Google DeepMind

Discrete Generative Modeling with Masked Diffusions

Jiaxin Shi
Google DeepMind
2024/9/18 @GenU 2024

jiaxins.io

Why Discrete Diffusion Models

- Generating discrete data with parallel sampling

Why Discrete Diffusion Models

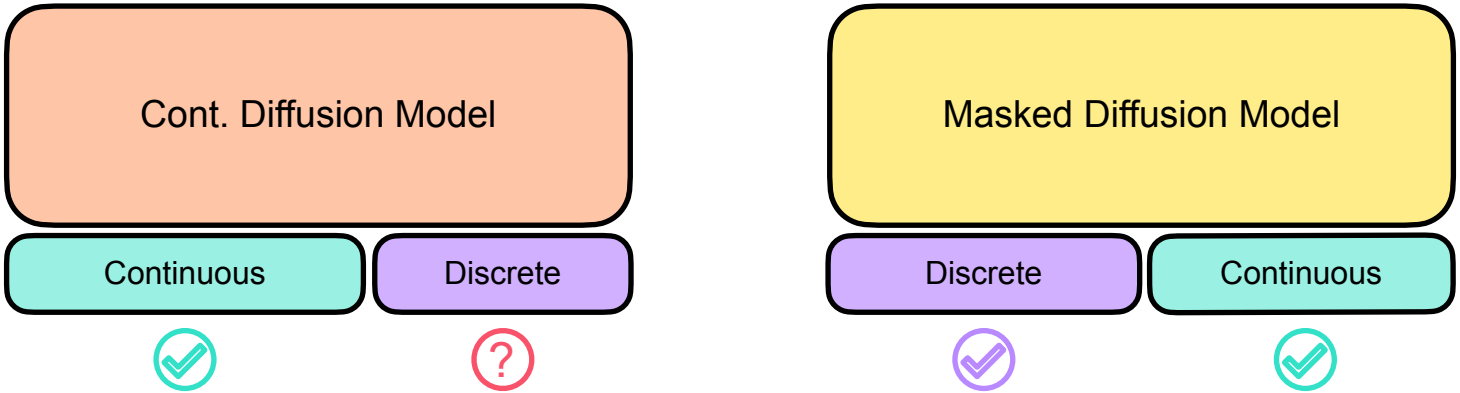
- Generating discrete data with parallel sampling
- Any-order reasoning
  - A world model should facilitate reasoning in any directions
  - AR models are built to follow a fixed order

This makes teaching Chinese literature, teaching Du Fu so much easier," Ling explained. "We all teach literature, but the tradition is different. Having that, it makes the teaching a much more collaborative idea." Wai-Yee Li, another professor of Chinese Literature, also lauded the translation. "It's definitely not normal for Chinese speakers to have the same chance to be exposed to books of a different language," he said.

Infilling with our 400M masked diffusion model

Why Discrete Diffusion Models

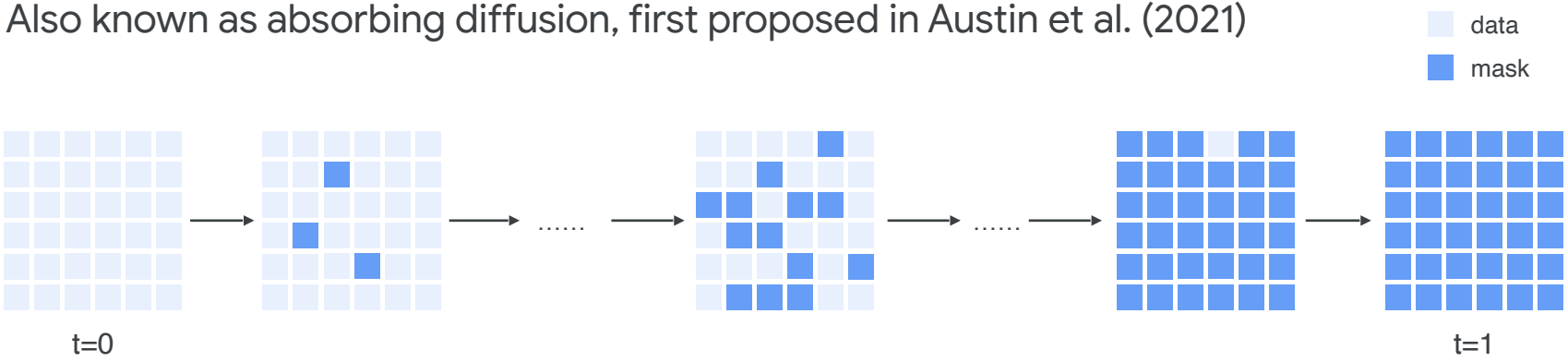
- Generating discrete data with parallel sampling
- Any-order reasoning
- Unification of modalities
  - Continuous diffusion suffers on discrete data [Dieleman et al., 22; Gulrajani et al., 23]
  - (We will show) discrete diffusion can work as well on inherently cont. data

