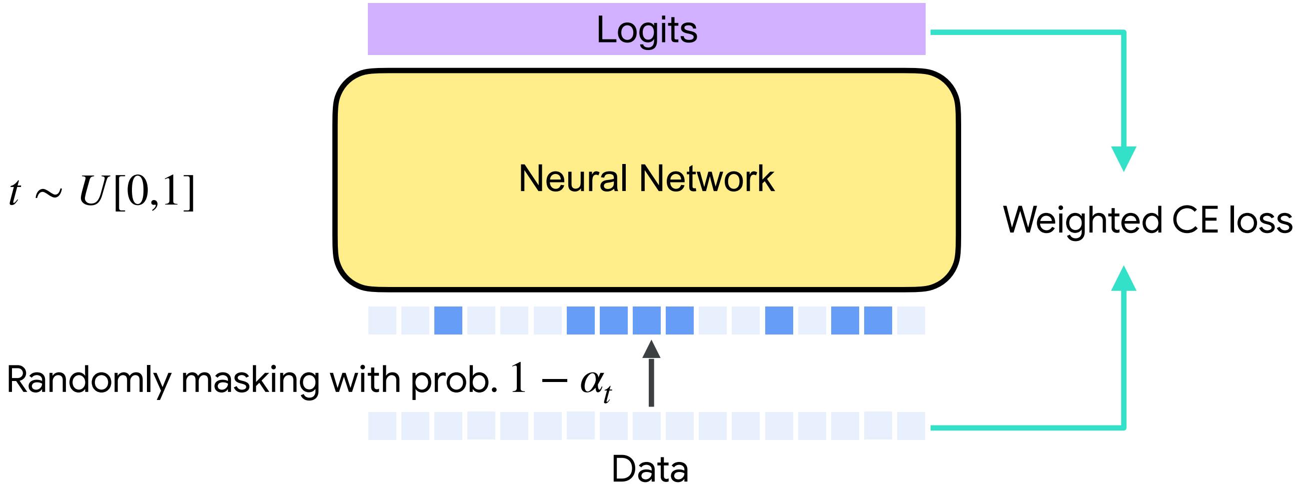
$\approx \mu_\theta(x_t)_j \triangleq \mathrm{softmax}(\mathrm{NN}_\theta(x_t))_j$

$$\begin{cases} \text{w/ prob. } \dfrac{\alpha_s - \alpha_t}{1-\alpha_t}\mathbb{E}[x_0^{(n)} = j \mid x_t], \text{ unmask to state } j \\[2em] \text{w/ prob. } \dfrac{1-\alpha_s}{1-\alpha_t}, \text{ remain masked} \end{cases}$$

MD4 Objective: Weighted Cross-Entropy Losses

Continuous-time Negative ELBO $(T \to \infty)$

$$\int_0^1 \frac{\alpha_t'}{1-\alpha_t} \mathbb{E}_{q(x_t|x_0)} \left[\sum_{n: x_t^{(n)} = m} (x_0^{(n)})^\top \log \mu_\theta^{(n)}(x_t, t) \right] \mathrm{d}t.$$



Three Interpretations of MD4

VDM (Kingma et al., 2021) version of D3PM (Austin et al., 2021)

- Continuous-time model
- Simplification as weighted cross-entropy loss

Adaptation of CTMC ELBO (Campbell et al., 2022) to enable low-variance estimate

- Campbell et al. (2022) requires multiple NN passes—estimation has high variance
- MD4 applies discrete "integration-by-part" to fix this

Mean parameterization counterpart of score parameterization (Lou et al., 2023)

- Score parameterization breaks consistency between forward & reverse processes

Kingma et al. (2021). Variational diffusion models.
Campbell et al. (2022). A continuous time framework for discrete denoising models.
Lou et al. (2023). Discrete diffusion language modeling by estimating the ratios of the data distribution.

