

Discrete Flow and Diffusion Models

Lecture 07

Qiang Sun

Discrete Diffusion Models in the news

Forbes

Subscribe To Newsletters

INNOVATION > AI

Generative AI Gets Shaken Up By Newly Announced Text-Producing Diffusion LLMs

SEARCH FORTUNE

HOME NEWS FORTUNE 500 TECH FINANCE LEADERSHIP LIFESTYLE MULTIMEDIA

TECH: GOOGLE

Gemini Diffusion was the sleeper hit of Google I/O and some say its blazing speed could reshape the AI model wars

BY SHAMON GOLDMAN
@SHAMONGOLDMAN

May 21, 2024 at 5:04 PM EST

Inception Labs @InceptionAILabs · Feb 26
We achieve over 1000 tokens/second on NVIDIA H100s. Blazing fast generations without specialized chips!

Output Speed: Smaller models
Model Tokens per Second (Higher is Better) | 1.8M Super Tokens, Output: Poached vermicelli

Model	Output Speed (Tokens/Second)
gpt-4o	1100
gpt-4o-mini	1100
gpt-4o-mini-2024-08-07	1100
gpt-4o-mini-2024-07-18	1100
gpt-4o-mini-2024-07-18-16k	1100
gpt-4o-mini-2024-07-18-32k	1100
gpt-4o-mini-2024-07-18-64k	1100
gpt-4o-mini-2024-07-18-128k	1100
gpt-4o-mini-2024-07-18-256k	1100
gpt-4o-mini-2024-07-18-512k	1100
gpt-4o-mini-2024-07-18-1024k	1100
gpt-4o-mini-2024-07-18-2048k	1100
gpt-4o-mini-2024-07-18-4096k	1100
gpt-4o-mini-2024-07-18-8192k	1100
gpt-4o-mini-2024-07-18-16384k	1100
gpt-4o-mini-2024-07-18-32768k	1100
gpt-4o-mini-2024-07-18-65536k	1100
gpt-4o-mini-2024-07-18-131072k	1100
gpt-4o-mini-2024-07-18-262144k	1100
gpt-4o-mini-2024-07-18-524288k	1100
gpt-4o-mini-2024-07-18-1048576k	1100
gpt-4o-mini-2024-07-18-2097152k	1100
gpt-4o-mini-2024-07-18-4194304k	1100
gpt-4o-mini-2024-07-18-8388608k	1100
gpt-4o-mini-2024-07-18-16777216k	1100
gpt-4o-mini-2024-07-18-33554432k	1100
gpt-4o-mini-2024-07-18-67108864k	1100
gpt-4o-mini-2024-07-18-134217728k	1100
gpt-4o-mini-2024-07-18-268435456k	1100
gpt-4o-mini-2024-07-18-536870912k	1100
gpt-4o-mini-2024-07-18-1073741824k	1100
gpt-4o-mini-2024-07-18-2147483648k	1100
gpt-4o-mini-2024-07-18-4294967296k	1100
gpt-4o-mini-2024-07-18-8589934592k	1100
gpt-4o-mini-2024-07-18-17179869184k	1100
gpt-4o-mini-2024-07-18-34359738368k	1100
gpt-4o-mini-2024-07-18-68719476736k	1100
gpt-4o-mini-2024-07-18-137438953472k	1100
gpt-4o-mini-2024-07-18-274877906944k	1100
gpt-4o-mini-2024-07-18-549755813888k	1100
gpt-4o-mini-2024-07-18-1099511627776k	1100
gpt-4o-mini-2024-07-18-2199023255552k	1100
gpt-4o-mini-2024-07-18-4398046511104k	1100
gpt-4o-mini-2024-07-18-8796093022208k	1100
gpt-4o-mini-2024-07-18-17592186044416k	1100
gpt-4o-mini-2024-07-18-35184372088832k	1100
gpt-4o-mini-2024-07-18-70368744177664k	1100
gpt-4o-mini-2024-07-18-140737488355328k	1100
gpt-4o-mini-2024-07-18-281474976710656k	1100
gpt-4o-mini-2024-07-18-562949953421312k	1100
gpt-4o-mini-2024-07-18-1125899906842624k	1100
gpt-4o-mini-2024-07-18-2251799813685248k	1100
gpt-4o-mini-2024-07-18-4503599627370496k	1100
gpt-4o-mini-2024-07-18-9007199254740992k	1100