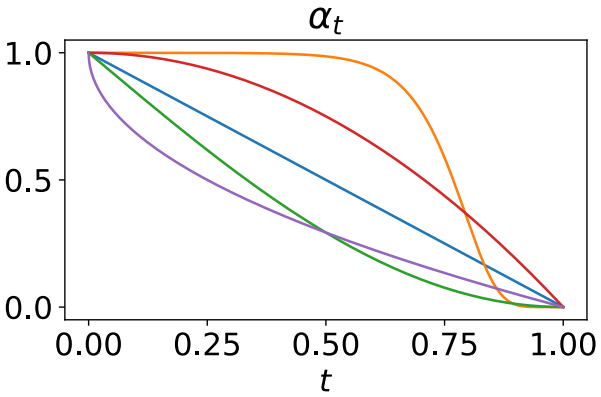


Dieleman et al. (2022). Continuous diffusion for categorical data.
Gulrajani & Hashimoto (2024). Likelihood-based diffusion language models.

Masked Diffusion

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



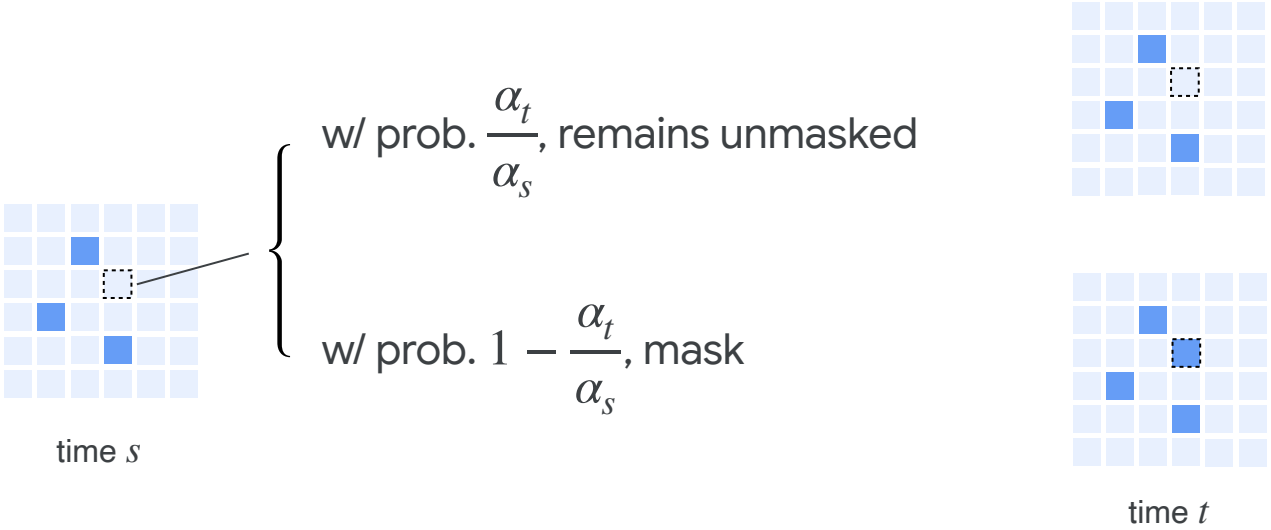


Masking schedule α_t: The expected proportion of unmasked tokens at t

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

Masked Diffusion

Forward process $q(x_t \mid x_s)$

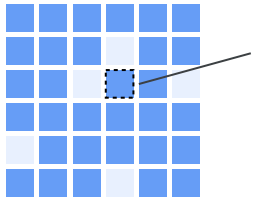


Transition matrix: $\bar{Q}(s,t)_{ij} \triangleq q(x_t = j \mid x_s = i)$

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

Masked Diffusion

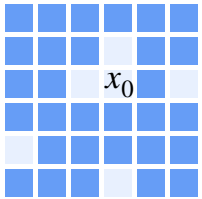
Reverse process $q(x_s \mid x_t, x_0)$ (knowing x_0)