Score v.s. Mean Parameterization

Proposition 1. The discrete score $s(x_t, t)_j = \frac{q_t(j)}{q_t(x_t)}$ for $x_t = m$ and $j \neq m$ can be expressed as

$$s(m, t)_j = \frac{\alpha_t}{1 - \alpha_t} \mathbb{E}[x_0 \,|\, x_t = m]^\top e_j$$

See also concurrent work based on this (Ou et al, 2024)

Implications

- True score satisfies the constraint $\sum_{j \neq m} s(m, t)_j = \frac{\alpha_t}{1 - \alpha_t}$
- Score parameterization breaks this and leads to inconsistency between forward & reverse processes

mean parameterization fixes the problem

$$s_\theta(m, t)_j = \frac{\alpha_t}{1 - \alpha_t} \mu_\theta(m, t)_j$$

GenMD4: State-dependent Schedules

Idea: Tokens are not created equal — make the probability of masking a token depend on the token value

Before	After
$\alpha_t : [0,1] \to [0,1]$	$\alpha_t : [0,1] \to [0,1]^{\lvert V \rvert}$
$\bar{Q}(s,t) = \frac{\alpha_t}{\alpha_s} I + \left(1 - \frac{\alpha_t}{\alpha_s}\right) \mathbf{1} e_m^\top$	$\bar{Q}(s,t) = \operatorname{diag}\left(\frac{\alpha_t}{\alpha_s}\right) + \left(I - \operatorname{diag}\left(\frac{\alpha_t}{\alpha_s}\right)\right) \mathbf{1} e_m^\top$

- ELBO is a bit complicated in discrete time
- Good news: it significantly simplifies as $T \to \infty$

$$\mathcal{L}_\infty = \int_0^1 \left(\frac{\alpha_t'}{\mathbf{1} - \alpha_t}\right)^\top \mathbb{E}_{q(x_t|x_0)} \left[\delta_{x_t,m} \cdot \left(x_0 - \mu_\theta(x_t,t) + x_0 x_0^\top \log \mu_\theta(x_t,t)\right)\right] \mathrm{d}t$$

GenMD4: Learned State-Dependent Schedules

$\alpha_t : [0,1] \to [0,1]^{|V|}$. Schedule for token type i: $(\alpha_t)_i = 1 - t^{w_i}$

Token types with largest ws (unmask first)

'<|endoftext|>',
'\n',
'.',
' (',
'_',
' "',
',',
' strutConnector',
' \xa0\xa0',
' DevOnline'

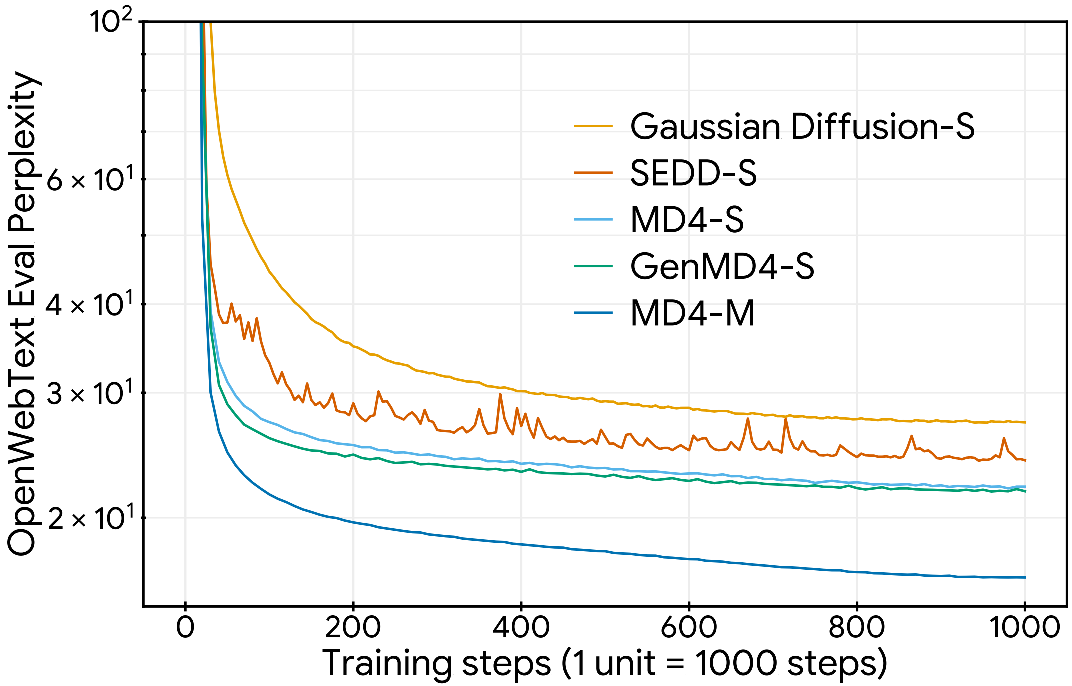
Token types with smallest ws

' diligently',
' unreliable',
' irresistible',
' dart',
' tracing',
' enlarged',
' playful',
' freeing',
' weighted',
'407'