gpt-4o-mini-2024-07-18-18014398519481984k	1100
gpt-4o-mini-2024-07-18-36028797038963968k	1100
gpt-4o-mini-2024-07-18-72057594077927936k	1100
gpt-4o-mini-2024-07-18-144115188155855872k	1100
gpt-4o-mini-2024-07-18-288230376311711744k	1100
gpt-4o-mini-2024-07-18-576460752623423488k	1100
gpt-4o-mini-2024-07-18-1152921505246846976k	1100
gpt-4o-mini-2024-07-18-2305843010493693952k	1100
gpt-4o-mini-2024-07-18-4611686020987387904k	1100
gpt-4o-mini-2024-07-18-9223372041974775808k	1100
gpt-4o-mini-2024-07-18-18446744083549551616k	1100
gpt-4o-mini-2024-07-18-36893488167099103232k	1100
gpt-4o-mini-2024-07-18-73786976334198206464k	1100
gpt-4o-mini-2024-07-18-147573952668396412928k	1100
gpt-4o-mini-2024-07-18-295147905336792825856k	1100
gpt-4o-mini-2024-07-18-590295810673585651712k	1100
gpt-4o-mini-2024-07-18-1180591621347171303424k	1100
gpt-4o-mini-2024-07-18-2361183242694342606848k	1100
gpt-4o-mini-2024-07-18-4722366485388685213696k	1100
gpt-4o-mini-2024-07-18-9444732970777370427392k	1100
gpt-4o-mini-2024-07-18-18889465941554740854784k	1100
gpt-4o-mini-2024-07-18-37778931883109481709568k	1100
gpt-4o-mini-2024-07-18-75557863766218963419136k	1100
gpt-4o-mini-2024-07-18-151115727532437926838272k	1100
gpt-4o-mini-2024-07-18-302231455064875853676544k	1100
gpt-4o-mini-2024-07-18-604462910129751707353088k	1100
gpt-4o-mini-2024-07-18-1208925820259503414706176k	1100
gpt-4o-mini-2024-07-18-2417851640519006829412352k	1100
gpt-4o-mini-2024-07-18-4835703281038013658824704k	1100
gpt-4o-mini-2024-07-18-9671406562076027317649408k	1100
gpt-4o-mini-2024-07-18-19342813124152054635298816k	1100
gpt-4o-mini-2024-07-18-38685626248304109270597632k	1100
gpt-4o-mini-2024-07-18-77371252496608218541195264k	1100
gpt-4o-mini-2024-07-18-154742504993216437082390528k	1100
gpt-4o-mini-2024-07-18-309485009986432874164781056k	1100
gpt-4o-mini-2024-07-18-618970019972865748329562112k	1100
gpt-4o-mini-2024-07-18-1237940039845731496659124224k	1100
gpt-4o-mini-2024-07-18-2475880079691462993318248448k	1100
gpt-4o-mini-2024-07-18-4951760159382925986636496896k	1100
gpt-4o-mini-2024-07-18-9903520318765851973272993792k	1100
gpt-4o-mini-2024-07-18-19807040637531703946545987584k	1100
gpt-4o-mini-2024-07-18-39614081275063407893091975168k	1100
gpt-4o-mini-2024-07-18-79228162550126815786183950336k	1100
gpt-4o-mini-2024-07-18-158456325100253631572367900672k	1100
gpt-4o-mini-2024-07-18-316912650200507263144735801344k	1100
gpt-4o-mini-2024-07-18-633825300401014526289471602688k	1100
gpt-4o-mini-2024-07-18-1267650600802029052578943205376k	1100
gpt-4o-mini-2024-07-18-2535301201604058105157886410752k	1100
gpt-4o-mini-2024-07-18-5070602403208116210315772821504k	1100
gpt-4o-mini-2024-07-18-1014120480641623242063154443008k	1100
gpt-4o-mini-2024-07-18-2028240961283246484126308886016k	1100
gpt-4o-mini-2024-07-18-4056481922566492968252617772032k	1100
