Yinuo Ren[1,2], Haoxuan Chen[1,2], Yuchen Zhu[1,3], Wei Guo[1,3]
Yongxin Chen[3], Grant M. Rotskoff[2], Molei Tao[3], Lexing Ying[2]
{yinuoren,haoxuanc,rotskoff,lexing}@stanford.edu,
{yzhu738, wei.guo, yongchen, mtao}@gatech.edu

[1]Equal Contribution [2]Stanford University [3]Georgia Institute of Technology

November 8, 2025

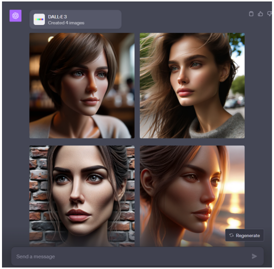
Fast Solvers for Discrete Diffusion Models:

Theory and Applications of High-Order Algorithms

Section 1: Discrete Diffusion Models (DMs): an Introduction and the Stochastic Integral Framework

Diffusion Models

Introduction



(a) DALL·E 3

(b) Stable Diffusion

(c) AI4Science

Figure: Diffusion and flow-based generative models have exerted huge impacts on scientific research in many fields.

Problem Setting: Discrete Diffusion Models

- **Task:** Sample from discrete data distribution $p_0 \in \Delta^{|\mathcal{X}|}$

Problem Setting: Discrete Diffusion Models

- **Task:** Sample from discrete data distribution $p_0 \in \Delta^{|\mathcal{X}|}$
- **Forward Continuous Time Markov Chain (CTMC):**

$$\frac{\mathrm{d}\boldsymbol{p}_t}{\mathrm{d}t} = \boldsymbol{Q}_t \boldsymbol{p}_t, \text{ with } Q_t(x,x) = -\sum_{y \neq x} Q_t(y,x) \text{ and } Q_t(x,y) \geq 0 \ (x \neq y)$$

Problem Setting: Discrete Diffusion Models

- **Task:** Sample from discrete data distribution $p_0 \in \Delta^{|\mathcal{X}|}$
- **Forward Continuous Time Markov Chain (CTMC):**

$$\frac{\mathrm{d}\boldsymbol{p}_t}{\mathrm{d}t} = \boldsymbol{Q}_t \boldsymbol{p}_t, \text{ with } Q_t(x,x) = -\sum_{y \neq x} Q_t(y,x) \text{ and } Q_t(x,y) \geq 0 \ (x \neq y)$$

- **Backward CTMC:**

$$\frac{\mathrm{d}\overleftarrow{\boldsymbol{p}}_s}{\mathrm{d}s} = \overline{\boldsymbol{Q}}_s \overleftarrow{\boldsymbol{p}}_s, \text{ with } \overline{Q}_s(y,x) = \begin{cases} \frac{\overleftarrow{p}_s(y)}{\overleftarrow{p}_s(x)} \overleftarrow{Q}_s(x,y), & \forall x \neq y \in \mathbb{X}, \\ -\sum_{y' \neq x} \overline{Q}_s(y',x), & \forall x = y \in \mathbb{X}. \end{cases}$$

with $\boldsymbol{p}_0 = \boldsymbol{p}_T$ (commonly uniform or Dirac-delta over Δ) and $\overleftarrow{p}_T = p_0$

Problem Setting: Discrete Diffusion Models

- **Score Function**: $\boldsymbol{s}_t(x) = (s_t(x,y))_{y \in \mathbb{X}} = \frac{\boldsymbol{p}_t}{p_t(x)}$ by optimizing

$$\int_0^T \psi_t \mathbb{E}_{x_t \sim p_t} \left[\sum_{y \neq x} \left(-\log \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} - 1 + \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} \right) s_t(x,y) Q_t(x,y) \right] \mathrm{d}t$$

Problem Setting: Discrete Diffusion Models

- **Score Function**: $\boldsymbol{s}_t(x) = (s_t(x, y))_{y \in \mathbb{X}} = \frac{\boldsymbol{p}_t}{p_t(x)}$ by optimizing

$$\int_0^T \psi_t \mathbb{E}_{x_t \sim p_t} \left[\sum_{y \neq x} \left(-\log \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} - 1 + \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} \right) s_t(x, y) Q_t(x, y) \right] \mathrm{d}t$$

