NeurIPS22

Diffusion-LM Improves Controllable Text Generation

Xiang Lisa Li
Stanford University
xlisali@stanford.edu

John Thickstun
Stanford University
jthickst@stanford.edu

Ishaan Gulrajani
Stanford Univeristy
igul@stanford.edu

Percy Liang
Stanford Univeristy
pliang@cs.stanford.edu

Tatsunori B. Hashimoto
Stanford Univeristy
thashim@stanford.edu

ECNU—纪焘

BG: Controllable Text Generation

- **Generative Models:** $p_{\text{lm}}(\mathbf{w})$

$$\mathbf{w} = [w_1 \cdots w_n]$$

$$p_{\text{lm}}(\mathbf{W}) = \prod_{i=1}^{n} p\left(w_i \mid w_0, \cdots, w_{i-1}\right)$$

BG: Controllable Text Generation

- **Generative Models:** $p_{\text{lm}}(\mathbf{w})$

$$\mathbf{w} = [w_1 \cdots w_n]$$

$$p_{\text{lm}}(\mathbf{W}) = \prod_{i=1}^{n} p\left(w_i \mid w_0, \cdots, w_{i-1}\right)$$

- **Controllable Generation:** $p(\mathbf{w} \mid \mathbf{c})$

BG: Controllable Text Generation

- **Generative Models:** $p_{\text{lm}}(\mathbf{w})$

$$\mathbf{w} = [w_1 \cdots w_n]$$

$$p_{\text{lm}}(\mathbf{W}) = \prod_{i=1}^{n} p\left(w_i \mid w_0, \cdots, w_{i-1}\right)$$

- **Controllable Generation:** $p(\mathbf{w} \mid \mathbf{c})$

Bayes rule $p(\mathbf{w} \mid \mathbf{c}) \propto p_{\text{lm}}(\mathbf{w}) \cdot p(\mathbf{c} \mid \mathbf{w})$

Fluency Controllable

BG: Controllable Text Generation

$p_{\text{lm}}(\mathbf{w})$

$p(\mathbf{w} \mid \mathbf{c})$

[–] The potato and cauliflower are both in season to make combo breads, mounds, or pads. For an added challenge, try some garlic mashed potatoes.

[Negative] The potato is a pretty bad idea. It can make you fat, it can cause you to have a terrible immune system, and it can even kill you.. . .

[Positive] The potato chip recipe you asked for! We love making these, and I've been doing so for years. I've always had a hard time keeping a recipe secret. I think it's the way our kids love to eat them – so many little ones.

[Science] The potato was once thought to have no health problems and has been promoted as a nutritious food source since the mid-1800s, but recent reports indicate that it has many harmful health issues. In fact, researchers from Johns Hopkins University. . .

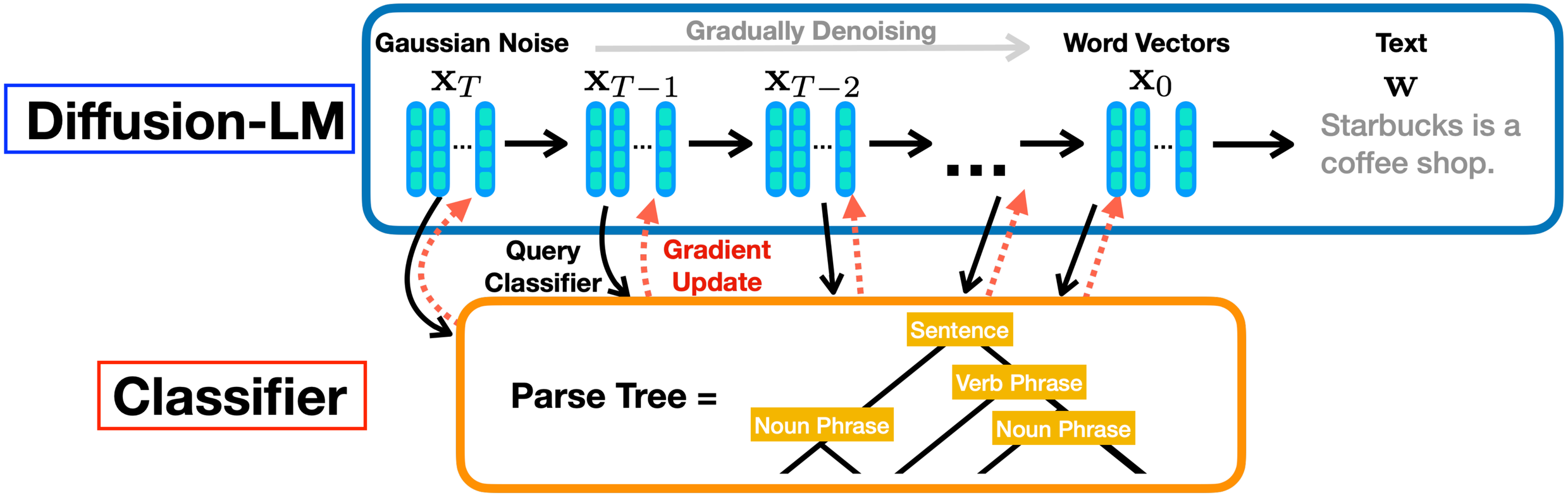
[Politics] **[Positive]** To conclude this series of articles, I will present three of the most popular and influential works on this topic. The first article deals with the role of women's political participation in building a political system that is representative of the will of the people.

[Politics] **[Negative]** To conclude, the most significant and lasting damage from the economic crisis in 2008 was that many governments, including those in the political center, lost power for the first time in modern history.

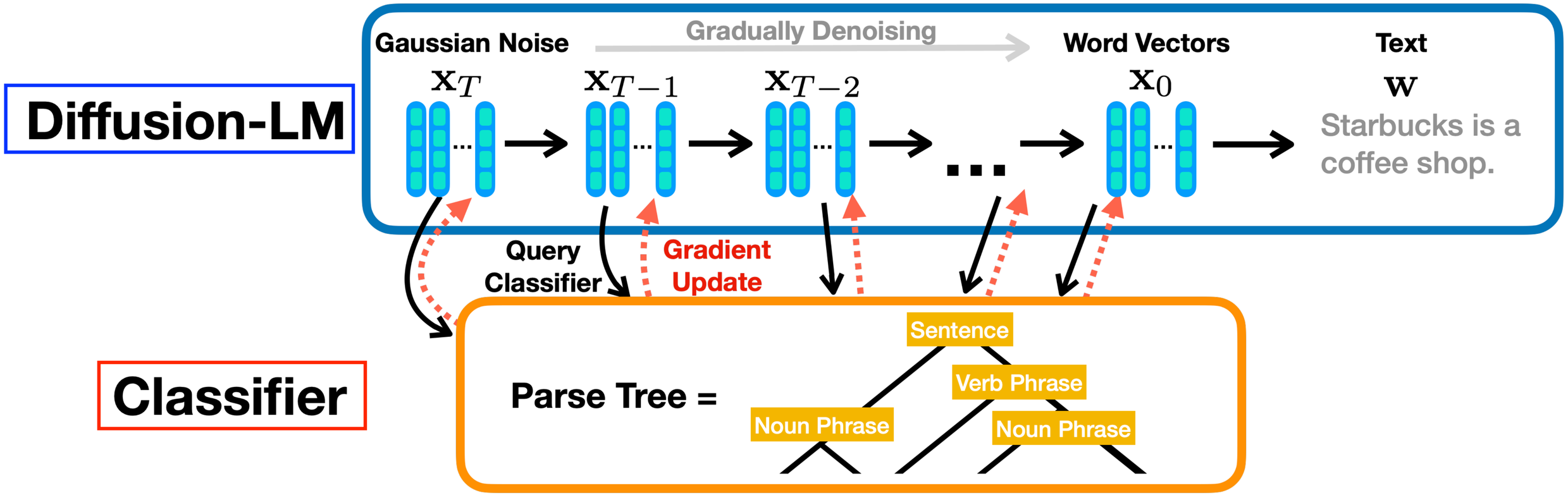
Main Q&C

- Propose Diffusion-LM
  - denoise Gaussian noise vectors into words → latent variable
  - add an embedding step and a rounding step → diffusion process
  - control it using a gradient-based method → objective
- 6 control targets (new SOTA)
- successfully compose multiple controls

Diffusion-LM



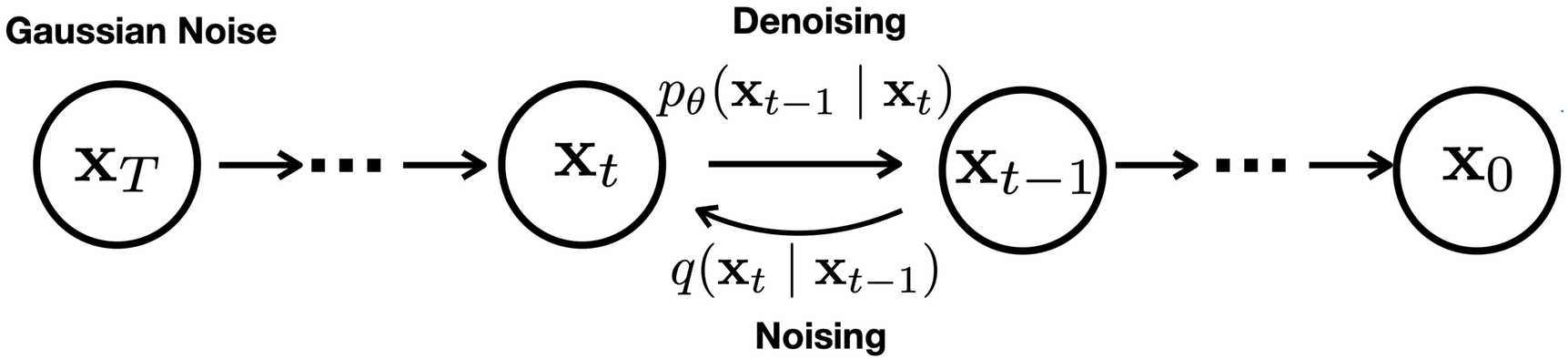
Diffusion-LM



Bayes rule $p(\mathbf{w} \mid \mathbf{c}) \propto p_{\text{lm}}(\mathbf{w}) \cdot p(\mathbf{c} \mid \mathbf{w})$