gpt-4o-mini-2024-07-18-8112963845132985936505235544064k	1100
gpt-4o-mini-2024-07-18-16225927690265971873010471088128k	1100
gpt-4o-mini-2024-07-18-32451855380531943746020942176256k	1100
gpt-4o-mini-2024-07-18-64903710761063887492041884352512k	1100
gpt-4o-mini-2024-07-18-129807421522127774984083768705024k	1100
gpt-4o-mini-2024-07-18-259614843044255549968167537410048k	1100
gpt-4o-mini-2024-07-18-519229686088511099936335074820096k	1100
gpt-4o-mini-2024-07-18-1038459372177022199872670149640192k	1100
gpt-4o-mini-2024-07-18-2076918744354044399745340299280384k	1100
gpt-4o-mini-2024-07-18-4153837488708088799490680598560768k	1100
gpt-4o-mini-2024-07-18-8307674977416177598981361197121536k	1100
gpt-4o-mini-2024-07-18-16615349954832355197962722384243072k	1100
gpt-4o-mini-2024-07-18-33230699909664710395925444768486144k	1100
gpt-4o-mini-2024-07-18-664613998193294207918508895369728k	1100
gpt-4o-mini-2024-07-18-1329227996386588415837017790739456k	1100
gpt-4o-mini-2024-07-18-2658455992773176831674035581478912k	1100
gpt-4o-mini-2024-07-18-5316911985546353663348071162957824k	1100
gpt-4o-mini-2024-07-18-10633823971092707326696142325915648k	1100
gpt-4o-mini-2024-07-18-21267647942185414653392284651831008k	1100
gpt-4o-mini-2024-07-18-42535295884370829306784569303662016k	1100
gpt-4o-mini-2024-07-18-85070591768741658613569138607324032k	1100
gpt-4o-mini-2024-07-18-170141183537483317227138277214648064k	1100
gpt-4o-mini-2024-07-18-340282367074966634454276554429296128k	1100
gpt-4o-mini-2024-07-18-680564734149933268908553108858592256k	1100
gpt-4o-mini-2024-07-18-1361129468299866537817106217717145024k	1100
gpt-4o-mini-2024-07-18-272225893659973307563421243543430048k	1100
gpt-4o-mini-2024-07-18-544451787319946615126842487086860096k	1100
gpt-4o-mini-2024-07-18-1088903574639933230253684974173720192k	1100
gpt-4o-mini-2024-07-18-2177807149279866460507369948347440384k	1100
gpt-4o-mini-2024-07-18-4355614298559732921014739896694880768k	1100
gpt-4o-mini-2024-07-18-8711228597119465842029479793389761536k	1100
gpt-4o-mini-2024-07-18-1742245719423893668405895958679523072k	1100
gpt-4o-mini-2024-07-18-3484491438847787336811791917359046144k	1100
gpt-4o-mini-2024-07-18-6968982877695574673623583834718092288k	1100
gpt-4o-mini-2024-07-18-139379657553911493472476676694361856k	1100
gpt-4o-mini-2024-07-18-2787593151078229869449533533887371104k	1100
gpt-4o-mini-2024-07-18-5575186302156459738899067067774742208k	1100
gpt-4o-mini-2024-07-18-11150372604312919477798134135548884416k	1100
gpt-4o-mini-2024-07-18-22300745208625838955596268271097768832k	1100
gpt-4o-mini-2024-07-18-44601490417251677911192536542195537664k	1100
gpt-4o-mini-2024-07-18-89202980834503355822385073084391075328k	1100
gpt-4o-mini-2024-07-18-178405961669006711644670146167922150656k	1100
gpt-4o-mini-2024-07-18-356811923338013423289340292335844301312k	1100
gpt-4o-mini-2024-07-18-713623846676026846578680584671688602624k	1100
