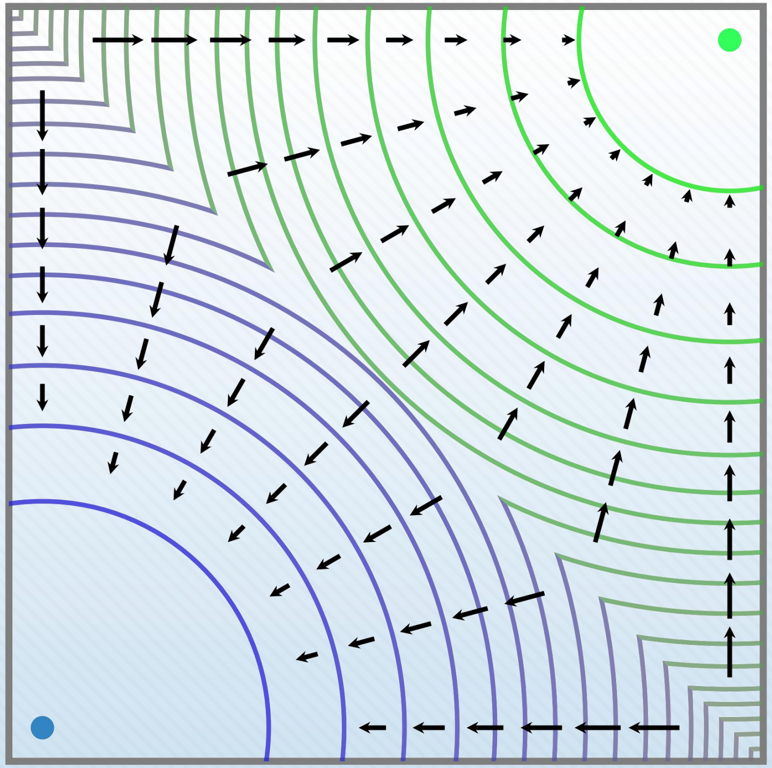
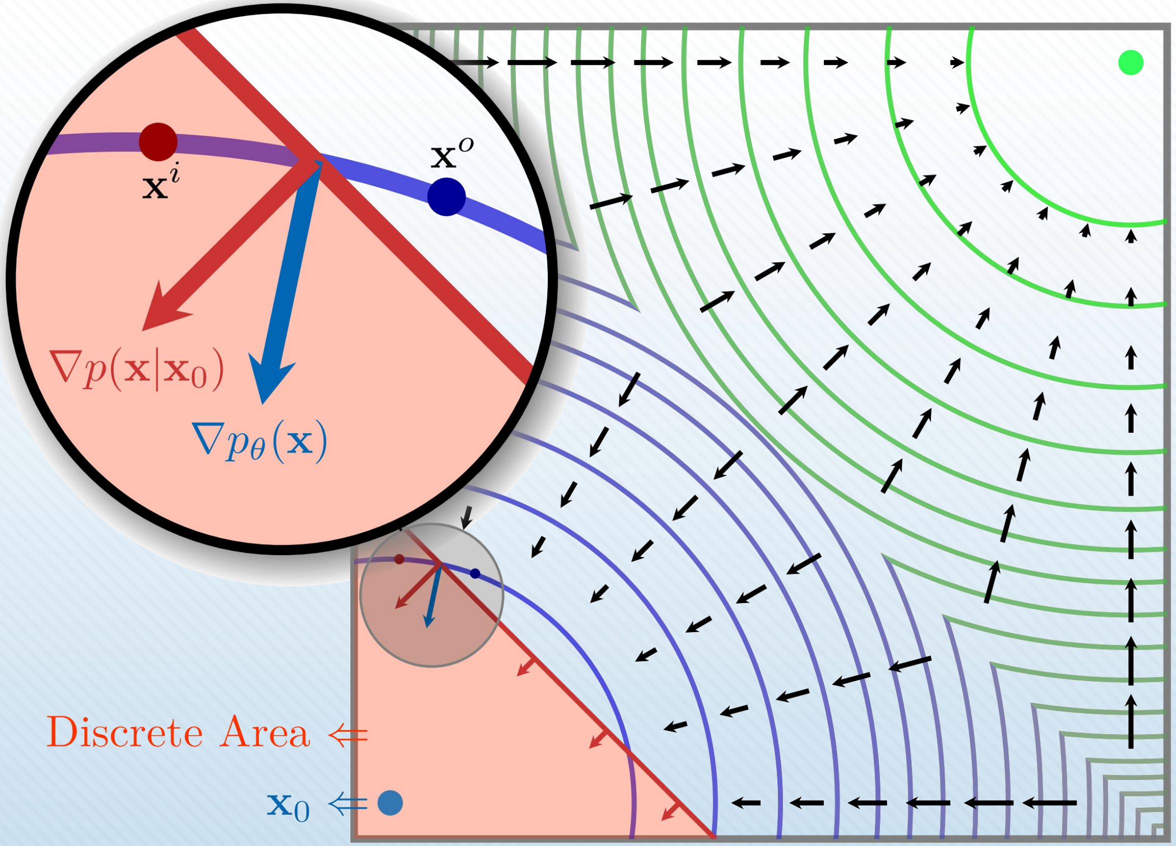
Discrete Modeling via Boundary Conditional Diffusion Processes