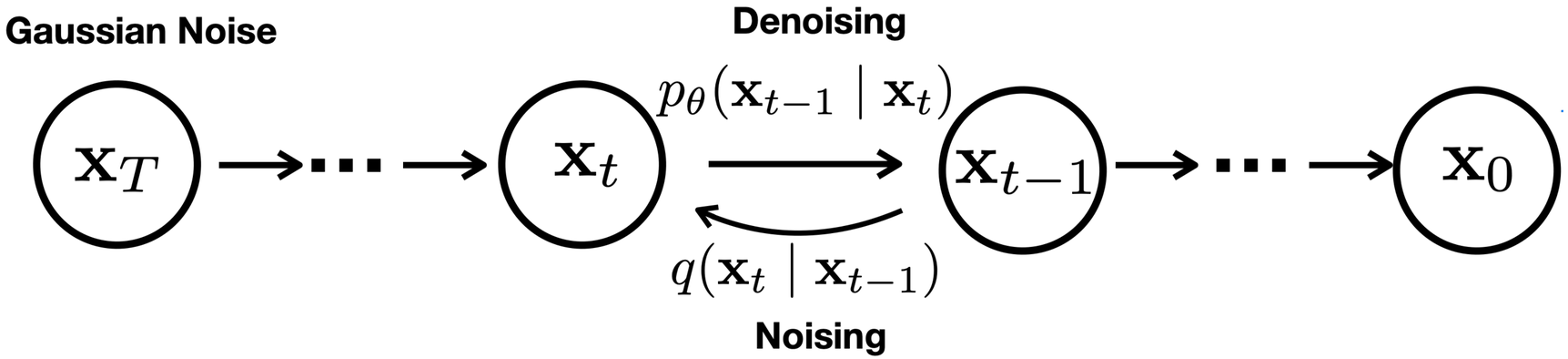
Fluency Controllable

Diffusion-LM



$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

Diffusion-LM

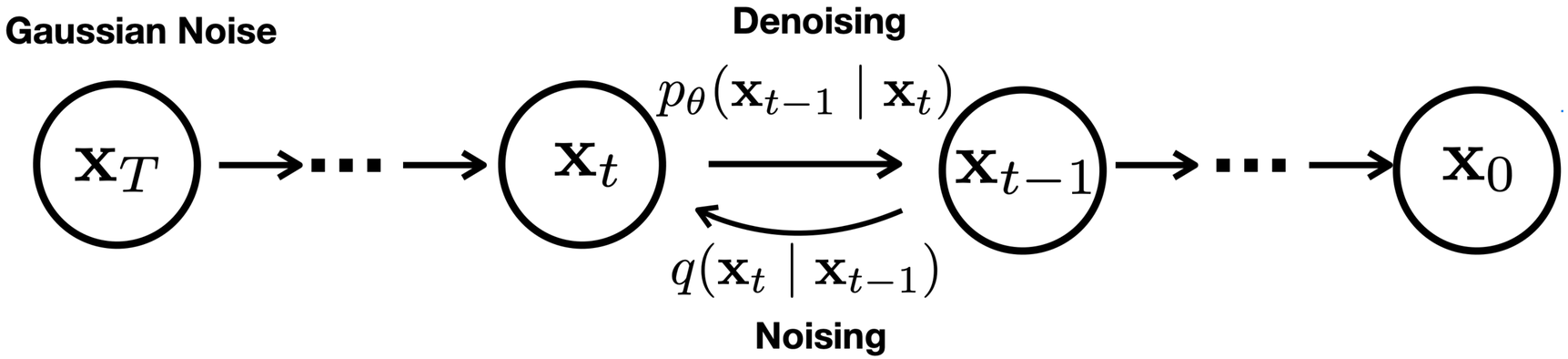


$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$ 2✅

$$\begin{aligned}
\boldsymbol{x}_t &= \alpha_t \boldsymbol{x}_{t-1} + \beta_t \boldsymbol{\varepsilon}_t \\
&= \alpha_t \left(\alpha_{t-1}\boldsymbol{x}_{t-2} + \beta_{t-1}\boldsymbol{\varepsilon}_{t-1}\right) + \beta_t \boldsymbol{\varepsilon}_t \\
&= \cdots \\
&= (\alpha_t \cdots \alpha_1)\boldsymbol{x}_0 + \underbrace{(\alpha_t \cdots \alpha_2)\beta_1 \boldsymbol{\varepsilon}_1 + (\alpha_t \cdots \alpha_3)\beta_2 \boldsymbol{\varepsilon}_2 + \cdots + \alpha_t \beta_{t-1}\boldsymbol{\varepsilon}_{t-1} + \beta_t \boldsymbol{\varepsilon}_t}_{\text{多个相互独立的正态噪声之和}}
\end{aligned}$$

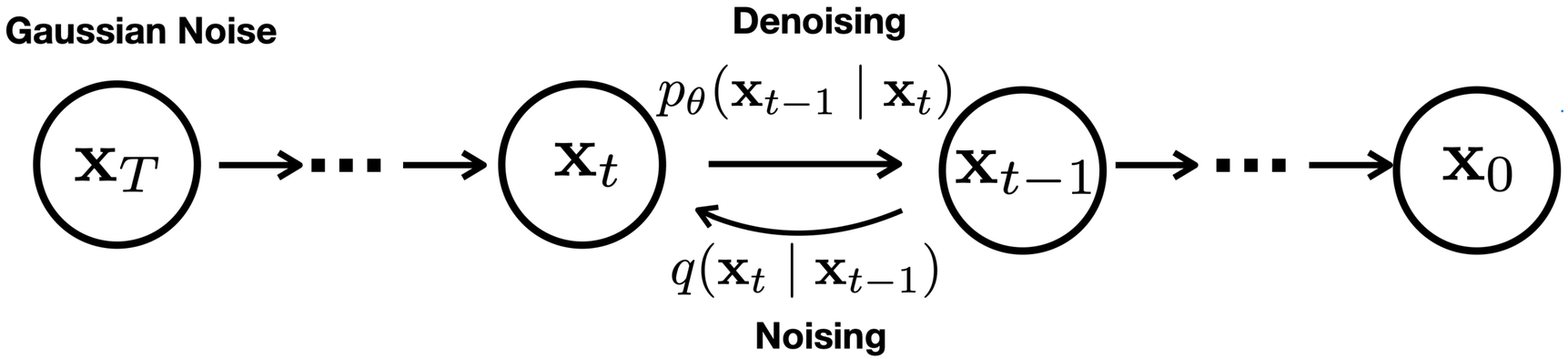
$$\boldsymbol{x}_t = \underbrace{(\alpha_t \cdots \alpha_1)}_{\text{记为}\bar{\alpha}_t}\boldsymbol{x}_0 + \underbrace{\sqrt{1-(\alpha_t \cdots \alpha_1)^2}}_{\text{记为}\bar{\beta}_t}\bar{\boldsymbol{\varepsilon}}_t, \quad \bar{\boldsymbol{\varepsilon}}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Diffusion-LM



Objective maximize the marginal likelihood of the data $\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}}[\log p_\theta(\mathbf{x}_0)]$

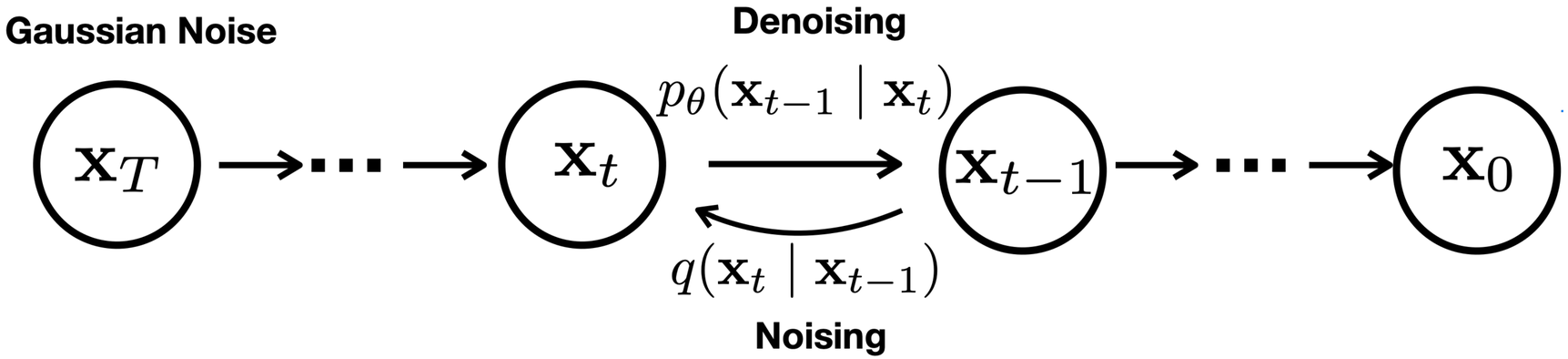
Diffusion-LM



Objective maximize the marginal likelihood of the data $\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}}[\log p_\theta(\mathbf{x}_0)]$

Expand $$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

Diffusion-LM



Objective maximize the marginal likelihood of the data $\mathbb{E}_{\mathbf{x}_0 \sim p_{\text{data}}}[\log p_\theta(\mathbf{x}_0)]$

Expand
$$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$$

Intro q()
$$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \frac{q(x_{1:T}|x_0)}{q(x_{1:T}|x_0)}$$

Diffusion-LM

Diffusion-LM

Exchange

$$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} q(x_{1:T}|x_0) \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Diffusion-LM

Exchange

$$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} q(x_{1:T}|x_0) \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Expectation

$$p_\theta(x_0) = \mathbb{E}_{q(x_{1:T}|x_0)} p_\theta(x_T) \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Diffusion-LM

Exchange

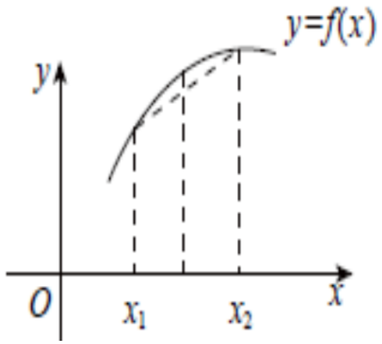
$$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} q(x_{1:T}|x_0) \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Expectation

$$p_\theta(x_0) = \mathbb{E}_{q(x_{1:T}|x_0)} p_\theta(x_T) \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Jensen **Objective**

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)} \log p_\theta(x_T) \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$