gpt-4o-mini-2024-07-18-142724769335205369315737116934337720512k	1100
gpt-4o-mini-2024-07-18-285449538670410738631474233868675441024k	1100
gpt-4o-mini-2024-07-18-570899077340821477262948467737350882048k	1100
gpt-4o-mini-2024-07-18-114179815468164295452589695447470172992k	1100
gpt-4o-mini-2024-07-18-228359630936328590905179390894940345984k	1100
gpt-4o-mini-2024-07-18-456719261872657181810358781789806711968k	1100
gpt-4o-mini-2024-07-18-91343852374531436362071756357961343936k	1100
gpt-4o-mini-2024-07-18-182687704749062872724143512715922687872k	1100
gpt-4o-mini-2024-07-18-365375409498125745448287025431845375648k	1100
gpt-4o-mini-2024-07-18-730750818996251490896574050863690751392k	1100
gpt-4o-mini-2024-07-18-1461501637992502981793148101727381502784k	1100
gpt-4o-mini-2024-07-18-2923003275985005963586296203454763005568k	1100
gpt-4o-mini-2024-07-18-5846006551970011927172592406909526011136k	1100
gpt-4o-mini-2024-07-18-11692013103940023854345184813819052022272k	1100
gpt-4o-mini-2024-07-18-23384026207880047708690369627638104044544k	1100
gpt-4o-mini-2024-07-18-46768052415760095417380739255276208089088k	1100
gpt-4o-mini-2024-07-18-93536104831520190834761478510552416171776k	1100
gpt-4o-mini-2024-07-18-187072209663040381669522957021104834343552k	1100
gpt-4o-mini-2024-07-18-374144419326080763339045914042208668687104k	1100
gpt-4o-mini-2024-07-18-748288838652161526678091828084417337374208k	1100
gpt-4o-mini-2024-07-18-1496577677304323053356183656168834674644416k	1100
gpt-4o-mini-2024-07-18-299315535460864610671236731233769313128896k	1100
gpt-4o-mini-2024-07-18-598631070921729221342473462467538626257792k	1100
gpt-4o-mini-2024-07-18-119726214184345844268494692493517325251536k	1100
gpt-4o-mini-2024-07-18-239452428368691688536989384987034650502784k	1100
gpt-4o-mini-2024-07-18-478904856737383377073978769974069301005568k	1100
gpt-4o-mini-2024-07-18-9578097134747667541479575399481386020011136k	1100
gpt-4o-mini-2024-07-18-191561942694953350829591507989677200022272k	1100
gpt-4o-mini-2024-07-18-383123885389906701659183015979354400044544k	1100
gpt-4o-mini-2024-07-18-766247770779813403318366031958708800089088k	1100
gpt-4o-mini-2024-07-18-1532495441559626806636732039177417600178176k	1100
gpt-4o-mini-2024-07-18-3064990883119253613273464078354835200356352k	1100
gpt-4o-mini-2024-07-18-6129981766238507226546928156709670400712704k	1100
gpt-4o-mini-2024-07-18-12259963532477014453093856313419340801425408k	1100
gpt-4o-mini-2024-07-18-24519927064954028906187712626838681602850816k	1100
gpt-4o-mini-2024-07-18-49039854129908057812375425253677363205701632k	1100
gpt-4o-mini-2024-07-18-98079708259816115624750850507354726411403264k	1100
gpt-4o-mini-2024-07-18-196159416519632231249501701014709528022806528k	1100
gpt-4o-mini-2024-07-18-392318833039264462499003402029419056045613056k	1100
gpt-4o-mini-2024-07-18-7846376660785289249980068040588381	