- **Implementations:**

Problem Setting: Discrete Diffusion Models

- **Score Function**: $\boldsymbol{s}_t(x) = (s_t(x,y))_{y\in\mathbb{X}} = \frac{\boldsymbol{p}_t}{p_t(x)}$ by optimizing

$$\int_0^T \psi_t \mathbb{E}_{x_t\sim p_t}\left[\sum_{y\neq x}\left(-\log\frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} - 1 + \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)}\right) s_t(x,y)Q_t(x,y)\right]\mathrm{d}t$$

- **Implementations:**
  - Approximate samplers like τ-leaping scheme [Gil01, CBDB+22], Euler sampler and Tweedie τ-leaping sampler [LME24]

Problem Setting: Discrete Diffusion Models

- **Score Function**: $\boldsymbol{s}_t(x) = (s_t(x,y))_{y\in\mathbb{X}} = \frac{\boldsymbol{p}_t}{p_t(x)}$ by optimizing

$$\int_0^T \psi_t \mathbb{E}_{x_t\sim p_t}\left[\sum_{y\neq x}\left(-\log\frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)} - 1 + \frac{\widehat{s}_t^\theta(x,y)}{s_t(x,y)}\right)s_t(x,y)Q_t(x,y)\right]\mathrm{d}t$$

- **Implementations:**
  - Approximate samplers like τ-leaping scheme [Gil01, CBDB+22], Euler sampler and Tweedie τ-leaping sampler [LME24]
  - Exact samplers like uniformization [VD92, CY24] and first-hitting sampler [ZCM+24]

Mathematical Preliminaries

Poisson Random Measure with Evolving Intensity

$(\Omega, \mathcal{F}, \mathcal{P})$: probability space
$(\mathbb{X}, \mathcal{B}, \nu)$: measure space
$\lambda_t(y)$: a non-negative predictable process on $\mathbb{R}^+ \times \mathbb{X} \times \Omega$ satisfying

$$\int_0^T \int_{\mathbb{X}} 1 \vee |y| \vee |y|^2 \lambda_t(y) \nu(\mathrm{d}y)\mathrm{d}t < \infty, \text{ a.s..}$$

for any $T > 0$, The random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ on $\mathbb{R}^+ \times \mathbb{X}$ is called a *Poisson random measure with evolving intensity* $\lambda_t(y)$ if

Mathematical Preliminaries

Poisson Random Measure with Evolving Intensity

$(\Omega, \mathcal{F}, \mathcal{P})$: probability space
$(\mathbb{X}, \mathcal{B}, \nu)$: measure space
$\lambda_t(y)$: a non-negative predictable process on $\mathbb{R}^+ \times \mathbb{X} \times \Omega$ satisfying

$$\int_0^T \int_{\mathbb{X}} 1 \vee |y| \vee |y|^2 \lambda_t(y) \nu(\mathrm{d}y)\mathrm{d}t < \infty, \text{ a.s..}$$

for any $T > 0$, The random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ on $\mathbb{R}^+ \times \mathbb{X}$ is called a *Poisson random measure with evolving intensity* $\lambda_t(y)$ if

1. For any $B \in \mathcal{B}$ and $0 \le s < t$, $N[\lambda]((s,t] \times B) \sim \mathcal{P}\left(\int_s^t \int_B \lambda_\tau(y)\nu(\mathrm{d}y)\mathrm{d}\tau\right)$;
   Interpretation: the number of jumps of magnitude y during the infinitesimal time interval $(t, t+\mathrm{d}t]$ is Poisson distributed with mean $\lambda_t(\nu)\gamma(\mathrm{d}\nu)\mathrm{d}t$.

Mathematical Preliminaries

Poisson Random Measure with Evolving Intensity

$(\Omega, \mathcal{F}, \mathcal{P})$: probability space
$(\mathbb{X}, \mathcal{B}, \nu)$: measure space
$\lambda_t(y)$: a non-negative predictable process on $\mathbb{R}^+ \times \mathbb{X} \times \Omega$ satisfying

$$\int_0^T \int_{\mathbb{X}} 1 \vee |y| \vee |y|^2 \lambda_t(y) \nu(\mathrm{d}y) \mathrm{d}t < \infty, \text{ a.s..}$$

for any $T > 0$, The random measure $N[\lambda](\mathrm{d}t, \mathrm{d}y)$ on $\mathbb{R}^+ \times \mathbb{X}$ is called a *Poisson random measure with evolving intensity* $\lambda_t(y)$ if

