Discrete Diffusion Modeling by Estimating the Ratios of the Data Distribution

Aaron Lou [1] Chenlin Meng [1 2] Stefano Ermon [1]

ICML 2024 Oral

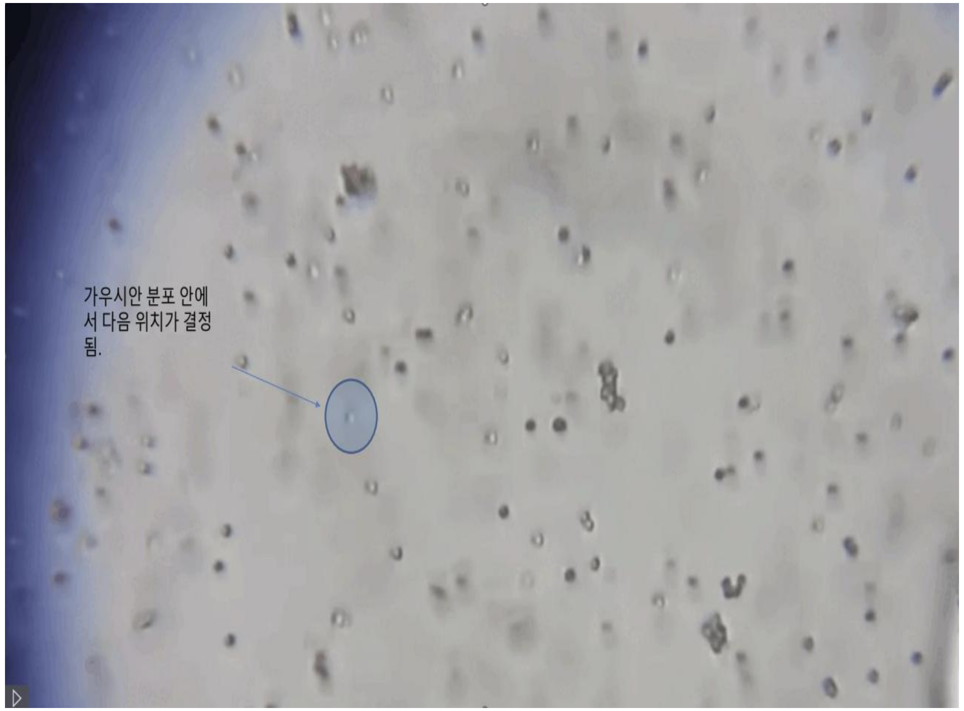
고경빈

2025.04.03

Background

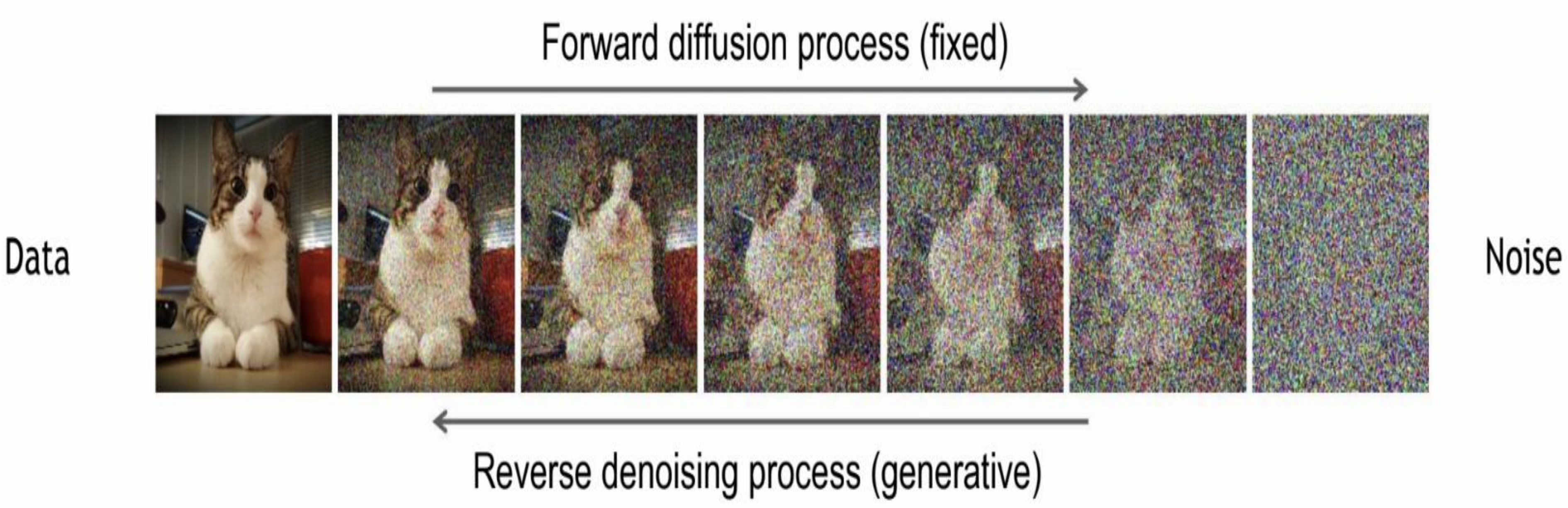
- Autoregressive models achieve impressive results, but face key limitations
  - Slow sequential token sampling
  - Limited controllability
  - Performance degradation without distribution annealing
- Many works have extended diffusion models to language domains
  - No approach yet matches autoregressive models without heavy tuning