Perplexity on GPT-2 Zero-Shot Eval

Size	Method	LAMBADA	WikiText2	PTB	WikiText103	IBW
Small	GPT-2 (WebText)*	**45.04**	42.43	138.43	41.60	75.20
	D3PM	$\leq$ 93.47	$\leq$ 77.28	$\leq$ 200.82	$\leq$ 75.16	$\leq$ 138.92
	Plaid	$\leq$ 57.28	$\leq$ 51.80	$\leq$ 142.60	$\leq$ 50.86	$\leq$ 91.12
	SEDD Absorb	$\leq$ 50.92	$\leq$ 41.84	$\leq$ 114.24	$\leq$ 40.62	$\leq$ 79.29
	SEDD Absorb (reimpl.)	$\leq$ 49.73	$\leq$ 38.94	$\leq$ 107.54	$\leq$ 39.15	$\leq$ 72.96
	MD4 (Ours)	$\leq$ 48.43	$\leq$ **34.94**	$\leq$ **102.26**	$\leq$ **35.90**	$\leq$ **68.10**
Medium	GPT-2 (WebText)*	**35.66**	31.80	123.14	31.39	55.72
	SEDD Absorb	$\leq$ 42.77	$\leq$ 31.04	$\leq$ 87.12	$\leq$ 29.98	$\leq$ 61.19
	MD4 (Ours)	$\leq$ 44.12	$\leq$ **25.84**	$\leq$ **66.07**	$\leq$ **25.84**	$\leq$ **51.45**

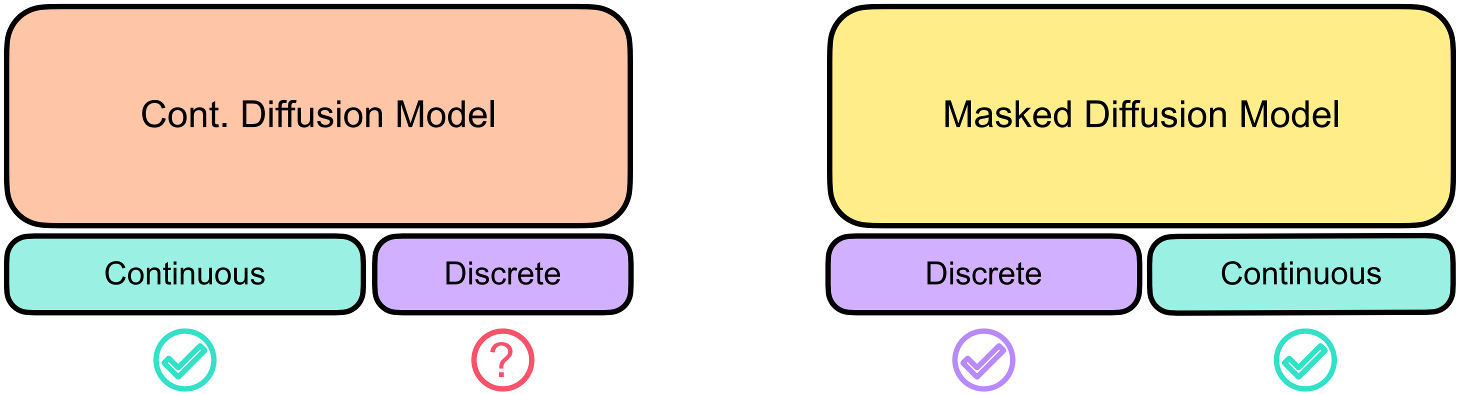
Perplexity on OpenWebText Validation Set



