Google DeepMind

Discrete Diffusion Models: An Introduction

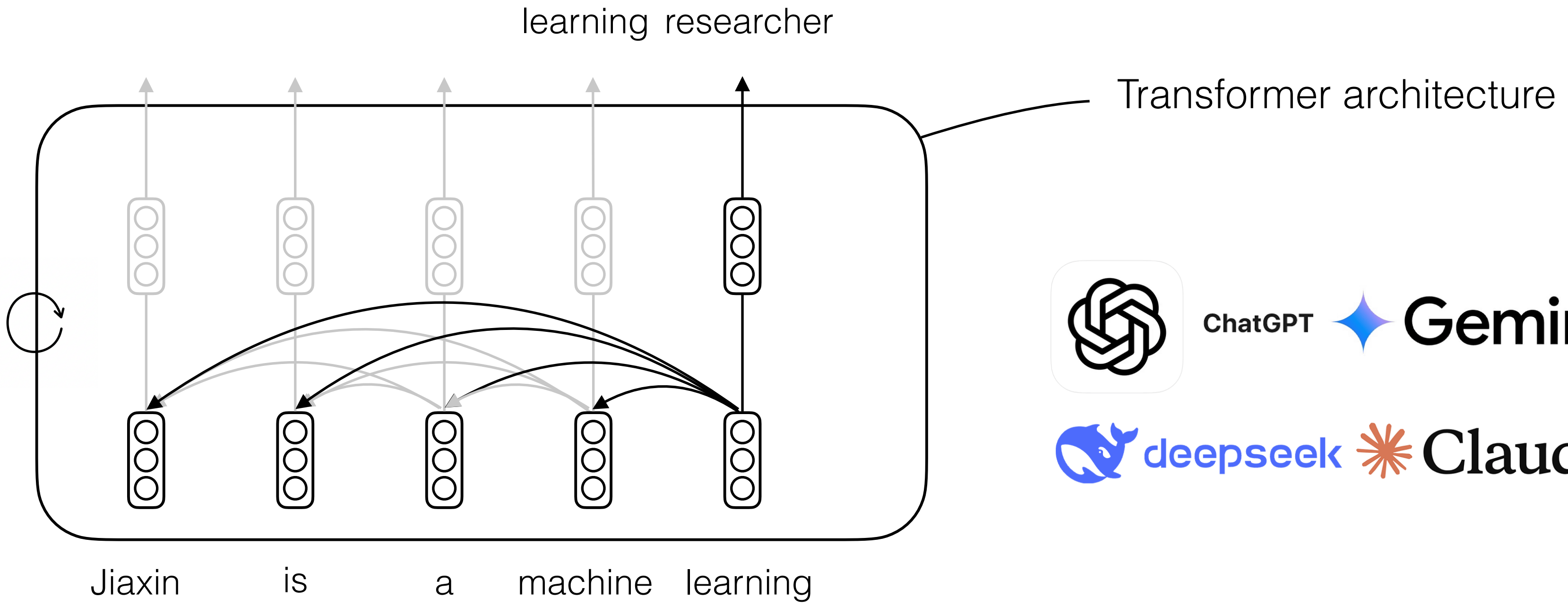
Jiaxin Shi
2025/10/30 @Oxford

jiaxins.io

Autoregressive (AR) Models for Discrete Data

Decompose the joint distribution into conditional distributions following a specified order.

$$p(x_1, x_2, \ldots, x_6) = p(x_1)p(x_2 | x_1) \cdots p(x_6 | x_1, x_2, \ldots, x_5)$$



ChatGPT Gemini deepseek Claude

Why Diffusion Models for Discrete Data

- Generating discrete data with parallel sampling

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

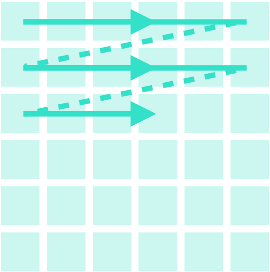
Mayor � Bowser said � meetings � Commissioner � on
Thursday that � new plan will be � board in � �

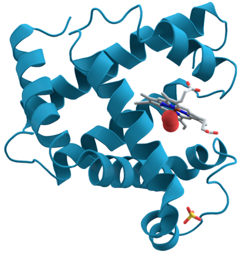
Mayor Muriel Bowser said after meetings � Commissioner � on
Thursday that � new plan will be � board in December �

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





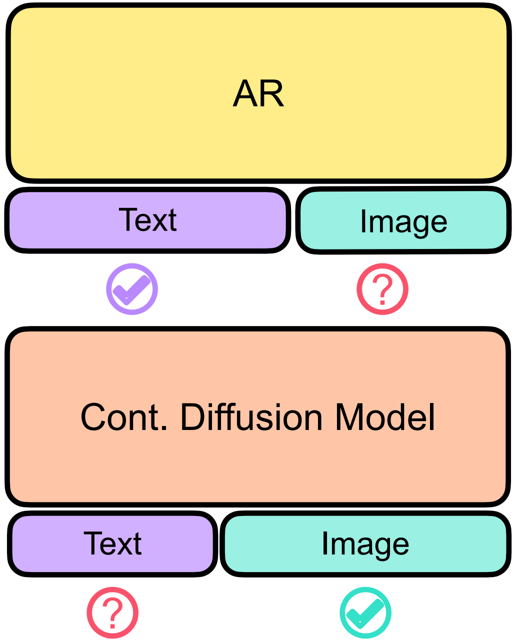
HGFGTLEHPIYKVAKQWSMVHDTTVYFSCGLHVAAHPATYVSM
TMLYHINMESFVNLEFCNFQTDDKYLEDPWARHEKYPIRKAIK
VSMDPNHGPVYCAKWDTILYMGKDGKERRTSAYMFTGVDEQHC
GRLFRITKSCWWGCCTLDNMKPDKAKACAEDMRRCRNIPVVQN
RNSKCRAIEWEIFQYWINCSTVVKTFAPCMFGFQFRFHYGYNY
DRETPVHAVNIINIWSAYKMTRYWCRIQCDSYWLWSGMTWRWC
CWEGSYKLMFCGWWRHFISKSMVTLGGHKKDDGRRWMLQSTHH

