Google DeepMind

Simplified and Generalized Masked Diffusion for Discrete Data

Jiaxin Shi

Kehang Han

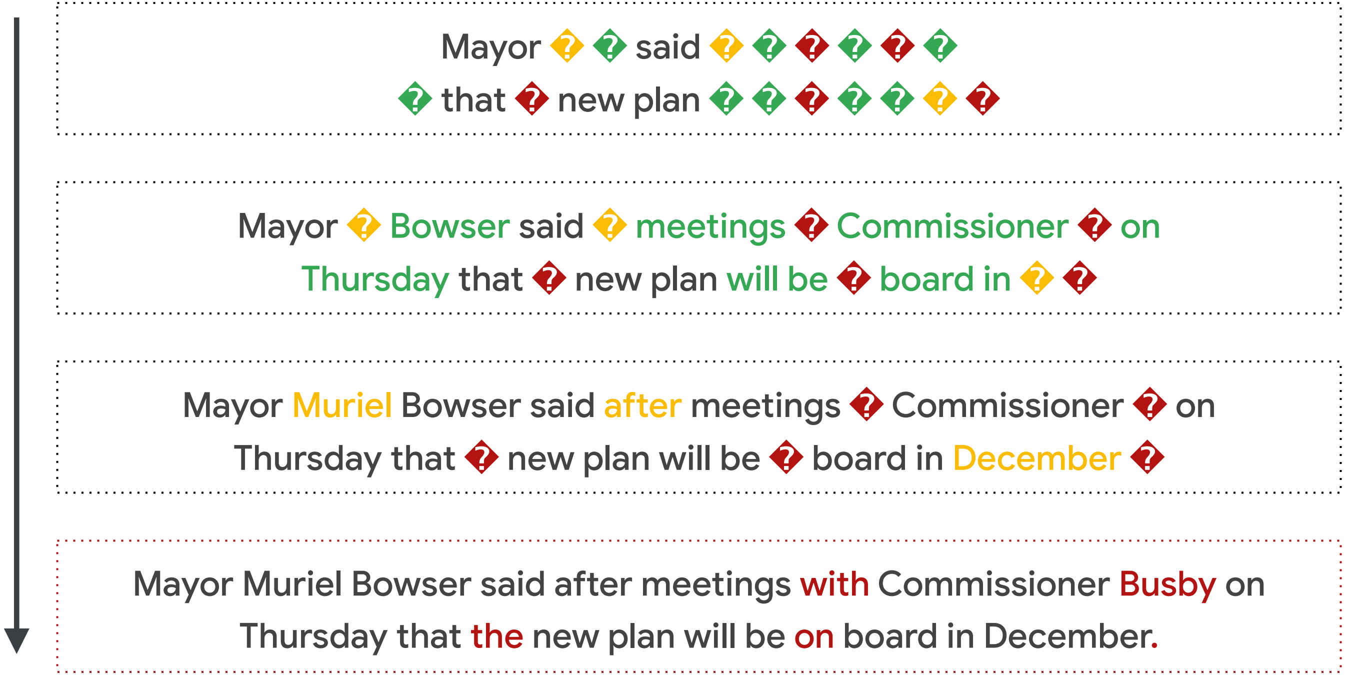
Zhe Wang

Arnaud Doucet

Michalis K. Titsias

Why Diffusion Models for Discrete Data

- Generating discrete data with parallel sampling



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

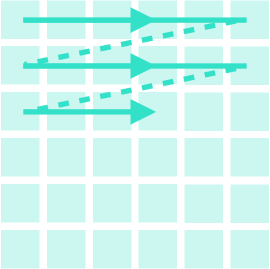
Mayor � Bowser said � meetings � Commissioner � on
Thursday that � new plan will be � board in � �

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

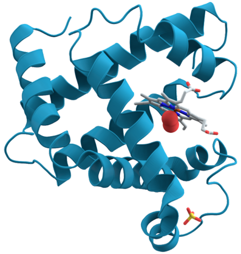
Mayor Muriel Bowser said after meetings with Commissioner Busby on
Thursday that the new plan will be on board in December.