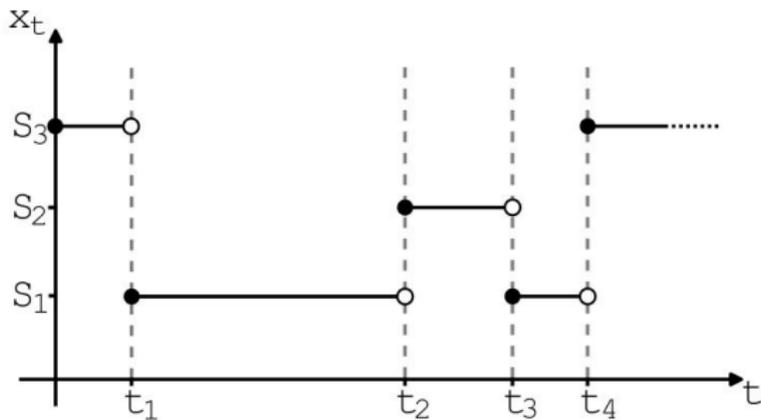
Diffusion LLMs generate text in arbitrary order

Austin et al., “Structured Denoising Diffusion Models in Discrete State-Spaces”, NeurIPS 2022

Discrete diffusion models and discrete flow matching

- Models for **discrete sequence data**: Language, protein sequences, etc.
- *Note: There is no diffusion/SDE and also no flow/ODE in discrete space.*
- Rather: Learning principles of flow matching and denoising can be generalized to discrete data!
- Mathematical model: **Continuous-time Markov chains (CTMCs)**
- Today:
 - CTMC Models
 - Discrete Flow Matching:
 - Discrete Probability Paths
 - Discrete Marginalization Trick
 - Discrete FM objective

Continuous Time Markov Chains



Figures: Andrew Campbell

Example - Continuous-time Markov chain

State space $S = \{a, b\}$

Rate matrix

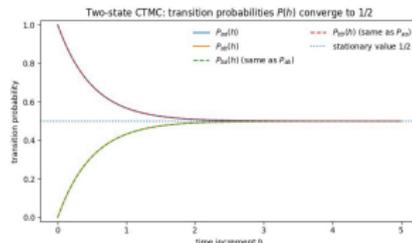
$$Q = \begin{array}{c|cc} & a & b \\ \hline a & -\lambda & \lambda \\ b & \lambda & -\lambda \end{array}$$

By showing evolution equation (taking derivatives), one obtains:

$$\begin{pmatrix} p(X_{t+h} = a | X_t = a) & p(X_{t+h} = a | X_t = b) \\ p(X_{t+h} = b | X_t = a) & p(X_{t+h} = b | X_t = b) \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 1 + e^{-2\lambda h} & 1 - e^{-2\lambda h} \\ 1 - e^{-2\lambda h} & 1 + e^{-2\lambda h} \end{pmatrix}$$

Convergence for h to infinity:

$$\rightarrow \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$



Examples of neighbors

$$Q_t^\theta(z|x) = 0$$

X and y are neighbors



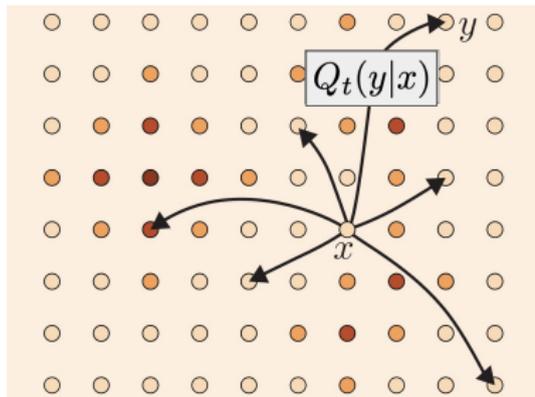
Y and z are neighbors



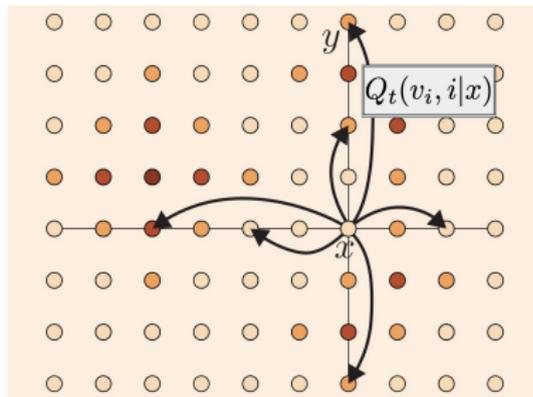
Z and x are not neighbors



General CTMC



Factorized CTMC



Figures: Yaron Lipman

Sampling from factorized CTMC models

Algorithm 7 Sampling from a Factorized CTMC Model (Euler / τ -leaping)

Require: Rate network Q_t^θ (factorized), initial distribution p_{init} , number of steps n

- 1: Set $t \leftarrow 0$
- 2: Set step size $h \leftarrow \frac{1}{n}$
- 3: Draw a sample $X_0 \sim p_{\text{init}}$, where $X_0 = (X_0^{(1)}, \dots, X_0^{(d)}) \in \mathcal{V}^d$
- 4: **for** $i = 1, \dots, n$ **do**
- 5: Compute factorized jump rates $\{q_j(v)\}_{j=1..d, v \in \mathcal{V}} \leftarrow Q_t^\theta(\cdot | X_t)$
- 6: **for** $j = 1, \dots, d$ (**in parallel**) **do**
- 7: $x \leftarrow X_t^{(j)}$ {current token at position j }
- 8: Define the per-position Euler transition probabilities $\tilde{p}_{j,t}(\cdot | X_t^{(j)} = x)$ by

$$\tilde{p}_{j,t}(v | x) = \begin{cases} h q_j(v), & v \neq x, \\ 1 - h \sum_{v' \in \mathcal{V} \setminus \{x\}} q_j(v'), & v = x. \end{cases}$$