Physical Intuition



Denoising Diffusion Models

- Forward diffusion process that gradually adds noise to input
- Reverse denoising process that learns to generate data by denoising



[Sohl-Dickstein et al., Deep Unsupervised Learning using Nonequilibrium Thermodynamics, ICML 2015](#)
[Ho et al., Denoising Diffusion Probabilistic Models, NeurIPS 2020](#)

Discrete Diffusion Process

- Forward Discrete Diffusion Process

$$\frac{dp_t}{dt} = Q_t p_t \quad p_0 \approx p_{\text{data}} \tag{1}$$

- Euler Approximation of Forward Transition Probability

$$p(x_{t+\Delta t} = y | x_t = x) = \delta_{xy} + Q_t(y, x)\Delta t + O(\Delta t^2) \tag{2}$$

- Reverse Diffusion Process via Score-Weighted Transitions

$$\frac{dp_{T-t}}{dt} = \overline{Q}_{T-t} p_{T-t} \quad \overline{Q}_t(y, x) = \frac{p_t(y)}{p_t(x)} Q_t(x, y)$$
$$\overline{Q}_t(x, x) = -\sum_{y \neq x} \overline{Q}_t(y, x) \tag{3}$$

Discrete Diffusion Models

- Concrete Score Matching

$$\mathcal{L}_{\mathrm{CSM}} = \frac{1}{2}\mathbb{E}_{x \sim p_t}\left[\sum_{y \neq x}\left(s_\theta(x_t, t)_y - \frac{p_t(y)}{p_t(x)}\right)^2\right] \quad (4)$$

  - L2 loss allows zero/negative values for $\frac{p_t(y)}{p_t(x)}$, causing divergence
  - Training fails in practice (10,000× worse perplexity)

Score Entropy Discrete Diffusion Models

- What we want?

$$s_\theta(x,t) \approx \left[\frac{p_t(y)}{p_t(x)}\right]_{y\neq x}$$

- Score Entropy Loss

$$\mathbb{E}_{x\sim p}\left[\sum_{y\neq x} w_{xy}\left(s_\theta(x)_y - \frac{p(y)}{p(x)}\log s_\theta(x)_y + K\left(\frac{p(y)}{p(x)}\right)\right)\right] \quad (5)$$

  - Bregman Divergence

$$D_F\left(s(x)_y, \frac{p(y)}{p(x)}\right) \; when \; F = -\log \rightarrow D_{-log}(u,v) = u - vlog + v(logv - 1)$$

    - Non-negative
    - Symmeric
    - Convex
    - Generalize standard cross entropy to general positive values

Score Entropy Properties

- Consistency
  - With infinite data and model size, optimal θ* makes $s_{\theta^*}(x)_y = \frac{p(y)}{p(x)}$ and $L_{SE} = 0$
- Improves concrete score matching by rescaling problematic gradients

$$w_{xy} = 1, \nabla_{s_\theta(x)_y} \mathcal{L}_{\text{SE}} = \frac{1}{s_\theta(x)_y} \nabla_{s_\theta(x)_y} \mathcal{L}_{\text{CSM}}$$