Why Diffusion Models for Discrete Data

- AR models require imposing an ordering which may be unnatural for many data types

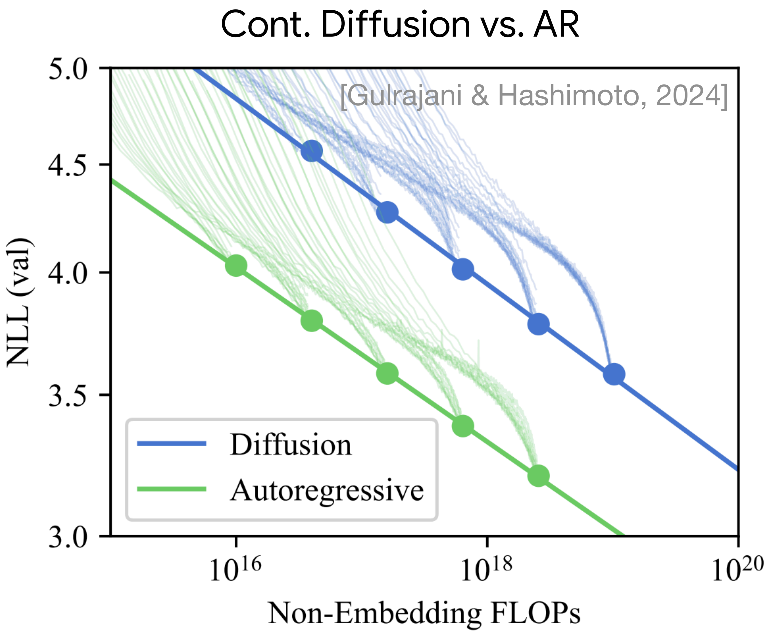


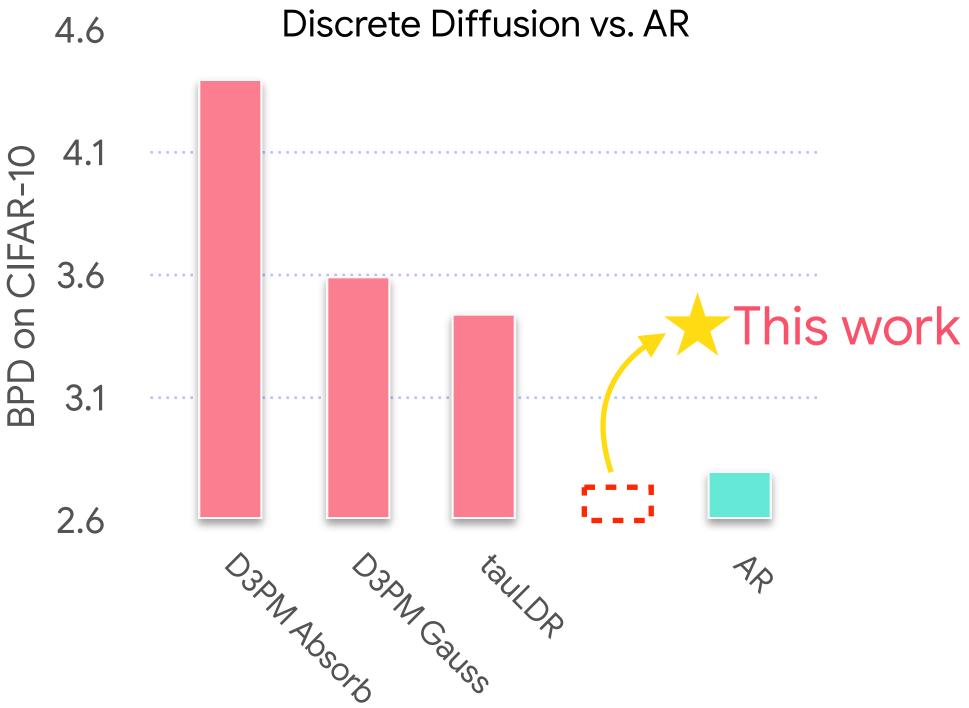
```
HGFGTLEHPIYKVAKQWSMVHDTTVYFSCGLHVAAHPATYVSMQAWKSTNDFPCRQHDNI
TMLYHINMESFVNLEFCNFQTDDKYLEDPWARHEKYPIRKAIKWEGLPNMQRLHMLHWIN
VSMDPNHGPVYCAKWDTILYMGKDGKERRTSAYMFTGVDEQHCRYEYRKFCGKHKAPKLM
GRLFRITKSCWWGCCTLDNMKPDKAKACAEDMRRCRNIPVVQNFQQCGKYWKATSQDNTK
RNSKCRAIEWEIFQYWINCSTVVKTFAPCMFGFQFRFHYGYNYMFWVTIKLSVYRWMPGV
DRETPVHAVNIINIWSAYKMTRYWCRIQCDSYWLWSGMTWRWCRWNREQPEWLSHDDMVQ
CWEGSYKLMFCGWWRHFISKSMVTLGGHKKDDGRRWMLQSTHHLHFPATINIHDDWFPHG
```