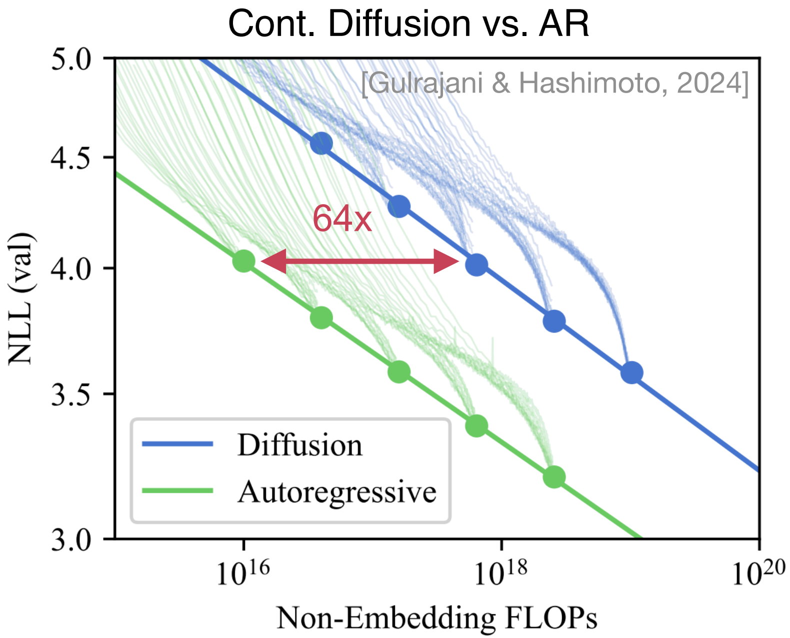
Duration (min)	IMDB Rating	Genre	Award
150	6.5	Action	Nominated
95	8.3	Romantic	Won
120	5.2	Horror	None

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

Why Discrete Diffusion Models

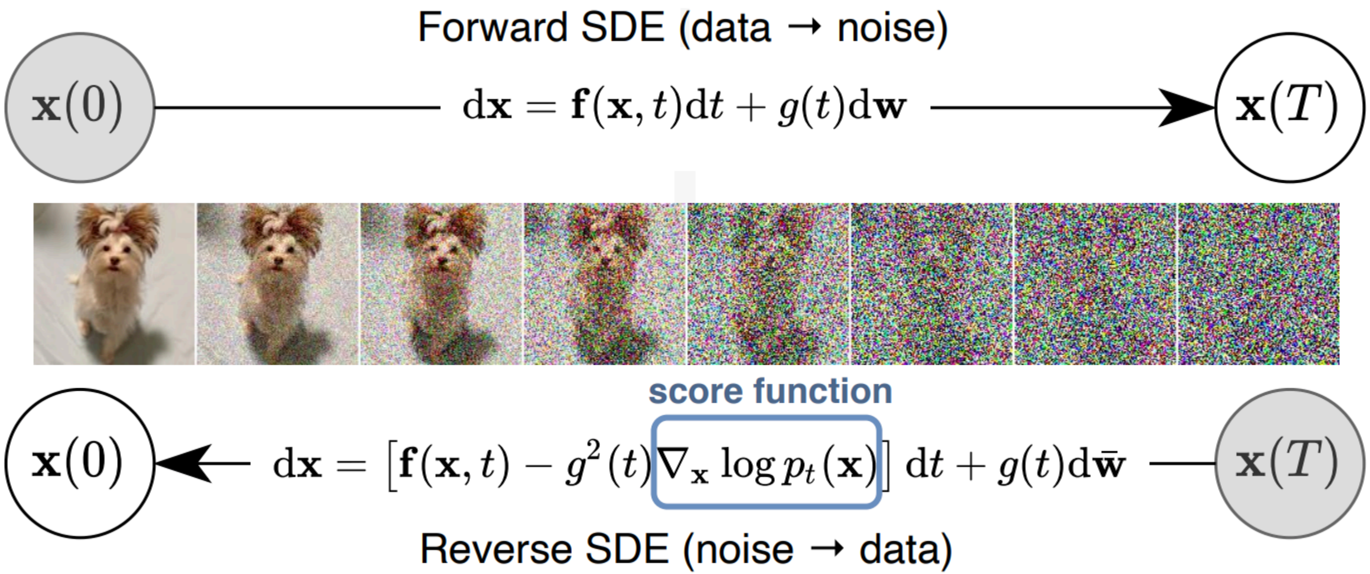
- Generating discrete data with parallel sampling
- AR models require imposing an ordering which may be unnatural for many data types
- Continuous diffusion is not great for discrete data





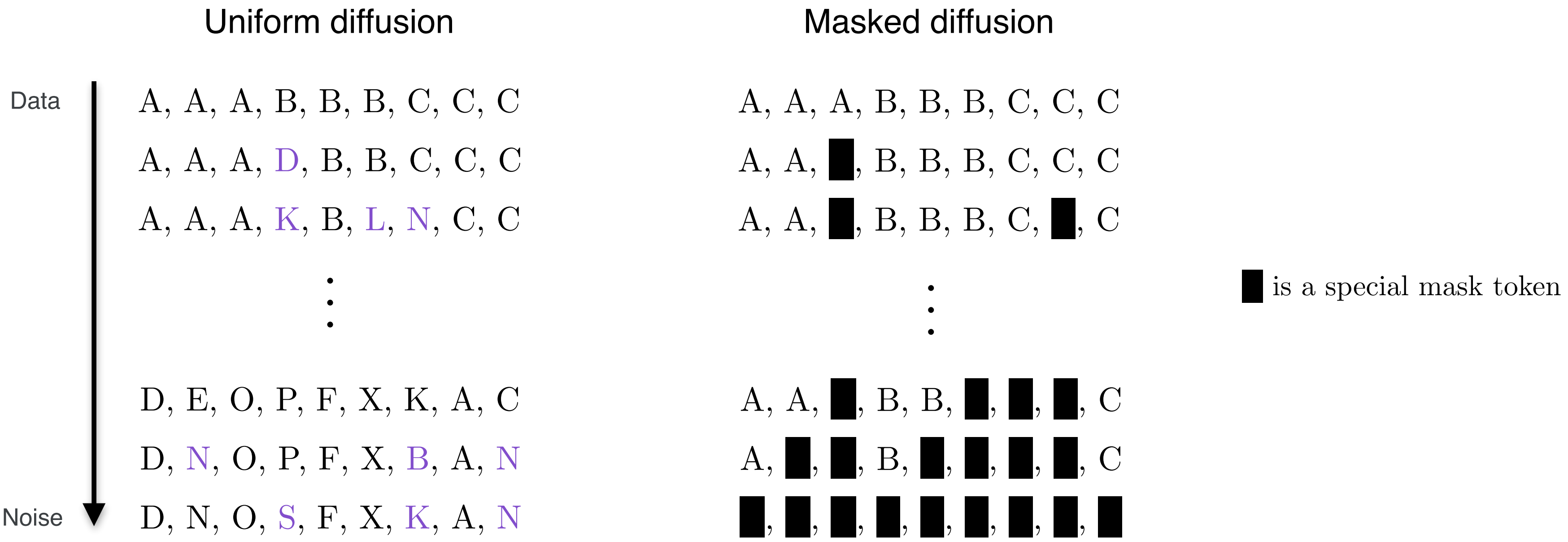
Gulrajani & Hashimoto (2024). Likelihood-based diffusion language models.