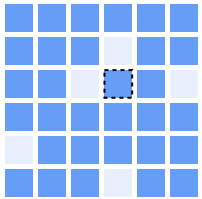


time t

w/ prob. $\dfrac{\alpha_s - \alpha_t}{1 - \alpha_t}$, unmask

w/ prob. $\dfrac{1 - \alpha_s}{1 - \alpha_t}$, remains masked



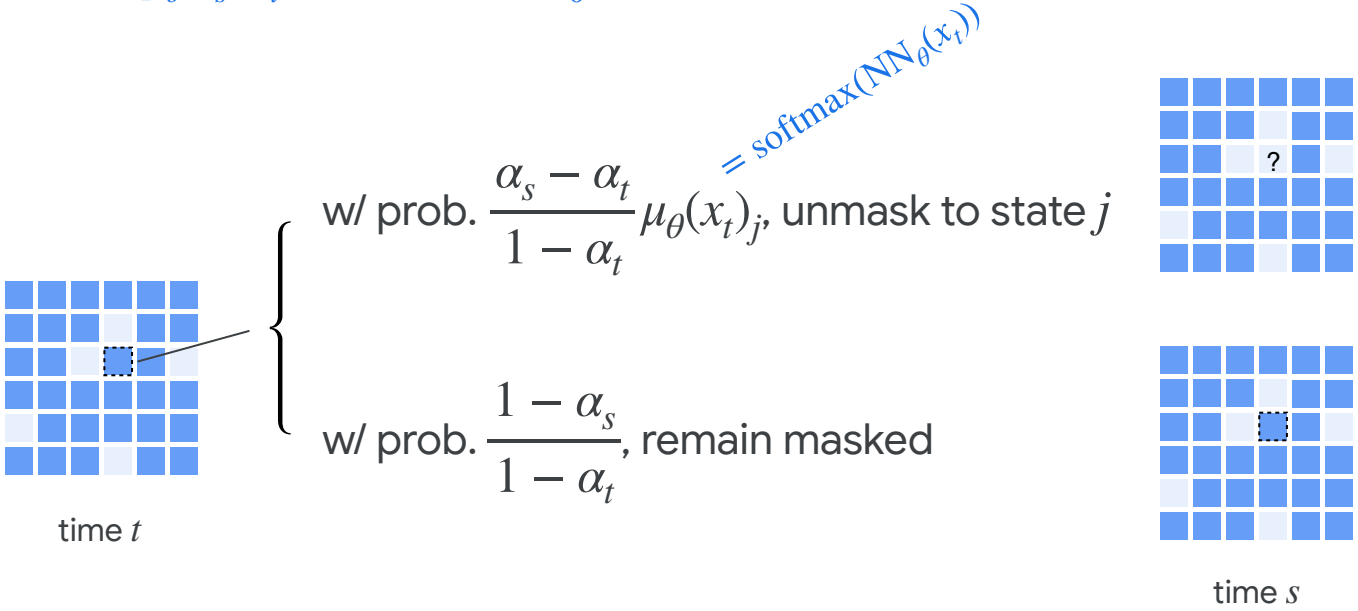


time s

Transition matrix $\bar{R}^{x_0}(t,s) = I + \dfrac{\alpha_s - \alpha_t}{1 - \alpha_t} e_m (x_0 - e_m)^\top$

Masked Diffusion Models

Generative model $p_\theta(x_s \mid x_t)$ (not knowing x_0)



Mean Parameterization: $\mu_\theta(x_t)$ as a prediction model for $\mathbb{E}[x_0 \mid x_t]$

MD4 Objective: Weighted Cross-Entropy Losses

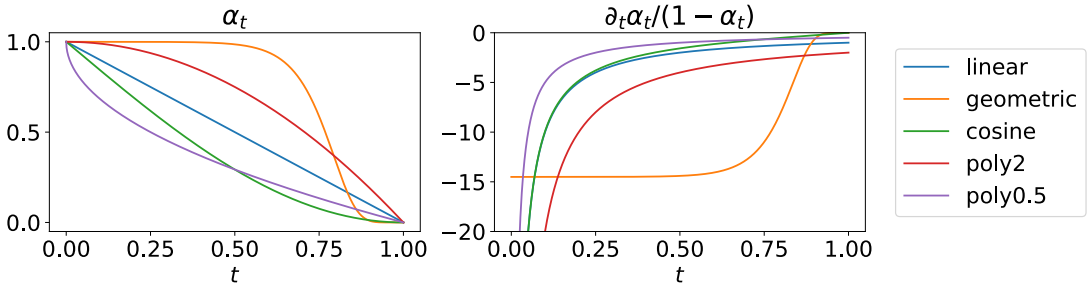
Discrete-time Evidence Lower Bound (ELBO)

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_{t(1)}|x_0)}[\log p(x_0 \,|\, x_{t(1)})] - \mathrm{KL}(q(x_{t(T)} \,|\, x_0) \| p(x_{t(T)}) - \mathscr{L}_T$$

$$\mathscr{L}_T = \sum_{i=2}^{T} \mathbb{E}_{q(x_{t(i)}|x_0)}[\mathrm{KL}(q(x_{s(i)} \,|\, x_{t(i)}, x_0) \| p_\theta(x_{s(i)} \,|\, x_{t(i)})]$$

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

$$\mathscr{L}_\infty = \int_0^1 \frac{\alpha_t'}{1-\alpha_t} \mathbb{E}_{q(x_t|x_0)}[\delta_{x_t,m} \cdot x_0^\top \log \mu_\theta(x_t, t)] dt$$