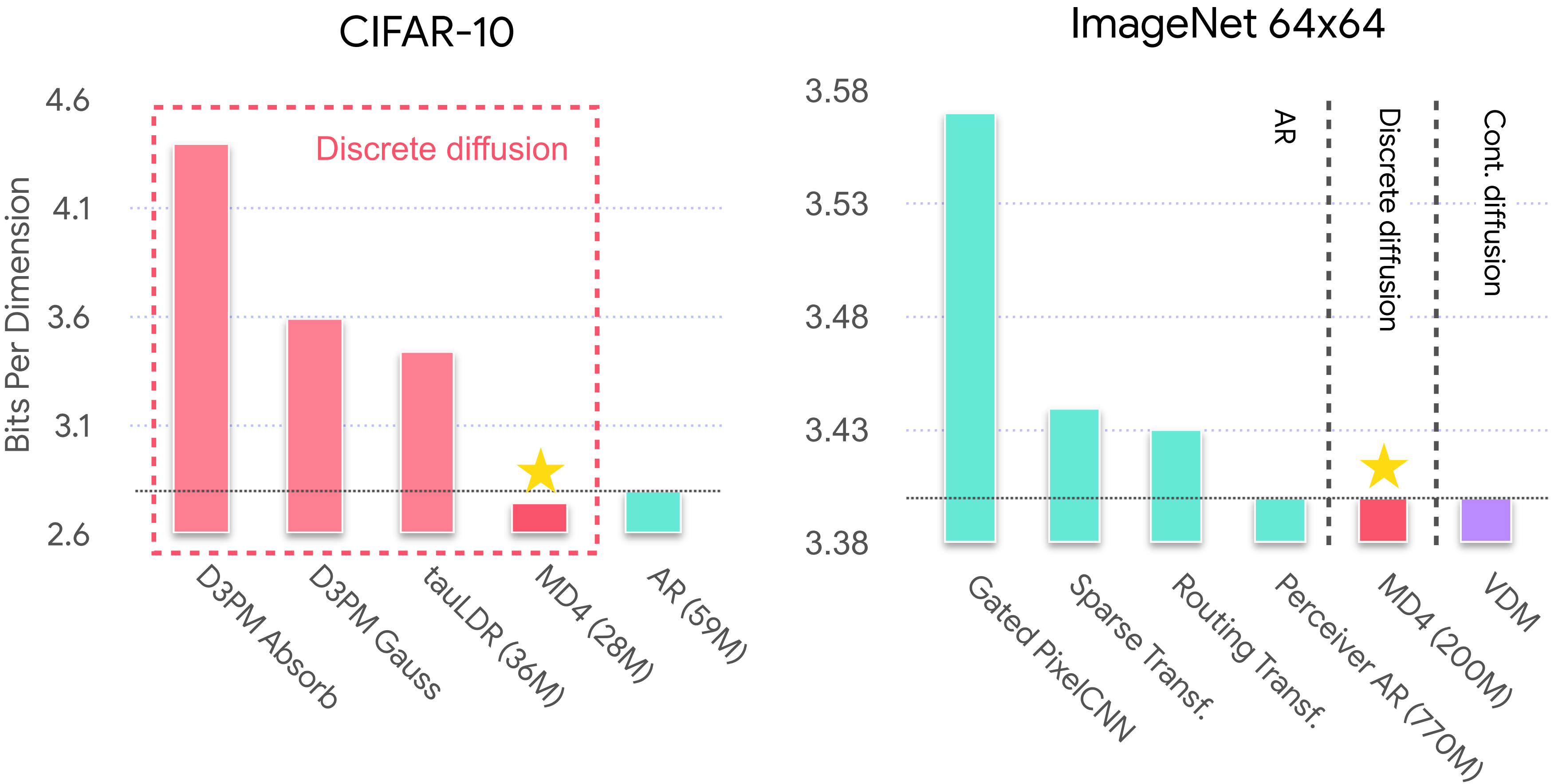
Size	Method	Perplexity (↓)
Small	Gaussian Diffusion	$\leq$ 27.28
	SEDD Absorb (reimpl.)	$\leq$ 24.10
	MD4 (Ours)	$\leq$ 22.13
	GenMD4 (Ours)	$\leq$ **21.80**
Medium	MD4 (Ours)	$\leq$ **16.64**