1. For any $B \in \mathcal{B}$ and $0 \leq s < t$, $N[\lambda]((s,t] \times B) \sim \mathcal{P}\left(\int_s^t \int_B \lambda_\tau(y) \nu(\mathrm{d}y) \mathrm{d}\tau\right)$;
   Interpretation: the number of jumps of magnitude y during the infinitesimal time interval $(t, t+\mathrm{d}t]$ is Poisson distributed with mean $\lambda_t(\nu)\gamma(\mathrm{d}\nu)\mathrm{d}t$.
2. For any $t \geq 0$ and disjoint sets $\{B_i\}_{i \in [n]} \subset \mathcal{B}$, $\{N_t[\lambda](B_i) := N[\lambda]((0,t] \times B_i)\}_{i \in [n]}$ are independent stochastic processes.

Stochastic Integral Formulation: True Processes

Forward Process:

$$x_t = x_0 + \int_0^t \int_{\mathbb{D}} \nu N[\lambda](\mathrm{d}s, \mathrm{d}\nu),$$

where $\mathbb{D}$ is the set of all possible jumps and the intensity function

$$\lambda_t(\nu, \omega) = Q_t^0(x_{t^-}(\omega) + \nu, x_{t^-}(\omega))$$

if $x_{t^-}(\omega) + \nu \in \mathbb{X}$ and 0 otherwise.
True Backward Process ($\overleftarrow{s}_t$ denotes the true score):

$$y_s = y_0 + \int_0^s \int_{\mathbb{D}} \nu N[\mu](\mathrm{d}s, \mathrm{d}\nu),$$

where the intensity function

$$\mu_\rho(\nu) = \overleftarrow{s}_\rho(y_{\rho^-}, y_{\rho^-} + \nu) \overleftarrow{Q}_\rho^0(y_{\rho^-}, y_{\rho^-} + \nu)$$

if $y_{s^-} + \nu \in \mathbb{X}$ and 0 otherwise.

Stochastic Integral Formulation: τ-Leaping Scheme

τ-leaping inference scheme (Euler):

$$\widehat{y}_{t+\Delta} = \widehat{y}_t + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\widehat{\mu}_t(\nu)\Delta\right)$$

with Δ being the time step and the estimated intensity function

$$\widehat{\mu}_\rho(\nu) = \widehat{s}^\theta_\rho(\widehat{y}_{\rho^-}, \widehat{y}_{\rho^-} + \nu)\overleftarrow{Q}^0_\rho(\widehat{y}_{\rho^-}, \widehat{y}_{\rho^-} + \nu)$$

Stochastic Integral formulation (equivalent to τ-leaping):

$$\widehat{y}_{t+\Delta} = \widehat{y}_t + \int_t^{t+\Delta} \int_{\mathbb{D}} \nu N[\widehat{\mu}](\mathrm{d}s, \mathrm{d}\nu)$$

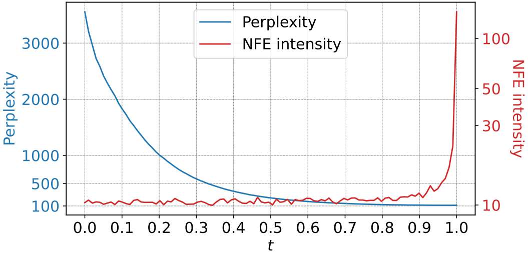
Error bound of τ-leaping [RCRY24]

$$D_{\mathrm{KL}}(p_0 \| \widehat{q}_T) \lesssim \underbrace{\exp(-T)}_{\text{truncation error}} + \underbrace{\epsilon}_{\text{score estimation error}} + \boxed{\underbrace{\kappa T}_{\text{numerical error}}}$$

Section 2:
High-Order Inference Algorithms for Discrete DMs: Formulation and Theory

Motivation: Redundancy of Exact Samplers

Discrete diffusion models with uniformization for text generation:



- x-axis: time of the backward process; y-axis: frequency of jumps (NFE).
- Perplexity convergence occurs before the NFE grows unbounded!

Methodology: High-Order Numerical Solvers

- Question: How to design faster **approximate** inference algorithms yielding better performance and error bound with the same NFE?