Challenge

Diffusion yet to match AR performance on discrete data

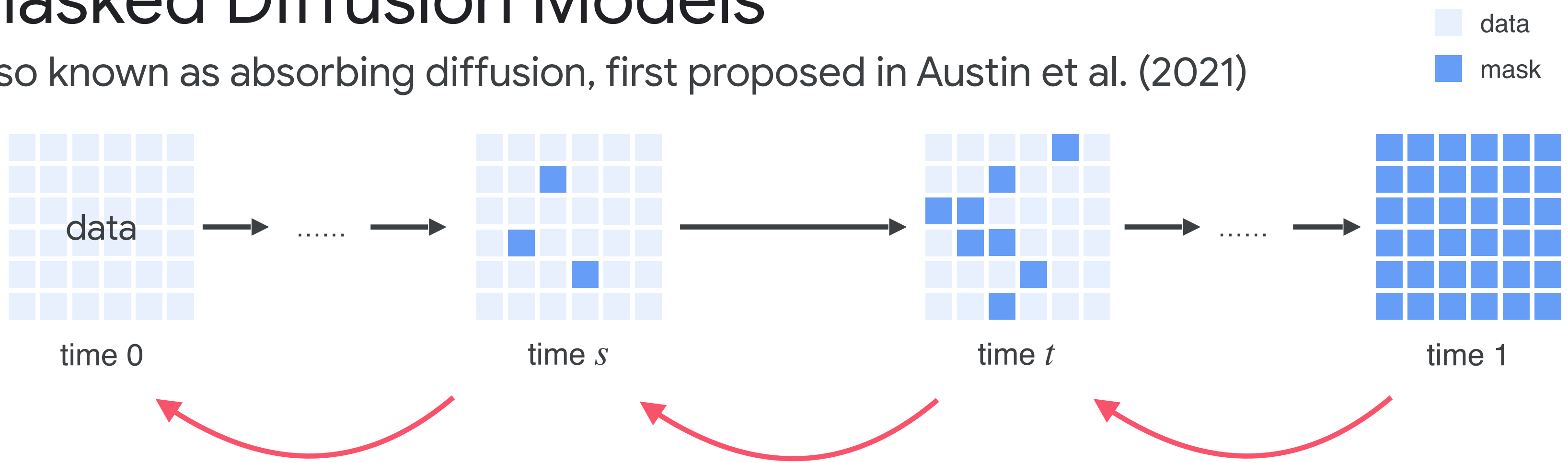


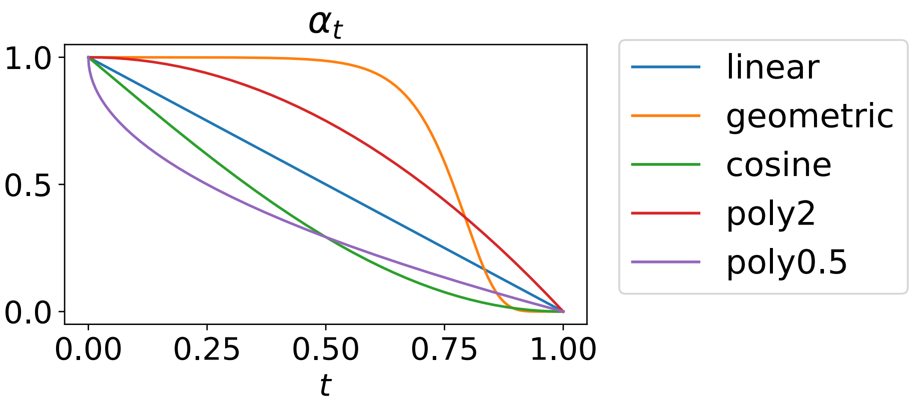


Gulrajani & Hashimoto (2024). Likelihood-based diffusion language models.