Yuxuan Gu, Xiaocheng Feng, Lei Huang, Yingsheng Wu, Zekun Zhou, Weihong Zhong, kun Zhu, Bing Qin
Harbin Institute of Technology

Diffusion Processes for Discrete Modeling

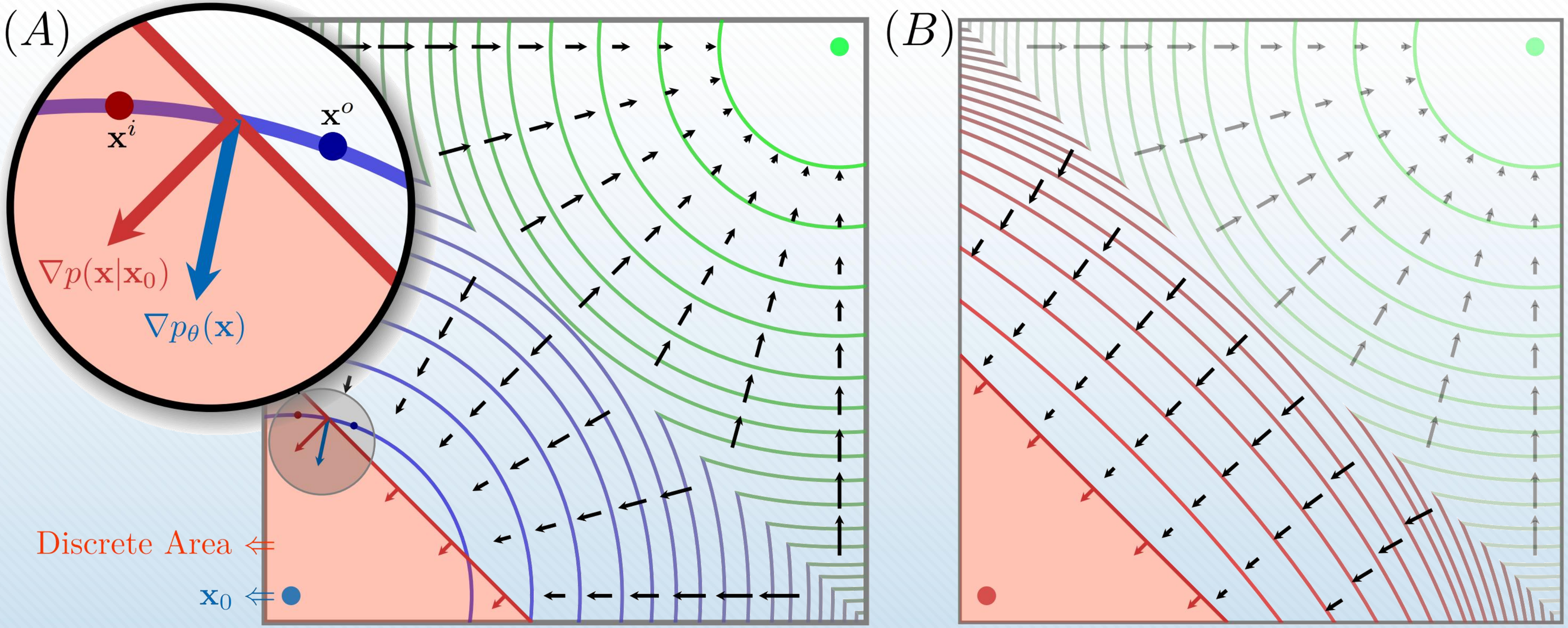
The gap between continuous diffusion processes and discrete modeling problems