Methodology: High-Order Numerical Solvers

- Question: How to design faster **approximate** inference algorithms yielding better performance and error bound with the same NFE?
- **Motivation: Runge-Kutta-2 methods for solving the ODE** $\mathrm{d}x_t = f_t(x_t)\mathrm{d}t$

$$\begin{aligned}
\widehat{x}^*_{t+\theta\Delta} &= \widehat{x}_t + f_t(\widehat{x}_t)\theta\Delta \\
\widehat{x}_{t+\Delta} &= \widehat{x}_t + \left[(1-\tfrac{1}{2\theta})f_t(\widehat{x}_t) + \tfrac{1}{2\theta}f_{t+\theta\Delta}(\widehat{x}^*_{t+\theta\Delta})\right]\Delta.
\end{aligned}$$

where $\theta \in (0,1)$.

Methodology: High-Order Numerical Solvers

- Question: How to design faster **approximate** inference algorithms yielding better performance and error bound with the same NFE?
- **Motivation: Runge-Kutta-2 methods for solving the ODE** $\mathrm{d}x_t = f_t(x_t)\mathrm{d}t$

$$\begin{aligned}\widehat{x}^*_{t+\theta\Delta} &= \widehat{x}_t + f_t(\widehat{x}_t)\theta\Delta\\ \widehat{x}_{t+\Delta} &= \widehat{x}_t + \left[(1-\tfrac{1}{2\theta})f_t(\widehat{x}_t) + \tfrac{1}{2\theta}f_{t+\theta\Delta}(\widehat{x}^*_{t+\theta\Delta})\right]\Delta.\end{aligned}$$

where $\theta \in (0,1)$.

- Notations: *time discretization scheme* $(s_i)_{i\in[0:N]}$ and θ*-section points* $(\rho_n)_{n\in[0:N]}$ are denoted by

$$0 = s_0 < s_1 < \cdots < s_N = T-\delta,\ \rho_n = (1-\theta)s_n + \theta s_{n+1}$$

Take $\alpha_1 = \frac{1}{2\theta(1-\theta)}$ and $\alpha_2 = \frac{(1-\theta)^2+\theta^2}{2\theta(1-\theta)}$ with $\alpha_1 - \alpha_2 = 1$

Methodology: High-Order Numerical Solvers

- **θ-RK-2 Method**: Interpolation between s_n and ρ_n

$$\widehat{y}^*_{\rho_n} \leftarrow \widehat{y}_{s_n} + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\widehat{\mu}_{s_n}(\nu)\theta\Delta_n\right),$$

$$\widehat{y}_{s_{n+1}} \leftarrow \widehat{y}_{s_n} + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\mathbf{1}_{\widehat{\mu}_{s_n} > 0}\left[\left(1 - \tfrac{1}{2\theta}\right)\widehat{\mu}_{s_n} + \tfrac{1}{2\theta}\widehat{\mu}^*_{\rho_n}\right]_+(\nu)\Delta_n\right).$$

Analogy: SDE [BB96], continuous diffusion model [KAAL22, WCW24] and chemical reaction [BT04]

Methodology: High-Order Numerical Solvers

- **θ-RK-2 Method**: Interpolation between s_n and ρ_n

$$\widehat{y}^*_{\rho_n} \leftarrow \widehat{y}_{s_n} + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\widehat{\mu}_{s_n}(\nu)\theta\Delta_n\right),$$
$$\widehat{y}_{s_{n+1}} \leftarrow \widehat{y}_{s_n} + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\mathbf{1}_{\widehat{\mu}_{s_n}>0}\left[\left(1-\tfrac{1}{2\theta}\right)\widehat{\mu}_{s_n} + \tfrac{1}{2\theta}\widehat{\mu}^*_{\rho_n}\right]_+(\nu)\Delta_n\right).$$

  Analogy: SDE [BB96], continuous diffusion model [KAAL22, WCW24] and chemical reaction [BT04]

- **θ-Trapezoidal Method**: Extrapolation beyond s_n and ρ_n

$$\widehat{y}^*_{\rho_n} \leftarrow \widehat{y}_{s_n} + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\widehat{\mu}_{s_n}(\nu)\theta\Delta_n\right),$$
$$\widehat{y}_{s_{n+1}} \leftarrow \widehat{y}^*_{\rho_n} + \sum_{\nu \in \mathbb{D}} \nu \mathcal{P}\left(\left(\alpha_1\widehat{\mu}^*_{\rho_n} - \alpha_2\widehat{\mu}_{s_n}\right)_+(\nu)(1-\theta)\Delta_n\right).$$