Diffusion-LM

Exchange

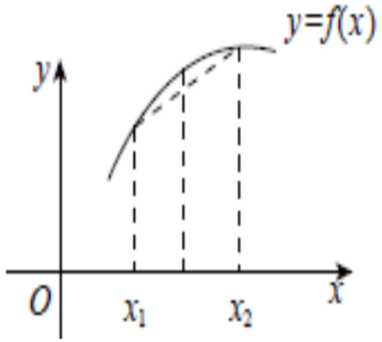
$$p_\theta(x_0) = p_\theta(x_T) \prod_{t=1}^{T} q(x_{1:T}|x_0) \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Expectation

$$p_\theta(x_0) = \mathbb{E}_{q(x_{1:T}|x_0)} p_\theta(x_T) \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$

Jensen **Objective**

$$\log p_\theta(x_0) \geq \mathbb{E}_{q(x_{1:T}|x_0)} \log p_\theta(x_T) \prod_{t=1}^{T} \frac{p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}$$



[ICML15] Deep Unsupervised Learning using Nonequilibrium Thermodynamics

B.1. Entropy of $p\left(\mathbf{X}^{(T)}\right)$

$$K = \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \sum_{t=1}^{T} \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] + \int d\mathbf{x}^{(T)} q\left(\mathbf{x}^{(T)}\right) \log p\left(\mathbf{x}^{(T)}\right) \tag{40}$$

$$= \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \sum_{t=1}^{T} \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] + \int d\mathbf{x}^{(T)} q\left(\mathbf{x}^{(T)}\right) \log \pi\left(\mathbf{x}^{T}\right) \tag{41}$$

$$. \tag{42}$$

By design, the cross entropy to $\pi\left(\mathbf{x}^{(t)}\right)$ is constant under our diffusion kernels, and equal to the entropy of $p\left(\mathbf{x}^{(T)}\right)$. Therefore,

$$K = \sum_{t=1}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] - H_p\left(\mathbf{X}^{(T)}\right). \tag{43}$$

B.1. Entropy of $p\left(\mathbf{X}^{(T)}\right)$

$$K = \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \sum_{t=1}^{T} \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] + \int d\mathbf{x}^{(T)} q\left(\mathbf{x}^{(T)}\right) \log p\left(\mathbf{x}^{(T)}\right) \tag{40}$$

$$= \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \sum_{t=1}^{T} \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] + \int d\mathbf{x}^{(T)} q\left(\mathbf{x}^{(T)}\right) \log \pi\left(\mathbf{x}^{T}\right) \tag{41}$$

$$. \tag{42}$$

By design, the cross entropy to $\pi\left(\mathbf{x}^{(t)}\right)$ is constant under our diffusion kernels, and equal to the entropy of $p\left(\mathbf{x}^{(T)}\right)$. Therefore,

$$K = \sum_{t=1}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] - H_p\left(\mathbf{X}^{(T)}\right). \tag{43}$$

Objective $\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=1}^{T} \frac{p_\theta\left(x_{t-1} \mid x_t\right)}{q\left(x_{1:T} \mid x_0\right)}$

B.2. Remove the edge effect at $t = 0$

In order to avoid edge effects, we set the final step of the reverse trajectory to be identical to the corresponding forward diffusion step,

$$p\left(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}\right) = q\left(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}\right)\frac{\pi\left(\mathbf{x}^{(0)}\right)}{\pi\left(\mathbf{x}^{(1)}\right)} = T_\pi\left(\mathbf{x}^{(0)}|\mathbf{x}^{(1)};\beta_1\right). \tag{44}$$

We then use this equivalence to remove the contribution of the first time-step in the sum,

$$K = \sum_{t=2}^{T}\int d\mathbf{x}^{(0\cdots T)}q\left(\mathbf{x}^{(0\cdots T)}\right)\log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] + \int d\mathbf{x}^{(0)}d\mathbf{x}^{(1)}q\left(\mathbf{x}^{(0)},\mathbf{x}^{(1)}\right)\log\left[\frac{q\left(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}\right)\pi\left(\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}\right)\pi\left(\mathbf{x}^{(1)}\right)}\right] - H_p\left(\mathbf{X}^{(T)}\right) \tag{45}$$

$$= \sum_{t=2}^{T}\int d\mathbf{x}^{(0\cdots T)}q\left(\mathbf{x}^{(0\cdots T)}\right)\log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] - H_p\left(\mathbf{X}^{(T)}\right), \tag{46}$$

B.2. Remove the edge effect at $t = 0$

In order to avoid edge effects, we set the final step of the reverse trajectory to be identical to the corresponding forward diffusion step,

$$p\left(\mathbf{x}^{(0)}|\mathbf{x}^{(1)}\right) = q\left(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}\right)\frac{\pi\left(\mathbf{x}^{(0)}\right)}{\pi\left(\mathbf{x}^{(1)}\right)} = T_\pi\left(\mathbf{x}^{(0)}|\mathbf{x}^{(1)};\beta_1\right). \tag{44}$$

We then use this equivalence to remove the contribution of the first time-step in the sum,

$$K = \sum_{t=2}^{T}\int d\mathbf{x}^{(0\cdots T)}q\left(\mathbf{x}^{(0\cdots T)}\right)\log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] + \int d\mathbf{x}^{(0)}d\mathbf{x}^{(1)}q\left(\mathbf{x}^{(0)},\mathbf{x}^{(1)}\right)\log\left[\frac{q\left(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}\right)\pi\left(\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(1)}|\mathbf{x}^{(0)}\right)\pi\left(\mathbf{x}^{(1)}\right)}\right] - H_p\left(\mathbf{X}^{(T)}\right) \tag{45}$$

$$= \sum_{t=2}^{T}\int d\mathbf{x}^{(0\cdots T)}q\left(\mathbf{x}^{(0\cdots T)}\right)\log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(t-1)}\right)}\right] - H_p\left(\mathbf{X}^{(T)}\right), \tag{46}$$

Objective $\mathbb{E}_{q(x_{1:T}|x_0)}\prod_{t=2}^{T}\frac{p_\theta\left(x_{t-1}\mid x_t\right)}{q\left(x_{1:T}\mid x_0\right)}$

B.3. Rewrite in terms of posterior $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}\right)$

Using Bayes' rule we can rewrite this in terms of a posterior and marginals from the forward trajectory,

$$K = \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log \left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right)} \frac{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}\right)} \right] - H_p\left(\mathbf{X}^{(T)}\right).$$

$$\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=2}^{T} \frac{p_\theta\left(x_{t-1} \mid x_t\right)}{q\left(x_t \mid x_{t-1}\right)}$$

B.3. Rewrite in terms of posterior $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}\right)$

Using Bayes' rule we can rewrite this in terms of a posterior and marginals from the forward trajectory,

$$K = \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log \left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right)} \frac{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}\right)} \right] - H_p\left(\mathbf{X}^{(T)}\right).$$

$$\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=2}^{T} \frac{p_\theta\left(x_{t-1} \mid x_t\right)}{q\left(x_t \mid x_{t-1}\right)}$$

Bayes rule Objective

$$\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=2}^{T} \frac{p_\theta\left(x_{t-1} \mid x_t\right) q(x_{t-1}|x_0)}{q\left(x_{t-1} \mid x_t\right) q\left(x_t \mid x_0\right)}$$

B.3. Rewrite in terms of posterior $q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}\right)$

Using Bayes' rule we can rewrite this in terms of a posterior and marginals from the forward trajectory,

$$K = \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log \left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}, \mathbf{x}^{(0)}\right)} \frac{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(0)}\right)}{q\left(\mathbf{x}^{(t)}|\mathbf{x}^{(0)}\right)} \right] - H_p\left(\mathbf{X}^{(T)}\right).$$

$$\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=2}^{T} \frac{p_\theta\left(x_{t-1} \mid x_t\right)}{\boxed{q\left(x_t \mid x_{t-1}\right)}}$$

Bayes rule **Objective**

$$\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=2}^{T} \frac{p_\theta\left(x_{t-1} \mid x_t\right) q(x_{t-1}|x_0)}{\boxed{q\left(x_{t-1} \mid x_t\right)} q\left(x_t \mid x_0\right)}$$

B.4. Rewrite in terms of KL divergences and entropies

$$K = \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)}\right] + \sum_{t=2}^{T}\left[H_q\left(\mathbf{X}^{(t)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(t-1)}|\mathbf{X}^{(0)}\right)\right] - H_p\left(\mathbf{X}^{(T)}\right) \tag{49}$$

$$= \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)}\right] + H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right). \tag{50}$$

Finally we transform the log ratio of probability distributions into a KL divergence,

$$\begin{aligned} K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)},\mathbf{x}^{(t)}\right) D_{KL}\left(q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)\middle\|p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)\right) \\ + H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right). \end{aligned} \tag{51}$$

B.4. Rewrite in terms of KL divergences and entropies

$$K = \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)}\right] + \sum_{t=2}^{T}\left[H_q\left(\mathbf{X}^{(t)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(t-1)}|\mathbf{X}^{(0)}\right)\right] - H_p\left(\mathbf{X}^{(T)}\right) \tag{49}$$

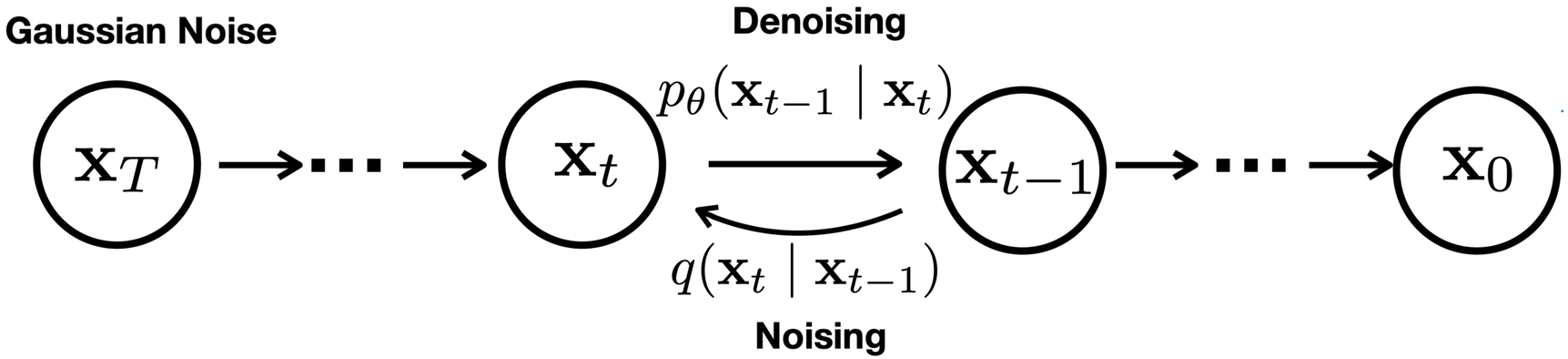
$$= \sum_{t=2}^{T} \int d\mathbf{x}^{(0\cdots T)} q\left(\mathbf{x}^{(0\cdots T)}\right) \log\left[\frac{p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)}{q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)}\right] + H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right). \tag{50}$$