Unifying Discrete & Continuous Modalities

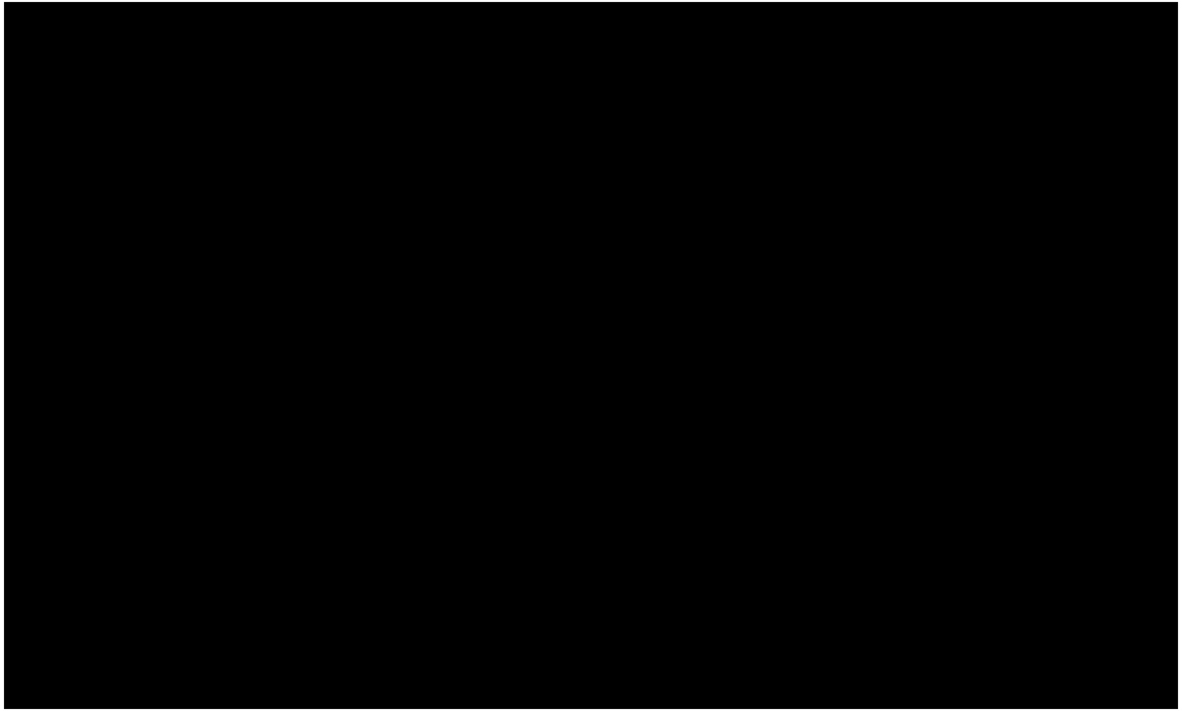
- Continuous diffusion suffers on discrete data [Dieleman et al., 22; Gulrajani et al., 23]
- (We will show) discrete diffusion models are effective for inherently continuous data