  Analogy: SDE [AM11] and chemical reaction [HLM11]

Theoretical Analysis: Assumptions

- **Exponential convergence of the forward process:** $D_{\mathrm{KL}}(p_T \| p_\infty) \lesssim \exp(-T)$.
- **Regularity of intensity:** Both the true intensity μ_s and the estimated intensity $\widehat{\mu}_s$ are in $\mathcal{C}^2$ and bounded for any $s \in [0, T-\delta]$.
- **Score estimation error:** The following error bounds hold for any grid point or θ-section point $s \in \cup_{n=0}^{N-1} \{s_n, \rho_n\}$:

$$\mathbb{E}\left[\sum_{\nu \in \mathbb{D}} \left(\mu_s(\nu) \left(\log \frac{\mu_s(\nu)}{\widehat{\mu}_s(\nu)} - 1\right) + \widehat{\mu}_s(\nu)\right)\right] \le \epsilon_{\mathrm{I}},$$

$$\mathbb{E}\left[\sum_{\nu \in \mathbb{D}} |\mu_s(\nu) - \widehat{\mu}_s(\nu)|\right] \le \epsilon_{\mathrm{II}}.$$

.

Theoretical Analysis: Informal Bounds

- **θ-RK-2 Method**: Suppose $\theta \in (0, \frac{1}{2}]$ and $(1 - \frac{1}{2\theta})\widehat{\mu}_{\lfloor s \rfloor} + \frac{1}{2\theta}\widehat{\mu}^*_{\rho_s} \geq 0$, then

$$D_{\mathrm{KL}}\left(p_0 \| \widehat{q}_T^{\mathrm{RK}}\right) \lesssim \exp(-T) + (\epsilon_{\mathrm{I}} + \epsilon_{\mathrm{II}})T + \boxed{\kappa^2 T}$$

- **θ-Trapezoidal Method**: Suppose $\theta \in (0, 1]$ and $\alpha_1 \widehat{\mu}^*_{\rho_s} - \alpha_2 \widehat{\mu}_{\lfloor s \rfloor} \geq 0$, then

$$D_{\mathrm{KL}}\left(p_0 \| \widehat{q}_T^{\mathrm{trap}}\right) \lesssim \exp(-T) + (\epsilon_{\mathrm{I}} + \epsilon_{\mathrm{II}})T + \boxed{\kappa^2 T}$$

Theoretical Analysis: Informal Bounds

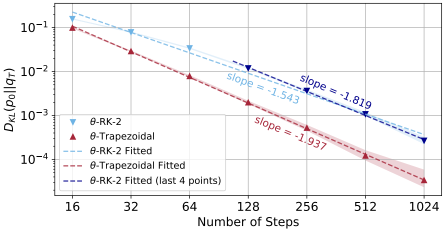
- **Proof Technique:** Change of measure [RCRY24] between stochastic integral with respect to *Poisson random measures*, error decomposition, and the usage of *Dynkin's Formula*;
- **Second-Order Numerical Error Guarantee:** $\mathcal{O}(\kappa^2 T)$ for both θ-RK-2 and θ-Trapezoidal methods vs. $\mathcal{O}(\kappa T)$ for τ-leaping method;
- **Range of θ:** $\theta \in (0, 1/2]$ for RK-2 method and $\theta \in (0, 1]$ for trapezoidal method, which is caused by application of *Jensen's Inequality*.

Section 3:

Experiments: Toy Model, Text and Image Generation, Math Reasoning

Experiment: Toy Model

Data: synthetic discrete distribution with $S = 15$ states, dimension $d = 1$



The figure above plots empirical KL divergence between the true and generated distribution of the toy model versus the number of steps. Data are fitted with linear regression with 95% confidence interval by bootstrapping.

Experiment: Text Generation

Data: OpenWebText dataset [GC19], number of states $S = 50258$, dimension $d = 1024$;
Model: RADD [ONX$^+$24] (pretrained score of a masked discrete DM, GPT2-level)

Table: Generative perplexity of texts generated by different sampling algorithms on GPT-2 large. Lower values are better, with the best in **bold**.