- Score entropy is tractable without $\frac{p(y)}{p(x)}$
  - Implicit Score Entropy

$$\mathcal{L}_{\text{ISE}} = \mathbb{E}_{x \sim p} \left[\sum_{y \neq x} w_{xy} s_\theta(x)_y - w_{yx} \log s_\theta(y)_x \right] \quad (6)$$

  - Denoising Score Entropy

$$\mathop{\mathbb{E}}_{\substack{x_0 \sim p_0 \\ x \sim p(\cdot|x_0)}} \left[\sum_{y \neq x} w_{xy} \left(s_\theta(x)_y - \frac{p(y|x_0)}{p(x|x_0)} \log s_\theta(x)_y \right) \right] \quad (7)$$

Likelihood Bound For Score Entropy Discrete Diffusion

- Score Entropy enables ELBO for likelihood-based training and evaluation
  - $Q_t^\theta(y,x) = \begin{cases} s_\theta(x,t)_y \cdot Q_t(x,y)_t, & x \neq y \\ -\sum_{z \neq x} Q_t^\theta(z,x), & x = y \end{cases}$ ➡ $\frac{dp_{T-t}^\theta}{dt} = \overline{Q}_{T-t}^\theta p_{T-t}^\theta \quad p_T^\theta = p_{\text{base}} \approx p_T \qquad (8)$
- Log-likelihood is upper-bounded by score entropy and KL

$$-\log p_0^\theta(x_0) \leq \mathcal{L}_{\text{DWDSE}}(x_0) + D_{KL}(p_{T|0}(\cdot|x_0) \parallel p_{\text{base}}) \qquad (9)$$

- DWDSE measures score error weighted by diffusion steps

$$\int_0^T \mathbb{E}_{x_t \sim p_{t|0}(\cdot|x_0)} \sum_{y \neq x_t} Q_t(x_t, y) \left(s_\theta(x_t, t)_y - \frac{p_{t|0}(y|x_0)}{p_{t|0}(x_t|x_0)} \log s_\theta(x_t, t)_y + K\left(\frac{p_{t|0}(y|x_0)}{p_{t|0}(x_t|x_0)}\right)\right) dt \qquad (10)$$

Practical Implementation

- Score entropy can be scaled to high dimensional tasks

$$Q_t(x^1 \dots x^i \dots x^d, x^1 \dots \widehat{x}^i \dots x^d) = Q_t^{\text{tok}}(x^i, \widehat{x}^i) \quad (11)$$

$$(s_\theta(x^1 \dots x^i \dots x^d, t))_{i,\widehat{x}^i} \approx \frac{p_t(x^1 \dots \widehat{x}^i \dots x^d)}{p_t(x^1 \dots x^i \dots x^d)} \quad (12)$$

- The sequence transition factorizes into independent token transitions

$$p_{t|0}^{\text{seq}}(\widehat{\mathbf{x}}|\mathbf{x}) = \prod_{i=1}^{d} p_{t|0}^{\text{tok}}(\widehat{x}^i|x^i) \quad (13)$$

- Token transitions come from the exponential of the noise-scaled matrix

$$p_{t|0}^{\text{tok}}(\cdot|x) = x\text{-th column of } \exp\left(\overline{\sigma}(t) Q^{\text{tok}}\right) \quad (14)$$

  - $\frac{dp_t}{dt} = Q_t p_t \rightarrow \frac{dp_t}{dt} = \sigma(t) Q p_t, \quad p_t = \exp(\int_0^t \sigma(s) ds \cdot Q) p_0$