Diffusion Processes for Discrete Modeling

Take the discrete boundary as a prior for the diffusion processes



Boundary Conditional Diffusion Processes

- The Forward Process

$$\tilde{p}_t(\mathbf{x}|\mathbf{x}_0) = \int \underbrace{\tilde{p}_t(\mathbf{x}|\mathbf{x}_{t_0}, t_0, \mathbf{x}_0)}_{\text{Trajectory Rescaling}} \quad \underbrace{p(\mathbf{x}_{t_0}, t_0|\mathbf{x}_0)}_{\text{Boundary Estimation}} \, \mathrm{d}\mathbf{x}_{t_0} \mathrm{d}t_0$$

  - Boundary Estimation

Diffusion Process For variance preserving, there is $\mathbf{u}_t^2 + \mathbf{v}_t^2 = 1$ and we have:

$$\mathbf{u}_{t_0} = 1 \Big/ \sqrt{1 + \left(\frac{f(\mathbf{x}_0, \mathcal{I}) - f(\mathbf{x}_0, \mathcal{J})}{f(\boldsymbol{\epsilon}, \mathcal{J}) - f(\boldsymbol{\epsilon}, \mathcal{I})}\right)^2} \text{ and } \mathbf{v}_{t_0} = 1 \Big/ \sqrt{1 + \left(\frac{f(\boldsymbol{\epsilon}, \mathcal{J}) - f(\boldsymbol{\epsilon}, \mathcal{I})}{f(\mathbf{x}_0, \mathcal{I}) - f(\mathbf{x}_0, \mathcal{J})}\right)^2}.$$

For variance exploding, there are $\mathbf{u}_t = 1$ and $\mathbf{v}_t = \sigma_t$. We can obtain:

$$\mathbf{u}_{t_0} = 1 \text{ and } \mathbf{v}_{t_0} = (f(\boldsymbol{\epsilon}, \mathcal{J}) - f(\boldsymbol{\epsilon}, \mathcal{I})) \,/\, (f(\mathbf{x}_0, \mathcal{I}) - f(\mathbf{x}_0, \mathcal{J})) .$$

Flow Matching For optimal transport, there is $\mathbf{u}_t + \mathbf{v}_t = 1$ and similarly we get:

$$\mathbf{u}_{t_0} = 1 \Big/ \left(1 + \frac{f(\mathbf{x}_0, \mathcal{I}) - f(\mathbf{x}_0, \mathcal{J})}{f(\boldsymbol{\epsilon}, \mathcal{J}) - f(\boldsymbol{\epsilon}, \mathcal{I})}\right) \text{ and } \mathbf{v}_{t_0} = 1 \Big/ \left(1 + \frac{f(\boldsymbol{\epsilon}, \mathcal{J}) - f(\boldsymbol{\epsilon}, \mathcal{I})}{f(\mathbf{x}_0, \mathcal{I}) - f(\mathbf{x}_0, \mathcal{J})}\right).$$

Boundary Conditional Diffusion Processes

- The Forward Process

$$\tilde{p}_t(\mathbf{x}|\mathbf{x}_0) = \int \underbrace{\tilde{p}_t(\mathbf{x}|\mathbf{x}_{t_0}, t_0, \mathbf{x}_0)}_{\text{Trajectory Rescaling}} \quad \underbrace{p(\mathbf{x}_{t_0}, t_0|\mathbf{x}_0)}_{\text{Boundary Estimation}} \; \mathrm{d}\mathbf{x}_{t_0}\mathrm{d}t_0$$