- 9: Sample $X_{t+h}^{(j)} \sim \text{CATEGORICAL}(\{\tilde{p}_{j,t}(v | x)\}_{v \in \mathcal{V}})$
 - 10: **end for**
 - 11: Set $t \leftarrow t + h$
 - 12: **end for**
 - 13: **return** X_1
-

Diffusion LLMs generate text in arbitrary order

Austin et al., “Structured Denoising Diffusion Models in Discrete State-Spaces”, NeurIPS 2022

Generative Modeling with CTMCs

Data distribution:

$$p_{\text{data}}(z) \quad (z \in S)$$

*E.g. distribution
of texts on the
internet*

Initial distribution:

$$p_{\text{init}}(z) \quad (z \in S)$$

*E.g. uniform
distribution*

$$p_{\text{init}}(z) = \frac{1}{|S|}$$

Goal: Convert “Noise” to Data with a CTMC

$$X_0 \sim p_{\text{init}} \xrightarrow{\text{CTMC}} X_1 \sim p_{\text{data}}$$

Continuous Flow Matching



The Discrete Flow Matching Matrix

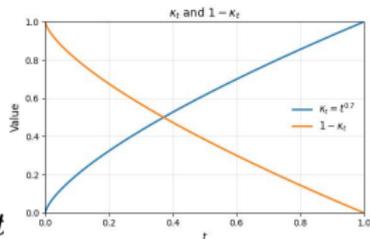


Example - Factorized Mixture Path

Scheduler: $0 \leq \kappa_t \leq 1$ such that

$$\kappa_0 = 0, \kappa_1 = 1$$

Idea: Noise each token independently with probability κ_t



$$p_t(x | z) = \prod_{j=1}^d \left[(1 - \kappa_t) \underset{\text{Downweight noise}}{p_{\text{init}}^{(j)}(x_j)} + \kappa_t \underset{\text{Upweight data}}{\delta_{z_j}(x_j)} \right]$$

Sampling procedure:

$$\begin{aligned} m_j &\sim \text{Bernoulli}(\kappa_t), \quad \xi_j \sim p_{\text{init}}^{(j)} \\ x_j &= m_j z_j + (1 - m_j) \xi_j, \quad j = 1, \dots, d \\ x &= (x_1, \dots, x_d) \end{aligned}$$

**Check for yourself
that this is a cond.
prob. path!**

Kolmogorov Forward Equation

Discrete analogue to the continuity equation.

A CTMC with rate matrix Q_t follows the probability path

$$X_t \sim p_t \quad (0 \leq t \leq 1)$$

if and only if the Kolmogorov Forward Equation (KFE) holds:

$$\frac{d}{dt} p_t(x) = \sum_{y \in S} Q_t(x|y) p_t(y)$$

change of probability

Net inflow

Conditional Rate Matrix for Factorized Mixture Path

The conditional rate matrix for factorized mixture path is factorized (i.e. rates only non-negative for one token updates) and given by

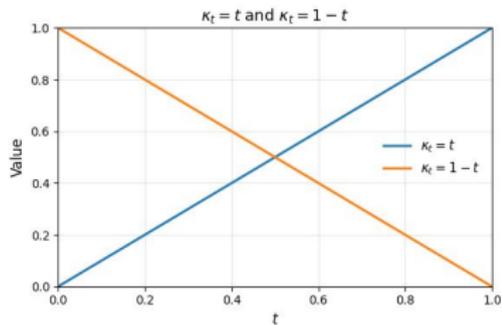
$$Q_t^z(y|x) = (Q_t^z(v_i, j|x_j))_{v_i, j}$$
$$Q_t^z(v_i, j|x_j) = \frac{\dot{\kappa}_t}{1 - \kappa_t} (\delta_{z_j}(v_i) - \delta_{x_j}(v_i))$$
$$= \frac{\dot{\kappa}_t}{1 - \kappa_t} \begin{cases} 0 & \text{if } x_j = z_j \\ 1 & \text{if } v_i = z_j, x_j \neq z_j \\ 0 & \text{if } v_i \neq z_j, x_j \neq z_j \\ -1 & \text{if } v_i = x_j, x_j \neq z_j \end{cases}$$