GenMD4: State-dependent Schedules

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

Before

$$\alpha_t : [0,1] \to [0,1]$$

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

After

$$\alpha_t : [0,1] \to [0,1]^{|V|}$$

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

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

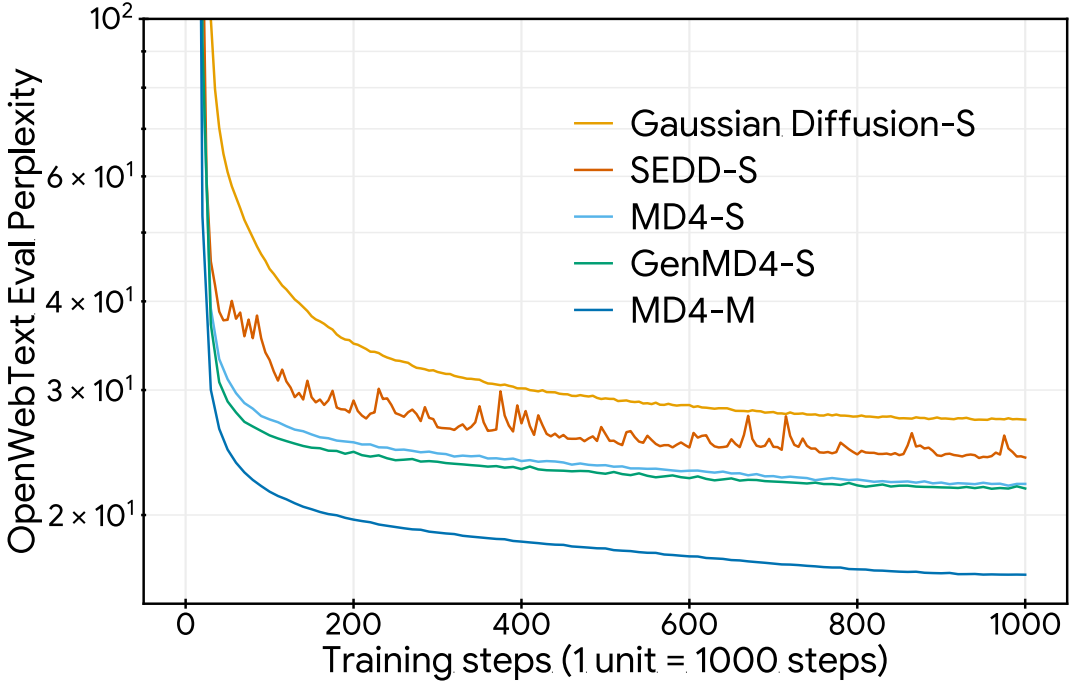
$$\mathcal{L}_\infty = \int_0^1 \left(\frac{\alpha_t'}{\mathbf{1} - \alpha_t}\right)^\top \mathbb{E}_{q(x_t|x_0)} \left[\delta_{x_t,m} \cdot (x_0 - \mu_\theta(x_t,t) + x_0 x_0^\top \log \mu_\theta(x_t,t))\right] \mathrm{d}t$$

Perplexity on GPT-2 Zero-Shot Eval

Table 1: Zero-shot unconditional perplexity on five benchmark datasets from Radford et al. [43]. The numbers for other methods are from Lou et al. [32] except our reimplementation of SEDD Absorb. Our MD4 model achieves the best result on all benchmarks except LAMBADA where it is the second best. *The GPT-2 numbers are reported for the GPT-2 checkpoint pretrained on WebText instead of OWT thus is not a direct comparison.

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**

- Gaussian diffusion (Plaid reimpl.) is worse than discrete diffusion
- Training with SEDD [Lou et al., 23] is unstable due to the inconsistency of forward/ backward processes

Masking Schedules Learned by GenMD4

Schedule for token type i: $(\alpha_t)_i = 1 - t^{w_i}$

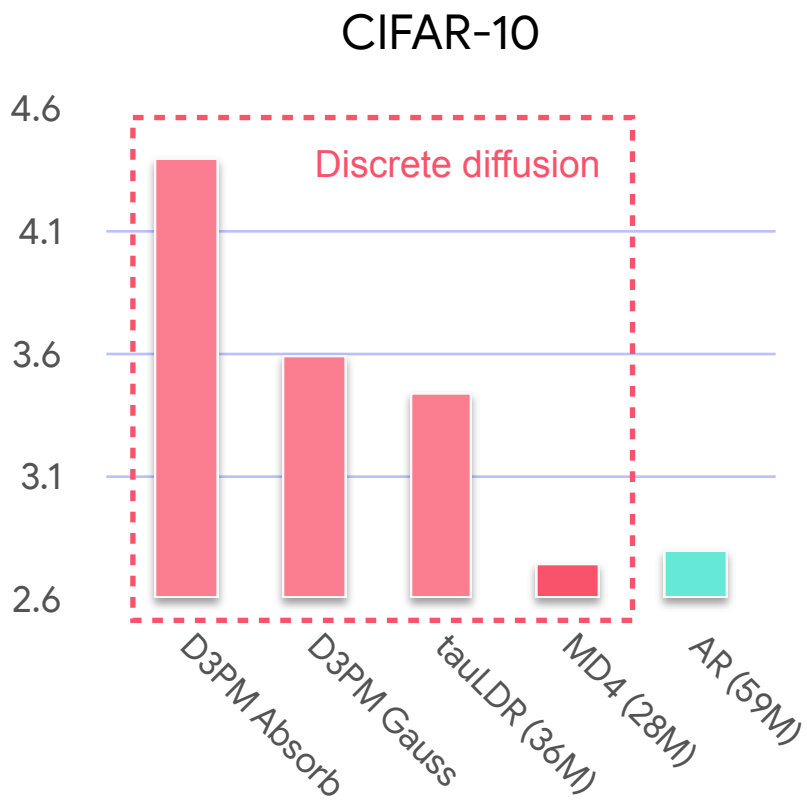
Token types with largest ws (unmask first)