  - Boundary Estimation

$$p(\mathbf{x}_{t_0}, t_0|\mathbf{x}_0) = [\Psi]_* \pi([\mathbf{x}_{t_0}; t_0])$$

$$\Psi(\boldsymbol{\epsilon}) = [\psi_{G(\mathbf{x}_0,\boldsymbol{\epsilon})}(\boldsymbol{\epsilon}); G(\mathbf{x}_0, \boldsymbol{\epsilon})] \qquad \Psi^{-1}([\mathbf{x}_{t_0}; t_0]) = (\mathbf{x}_{t_0} - \mathbf{u}(\mathbf{x}_0, t_0)\mathbf{x}_0)/\mathbf{v}(\mathbf{x}_0, t_0)$$

$$t_0 = G(\mathbf{x}_0, \boldsymbol{\epsilon}), \text{ where } \mathbf{u}(\mathbf{x}_0, G(\mathbf{x}_0, \boldsymbol{\epsilon})) = \mathbf{u}_{t_0} \text{ and } \mathbf{v}(\mathbf{x}_0, G(\mathbf{x}_0, \boldsymbol{\epsilon})) = \mathbf{v}_{t_0}$$

Boundary Conditional Diffusion Processes

- The Forward Process

$$\tilde{p}_t(\mathbf{x}|\mathbf{x}_0) = \int \underbrace{\tilde{p}_t(\mathbf{x}|\mathbf{x}_{t_0}, t_0, \mathbf{x}_0)}_{\text{Trajectory Rescaling}} \quad \underbrace{p(\mathbf{x}_{t_0}, t_0|\mathbf{x}_0)}_{\text{Boundary Estimation}} \; \mathrm{d}\mathbf{x}_{t_0}\mathrm{d}t_0$$