If current token correct, zero rate

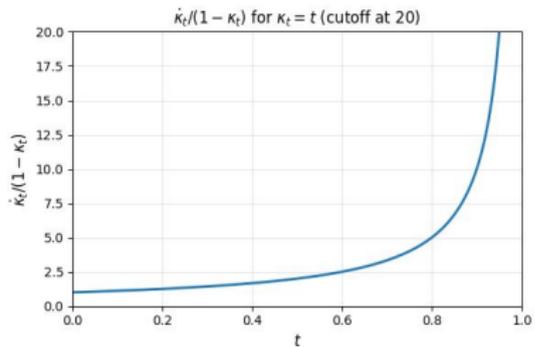
If incorrect, jump to correct token

Outgoing rate from current token, if incorrect

$$\kappa_t$$



$$\frac{\dot{\kappa}_t}{1 - \kappa_t}$$



Conditional Rate Matrix for Factorized Mixture Path

$$Q_t^z(y|x) = (Q_t^z(v_i, j|x_j))_{v_i, j}$$

$$Q_t^z(v_i, j|x_j) = \frac{\dot{\kappa}_t}{1 - \kappa_t} (\delta_{z_j}(v_i) - \delta_{x_j}(v_i))$$

$$= \frac{\dot{\kappa}_t}{1 - \kappa_t} \begin{cases} 0 & \text{if } x_j = z_j \\ 1 & \text{if } v_i = z_j, x_j \neq z_j \\ 0 & \text{if } v_i \neq z_j, x_j \neq z_j \\ -1 & \text{if } v_i = x_j, x_j \neq z_j \end{cases}$$

Rates explode at
 $t=1$

If current token correct, zero rate

If incorrect, jump to correct
token

Outgoing rate from current token, if
incorrect

Conditional Prob. Path, Vector Field, and Score

Notation

Key property

Factorized mixture

Conditional Probability Path $p_t(x|z)$ Interpolates p_{init} and a data point z $\prod_{j=1}^d [(1 - \kappa_t) p_{\text{init}}^{(j)}(x_j) + \kappa_t \delta_{z_j}(x_j)]$

Conditional Rate Matrix $Q_t^z(y|x)$ CTMC follows conditional path

$$Q_t^z(y|x) = (Q_t^z(v_i, j|x_j))_{v_i, j}$$
$$Q_t^z(v_i, j|x_j) = \frac{\dot{\kappa}_t}{1 - \kappa_t} (\delta_{z_j}(v_i) - \delta_{x_j}(v_i))$$

Marginal Prob. Path, Vector Field, and Score

	Notation	Key property	Formula
Marginal Probability Path	p_t	Interpolates p_{init} and p_{data}	$\sum_{z \in S} p_t(x z) p_{\text{data}}(z)$
Marginal Vector Field	$Q_t(y x)$	CTMC follows marginal path	$\sum_{z \in S} Q_t^z(y x) \frac{p_t(x z) p_{\text{data}}(z)}{p_t(x)}$

Marginal Rate Matrix for Factorized Mixture Path

Conditional
Rate Matrix

$$Q_t^z(y|x) = (Q_t^z(v_i, j|x_j))_{v_i, j}$$
$$Q_t^z(v_i, j|x_j) = \frac{\dot{\kappa}_t}{1 - \kappa_t} (\delta_{z_j}(v_i) - \delta_{x_j}(v_i))$$

Known
terminal
point!

Marginal
Rate Matrix

$$Q_t(y|x) = (Q_t(v_i, j|x))_{v_i, j}$$
$$Q_t(v_i, j|x) = \frac{\dot{\kappa}_t}{1 - \kappa_t} (p_{1|t}(z_j = v_i|x) - \delta_{x_j}(v_i))$$

Conditional probability!

Only unknown!

Discrete Flow Matching loss

Posterior probability network

$$p_{1|t}^{\theta}(z_j|x)$$

Learn via classification:

$$\mathcal{L}_{\text{DFM}}(\theta) = \mathbb{E}_{z \sim p_{\text{data}}, t \sim \text{Unif}_{[0,1]}, x \sim p_t(\cdot|z)} \left[\sum_{j=1}^d -\log p_{1|t}^{\theta}(z_j|x) \right]$$

Cross-entropy loss for every dimension!