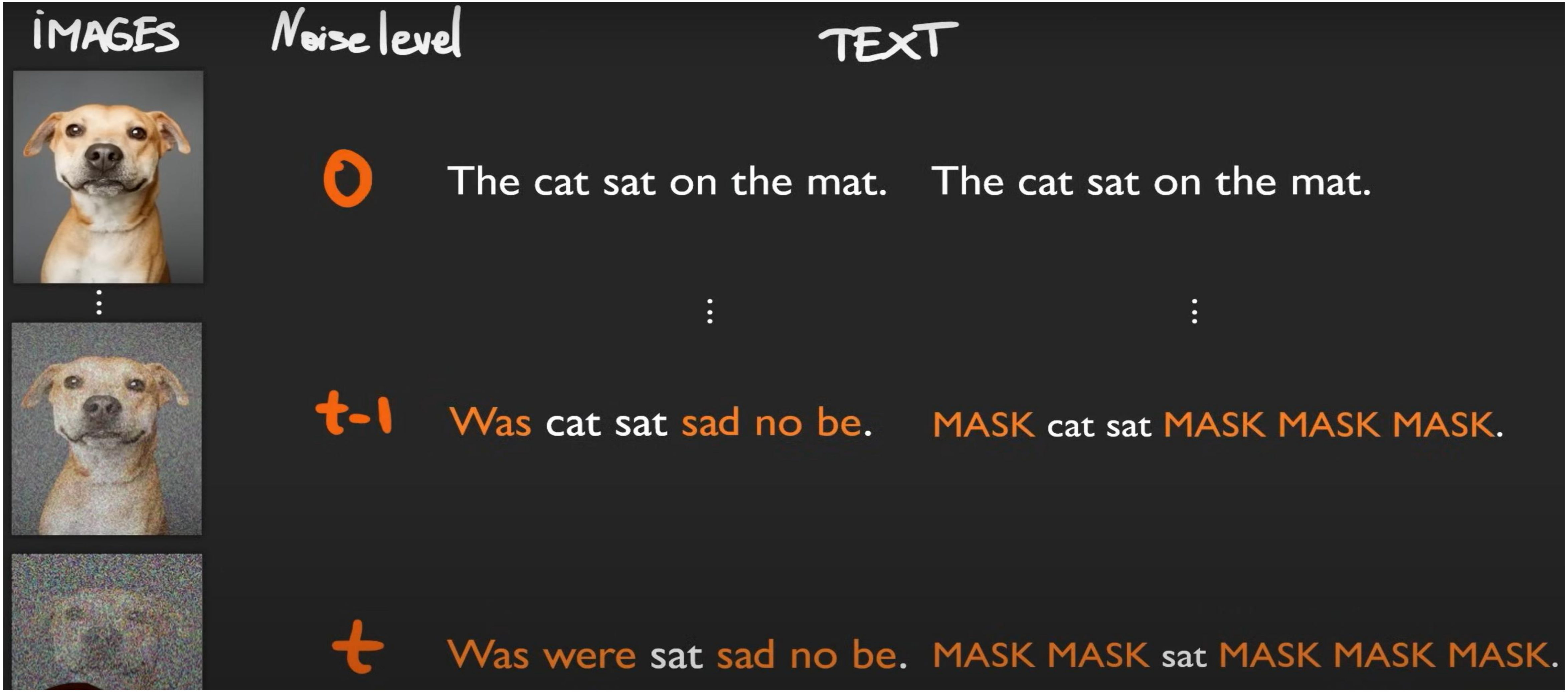
Practical Implementation

- But, most Q^{tok} unusable for large scale experiment
  - Not able to store all edge weights $\mathrm{Q}^{tok}(i,j)$
  - Extremely slow to access
  - Avoid matrix-matrix multiplication in computing exp columns
- Solutions

$$Q^{\text{uniform}} = \begin{bmatrix} 1-N & 1 & \cdots & 1 \\ 1 & 1-N & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1-N \end{bmatrix} \quad (15)$$

$$Q^{\text{absorb}} = \begin{bmatrix} -1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 0 \\ 1 & 1 & \cdots & 1 & 0 \end{bmatrix} \quad (16)$$

Practical Implementation



Simulating Reverse Diffusion with Concrete Scores

- Time-Reversal Strategies
  - τ_leaping: $x_{t-\Delta t} \sim \delta_{x_t^i}(x_{t-\Delta t}^i) + \Delta t Q_t^{\mathrm{tok}}(x_t^i, x_{t-\Delta t}^i) s_\theta(\mathbf{x}_t, t)_{i, x_{t-\Delta t}^i}$ (17)
  - Discrete Tweedie's theorem

$$p_{0|t}(x_0|x_t) = \left(\exp(-tQ) \left[\frac{p_t(i))}{p_t(x_t)} \right]_{i=1}^N \right)_{x_0} \exp(tQ)(x_t, x_0) \quad (18)$$