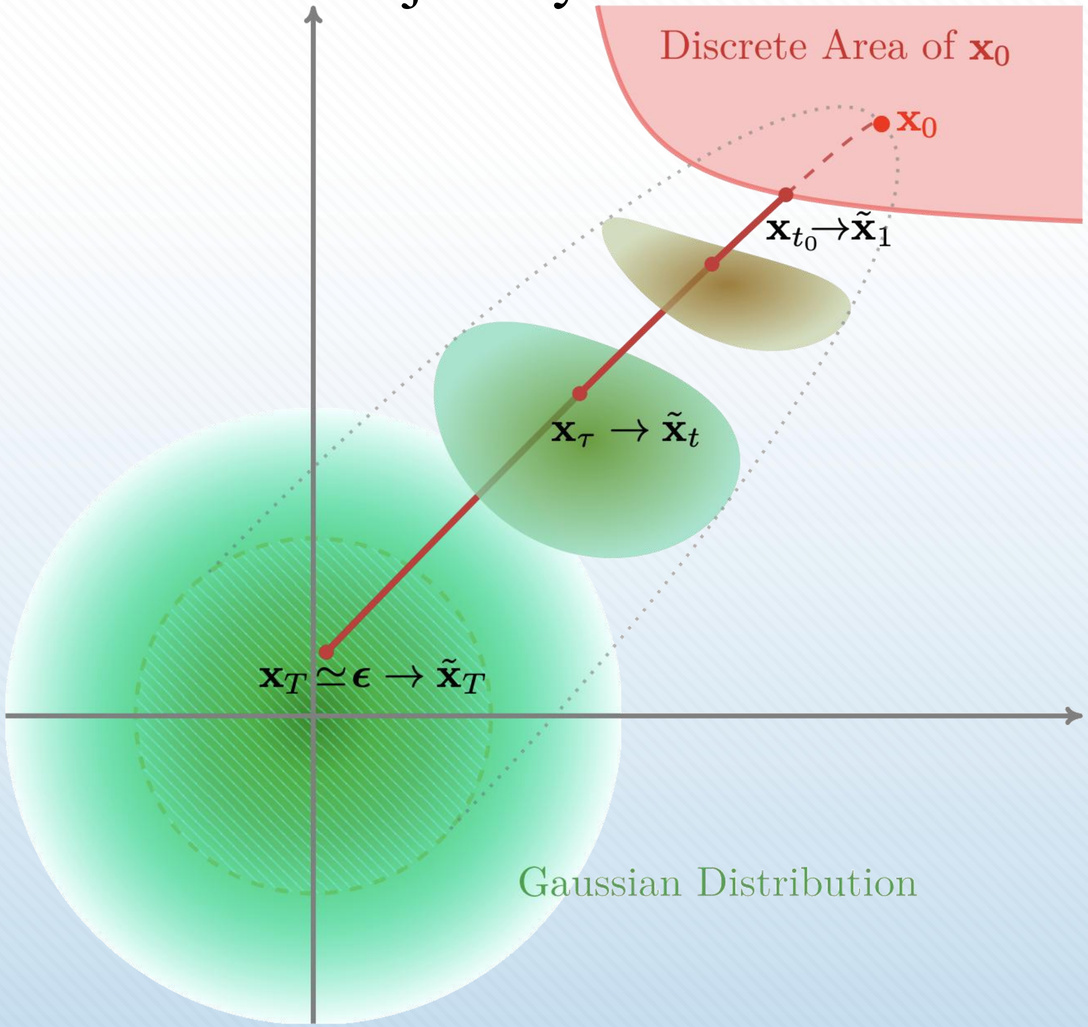
  - Trajectory Rescaling

$$\tilde{p}_t(\tilde{\mathbf{x}}_t|\mathbf{x}_0) = \int \tilde{p}_t(\tilde{\mathbf{x}}_t, \tau|\mathbf{x}_0)\mathrm{d}\tau = \int \tilde{p}_t(\tilde{\mathbf{x}}_t, \tau|\mathbf{x}_{t_0}, t_0, \mathbf{x}_0)p(\mathbf{x}_{t_0}, t_0|\mathbf{x}_0)\mathrm{d}[\mathbf{x}_{t_0}; t_0]\mathrm{d}\tau$$
$$= \int [\psi_\tau \circ \Psi^{-1} \circ \Psi]_* \pi([\tilde{\mathbf{x}}_t; \tau])\mathrm{d}\tau = \int [\psi_\tau]_* \pi([\tilde{\mathbf{x}}_t; \tau])\mathrm{d}\tau.$$

$$\tilde{\mathbf{x}}_t = \mathbf{x}_\tau, \quad \tau = \mathcal{T}(t, t_0) = r \times t_0 + t \times (T - r \times t_0)/T$$
$$= \mathbf{u}(\mathbf{x}_0, \mathcal{T}(t, t_0))\mathbf{x}_0 + \mathbf{v}(\mathbf{x}_0, \mathcal{T}(t, t_0))\Psi^{-1}([\mathbf{x}_{t_0}; t_0]).$$

Boundary Conditional Diffusion Processes

- Rescaled Trajectory



Algorithm 1 Training

repeat
  $\mathbf{x}_0 \sim q(\mathbf{x}_0), \boldsymbol{\epsilon} \sim \pi(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $t \sim \text{Uniform}(\{1, \dots, T\})$
  $\tau := \mathcal{T}(t, G(\mathbf{x}_0, \boldsymbol{\epsilon}))$ // eqs. (12) and (19)
  $\tilde{\mathbf{x}}_t := \mathbf{u}(\mathbf{x}_0, \tau)\mathbf{x}_0 + \mathbf{v}(\mathbf{x}_0, \tau)\boldsymbol{\epsilon}$ // eq. (22)
  Take gradient descent step on
   $\nabla_\theta \|\mathbf{x}_0 - \mathbf{x}_\theta(\tilde{\mathbf{x}}_t, t)\|^2$ // eq. (23)
until converged