Finally we transform the log ratio of probability distributions into a KL divergence,

$$\begin{aligned} K = -\sum_{t=2}^{T} \int d\mathbf{x}^{(0)} d\mathbf{x}^{(t)} q\left(\mathbf{x}^{(0)},\mathbf{x}^{(t)}\right) D_{KL}\left(q\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)},\mathbf{x}^{(0)}\right)\middle|\middle|p\left(\mathbf{x}^{(t-1)}|\mathbf{x}^{(t)}\right)\right) \\ + H_q\left(\mathbf{X}^{(T)}|\mathbf{X}^{(0)}\right) - H_q\left(\mathbf{X}^{(1)}|\mathbf{X}^{(0)}\right) - H_p\left(\mathbf{X}^{(T)}\right). \end{aligned} \tag{51}$$

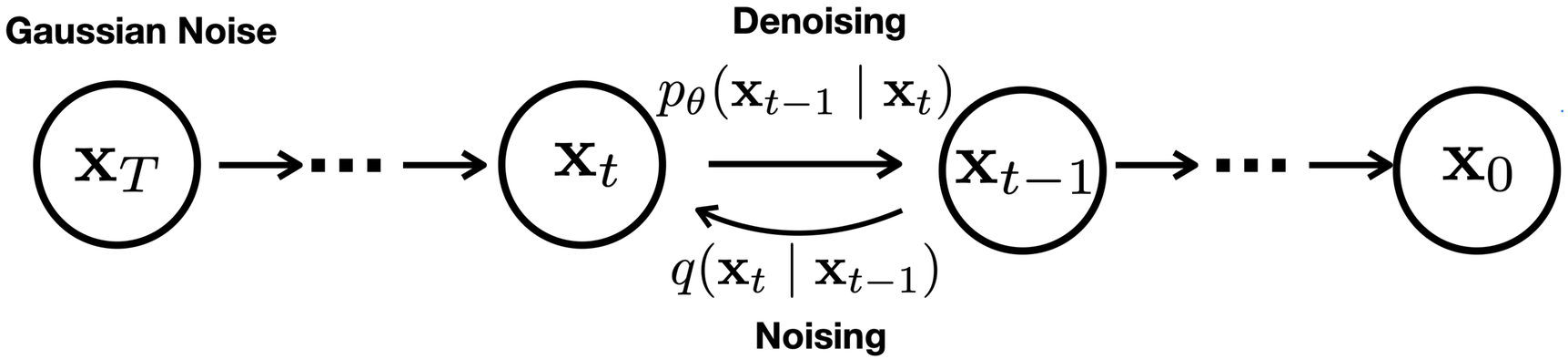
Objective $-\mathbb{E}_{q(x_{1:T}|x_0)} \prod_{t=2}^{T} KL\left(q\left(x_{t-1} \mid x_t\right)||p_\theta\left(x_{t-1} \mid x_t\right)\right)$

Diffusion-LM



$$\mathcal{L}_{\text{vlb}}(\mathbf{x}_0) = \mathop{\mathbb{E}}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)} + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right]. \qquad (1)$$

Diffusion-LM

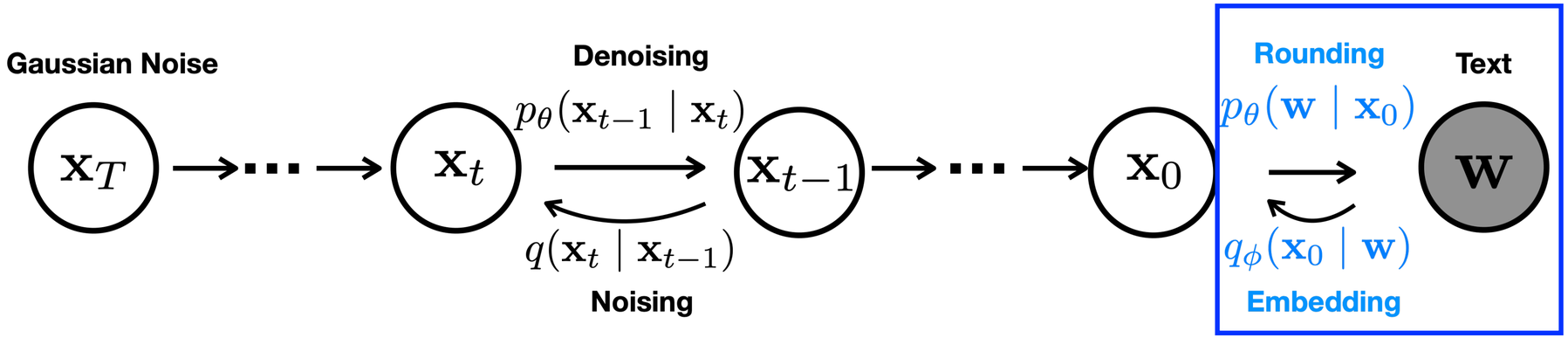


$$\mathcal{L}_{\text{vlb}}(\mathbf{x}_0) = \mathop{\mathbb{E}}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{q(\mathbf{x}_T|\mathbf{x}_0)}{p_\theta(\mathbf{x}_T)} + \sum_{t=2}^{T} \log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1) \right]. \qquad (1)$$

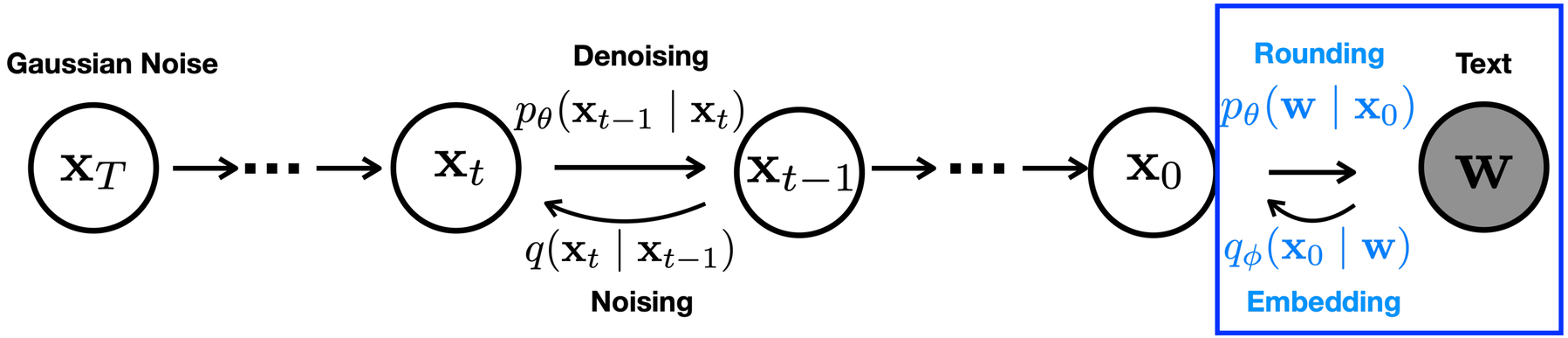
$$\mathcal{L}_{\text{simple}}(\mathbf{x}_0) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\mathbf{x}_t|\mathbf{x}_0)} ||\mu_\theta(\mathbf{x}_t, t) - \hat{\mu}(\mathbf{x}_t, \mathbf{x}_0)||^2,$$

$$\mathop{\mathbb{E}}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\log \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{x}_t)}{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)} \right] = \mathop{\mathbb{E}}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \left[\frac{1}{2\sigma_t^2} ||\mu_\theta(\mathbf{x}_t, t) - \hat{\mu}(\mathbf{x}_t, \mathbf{x}_0)||^2 \right] + C, \qquad (3)$$

Continuous Diffusion-LM



Continuous Diffusion-LM

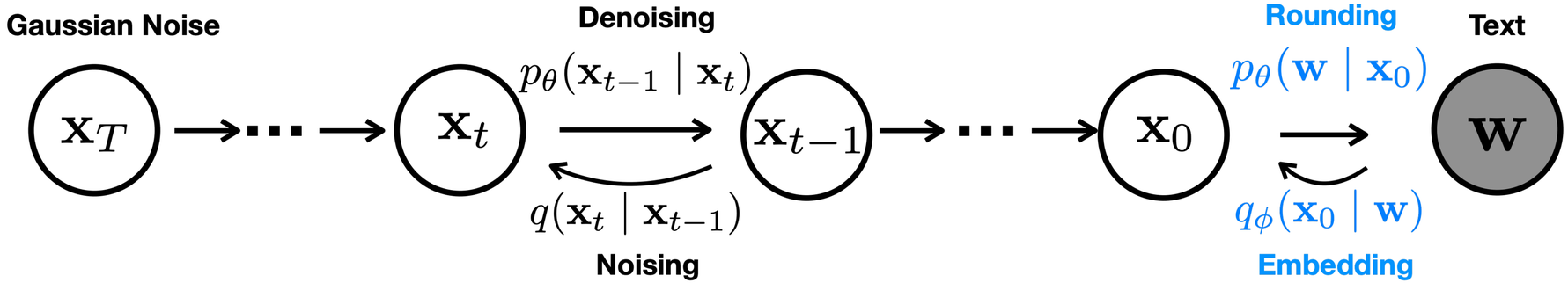


$$\textsc{Emb}(\mathbf{w}) = [\textsc{Emb}(w_1), \ldots, \textsc{Emb}(w_n)] \in \mathbb{R}^{nd}.$$

$$q_\phi(\mathbf{x}_0|\mathbf{w}) = \mathcal{N}(\textsc{Emb}(\mathbf{w}), \sigma_0 I)$$

$$p_\theta(\mathbf{w} \mid \mathbf{x}_0) = \prod_{i=1}^{n} p_\theta(w_i \mid x_i)$$

Continuous Diffusion-LM



$$\mathcal{L}_{\text{vlb}}^{\text{e2e}}(\mathbf{w}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{x}_0|\mathbf{w})} [\mathcal{L}_{\text{vlb}}(\mathbf{x}_0) + \log q_\phi(\mathbf{x}_0|\mathbf{w}) - \log p_\theta(\mathbf{w}|\mathbf{x}_0)]] \,,$$
$$\mathcal{L}_{\text{simple}}^{\text{e2e}}(\mathbf{w}) = \mathop{\mathbb{E}}_{q_\phi(\mathbf{x}_{0:T}|\mathbf{w})} \left[\mathcal{L}_{\text{simple}}(\mathbf{x}_0) + ||\text{EMB}(\mathbf{w}) - \mu_\theta(\mathbf{x}_1, 1)||^2 - \log p_\theta(\mathbf{w}|\mathbf{x}_0)\right] . \quad (2)$$