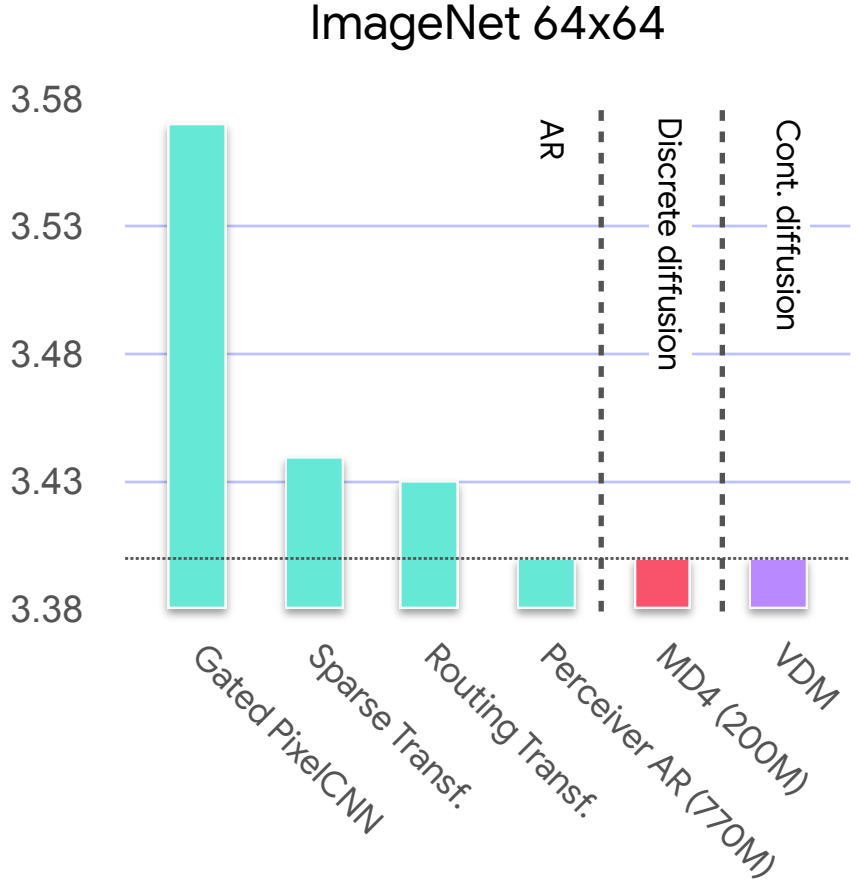
'<|endoftext|>',
'\n',
'.',
' (',
'-',
' "',
',',
' strutConnector',
' \xa0\xa0',
' DevOnline'

Token types with smallest ws

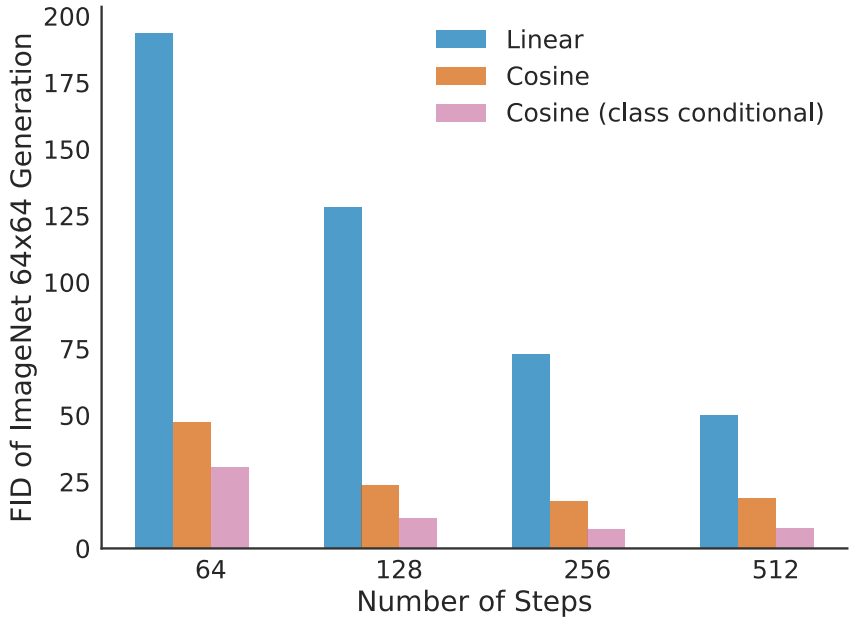
' diligently',
' unreliable',
' irresistible',
' dart',
' tracing',
' enlarged',
' playful',
' freeing',
' weighted',
'407'