Recap: Diffusion Models



Song et al., (2020). Score-based generative modeling through stochastic differential equations.

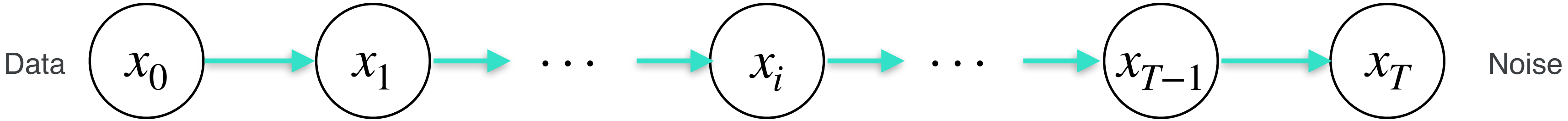
Discrete Noising Processes



It is empirically observed that masked diffusion generally works better than uniform diffusion in discrete generative modeling.

Hoogeboom et al., (2021). Argmax flows and multinomial diffusion: Learning categorical distributions.
Austin et al. (2021). Structured denoising diffusion models in discrete state-spaces.

Discrete-time Markov Chains



- x_0: clean data, x_T: noise. Finite state space of size M.
- Each (forward) transition follows the distribution $q(x_i \mid x_{i-1}) = \mathrm{Cat}(x_i; Q_i^\top x_{i-1})$
- Q_i is called the transition matrix: $[Q_i]_{jk} = q(x_i = k \mid x_{i-1} = j)$

$M \times M$

$$Q_i^{\text{uniform}} = \begin{bmatrix} 1-\beta_i+\beta_i/M & \beta_i/M & \cdots & \beta_i/M \\ \beta_i/M & 1-\beta_i+\beta_i/M & \cdots & \beta_i/M \\ \vdots & \vdots & \ddots & \beta_i/M \\ \beta_i/M & \beta_i/M & \cdots & 1-\beta_i+\beta_i/M \end{bmatrix}$$

$$(1-\beta_i)I + \frac{\beta_i}{M}\mathbf{1}\mathbf{1}^\top$$

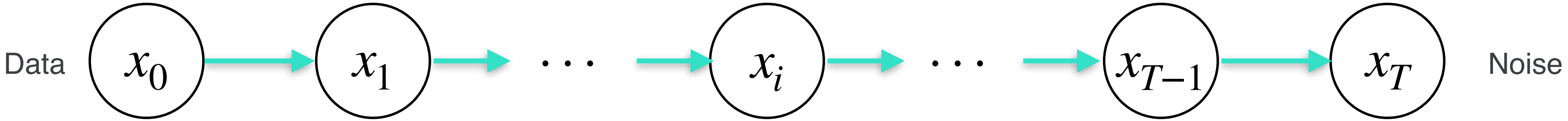
$(M+1) \times (M+1)$

$$Q_i^{\text{mask}} = \begin{bmatrix} 1-\beta_i & 0 & \cdots & 0 & \beta_i \\ 0 & 1-\beta_i & \cdots & 0 & \beta_i \\ \vdots & \vdots & \ddots & 0 & \beta_i \\ 0 & 0 & \cdots & 1-\beta_i & \beta_i \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}$$

$$(1-\beta_i)I + \beta_i \mathbf{1} e_M^\top$$

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

Discrete-time Markov Chains



- Product of transition matrices is transition matrix

$$q(x_2 | x_0) = \sum_{x_1} q(x_2 | x_1) q(x_1 | x_0) = \mathrm{Cat}(x_2; (Q_1 Q_2)^\top x_0)$$

- Marginal distribution at step i:

$$q(x_i | x_0) = \mathrm{Cat}(x_i; \bar{Q}_i^\top x_0), \text{ where } \bar{Q}_i = Q_1 Q_2 \cdots Q_i$$

- Take the masked diffusion as an example:

$$\bar{Q}_1 = (1 - \beta_1) I + \beta_1 \mathbf{1} e_M^\top$$

$$\bar{Q}_2 = (1 - \beta_1)(1 - \beta_2) I + (1 - (1 - \beta_1)(1 - \beta_2)) \mathbf{1} e_M^\top$$

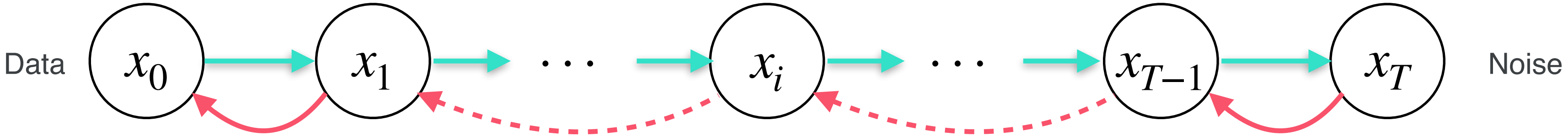
$$\vdots$$

$$\bar{Q}_i = \prod_{j=1}^{i} (1 - \beta_j) I + \left(1 - \prod_{j=1}^{i} (1 - \beta_j)\right) \mathbf{1} e_M^\top$$