Continuous Diffusion-LM

- **Word Vector**
  - **Random init ✅**
  - **Pretrained ❎**
- **backpropagate**
  - **reparametrization trick**

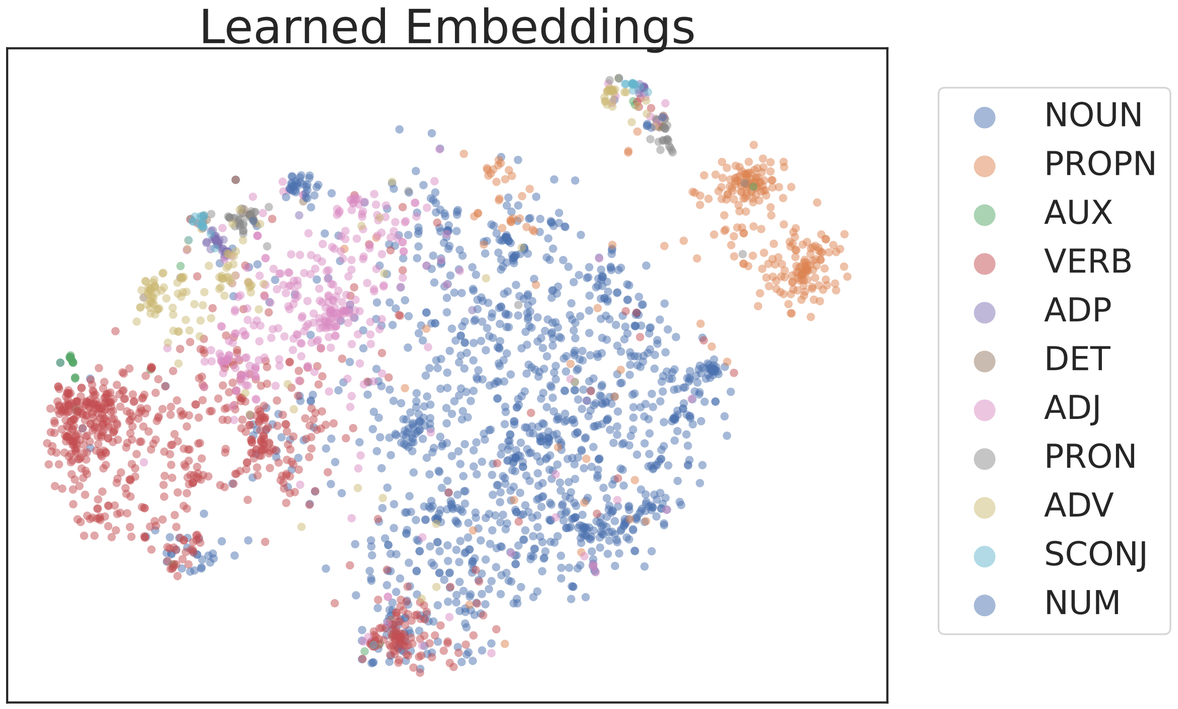


Figure 3: A t-SNE [41] plot of the learned word embeddings. Each word is colored by its POS.

Continuous Diffusion-LM

- **Word Vector**
  - **Random init ✅**
  - **Pretrained ❎**
- **backpropagate**
  - **reparametrization trick**

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}} \cdot \text{Clamp}(f_\theta(\mathbf{x}_t, t)) + \sqrt{1-\bar{\alpha}}\epsilon$$
$$\epsilon \sim \mathcal{N}(0, I)$$

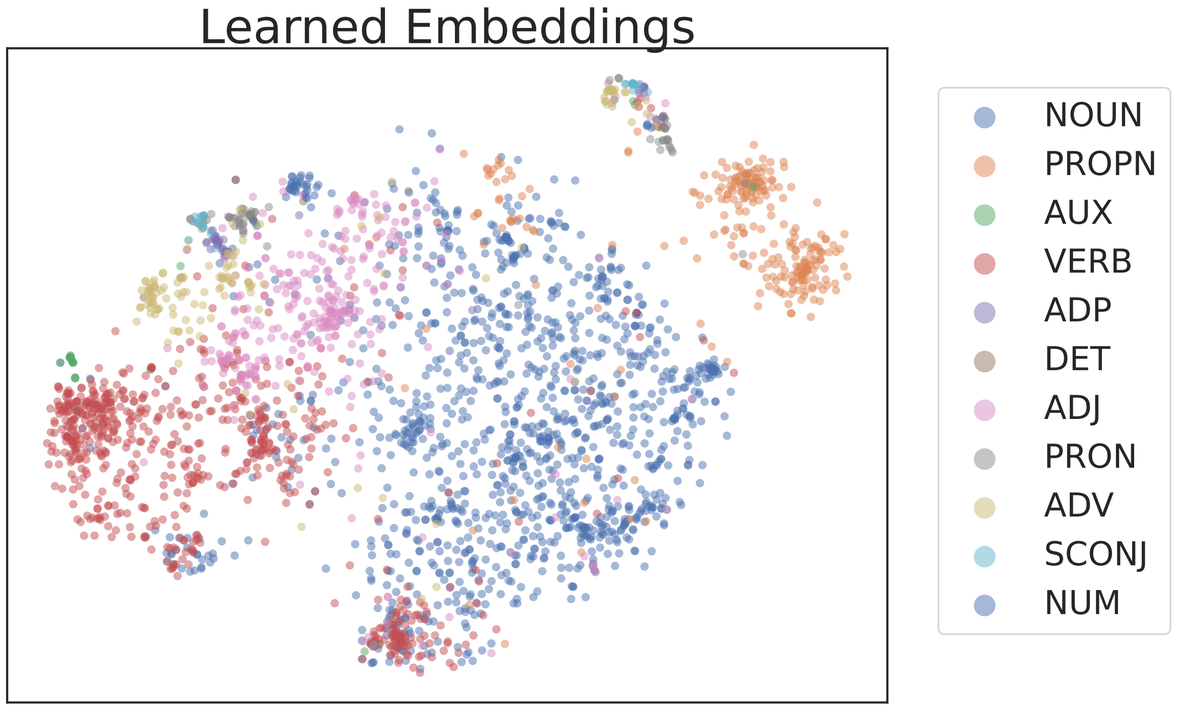
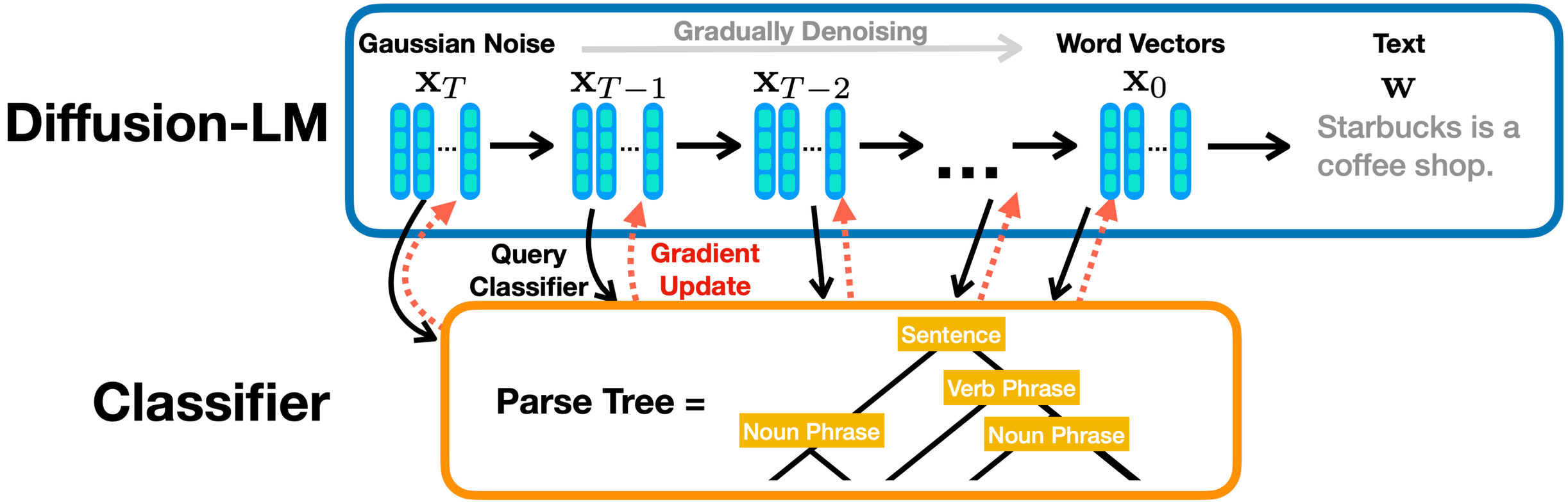


Figure 3: A t-SNE [41] plot of the learned word embeddings. Each word is colored by its POS.

Diffusion-LM



Controllable Text Generation

$$p(\mathbf{x}_{0:T}|\mathbf{c}) = \prod_{t=1}^{T} p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$$

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) \propto p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \cdot p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{x}_t)$$

$$\nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) = \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) + \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{c} \mid \mathbf{x}_{t-1}),$$

Controllable Text Generation

$$p(\mathbf{x}_{0:T}|\mathbf{c}) = \prod_{t=1}^{T} p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c})$$

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) \propto p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \cdot p(\mathbf{c} \mid \mathbf{x}_{t-1}, \mathbf{x}_t)$$

$$\nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{c}) = \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{x}_{t-1} \mid \mathbf{x}_t) + \nabla_{\mathbf{x}_{t-1}} \log p(\mathbf{c} \mid \mathbf{x}_{t-1}),$$

- **downsample the diffusion steps from 2000 to 200**
- **E2E dataset** consists of 50K restaurant reviews
- **ROCStories dataset** consists of 98K five-sentence stories
- **80M Transformer for LM**
- **1.5x slower than FUDGE but 60x faster than PPLM**

Classifier-Guided Controls

- **Semantic Content:** autoregressive LM (GPT-2 small architecture) to predict the (field, value) pair conditioned on text.
- **Parts-of-speech:** BERT-base tagger
- **Syntax Tree:** Transformer-based constituency parser
- **Syntax Spans:** same with **Syntax Span.**

Task

input (Semantic Content) output text	food : Japanese Browns Cambridge is good for Japanese food and also children friendly near The Sorrento .
input (Parts-of-speech) output text	PROPN AUX DET ADJ NOUN NOUN VERB ADP DET NOUN ADP DET NOUN PUNCT Zizzi is a local coffee shop located on the outskirts of the city .
input (Syntax Tree) output text	(TOP (S (NP (*) (*) (*)) (VP (*) (NP (NP (*) (*))))) The Twenty Two has great food
input (Syntax Spans) output text	(7, 10, VP) Wildwood pub serves multicultural dishes and is ranked 3 stars
input (Length) output text	14 Browns Cambridge offers Japanese food located near The Sorrento in the city centre .
input (left context) input (right context) output text	My dog loved tennis balls. My dog had stolen every one and put it under there. One day, I found all of my lost tennis balls underneath the bed.