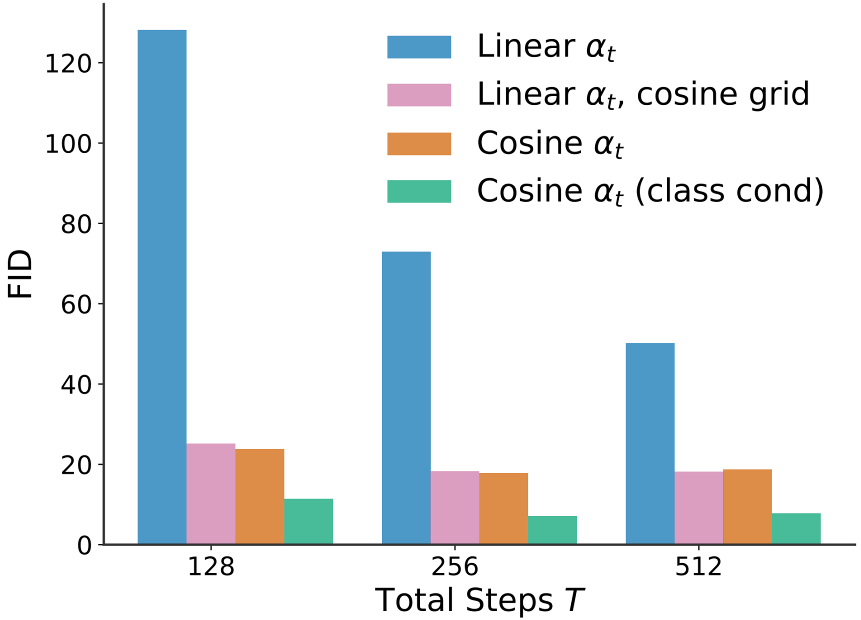
Pixel-level Image Modeling

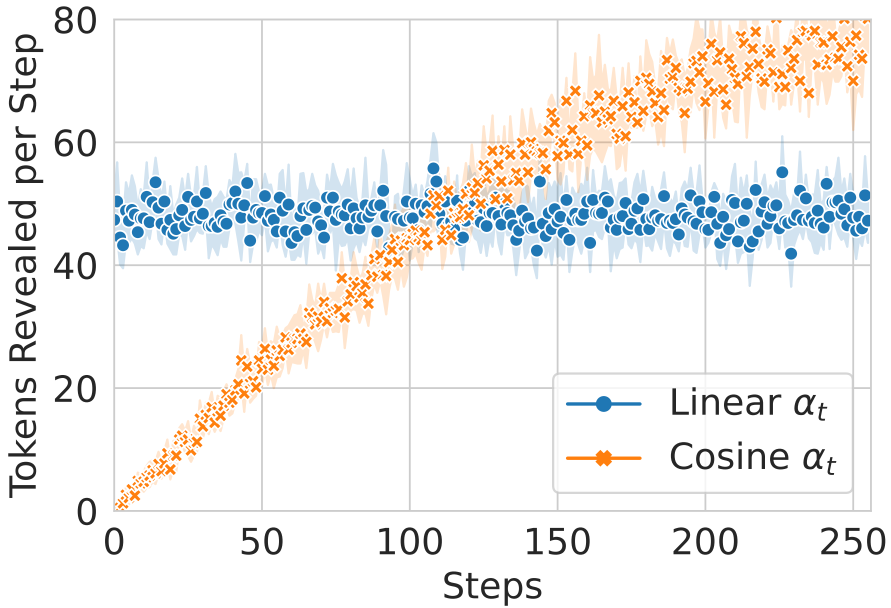


Pixel-level Image Modeling



Sampling





- The masking schedule controls the the quantity of simultaneously predicted tokens.
- The cosine schedule that gradually increases parallel predictions works best.
- For linear schedule, using the cosine grid has the same effect: $t(i) = \cos\left(\frac{\pi}{2}(1 - \frac{i}{T})\right)$

Any-order Generation

Conditional text generation

MD4-M linear schedule	skydiving is a fun sport, but it’s pretty risky. You’re getting is one to get last one for the season if something goes wrong and it can happen you know, we know about season, especially in Skydiving, but anybody that wins this year	Then some time on Saturday you should pretty much say: "This is what I am going to be doing right now." It's just the simplest thing—that is why I always shampoo twice a day and shower three times a day.
MD4-M cosine schedule	skydiving is a fun sport, but it’s extremely risky. You can have so many injuries one time and then one next time. There are so many ways you can hurt, so, neuroconcussions, especially from Skydiving, are continuing to rise every year	Though antibacterial products are a poison, the skin needs a chemical solution that protects it from bacteria and spots that form within it — that is why I always shampoo twice a day and shower three times a day.

Concurrent Work

Simple and Effective Masked Diffusion Language Models

Subham Sekhar Sahoo
Cornell Tech, NYC, USA.
ssahoo@cs.cornell.edu

Marianne Arriola
Cornell Tech, NYC, USA.
ma2238@cornell.edu

Yair Schiff
Cornell Tech, NYC, USA
yzs2@cornell.edu

Aaron Gokaslan
Cornell Tech, NYC, USA.
akg87@cs.cornell.edu

Edgar Marroquin
Cornell Tech, NYC, USA.
emm392@cornell.edu

Justin T Chiu
Cornell Tech, NYC, USA.
jtc257@cornell.edu

Alexander Rush
Cornell Tech, NYC, USA.
ar459@cornell.edu

Volodymyr Kuleshov
Cornell Tech, NYC, USA.
kuleshov@cornell.edu

Your Absorbing Discrete Diffusion Secretly Models the Conditional Distributions of Clean Data