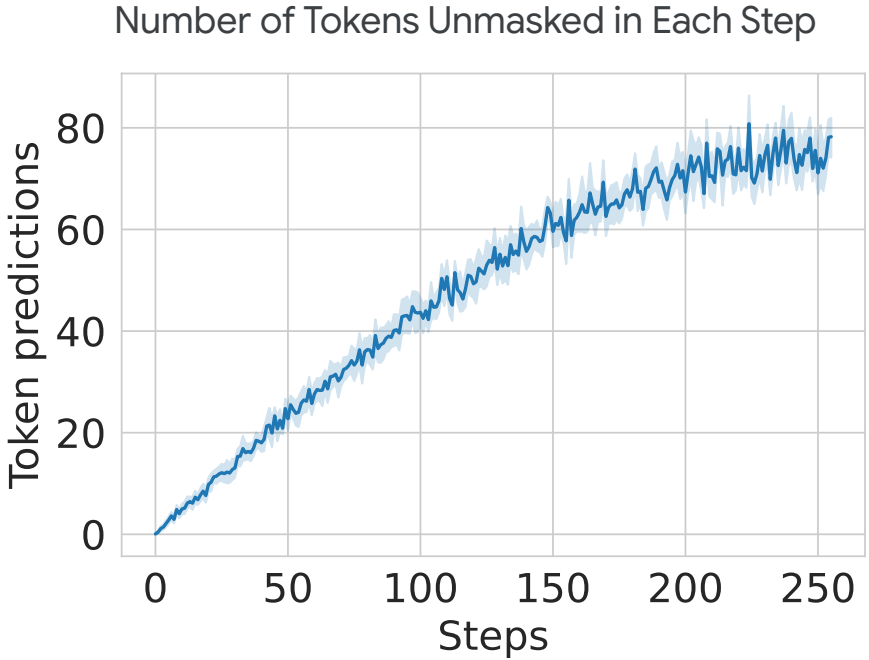
Pixel-level Image Modeling





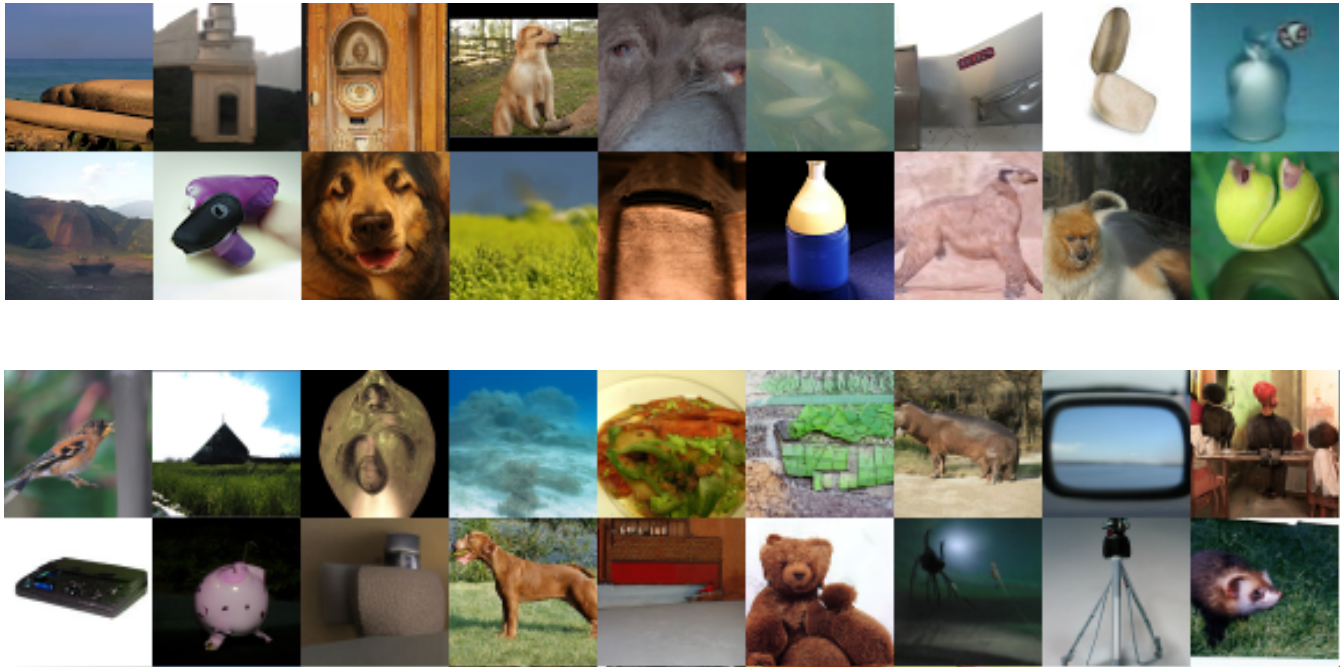
Sampling





- The masking schedule controls the the quantity of simultaneously predicted tokens.
- The cosine schedule that gradually increases parallel predictions works best.

ImageNet 64x64 Samples



Conditional Sampling (400M)

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.