Table 1: Example input control and output text for each control tasks.

Task

- **lm score: PPL @ Finetune GPT2**
- **Classifier-Guided Baseline:**
  - **PPLM:** only semantic content
  - **FUDEG:** future discriminator + reweights
  - **FT:** finetune
- **LM NLL (vs. AR LM):**
  - **2.28 vs. 1.77 @ E2E**
  - **3.88 vs. 3.05 @ ROCStories**

Classifier-Guided

	Semantic Content		Parts-of-speech		Syntax Tree		Syntax Spans		Length	
	ctrl ↑	lm ↓	ctrl ↑	lm ↓	ctrl ↑	lm ↓	ctrl ↑	lm ↓	ctrl ↑	lm ↓
PPLM	9.9	5.32	-	-	-	-	-	-	-	-
FUDGE	69.9	2.83	27.0	7.96	17.9	**3.39**	54.2	4.03	46.9	3.11
Diffusion-LM	**81.2**	**2.55**	**90.0**	**5.16**	**86.0**	3.71	**93.8**	**2.53**	**99.9**	**2.16**
FT-sample	72.5	2.87	89.5	4.72	64.8	5.72	26.3	2.88	98.1	3.84
FT-search	89.9	1.78	93.0	3.31	76.4	3.24	54.4	2.19	100.0	1.83

Classifier-Guided

	Semantic Content		Parts-of-speech		Syntax Tree		Syntax Spans		Length	
	ctrl ↑	lm ↓	ctrl ↑	lm ↓	ctrl ↑	lm ↓	ctrl ↑	lm ↓	ctrl ↑	lm ↓
PPLM	9.9	5.32	-	-	-	-	-	-	-	-
FUDGE	69.9	2.83	27.0	7.96	17.9	**3.39**	54.2	4.03	46.9	3.11
Diffusion-LM	**81.2**	**2.55**	**90.0**	**5.16**	**86.0**	3.71	**93.8**	**2.53**	**99.9**	**2.16**
FT-sample	72.5	2.87	89.5	4.72	64.8	5.72	26.3	2.88	98.1	3.84
FT-search	89.9	1.78	93.0	3.31	76.4	3.24	54.4	2.19	100.0	1.83

Diffusion-LM 🏆 (@Syntax)

Classifier-Guided

Syntactic Parse	(S (S (NP *) (VP * (NP (NP * *) (VP * **(NP (ADJP * *) *)**)))) * (S (NP * * *) (VP * (ADJP (ADJP *)))))
FUDGE	Zizzi is a cheap restaurant . **[incomplete]**
Diffusion-LM	Zizzi is a pub providing **family friendly Indian food** Its customer rating is low
FT	Cocum is a Pub serving **moderately priced meals** and the customer rating is high
Syntactic Parse	(S (S (VP * (PP * (NP * *)))) * **(NP * * *)** (VP * (NP (NP * *) (SBAR (WHNP *) (S (VP * (NP * *)))))) *)
FUDGE	In the city near The Portland Arms is a coffee and fast food place named The Cricketers which is not family - friendly with a customer rating of 5 out of 5 .
Diffusion-LM	Located on the riverside , **The Rice Boat** is a restaurant that serves Indian food .
FT	Located near The Sorrento, **The Mill** is a pub that serves Indian cuisine.

Classifier-Guided

Syntactic Parse	(S (S (NP *) (VP * (NP (NP * *) (VP * **(NP (ADJP * *) *)**)))) * (S (NP * * *) (VP * (ADJP (ADJP *)))))
FUDGE	Zizzi is a cheap restaurant . **[incomplete]**
Diffusion-LM	Zizzi is a pub providing **family friendly Indian food** Its customer rating is low
FT	Cocum is a Pub serving **moderately priced meals** and the customer rating is high
Syntactic Parse	(S (S (VP * (PP * (NP * *)))) * **(NP * * *)** (VP * (NP (NP * *) (SBAR (WHNP *) (S (VP * (NP * *)))))) *)
FUDGE	In the city near The Portland Arms is a coffee and fast food place named The Cricketers which is not family - friendly with a customer rating of 5 out of 5 .
Diffusion-LM	Located on the riverside , **The Rice Boat** is a restaurant that serves Indian food .
FT	Located near The Sorrento, **The Mill** is a pub that serves Indian cuisine.

Diffusion-LM 🏆 (@Syntax)

Classifier-Guided

	Semantic Content + Syntax Tree			Semantic Content + Parts-of-speech		
	semantic ctrl ↑	syntax ctrl ↑	lm ↓	semantic ctrl ↑	POS ctrl ↑	lm ↓
FUDGE	61.7	15.4	3.52	64.5	24.1	3.52
Diffusion-LM	**69.8**	**74.8**	5.92	**63.7**	**69.1**	3.46
FT-PoE	61.7	29.2	**2.77**	29.4	10.5	**2.97**

Classifier-Guided

	Semantic Content + Syntax Tree			Semantic Content + Parts-of-speech		
	semantic ctrl ↑	syntax ctrl ↑	lm ↓	semantic ctrl ↑	POS ctrl ↑	lm ↓
FUDGE	61.7	15.4	3.52	64.5	24.1	3.52
Diffusion-LM	**69.8**	**74.8**	5.92	**63.7**	**69.1**	3.46
FT-PoE	61.7	29.2	**2.77**	29.4	10.5	**2.97**

Diffusion-LM 🏆 (@Syntax)

Classifier-Free

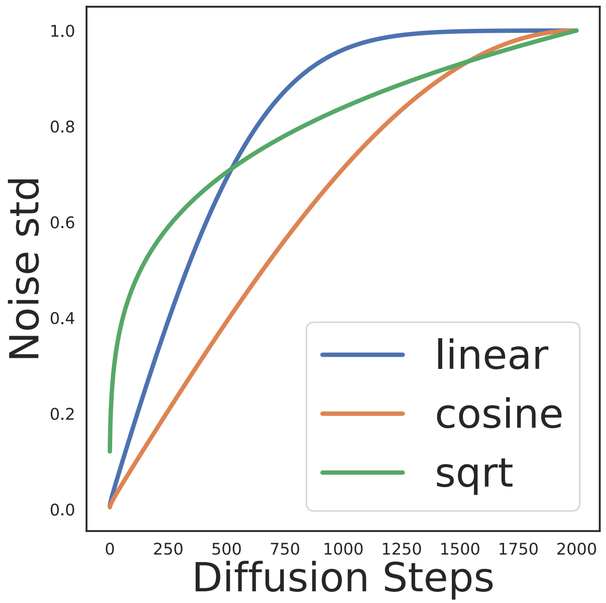
Sentence Infilling Task	Automatic Eval				Human Eval
	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑	BERTScore ↑	
Left-only	0.9	16.3	3.5	38.5	n/a
DELOREAN	1.6	19.1	7.9	41.7	n/a
COLD	1.8	19.5	10.7	42.7	n/a
Diffusion	**7.1**	**28.3**	**30.7**	**89.0**	$\mathbf{0.37}^{+0.03}_{-0.02}$
AR	6.7	27.0	26.9	**89.0**	$\mathbf{0.39}^{+0.02}_{-0.03}$