Jingyang Ou[1] **Shen Nie**[1] **Kaiwen Xue**[1] **Fengqi Zhu**[1]
Jiacheng Sun[2] **Zhenguo Li**[2] **Chongxuan Li** [1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2] Huawei Noah's Ark Lab
{oujingyang, nieshen,kaiwenxue,chongxuanli}@ruc.edu.cn;
fengqizhu@whu.edu.cn;{sunjiacheng1,li.zhenguo}@huawei.com;

Takeaways

- Masked diffusion model is a promising candidate for wc models that can reason in any modality and direction
- MD4 is as simple as training an ensemble of BERTs.
- GenMD4 allows state-dependent unmasking behaviors
- Many exciting avenues for future research (e.g., improv sampling speed & quality)

Simplified and Generalized Masked Diffusion for Discrete Data

Jiaxin Shi*, Kehang Han*, Zhe Wang, Arnaud Doucet, Michalis K. Titsias
Google DeepMind

Abstract

Masked (or absorbing) diffusion is actively explored as an alternative to autoregressive models for generative modeling of discrete data. However, existing work in this area has been hindered by unnecessarily complex model formulations and unclear relationships between different perspectives, leading to suboptimal parameterization, training objectives, and ad hoc adjustments to counteract these issues. In this work, we aim to provide a simple and general framework that unlocks the full potential of masked diffusion models. We show that the continuous-time variational objective of masked diffusion models is a simple weighted integral of cross-entropy losses. Our framework also enables training generalized masked diffusion models with state-dependent masking schedules. When evaluated by perplexity, our models trained on OpenWebText surpass prior diffusion language models at GPT-2 scale and demonstrate superior performance on 4 out of 5 zero-shot language modeling tasks. Furthermore, our models vastly outperform previous discrete diffusion models on pixel-level image modeling, achieving 2.75 (CIFAR-10) and 3.40 (ImageNet 64×64) bits per dimension that are better than autoregressive models of similar sizes. Our code is available at https://github.com/google-deepmind/md4.