Method	NFE $= 16$	NFE $= 32$	NFE $= 64$	NFE $= 128$	NFE $= 256$	NFE $= 512$	NFE $= 1024$
FHS	≤ 307.425	≤ 186.594	≤ 141.625	≤ 122.732	≤ 113.310	≤ 113.026	≤ 109.406
Euler	≤ 277.962	≤ 160.586	≤ 111.597	≤ 86.276	≤ 68.092	≤ 55.622	≤ 44.686
Tweedie τ-leaping	≤ 277.133	≤ 160.248	≤ 110.848	≤ 85.738	≤ 70.102	≤ 55.194	≤ 44.257
τ-leaping	≤ 126.835	≤ 96.321	≤ 69.226	≤ 52.366	≤ 41.694	≤ 33.789	≤ 28.797
Semi-AR	≤ 2857.469	≤ 1543.302	≤ 741.184	≤ 360.793	≤ 222.303	≤ 164.162	≤ 147.406
θ-RK-2	≤ 127.363	≤ 109.351	≤ 86.102	≤ 64.317	≤ 49.816	≤ 40.375	≤ 33.971
θ-Trapezoidal	$\mathbf{\leq 123.585}$	$\mathbf{\leq 89.912}$	$\mathbf{\leq 66.549}$	$\mathbf{\leq 49.051}$	$\mathbf{\leq 39.959}$	$\mathbf{\leq 32.456}$	$\mathbf{\leq 27.553}$

Experiment: Text Generation

Table: Generative perplexity of texts generated by different sampling algorithms on LLaMA 3. Lower values are better, with the best in **bold**.

Method	NFE $= 16$	NFE $= 32$	NFE $= 64$	NFE $= 128$	NFE $= 256$	NFE $= 512$	NFE $= 1024$
FHS	≤ 342.498	≤ 210.742	≤ 155.258	≤ 132.135	≤ 127.526	≤ 123.013	≤ 120.791
Euler	≤ 318.413	≤ 175.555	≤ 125.955	≤ 91.051	≤ 75.245	≤ 59.971	≤ 49.406
Tweedie τ-leaping	≤ 316.744	≤ 172.941	≤ 121.248	≤ 94.253	≤ 75.403	≤ 59.943	≤ 49.239
τ-leaping	≤ 152.867	≤ 117.930	≤ 86.980	≤ 68.090	$\boldsymbol{\leq 53.664}$	≤ 44.676	≤ 38.293
Semi-AR	≤ 2696.883	≤ 1684.973	≤ 829.391	≤ 410.177	≤ 251.963	≤ 166.927	≤ 162.093
θ-RK-2	≤ 150.439	≤ 132.090	≤ 107.066	≤ 80.742	≤ 63.277	≤ 52.563	≤ 44.687
θ-Trapezoidal	$\boldsymbol{\leq 146.027}$	$\boldsymbol{\leq 113.260}$	$\boldsymbol{\leq 83.456}$	$\boldsymbol{\leq 66.071}$	≤ 54.307	$\boldsymbol{\leq 44.293}$	$\boldsymbol{\leq 35.524}$

Experiment: Image Generation

Data: ImageNet [DDS+09], number of states $S = 1025$, dimension $d = 256$, (VQ-GAN based tokenization)
Model: MaskGIT [CZJ+22]

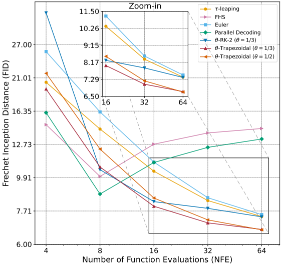


Figure: FID of images generated vs. NFEs with different parameter choices.