$\triangleq \alpha_i$

$$\bar{Q}_i = \alpha_i I + (1 - \alpha_i) \mathbf{1} e_M^\top$$

Discrete-time Model



- We learn a reverse generative model (decoder) p to approximate q:

$$p(x_{0:T}) = p(x_0 \,|\, x_1) p(x_1 \,|\, x_2) \cdots p(x_{T-1} \,|\, x_T)$$

- Recall the diffusion model ELBO

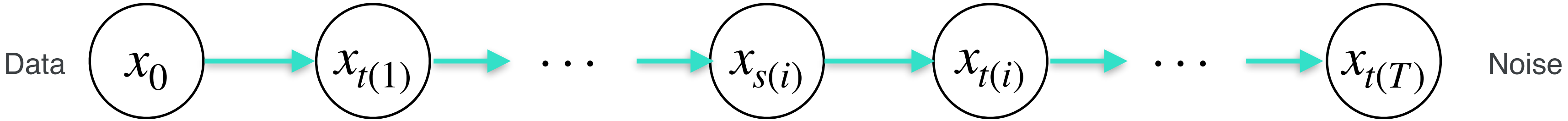
$$\log p(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)} \left[\log p(x_0 \,|\, x_1) - \mathrm{KL}(q(x_T \,|\, x_0) \| p(x_T)) - \sum_{i=2}^{T} \mathrm{KL}(q(x_{i-1} \,|\, x_i, x_0) \| p(x_{i-1} \,|\, x_i)) \right]$$

- $q(x_{i-1} \,|\, x_i, x_0)$ can be computed analytically via Bayes' rule

$$q(x_{i-1} \,|\, x_i, x_0) = \frac{q(x_i \,|\, x_{i-1}) q(x_{i-1} \,|\, x_0)}{q(x_i \,|\, x_0)} = \mathrm{Cat}\left(x_i; \frac{Q_i x_i \odot \bar{Q}_{i-1}^{\top} x_0}{x_0^{\top} \bar{Q}_i x_i} \right) \qquad p(x_{i-1} \,|\, x_i) \triangleq q(x_{i-1} \,|\, x_i, \mu_\theta(x_i))$$

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

From Discrete-time to Continuous-time



- We divide time between $[0,1]$ into T intervals: $s(i) = (i-1)/T$, $t(i) = i/T$
- Transition matrix Q_i: $[Q_i]_{jk} = q(x_{s(i)} = k \mid x_{t(i)} = j)$
- **Example** (masked diffusion):

$$\bar{Q}_i = \prod_{j=1}^{i} Q_i = \alpha_i I + (1-\alpha_i)\mathbf{1}e_M^\top, \quad \text{where } \alpha_i = \prod_{j=1}^{i}(1-\beta_j)$$

- Let $\beta_i = \dfrac{\beta(t(i))}{T}$ and $T \to \infty$ (cont. time limit)

$$\bar{Q}(t) \triangleq \lim_{T\to\infty} \bar{Q}_i = \alpha_t I + (1-\alpha_t)\mathbf{1}e_M^T, \quad \text{where } \alpha_t \triangleq \exp\left(-\int_0^t \beta(s)ds\right)$$

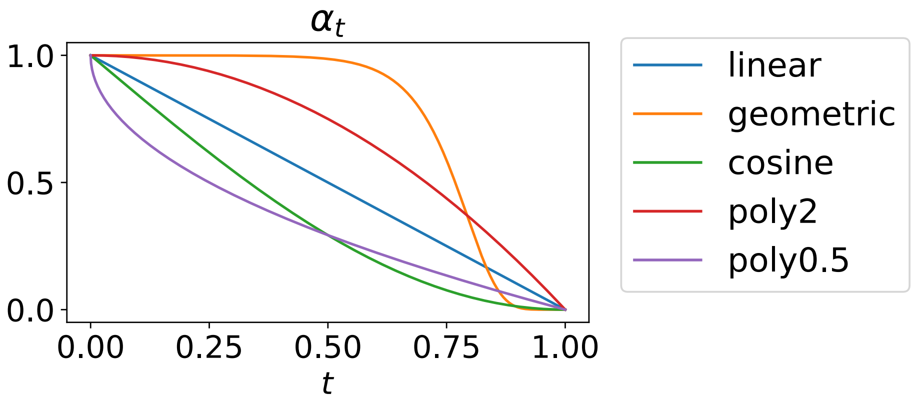
Derivation follows Shi et al. (2024). Simplified and generalized masked diffusion for discrete data.

From Discrete-time to Continuous-time

- The marginal distribution at time t:

Clean Masked

$$q(x_t | x_0) = \mathrm{Cat}(x_t; \bar{Q}(t)^\top x_0) = \mathrm{Cat}(x_t; \alpha_t x_0 + (1 - \alpha_t) e_M)$$



Masking schedule α_t: The probability of being unmasked at time t

- Assume the transition distribution from time s to time t is $q(x_t | x_s) = \mathrm{Cat}(x_t; \bar{Q}(s,t)^\top x_s)$
- Recall that transition matrix satisfies $\bar{Q}(t) = \bar{Q}(s)\bar{Q}(s,t)$, we can solve for $\bar{Q}(s,t)$:

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

Derivation follows Shi et al. (2024). Simplified and generalized masked diffusion for discrete data.

Continuous-time Model

- True reverse transition (knowing x_0):

$$q(x_s|x_t, x_0) \triangleq \frac{q(x_t|x_s)q(x_s|x_0)}{q(x_t|x_0)} = \mathrm{Cat}(x_s; \bar{R}(t,s)^\top x_t), \text{ where } \bar{R}(t,s) = I + \frac{\alpha_s - \alpha_t}{1-\alpha_t} e_M (x_0 - e_M)^\top$$

- or equivalently:

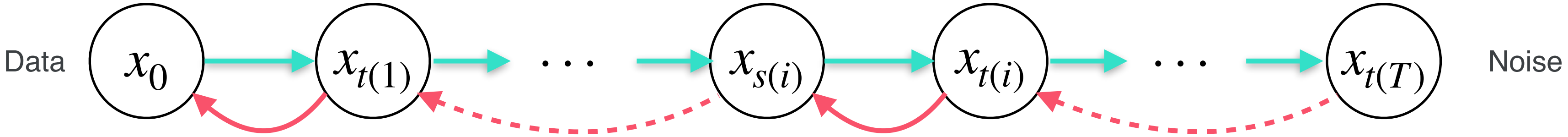
$$q(x_s|x_t, x_0) = \begin{cases} \mathrm{Cat}(x_s; x_t) & x_t \neq e_M \\ \mathrm{Cat}\left(x_s; \frac{1-\alpha_s}{1-\alpha_t} e_M + \frac{\alpha_s - \alpha_t}{1-\alpha_t} x_0\right) & x_t = e_M \end{cases}$$

- True reverse (unknown x_0): $q(x_s|x_t) = \sum_{x_0} q(x_s|x_t, x_0) q(x_0|x_t) = q(x_s|x_t, \boxed{\mathbb{E}[x_0|x_t]})$
- Reverse model: $p_\theta(x_s|x_t) \triangleq q(x_s|x_t, \boxed{\mu_\theta(x_t, t)})$.