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
- Applying discrete “integration-by-part” fixes 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.

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 world models that can reason in any modality & order.
- MD4 & GenMD4 make it simple, performant and scalable.
- MD4 provides a new perspective on discrete diffusion & any-order AR models
  - Masking schedule as a new degree of freedom that enables effective parallel sampling

Paper:

Simplified and Generalized Masked Diffusion for Discrete Data

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

Slides: jiaxins.io

Kehang Han

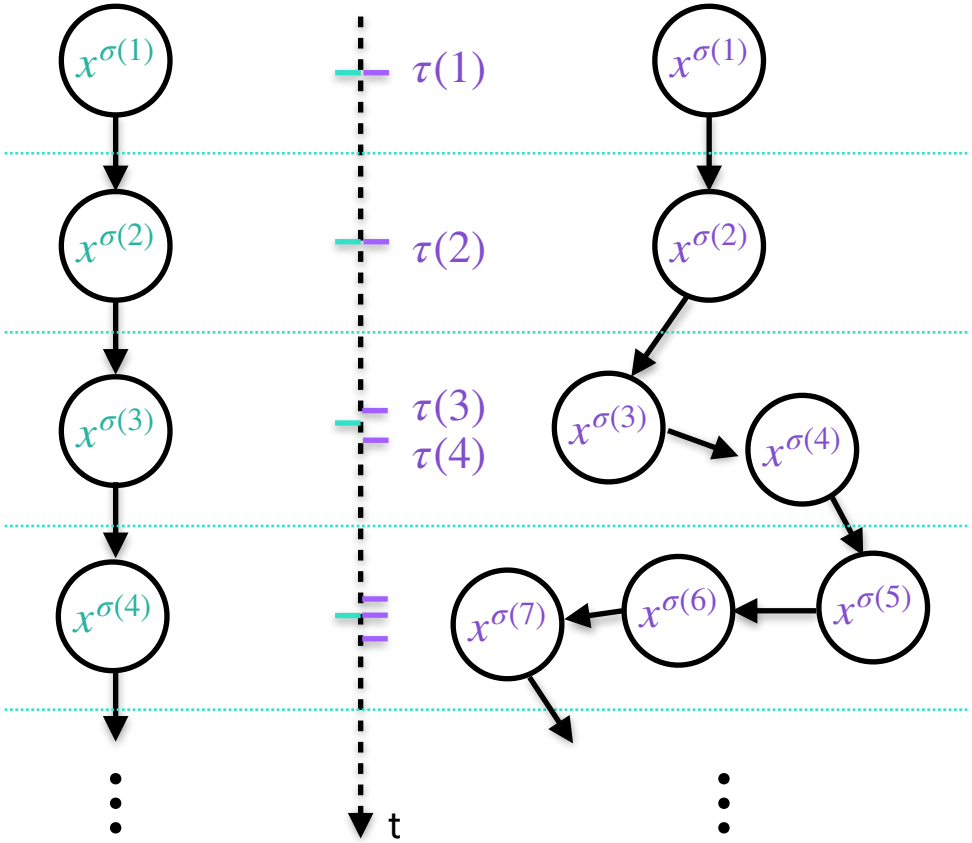
Zhe Wang

Arnaud Doucet

Michalis K. Titsias

Appendix

Are Masked Diffusion Just Any-Order AR Models?