  - Tweedie τ_leaping
    - $p_{t-\Delta t|t}^{tweedie}(x_{t-\Delta t}|x_t) = \big(\exp(-\sigma_t^{\Delta t} Q) s_\theta(\mathbf{x}_t, t)_i \big)_{x_{t-\Delta t}^i} \exp(\sigma_t^{\Delta t} Q)(x_t^i, x_{t-\Delta t}^i)$ (19)

$$\text{where } \sigma_t^{\Delta t} = (\overline{\sigma}(t) - \overline{\sigma}(t - \Delta t)) \quad (20)$$

Arbitrary Prompting and Infilling

- Unconditionally trained models can support arbitrary position conditioning
- Infilling: $p_t(\mathrm{x}^{\Omega} | \mathrm{x}^{\overline{\Omega}} = \mathrm{y})$ Ω unfilled indices $\overline{\Omega}$ filled (21)
- Apply Bayes' rule

$$\frac{p_t(\mathbf{x}^{\Omega} = \mathbf{z}' | \mathbf{x}^{\overline{\Omega}} = \mathbf{y})}{p_t(\mathbf{x}^{\Omega} = \mathbf{z} | \mathbf{x}^{\overline{\Omega}} = \mathbf{y})} = \frac{p_t(\mathbf{x} = \mathbf{z}' \oplus_{\Omega} \mathbf{y})}{p_t(\mathbf{x} = \mathbf{z} \oplus_{\Omega} \mathbf{y})} \qquad (22)$$

Language Modeling Comparison

- Text 8 Dataset

Type	Method	BPC (↓)
Autoregressive Backbone	IAF/SCF	1.88
	AR Argmax Flow	1.39
	Discrete Flow	**1.23**
	Autoregressive	**1.23**
Non-autoregressive	Mult. Diffusion	$\leq$ 1.72
	MAC	$\leq$ 1.40
	BFN	$\leq$ 1.41
	D3PM Uniform	$\leq$ 1.61
	D3PM Absorb	$\leq$ 1.45
Ours (NAR)	SEDD Uniform	$\leq$ 1.47
	SEDD Absorb	$\leq$ **1.39**

  - SEDD outperforms D3PM and approaches autoregressive performance

Language Modeling Comparison

- One Billion Words Dataset

Type	Method	Perplexity (↓)
Autoregressive	Transformer	**31.98**
Diffusion	D3PM Absorb	$\leq$ 77.50
	Diffusion-LM	$\leq$ 118.62
	BERT-Mouth	$\leq$ 142.89
	DiffusionBert	$\leq$ 63.78
Ours (Diffusion)	SEDD Uniform	$\leq$ 40.25
	SEDD Absorb	$\leq$ **32.79**

- SEDD shows 50–75% lower perplexity than other diffusion models
- Matches autoregressive models, proving non-autoregressive can compete