Denoiser: Mean Parameterization

Derivation follows Shi et al. (2024). Simplified and generalized masked diffusion for discrete data.

Continuous-time ELBO



- Start with the discrete-time ELBO

$$\log p(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)}[\log p(x_0 \,|\, x_{t(1)})] - \mathrm{KL}(q(x_{t(T)} \,|\, x_0) \| p(x_{t(T)})) - \underbrace{\sum_{i=2}^{T} \mathbb{E}_{q(x_{t(i)}|x_0)} \Big[\mathrm{KL}(q(x_{s(i)} \,|\, x_{t(i)}, x_0) \| p(x_{s(i)} \,|\, x_{t(i)}))\Big]}_{\triangleq \mathscr{L}_T}$$

- For masked diffusion, $\mathrm{KL}(q(x_{t(T)} \,|\, x_0) \| p(x_{t(T)})) = 0$ as both are delta mass at mask state

$$\mathrm{KL}(q(x_s \,|\, x_t, x_0) \| p(x_s \,|\, x_t)) = -\frac{\alpha_s - \alpha_t}{1 - \alpha_t} \delta_{x_t, M} \cdot x_0^\top \log \mu_\theta(x_t, t)$$

$$\lim_{T \to \infty} \mathscr{L}_T = -\lim_{T \to \infty} \sum_{i=2}^{T} \frac{\alpha_{s(i)} - \alpha_{t(i)}}{1 - \alpha_{t(i)}} \mathbb{E}_{q(x_{t(i)}|x_0)}[\delta_{x_t, M} \cdot x_0^\top \log \mu_\theta(x_{t(i)}, t(i))] = \int_0^1 \frac{\alpha_t'}{1 - \alpha_t} \mathbb{E}_{q(x_t|x_0)}[\delta_{x_t, M} \cdot x_0^\top \log \mu_\theta(x_t, t)] dt$$

Derivation follows Shi et al. (2024). Simplified and generalized masked diffusion for discrete data.

Invariance

Optimal denoiser is time-independent (Ou et al. 2024)

- we can use $\mu_\theta(x_t, t) = \mu_\theta(x_t)$ to approximate $\mathbb{E}[x_0 | x_t]$.
- **Proof**: Write out the form of $q(x_0 | x_t)$ via Bayes' rule and observe it's independent of α_t.

ELBO is invariant to masking schedule (Shi et al. 2024)

- **Proof**: Define log signal-to-noise ratio (log-SNR): $\lambda_t = \log \frac{\alpha_t}{1 - \alpha_t}$. Rewrite the ELBO as

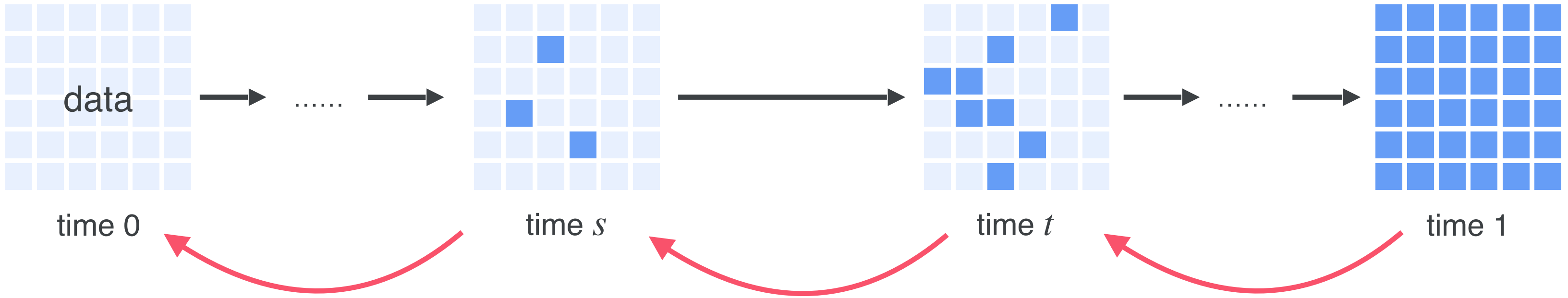
$$\log p_\theta(x_0) \geq \int_{-\infty}^{\infty} \sigma(\lambda) \mathbb{E}_{\tilde{q}(x_\lambda | x_0)}[\delta_{x_\lambda, M} \cdot x_0^\top \log \mu_\theta(x_\lambda)] d\lambda$$

Ou et al. (2024). Your absorbing discrete diffusion secretly models the conditional distributions of clean data.
Shi et al. (2024). Simplified and generalized masked diffusion for discrete data.