Experiment: Image Generation

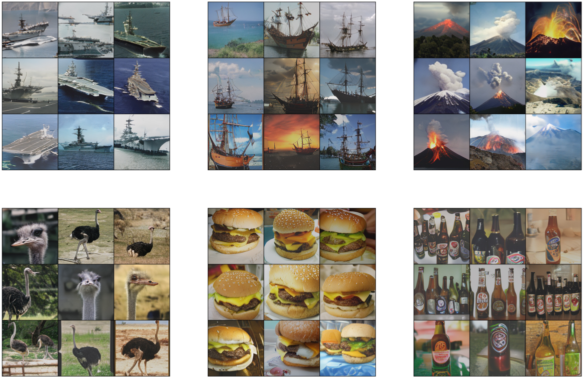


Figure: Visualization of samples generated by θ-Trapezoidal. **Upper Left**: Aircraft carrier (ImageNet-1k class: **933**); **Upper Middle**: Pirate; **Upper Right**: Volcano ; **Lower Left**: Ostrich; **Lower Middle**: Cheeseburger ; **Lower Right**: Beer bottle

Experiment: Math Reasoning

Data: GSM8K dataset [CKB+21]
Model: Large Language Diffusion Model (LLaDA)-Instruct8B [NZY+25]

Table: Response accuracy on GSM8K with different NFEs.

Accuracy (%)	NFE $= 64$	NFE $= 128$	NFE $= 256$
Semi-AR (Conf.)	33.6	32.0	39.1
Semi-AR (Rand.)	33.8	34.3	40.3
θ-Trapezoidal	35.1	38.4	39.7

Thank you for your attention!

References I

David F Anderson and Jonathan C Mattingly, *A weak trapezoidal method for a class of stochastic differential equations*, Communications in Mathematical Sciences **9** (2011), no. 1, 301–318.

Kevin Burrage and Pamela Marion Burrage, *High strong order explicit runge-kutta methods for stochastic ordinary differential equations*, Applied Numerical Mathematics **22** (1996), no. 1-3, 81–101.

K Burrage and T Tian, *Poisson runge-kutta methods for chemical reaction systems in advances in scientific computing and applications*, 2004.

Andrew Campbell, Joe Benton, Valentin De Bortoli, Thomas Rainforth, George Deligiannidis, and Arnaud Doucet, *A continuous time framework for discrete denoising models*, Advances in Neural Information Processing Systems **35** (2022), 28266–28279.

References II

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al., *Training verifiers to solve math word problems*, arXiv preprint arXiv:2110.14168 (2021).

Hongrui Chen and Lexing Ying, *Convergence analysis of discrete diffusion model: Exact implementation through uniformization*, arXiv preprint arXiv:2402.08095 (2024).

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman, *Maskgit: Masked generative image transformer*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 11315–11325.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, *Imagenet: A large-scale hierarchical image database*, 2009 IEEE conference on computer vision and pattern recognition, Ieee, 2009, pp. 248–255.

Aaron Gokaslan and Vanya Cohen, *Openwebtext corpus*, `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

References III

- Daniel T Gillespie, *Approximate accelerated stochastic simulation of chemically reacting systems*, The Journal of chemical physics **115** (2001), no. 4, 1716–1733.
- Yucheng Hu, Tiejun Li, and Bin Min, *A weak second order tau-leaping method for chemical kinetic systems*, The Journal of chemical physics **135** (2011), no. 2.
- Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine, *Elucidating the design space of diffusion-based generative models*, Advances in Neural Information Processing Systems **35** (2022), 26565–26577.
- Aaron Lou, Chenlin Meng, and Stefano Ermon, *Discrete diffusion modeling by estimating the ratios of the data distribution*, Forty-first International Conference on Machine Learning, 2024.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li, *Large language diffusion models*, arXiv preprint arXiv:2502.09992 (2025).

References IV

- Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li, *Your absorbing discrete diffusion secretly models the conditional distributions of clean data*, arXiv preprint arXiv:2406.03736 (2024).
- Yinuo Ren, Haoxuan Chen, Grant M Rotskoff, and Lexing Ying, *How discrete and continuous diffusion meet: Comprehensive analysis of discrete diffusion models via a stochastic integral framework*, arXiv preprint arXiv:2410.03601 (2024).
- Nico M Van Dijk, *Uniformization for nonhomogeneous markov chains*, Operations research letters **12** (1992), no. 5, 283–291.
- Yuchen Wu, Yuxin Chen, and Yuting Wei, *Stochastic Runge-Kutta methods: Provable acceleration of diffusion models*, arXiv preprint arXiv:2410.04760 (2024).

References V

Kaiwen Zheng, Yongxin Chen, Hanzi Mao, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang, *Masked diffusion models are secretly time-agnostic masked models and exploit inaccurate categorical sampling*, arXiv preprint arXiv:2409.02908 (2024).