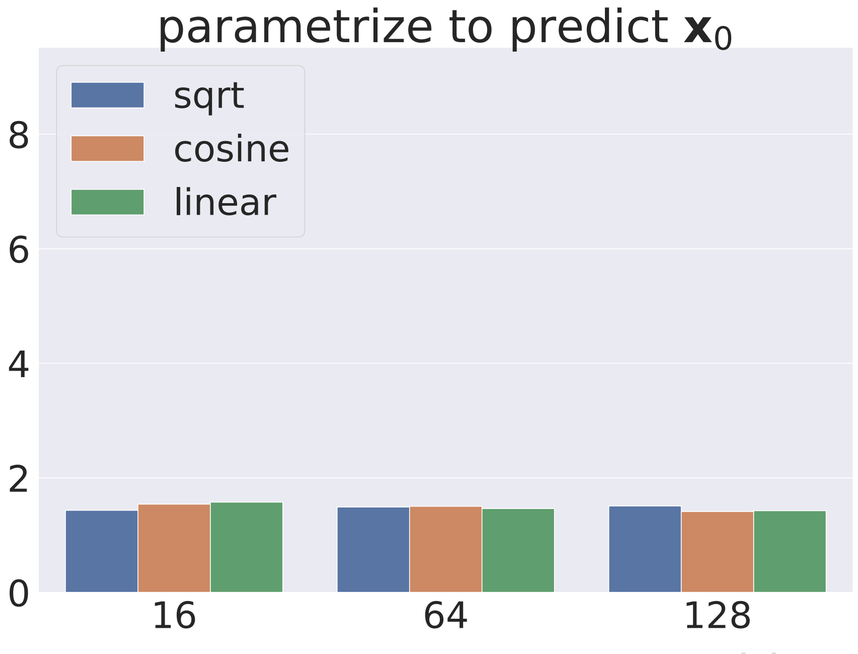
Classifier-Free

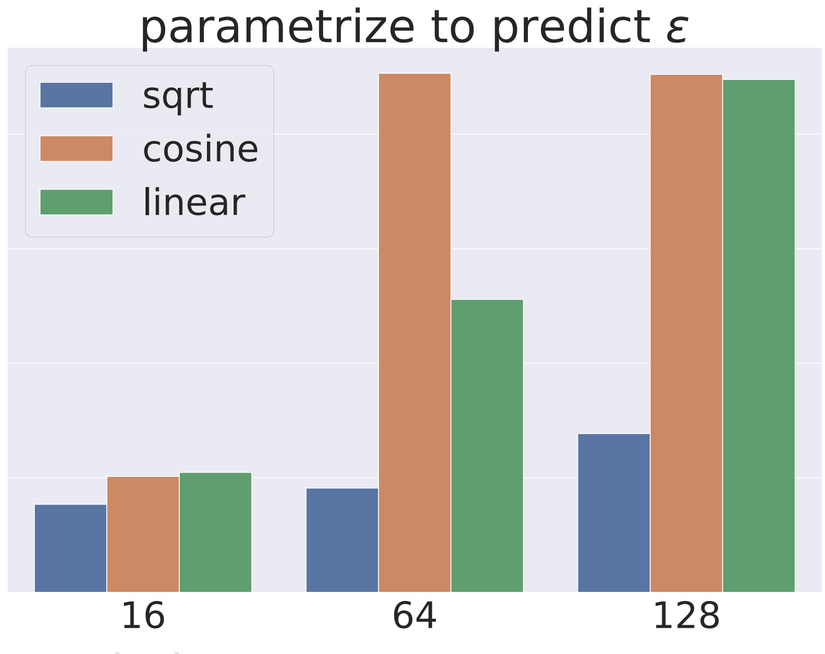
Sentence Infilling Task	Automatic Eval				Human Eval
	BLEU-4 ↑	ROUGE-L ↑	CIDEr ↑	BERTScore ↑	
Left-only	0.9	16.3	3.5	38.5	n/a
DELOREAN	1.6	19.1	7.9	41.7	n/a
COLD	1.8	19.5	10.7	42.7	n/a
Diffusion	**7.1**	**28.3**	**30.7**	**89.0**	$\mathbf{0.37}^{+0.03}_{-0.02}$
AR	6.7	27.0	26.9	**89.0**	$\mathbf{0.39}^{+0.02}_{-0.03}$