Masked Diffusion Models (multidimensional)

Each element is noised independently in the forward process

data
mask



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

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

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

$$\begin{cases} \text{w/ prob. } \dfrac{\alpha_s - \alpha_t}{1 - \alpha_t} \mathbb{E}[x_{0,j}^{(n)} \mid x_t], \text{ unmask to state } j \\ \text{w/ prob. } \dfrac{1 - \alpha_s}{1 - \alpha_t}, \text{ remains masked} \end{cases}$$

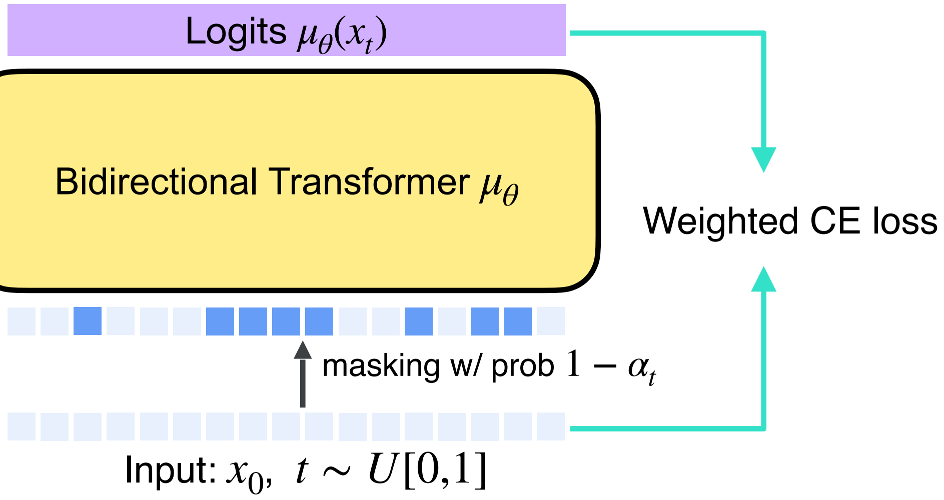
$\approx \mu_\theta^{(n)}(x_t)_j \triangleq \mathrm{softmax}(\mathrm{NN}_\theta(x_t))_j$

Masked Diffusion Models (multidimensional)

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

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

- Maximum likelihood = training a weighted ensemble of BERTs
- The simplified model and training objective lead to significant performance boost



Compressing Text

GPT2 zero-shot language modeling tasks

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**

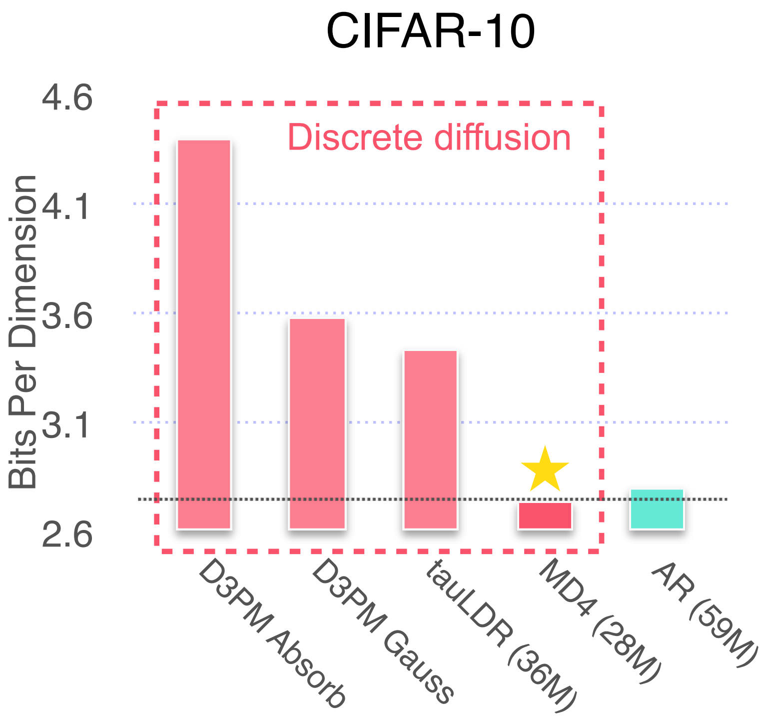
OpenWebText validation set

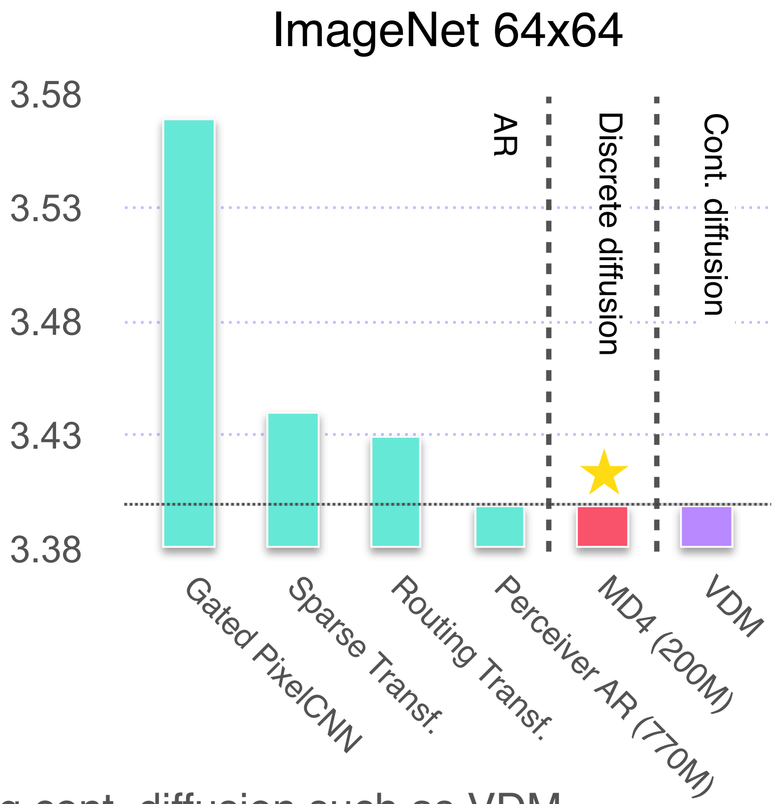
Size	Method	Perplexity ($\downarrow$)
Small	Gaussian Diffusion	$\leq$ 27.28
	SEDD Absorb (reimpl.)	$\leq$ 24.10
	MD4 (Ours)	$\leq$ 22.13
	GenMD4 (Ours)	$\leq$ **21.80**
Medium	MD4 (Ours)	$\leq$ **16.64**

- Many popular diffusion LLMs are now based on masked diffusion and this objective
- Concurrent work by Sahoo et al. (2024), Ou et al. (2024) studied similar losses for language

Ou et al. (2024). Your absorbing discrete diffusion secretly models the conditional distributions of clean data.
Sahoo et al. (2024). Simple and effective masked diffusion language models.