Masked Diffusion Models

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

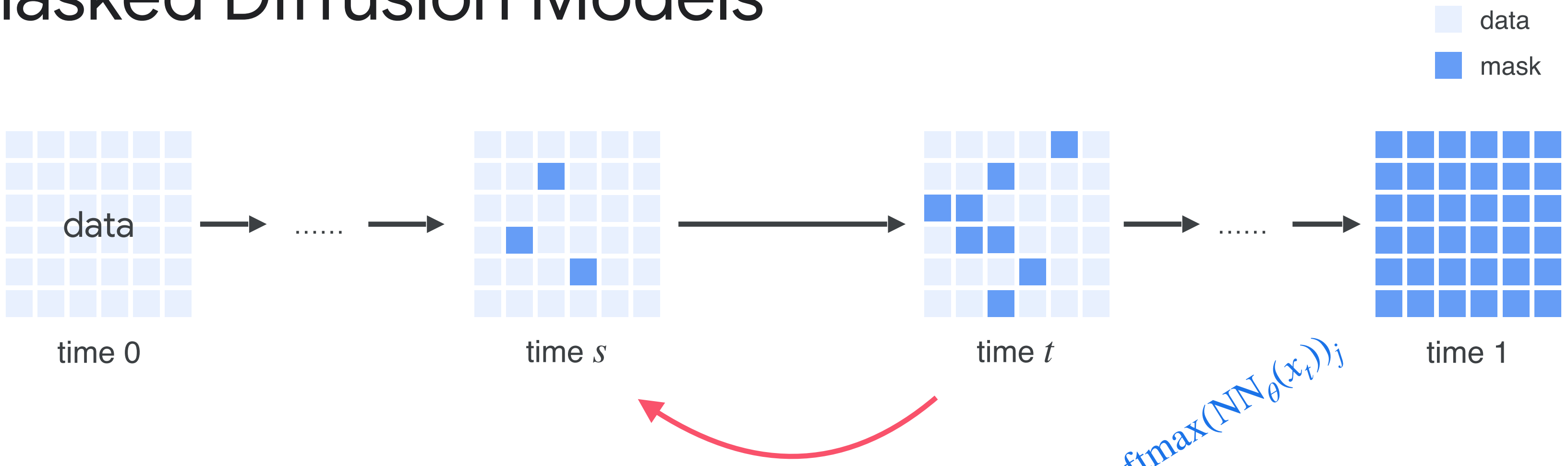




Masking schedule α_t: The expected proportion of unmasked tokens at t

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

Masked Diffusion Models

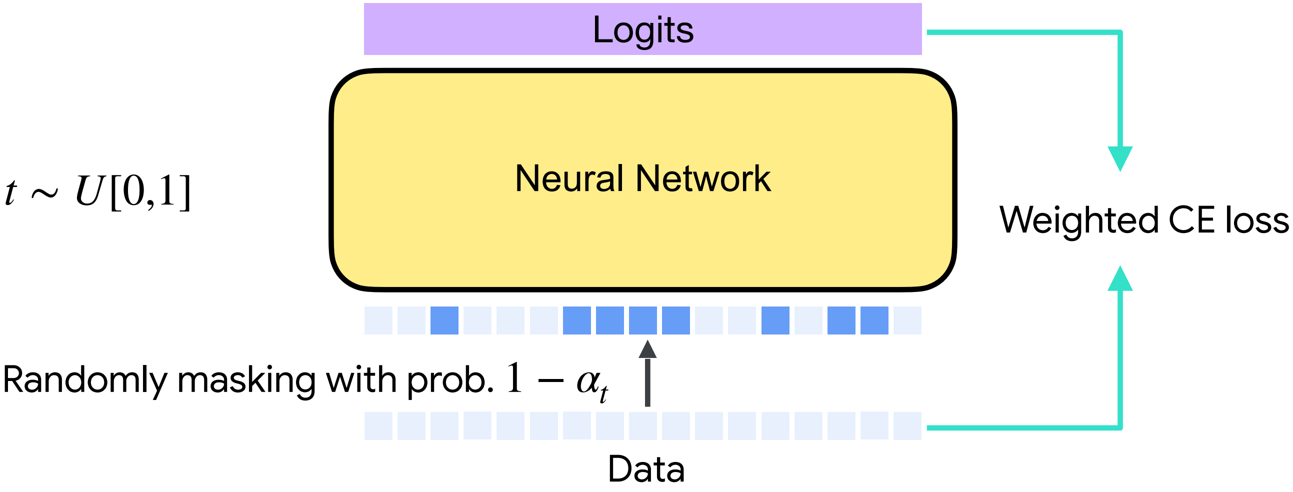


$$