Paper: arxiv.org/abs/2406.04329

Code: https://github.com/google-deepmind/md4

Slides: jiaxins.io

Kehang Han

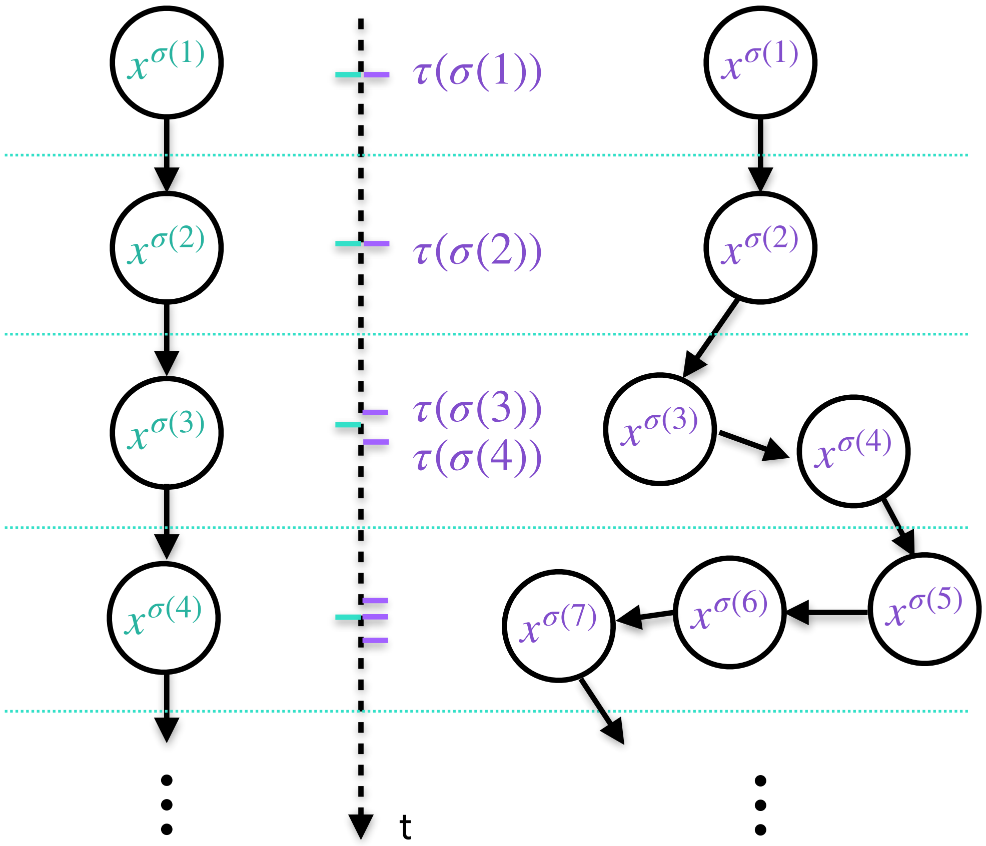
Zhe Wang

Arnaud Doucet

Michalis K. Titsias

Appendix

MD4 as Parallel Any-Order AR Models



A new dimension of freedom in AO-ARMs

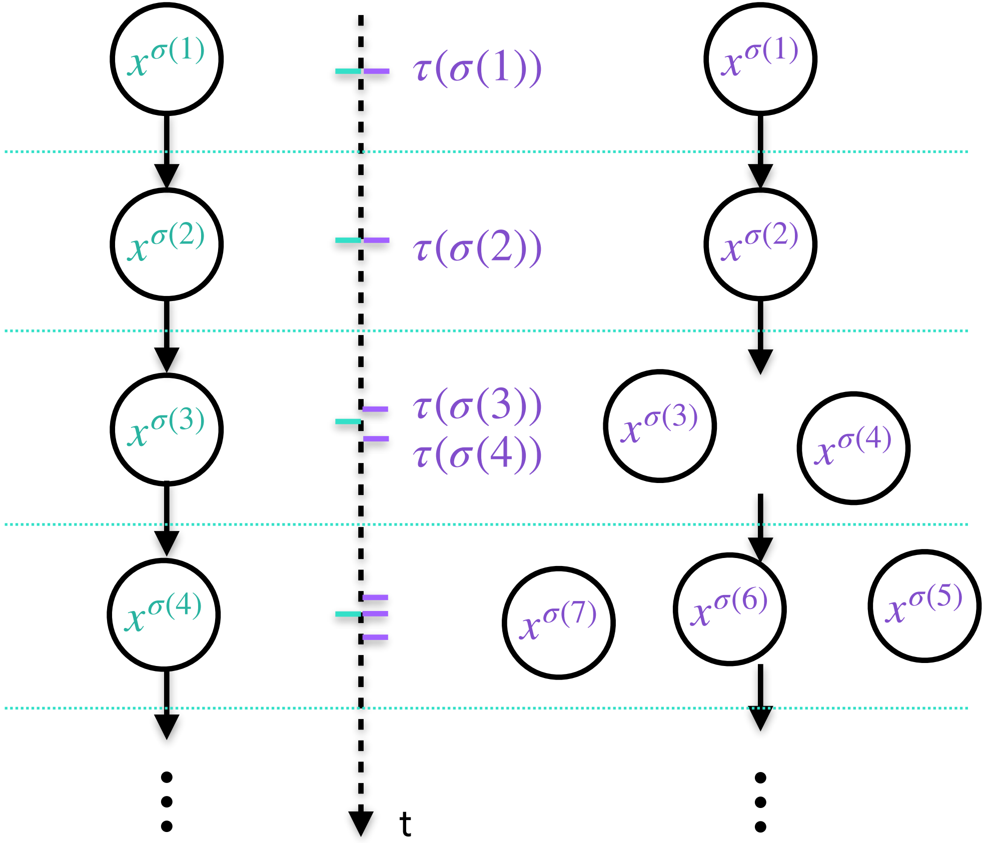
- Masking schedules control parallel sampling bandwidth

CDF of the jump times:

$$P(\tau(n) \leq t) = P(x_t^{(n)} = m) = 1 - \alpha_t$$

Uria, B. et al. (2014). A deep and tractable density estimator.
Hoogeboom et al. (2021). Autoregressive diffusion models.

MD4 as Parallel Any-Order AR Models



A new dimension of freedom in AO-ARMs

- Masking schedules control parallel sampling bandwidth

CDF of the jump times:

$$P(\tau(n) \leq t) = P(x_t^{(n)} = m) = 1 - \alpha_t$$

Uria, B. et al. (2014). A deep and tractable density estimator.
Hoogeboom et al. (2021). Autoregressive diffusion models.