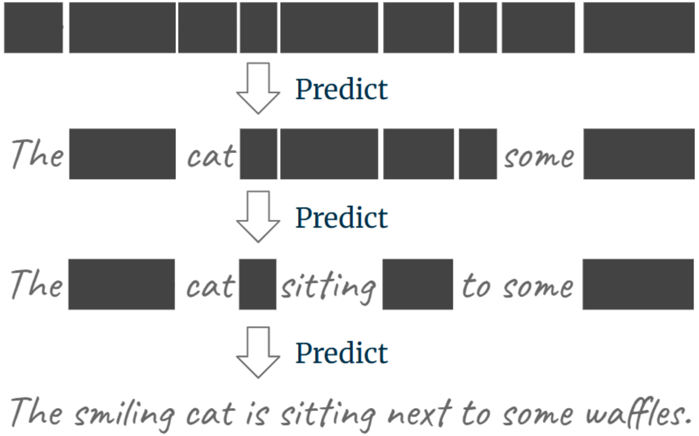
Compressing Image (Pixels)





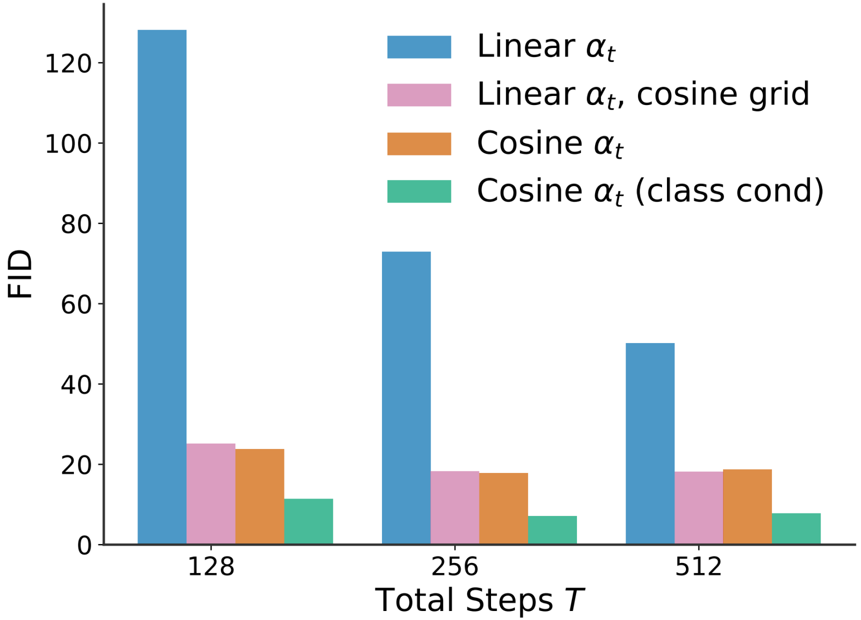
- Beating same-size AR & comparable with strong cont. diffusion such as VDM
- SoTA masked image models (MaskGIT/MAR) are non-likelihood-weighted MDMs

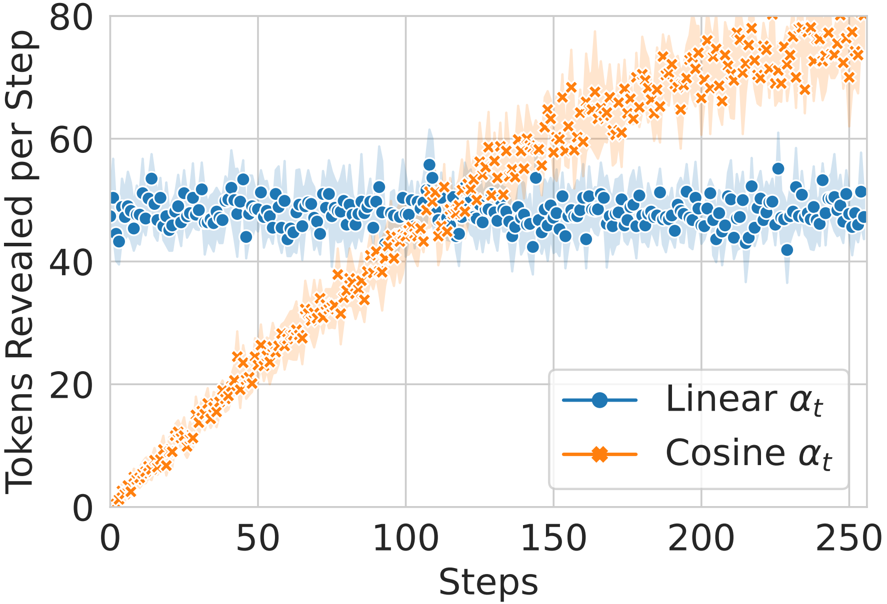
Faster Generation via Parallel Sampling



- Algorithm: Ancestral sampling from discrete-time reverse process
- Significant latency reduction, which has been validated by industry deployment (Google, Bytedance)
- Many details (e.g., schedules, JAX numerics, diversity/quality tradeoff) to get right in order to produce coherent samples with parallel sampling

Importance of Schedules





- 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.

Advanced Topics

An active area of research!

- Continuous-time Markov chain (CTMC) representation and transition rates
- Equivalence between cont. time masked diffusion models and any-order AR models
- Discrete “score function” and score parameterization
- Connection between uniform diffusion and masked diffusion (why mask works better?)
- Predictor-corrector sampling for discrete diffusion, remasking
- Hybrid autoregressive + discrete diffusion models
- Variable-length generation
- …

Campbell et al. (2022). A continuous time framework for discrete denoising models.
Hoogeboom et al. (2021). Autoregressive diffusion models.
Lou et al. (2023). Discrete diffusion modeling by estimating the ratios of the data distribution.
Amin et al. (2025). Why Masking Diffusion Works: Condition on the Jump Schedule for Improved Discrete Diffusion.
Zhao et al. (2024). Informed correctors for discrete diffusion models.
Wang et al. (2025). Remasking discrete diffusion models with inference-time scaling.
Arriola et al. (2025). Block diffusion: Interpolating between autoregressive and diffusion language models.
Kim et al. (2025). Any-Order Flexible Length Masked Diffusion.

Thanks

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

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

mean parameterization fixes the problem

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

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

Shi et al. (2025). Simplified and Generalized Masked Diffusion for Discrete Data.