\text{🟦}\ \begin{cases} \text{w/ prob. } \dfrac{\alpha_s - \alpha_t}{1-\alpha_t}\, \overset{\approx\, \mu_\theta(x_t)_j \,\triangleq\, \mathrm{softmax}(\mathrm{NN}_\theta(x_t))_j}{p(x_0 = j \mid x_t)}, \text{ unmask to state } j \ \text{⬜} \\[3em] \text{w/ prob. } \dfrac{1-\alpha_s}{1-\alpha_t}, \text{ remain masked } \text{🟦} \end{cases}
$$

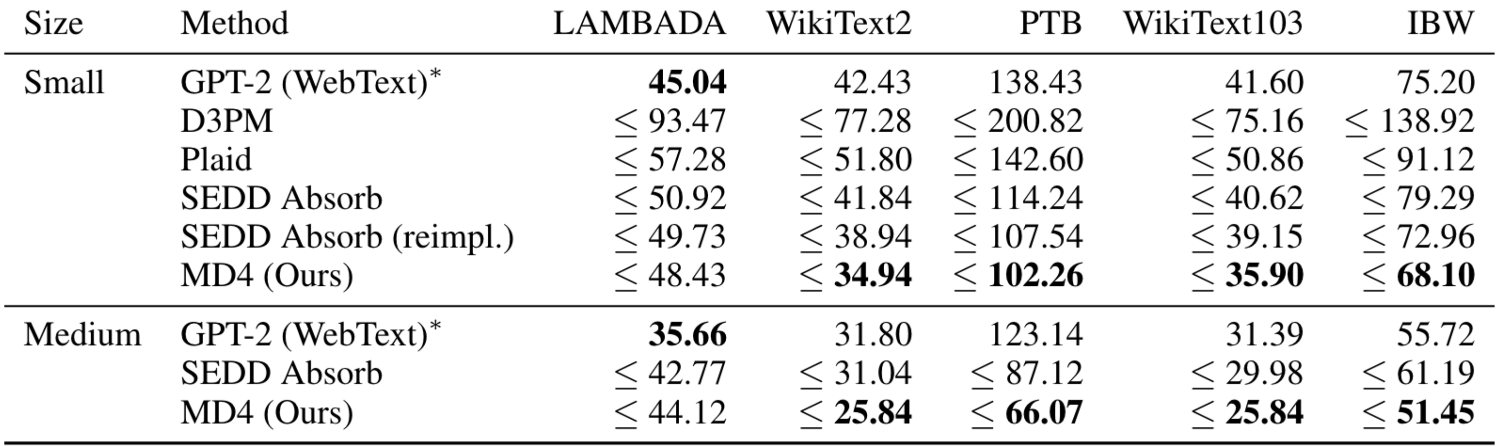
MD4 Objective: Weighted Cross-Entropy Losses