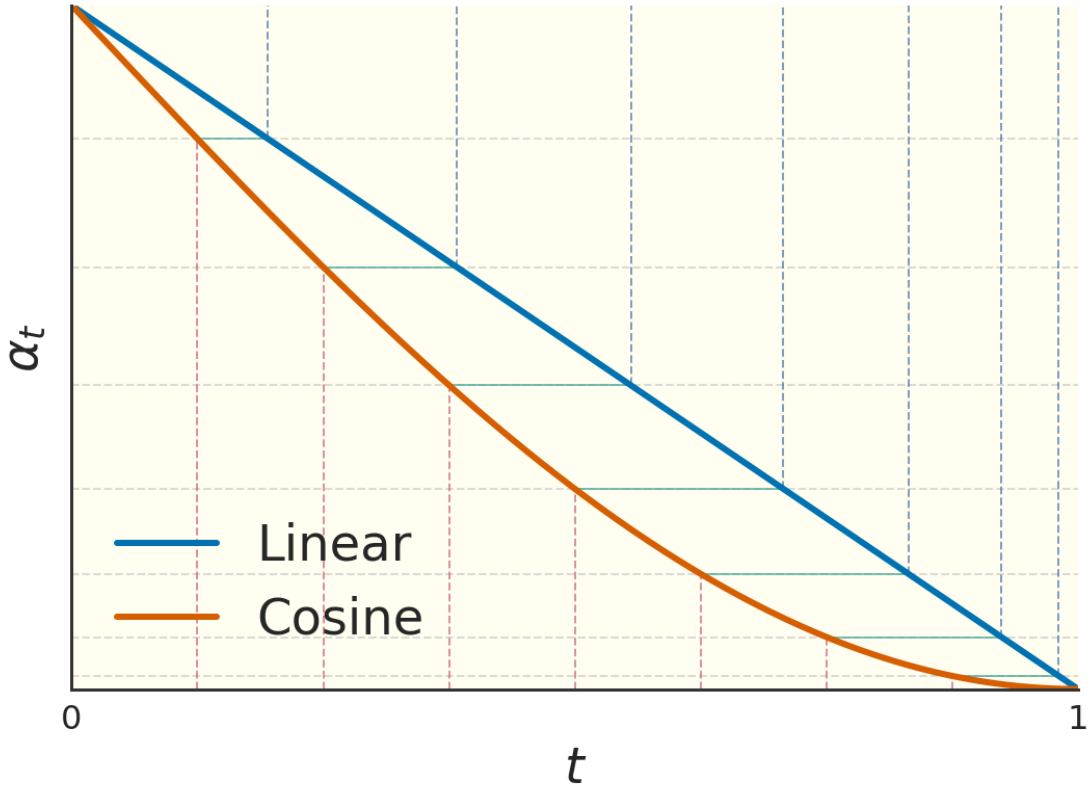
Yes, but a new dimension of freedom

- Masking schedules control parallel sampling bandwidth

CDF of the jump times:

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

Masking Schedules



Score v.s. Mean Parameterization

Proposition 1. The discrete score $s(x_t, t)_j = \dfrac{q_t(j)}{q_t(x_t)}$ 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)

mean parameterization fixes the problem

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

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 in GenMD4

- The reverse process now has a quadratic dependence on x_0

$$q(x_s|x_t = m, x_0, x_0 x_0^\top) = \left(\frac{\mathbf{1}-\alpha_s}{\mathbf{1}-\alpha_t}\right)^\top x_0 e_m^\top x_s + \left(\frac{\alpha_s-\alpha_t}{\mathbf{1}-\alpha_t}\right)^\top x_0 x_0^\top x_s$$

- Fortunately, $\mathbb{E}[x_0 x_0^\top | x_t] = \mathrm{diag}(\mathbb{E}[x_0|x_t])$ so we can reuse the mean prediction model to parameterize $p_\theta(x_s \,|\, x_t)$

Text8 Benchmark

Method	BPC (↓)
Continuous Diffusion	
Plaid [22] (Our impl.)	≤ 1.48
BFN [26]	≤ 1.41
Any-order Autoregressive	
ARDM [48]	≤ 1.43
MAC [49]	≤ 1.40
Autoregressive	
IAF/SCF [50]	1.88
AR Argmax Flow [15]	1.39
Discrete Flow [51]	**1.23**
Transformer AR [14]	**1.23**
Discrete Diffusion	
Mult. Diffusion [15]	≤ 1.72
D3PM Uniform [14]	≤ 1.61
D3PM Absorb [14]	≤ 1.45
SEDD Absorb [32]	≤ 1.41[5]
MD4 (Ours)	$\leq$ **1.37**
GenMD4 (Ours)	$\leq$ **1.34**

- A (very) challenging character-level language modeling benchmark
- SEDD previously reported 1.32 but turns out incorrect
- We were the best among diffusion and any-order AR
- Still a gap between diffusion and AR

The CTMC View

Get transition rate matrices from our results:

$$\bar{Q}(t, t+\Delta t) = I + Q(t)\Delta t + o(\Delta t) \quad \text{for} \quad Q(t) \triangleq \beta(t)(\mathbf{1}e_m^\top - I),$$

$$\bar{R}^{x_0}(t, t-\Delta t) = I + R^{x_0}(t)\Delta t + o(\Delta t) \quad \text{for} \quad R^{x_0}(t) \triangleq -\frac{\alpha_t'}{1-\alpha_t} e_m (x_0 - e_m)^\top$$

Plugging this into $\mathscr{L}_\infty$ and applying discrete "integration-by-part" recovers Campbell et al. (2022)'s loss

$$-\int_{t(1)}^{1} \mathbb{E}_{q_{t|0}(k|x_0)} \left[R_\theta(t)_{kk} + \sum_{j \neq k} Q(t)_{kj} \log R_\theta(t)_{jk} \right] \mathrm{d}t + \mathbf{C}$$

Problems of this expression:

- It needs $|V|$ NN passes to compute the inner sum
- Estimation via sampling j has high variance

Infilling Example (400M)

MD4-M linear schedule	This makes teaching Chinese literature, teaching Du Fu so much easier," Ling Fu said. "The vocabulary is different, the tradition is different. Having that, it gives the ability to teach Du Fu language." Wai-Yee Li, another professor of Chinese Literature, also lauded the translation. He recommended Du Fu books as a guest lecturer several years ago.
MD4-M cosine schedule	This makes teaching Chinese literature, teaching Du Fu so much easier," Ling explained. "We all teach literature, but the tradition is different. Having that, it makes the teaching a much more collaborative idea." Wai-Yee Li, another professor of Chinese Literature, also lauded the translation. "It's definitely not normal for Chinese speakers to have the same chance to be exposed to books of a different language," he said.