Diffusion-LM 🏆

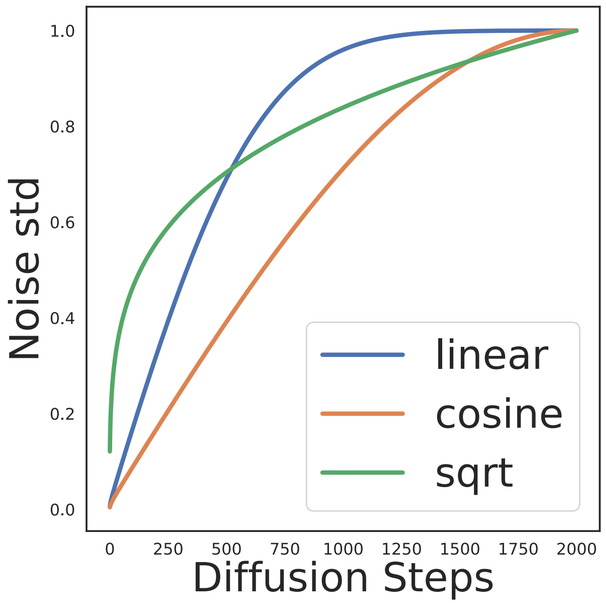
Noise Schedule

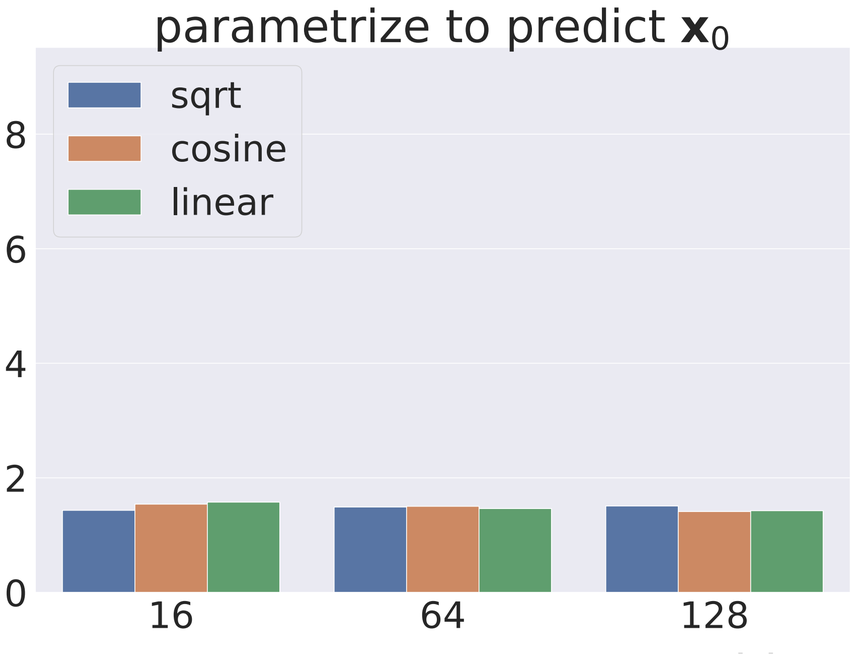


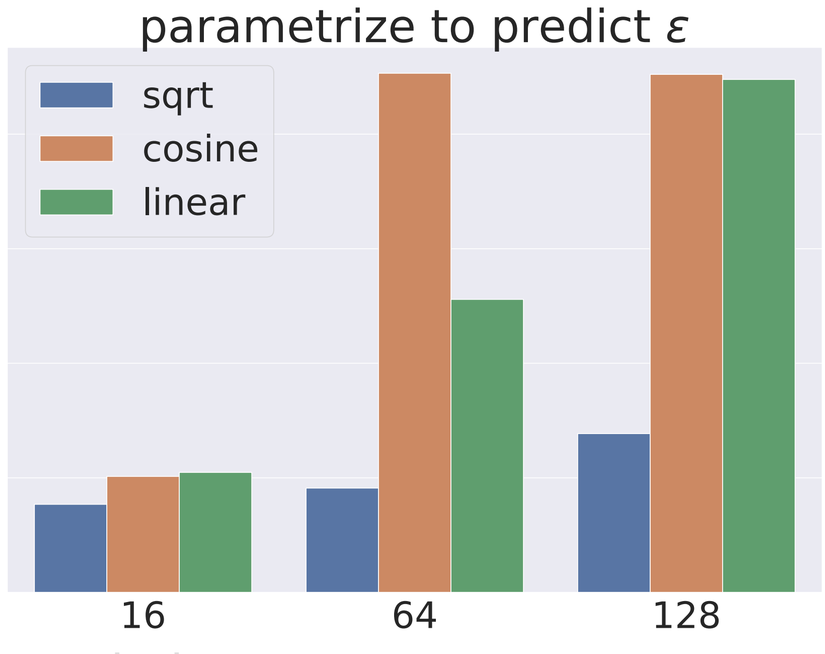




Noise Schedule

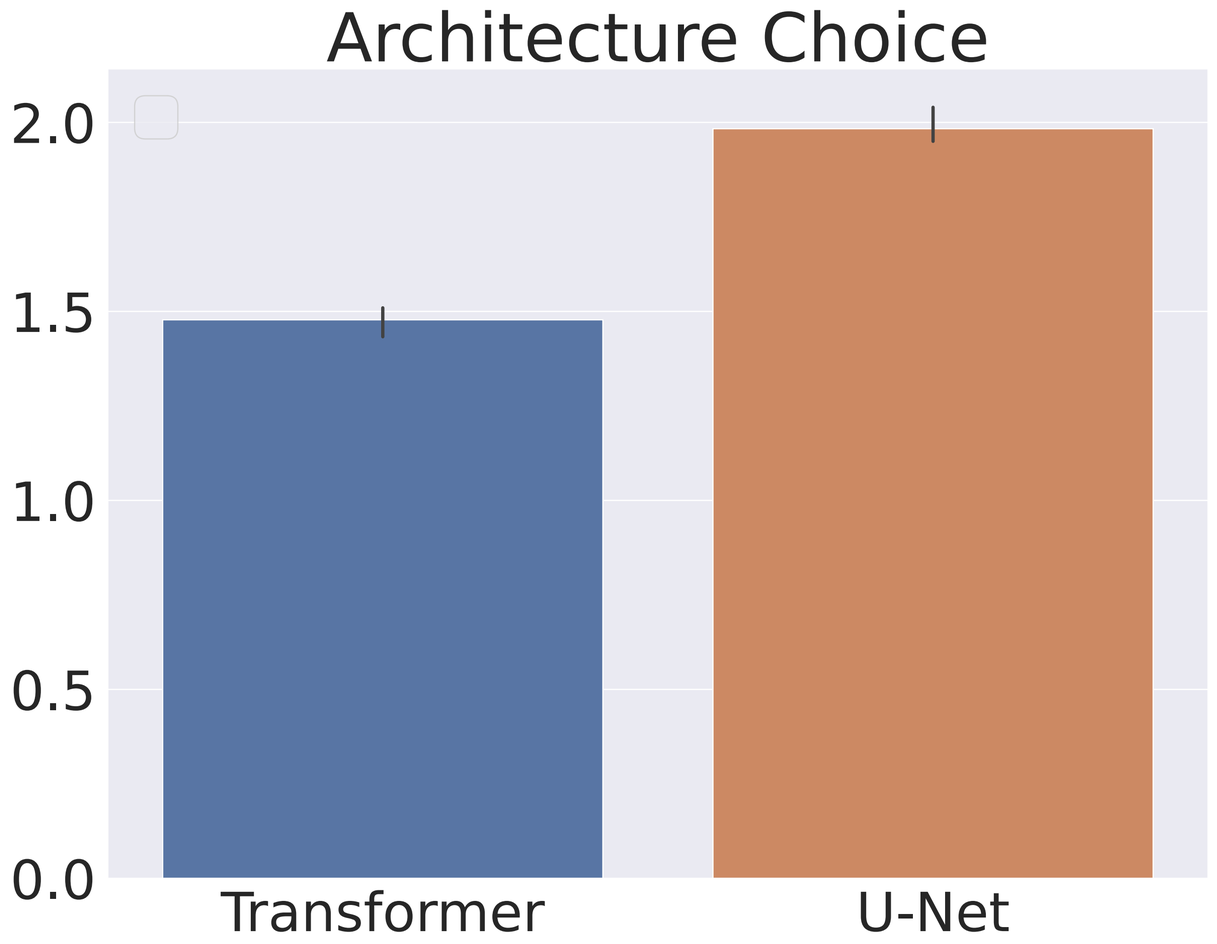




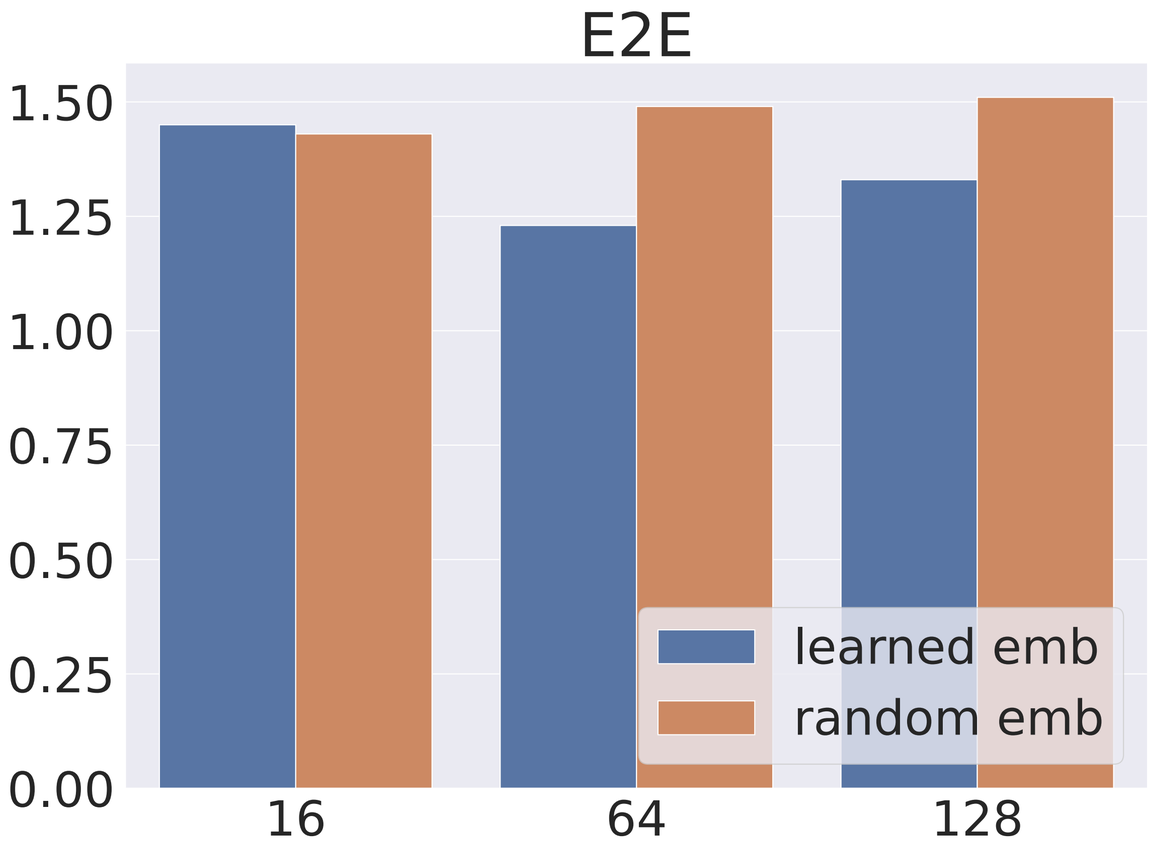


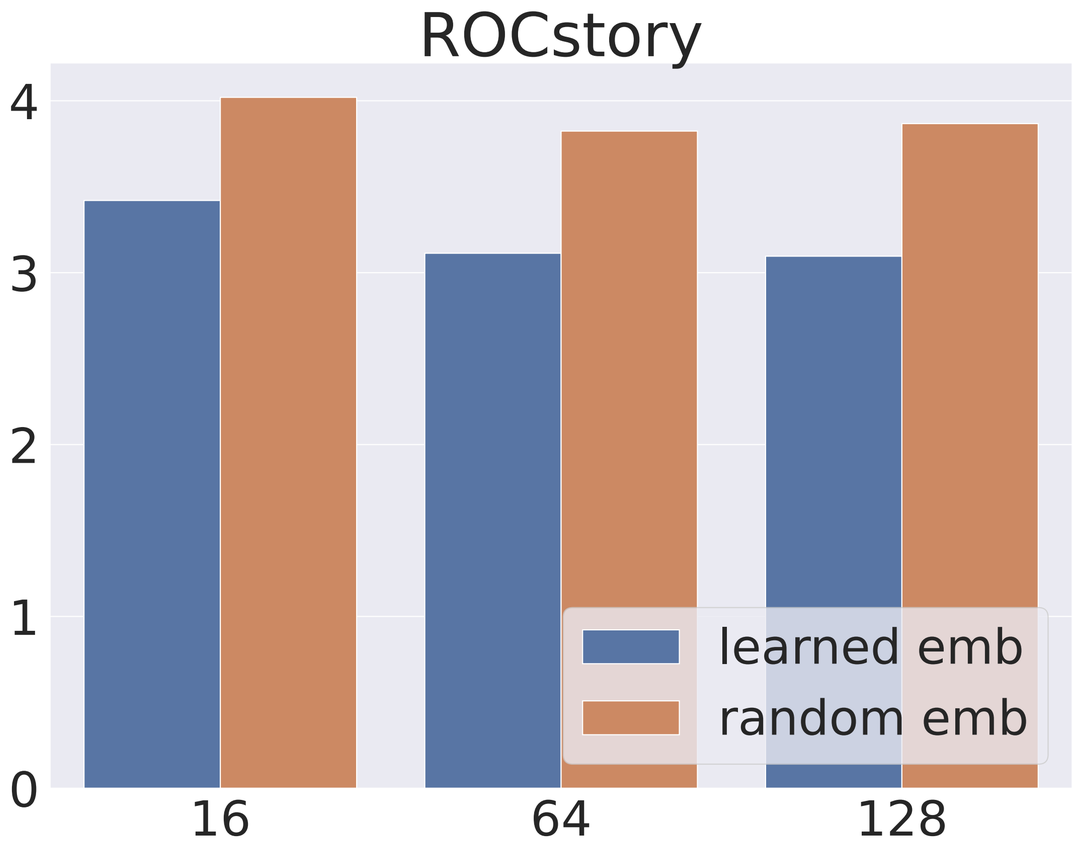
Sqrt 🏆

Net



EMB


E2E
1.50
1.25
1.00
0.75
0.50
0.25
0.00
16
64
128
learned emb
random emb


ROCstory
4
3
2
1
0
16
64
128
learned emb
random emb

target span	[3, 5, PP]
FUDGE	UNK the UNK for Italian food , The Eagle coffee shop is near Burger King in the riverside area . The Eagle has a customer rating of 5 out of 5 , and isn ’ t family - friendly . The Eagle has a cheap price range .
Diffusion-LM	The Plough , near Café Rouge , is a high priced fast food pub .
FT	Along the riverside near Café Rouge is The Golden Curry . It serves Italian food in a family - friendly environment . It has a low customer rating .
target span	[10, 12, PP]
FUDGE	Blue Spice is a high price range Fast food located in city centre .
Diffusion-LM	The Phoenix is a high priced food restaurant , located near the river .
FT	The Punter is a family restaurant with low prices and delicious sushi , located near the Café Sicilia
target span	[9, 14, S]
FUDGE	Zizzi pub serves Italian food for adults only . It has been rated average by customers .
Diffusion-LM	There is a Chinese restaurant called The Eagle , it has an average customer rating .
FT	On the riverside area are located Alimentum , has a very good French food for adults and kids , UNK price range are over 20 to 25 £ .
target span	[4, 16, VP]
FUDGE	The Cambridge Blue pub is near the Café Brazil and offers a high price range for their French food .
Diffusion-LM	On the Ranch there is a children friendly pub called The Cricketers with an average customer rating .
FT	The Travellers Rest Beefeater is an average rated restaurant located in the riverside area near Café Adriatic . Their price range is less than £ 20 .
target span	[0, 2, NP]
FUDGE	The Golden Palace is a cheap , 5 - star coffee shop , located on the river in the north of the city centre .
Diffusion-LM	The Olive Grove is a pub that provides Indian food in the high price range . It is in the city centre .
FT	The Golden Curry is located in city centre near Café Rouge which provides English food . Its customer rating is average and is not family - friendly .

target length	7
FUDGE	Wildwood is a cheap Japanese pub . Low rating .
Diffusion-LM	The Twenty Two serves Indian food .
FT	The Mill is an Indian restaurant .
target length	12
FUDGE	The Phoenix is an average Japanese restaurant that is in the City Centre .
Diffusion-LM	The Twenty Two serves Chinese food and is not family friendly .
FT	Green Man is an average priced restaurant located near All Bar One
target length	17
FUDGE	Fitzbillies is an expensive Italian coffee shop in the city centre . It is not child friendly. .
Diffusion-LM	The Twenty Two serves Indian food in the city centre . It is not family friendly .
FT	For low - priced food and a family - friendly atmosphere, visit Fitzbillies near Express by Holiday Inn
target length	22
FUDGE	The Golden Curry is an English food restaurant located near the Café Rouge in the Riverside area . The customer rating is average . Children are welcome .
Diffusion-LM	Strada is a fast food pub located near Yippee Noodle Bar and has a customer rating of 3 out of 5 .
FT	There is an Italian kid friendly restaurant in the riverside area near The Sorrento named Browns Cambridge in the riverside area .
target length	27
FUDGE	The Olive Grove is an expensive , children friendly , Fast food restaurant in the city centre . [missing 9 words]
Diffusion-LM	The Eagle is a family friendly coffee shop in the city centre near Burger King . It serves Italian food and has a low customer rating .
FT	A pub in the city centre near Yippee Noodle Bar is named Strada. It serves French food and has a customer rating of 3 out of 5
target length	32
FUDGE	The Golden Curry is a Japanese food restaurant with a high customer Rating , kid friendly and located along the riverside near Café Rouge . [missing 7 words]
Diffusion-LM	There is a family - friendly coffee shop in the city centre , it is called Zizzi . It is cheap and has a customer rating of 5 out of 5 .
FT	In the city centre is a kids friendly place called Green Man. It has Japanese food and is near All Bar One. It has a price range of £ 20 - 25

Main Q&C

- Propose Diffusion-LM （❓vs. *BERT GPT)
  - denoise Gaussian noise vectors into words → latent variable
  - add an embedding step and a rounding step → diffusion process
  - control it using a gradient-based method → objective
  - 6 control targets (new SOTA) (Speed 2000/200 * N)
  - successfully compose multiple controls
  - 👍 controllable, flexible, latent space
  - 👎 speed, ppl, pre-trained
- 推荐阅读：苏建林——生成扩散模型漫谈13讲

Q & A