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

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

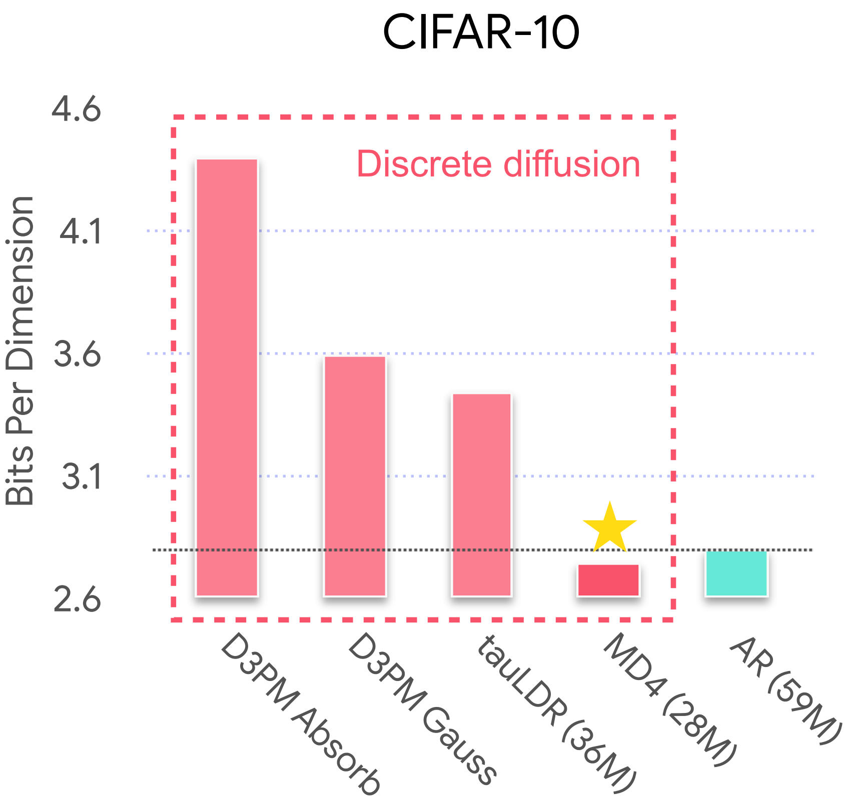


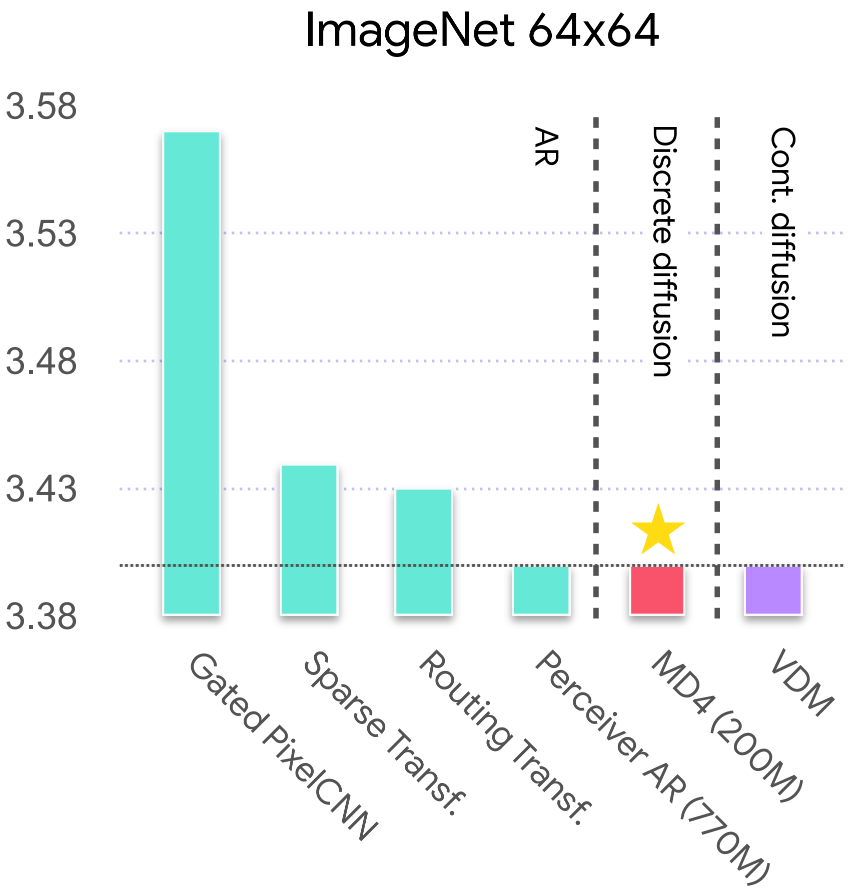
Perplexity on GPT-2 Zero-Shot Eval



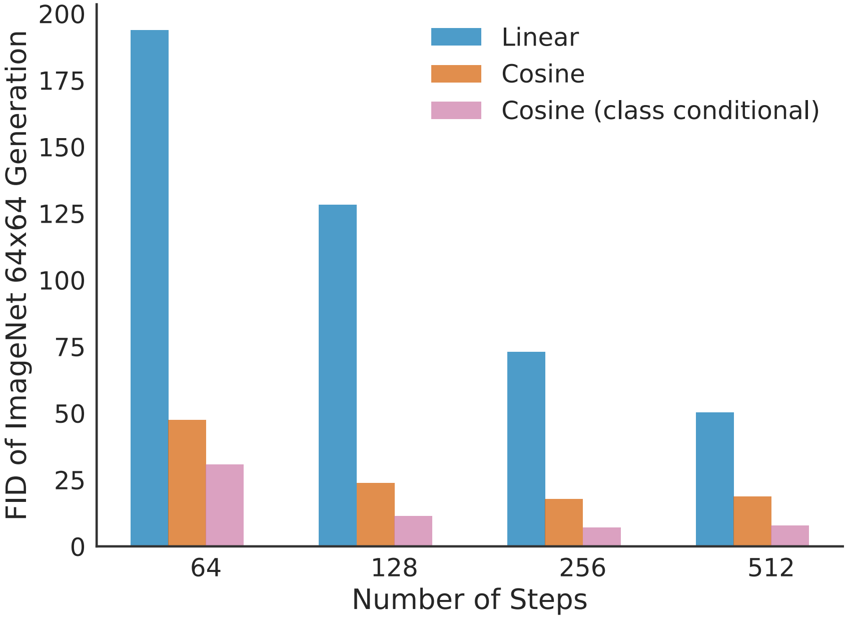
Size	Method	LAMBADA	WikiText2	PTB	WikiText103	IBW
Small	GPT-2 (WebText)*	**45.04**	42.43	138.43	41.60	75.20
	D3PM	$\leq$ 93.47	$\leq$ 77.28	$\leq$ 200.82	$\leq$ 75.16	$\leq$ 138.92
	Plaid	$\leq$ 57.28	$\leq$ 51.80	$\leq$ 142.60	$\leq$ 50.86	$\leq$ 91.12
	SEDD Absorb	$\leq$ 50.92	$\leq$ 41.84	$\leq$ 114.24	$\leq$ 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**