Relation to Score Entropy Loss

Score Entropy loss (Lou et al., 2024; Benton et al., 2024):

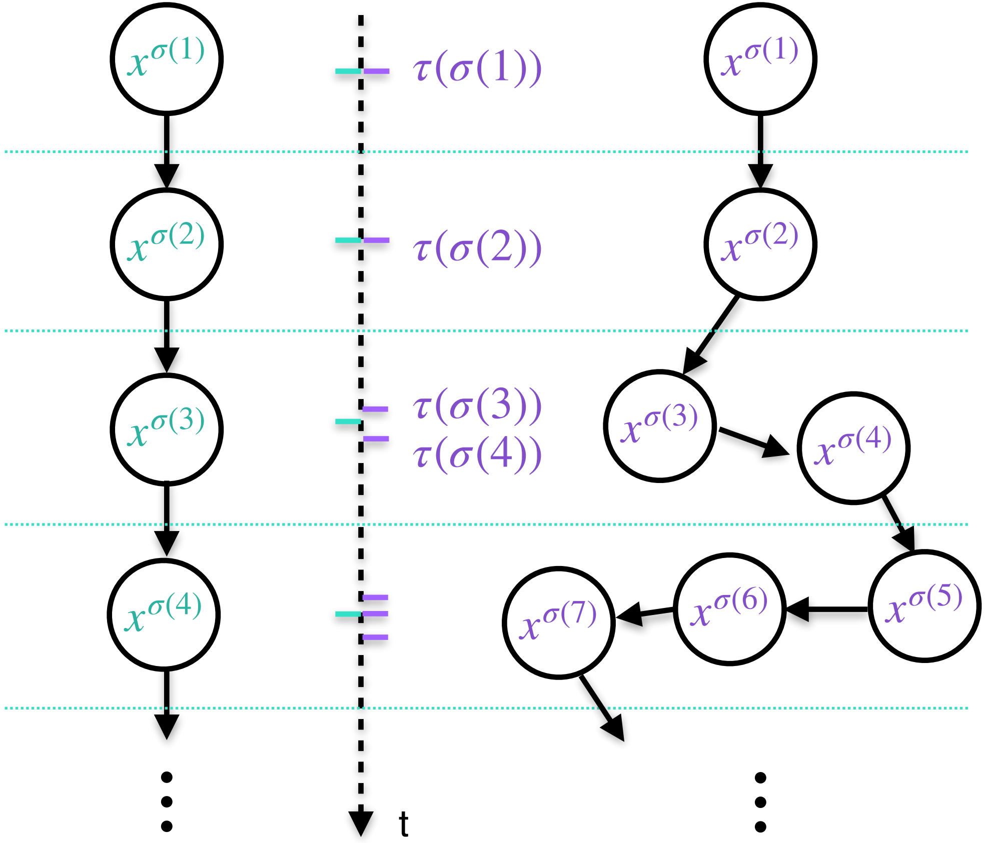
$$\mathscr{L}_{s_\theta} = \int_0^1 \mathbb{E}_{q_{t|0}(k|x_0)}\Big[\sum_{j \neq k} Q(t)_{jk}\Big(s_\theta(k,t)_j - \frac{q_{t|0}(j\,|\,x_0)}{q_{t|0}(k\,|\,x_0)}\log s_\theta(k,t)_j + \psi\Big(\frac{q_{t|0}(j\,|\,x_0)}{q_{t|0}(k\,|\,x_0)}\Big)\Big)\Big]dt$$

Plugging in the mean parameterization

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

recovers the MD4 objective.

MD4 as Parallel Any-Order AR Models



A new dimension of freedom in AO-ARMs

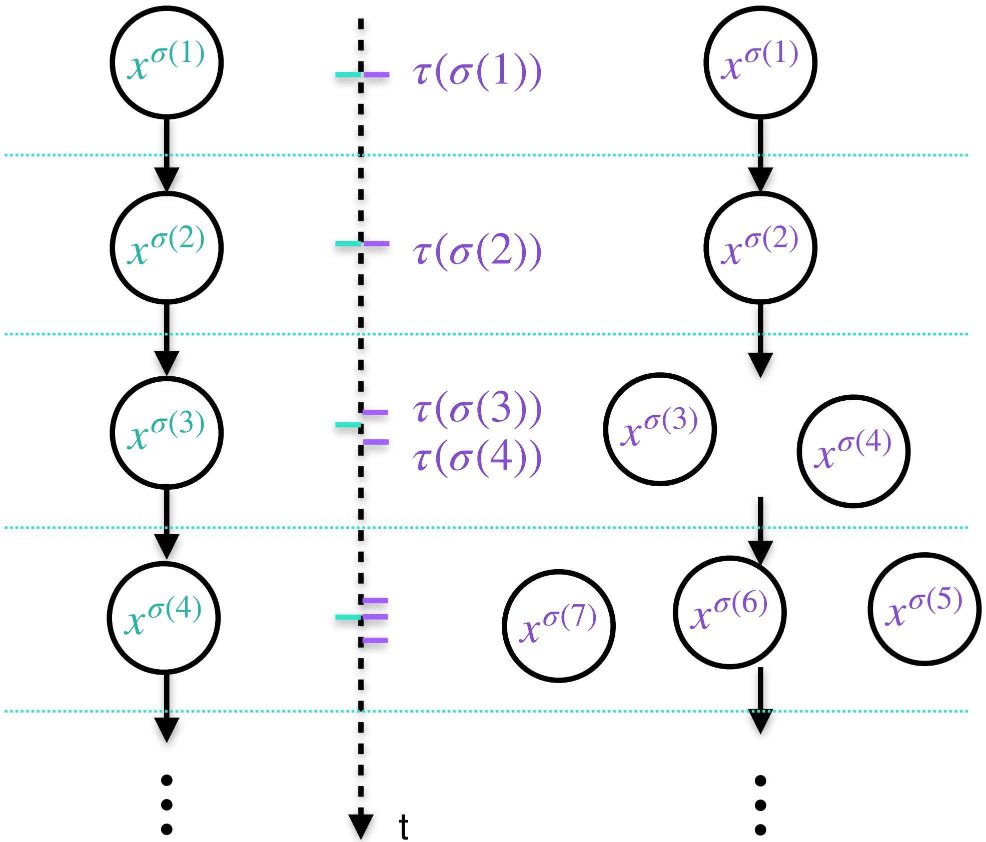
- Masking schedules control parallel sampling bandwidth

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.