Algorithm 8 Training factorized CTMC Model (Discrete Diffusion)

- Require:** Dataset of sequences $z \sim p_{\text{data}}$ with $z = (z_1, \dots, z_d) \in \mathcal{V}^d$;
initial (noise) token marginals $p_{\text{init}}^{(j)}$ on \mathcal{V} ; schedule $\kappa_t \in [0, 1]$;
posterior network f_θ returning per-position logits over \mathcal{V} ; optimizer OPT
- 1: **for** each training iteration **do**
 - 2: Sample a data point $z \sim p_{\text{data}}$
 - 3: Sample time $t \sim \text{Unif}[0, 1]$ and compute $\kappa \leftarrow \kappa_t$
 - 4: Sample a noisy state $x \sim p_t(\cdot | z)$ (factorized mixture path):
 - 5: **for** $j = 1, \dots, d$ (**in parallel**) **do**
 - 6: Sample mask $m_j \sim \text{Bernoulli}(\kappa)$
 - 7: Sample noise token $\xi_j \sim p_{\text{init}}^{(j)}$
 - 8: Set $x_j \leftarrow m_j z_j + (1 - m_j) \xi_j$
 - 9: **end for**
 - 10: $x \leftarrow (x_1, \dots, x_d)$
 - 11: Predict terminal-token posteriors via logits from the network:

$$\ell_j(\cdot) \leftarrow f_\theta(x, t)_j \quad \Rightarrow \quad p_{1|t}^\theta(v | x)_j = \text{Softmax}(\ell_j)(v)$$

- 12: Discrete Flow Matching loss (token-wise NLL of z):

$$\mathcal{L}_{\text{DFM}}(\theta) \leftarrow \sum_{j=1}^d \left[-\log p_{1|t}^\theta(z_j | x)_j \right]$$

- 13: Update parameters: $\theta \leftarrow \text{OPT.STEP}(\nabla_\theta \mathcal{L}_{\text{DFM}}(\theta))$
 - 14: **end for**
-

Mask Diffusion Language Models

Introduce new token into vocabulary: **[MASK]**

This token indicates that we masked the reference token

Initial distribution: $\delta_{[\text{MASK}]}$



Sampling from a Masked Language Model

Start with fully masked (initial distribution)

t=0.3

----- squad, -----
----- to remember ----- when -----
----- to -----
twenty adobe ----- of ----- water
----- along ----- polished ----- which -----
----- prehistoric ----- many
----- in -----



t=0.6

Many years later, as he faced --- ----- squad, -----
Buendía was to remember that distant ----- when his -----
took --- to discover ---- At that ----- was a village of
twenty adobe houses, built -- the bank of - river of clear water
that ran along - --- of polished stones, which were ----- and
----- like prehistoric eggs. --- ----- so recent ---- many
things lacked names, and in ----- -- ----- them it was
----- to -----



t=0.8

Many years later, as he faced the firing squad, Colonel Aureliano Buendía was to remember that distant ----- when his father took him to discover ice. At that time Macondo was a village of twenty adobe houses, built on the bank of - river of clear water that ran along a bed of polished stones, which were white and enormous, like prehistoric eggs. The world --- so recent that many things lacked names, and in order -- indicate them it was ----- to point.



t=1.0

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. The world was so recent that many things lacked names, and in order to indicate them it was necessary to point.



Diffusion LLMs generate text in arbitrary order

Austin et al., “Structured Denoising Diffusion Models in Discrete State-Spaces”, NeurIPS 2022

Discussion: Discrete Diffusion Models vs Autoregressive Models

Advantages

- Generate Multiple Tokens in Parallel → More **Speed?!**
- Generate Tokens in any order → **Text editing?!**
- New probability paths → **Can we design ones that make semantic sense?**

Disadvantages

- No KV caching → **Less Speed?!**
- Need to learn how to generate Tokens in *any* order → **Harder to learn?!**
- Autoregressive order (left-to-right) makes semantic sense → **Is it worth it?**

Continuous Flow Matching



The Discrete Flow Matching Matrix