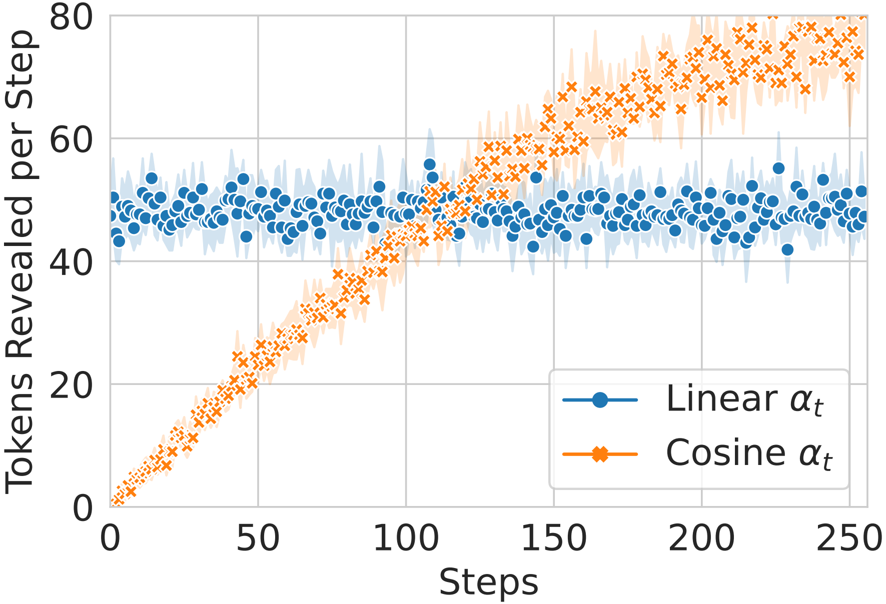
Pixel-level Image Modeling





Sampling

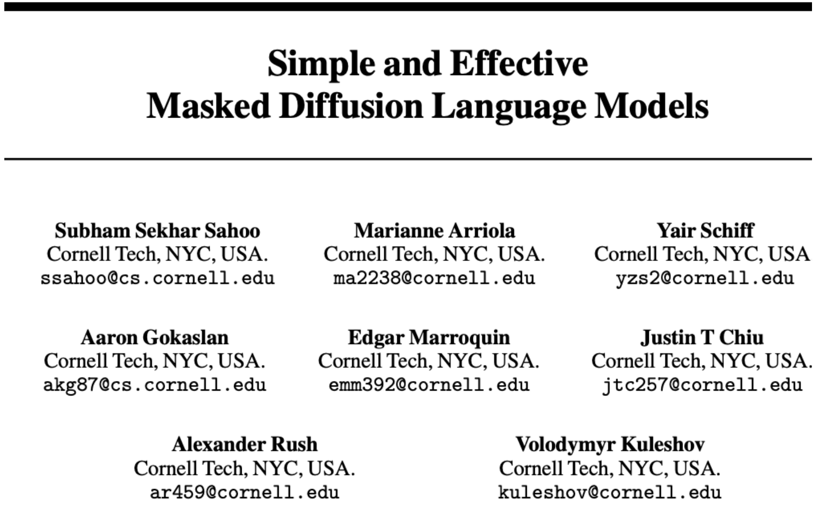




- The masking schedule controls the the quantity of simultaneously predicted tokens.

- The cosine schedule that gradually increases parallel predictions works best.

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

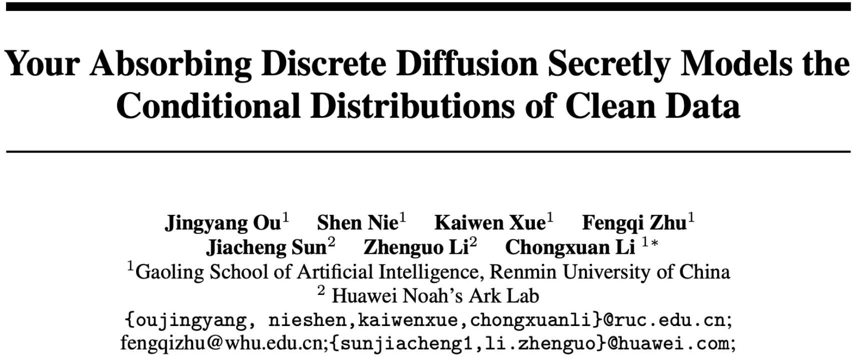
Aaron Gokaslan
Cornell Tech, NYC, USA.
akg87@cs.cornell.edu

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;

Thanks

ImageNet 64x64
unconditional generation

Conditional text generation

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.