Language Modeling Comparison

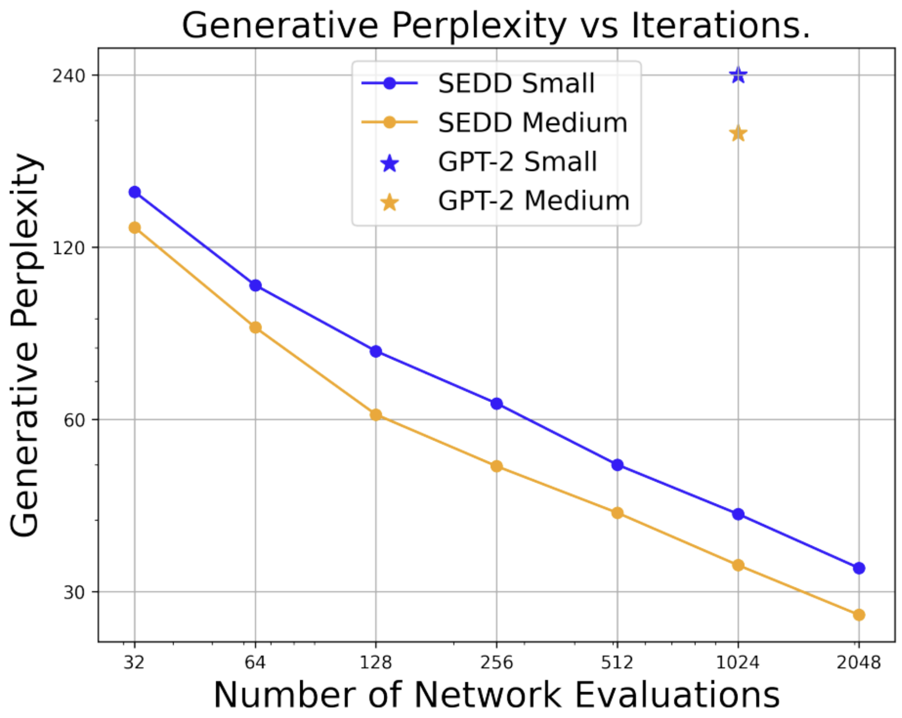
- GPT-2 Zero Shot Tasks

Size	Model	LAMBADA	WikiText2	PTB	WikiText103	1BW
Small	GPT-2	**45.04**	42.43	138.43	41.60	**75.20**
	SEDD Absorb	$\leq$50.92	$\leq$**41.84**	$\leq$**114.24**	$\leq$**40.62**	$\leq$79.29
	SEDD Uniform	$\leq$65.40	$\leq$50.27	$\leq$140.12	$\leq$49.60	$\leq$101.37
	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
Medium	GPT-2	**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
	SEDD Uniform	$\leq$51.28	$\leq$38.93	$\leq$102.28	$\leq$36.81	$\leq$79.12

  - SEDD Absorb achieves lower perplexity than GPT-2 on 3 out of 5 datasets
  - Best performance among all diffusion-based models
  - First non-autoregressive model to rival GPT-2

Language Generation Comparison

- Unconditional Generation



(a) Generative Perplexity (↓) vs. Sampling Iterations.

GPT-2 S	a hiring platform that "includes a fun club meeting place," says petitioner's AQQFredericks. They's the adjacent marijuana-hop. Others have allowed 3B Entertainment
GPT-2 M	misused, whether via Uber, a higher-order reality of quantified impulse or the No Mass Paralysis movement, but the most shamefully universal example is gridlock
SEDD S	As Jeff Romer recently wrote, "The economy has now reached a corner - 64% of household wealth and 80% of wealth goes to credit cards because of government austerity
SEDD M	Wyman worked as a computer science coach before going to work with the U.S. Secret Service in upstate New York in 2010. Without a license, the Secret Service will have to

(b) Generated Text (small models)

Infilling Conditional Generation

A bow and arrow is a traditional weapon that enables an attacker to attack targets at a range within a meter or maybe two meters. They have a range far longer than a human can walk, and they can be fired …
… skydiving is a fun sport that makes me feel incredibly silly. I think I may've spent too much, but it could've been amazing! While sky diving gives us exercise and fun, scuba diving is an act of physical fitness, …
… no one expected the results to much better than last year's one-sided endorsement. Nearly 90 percent of the results were surveyed as "independent," an promising result for school children across the country.
… results show that Donald Trump and Hillary Clinton are in 38 states combined with less than 1% of the national vote. In a way, it's Trump and Hillary Clinton who will work overtime to get people to vote this …

Method	Annealing	Mauve (↑)
GPT-2	Nucleus-0.95	0.955
	None	0.802
SSD-LM	Logit Threshold-0.95	0.919
	None	0.312
SEDD Standard	None	**0.957**
SEDD Infill	None	0.942

Conclusion

- Proposes Score Entropy for discrete diffusion via probability ratio
- SEDD outperforms D3PM, rivals GPT-2 on some tasks
- Enables high-quality text generation without annealing
- Supports infilling and flexible prompts

My Review

- I overlooked this paper, but that was a big mistake
- I learned a lot about diffusion and the difference between continuous and discrete data
- I think this paper breaks the belief that diffusion doesn't work in the natural language domain

Open Question

- Does the scaling law apply to diffusion models?
- How can we make diffusion language models smarter?