Algorithm 2 Sampling

$t := T, \tau := T$
$\hat{\epsilon} \simeq \tilde{\mathbf{x}}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ // Initialing
for $\Delta t := \Delta t_1, \dots, \Delta t_s$ **do** // $\sum_{\Delta t} = T$
  $\hat{\mathbf{x}}_0 := \mathbf{x}_\theta(\tilde{\mathbf{x}}_t, t)$ // Pseudo Target
  $t := t - \Delta t$ // Updating
  $\tau := \mathcal{T}(t, G(\hat{\mathbf{x}}_0, \hat{\boldsymbol{\epsilon}}))$ // eq. (25)
  $\tilde{\mathbf{x}}_t := \mathbf{u}(\hat{\mathbf{x}}_0, \tau)\hat{\mathbf{x}}_0 + \mathbf{v}(\hat{\mathbf{x}}_0, \tau)\hat{\boldsymbol{\epsilon}}$
  $\hat{\epsilon} := \Psi^{-1}([\tilde{\mathbf{x}}_t; \tau])$ // Trajectory Alteration
end for
$\mathbf{x}_0 := \mathbf{x}_\theta(\tilde{\mathbf{x}}_t, t)$ // $\mathbf{x}_1 \to \mathbf{x}_0$
return $\mathbf{x}_0$

Boundary Conditional Diffusion Processes

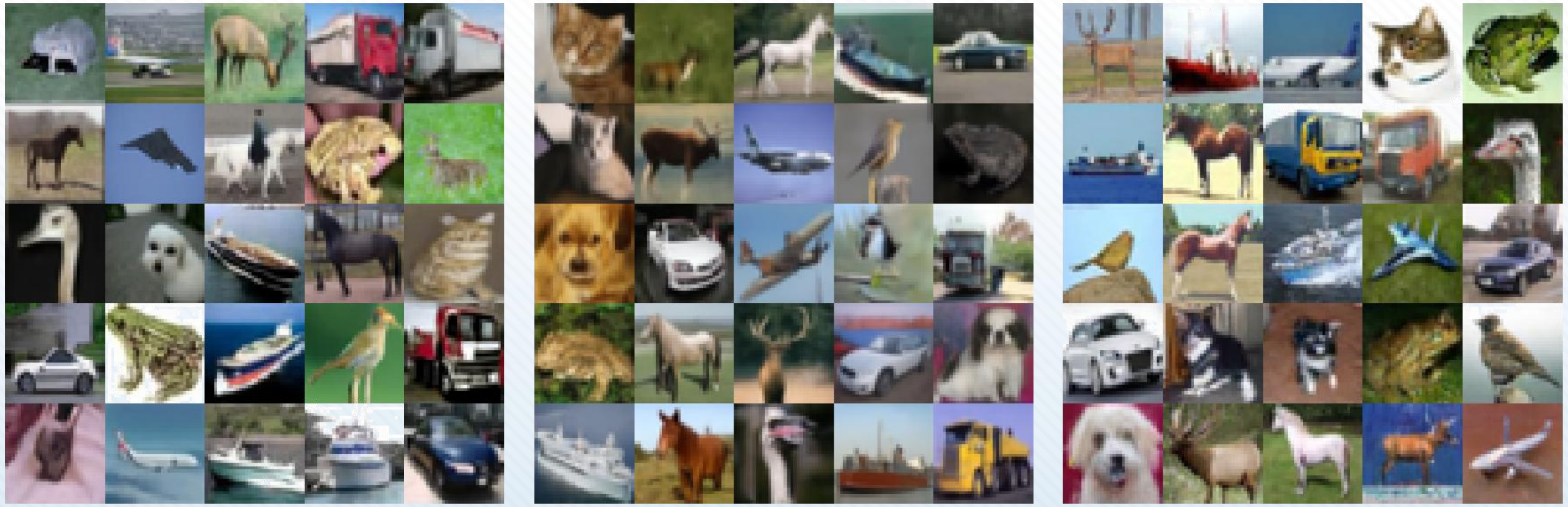
- Language Modeling

Table 1: Result of BLEU scores on machine translation and ROUGE scores on text summarization.

Models	IWSLT14 DE-EN BLEU (BLEU-1/2/3/4)⇑	WMT14 EN-DE BLEU (BLEU-1/2/3/4)⇑	WMT16 EN-RO BLEU (BLEU-1/2/3/4)⇑	GIGAWORD ROUGE-1/2/L⇑
Auto-Regressive Modeling				
Transformers	34.31 (67.3/41.6/27.9/19.1)	**28.01** (58.2/33.5/21.7/14.6)	34.05 (63.1/39.9/27.6/19.6)	**37.57**/**18.90**/34.69
Ours+Rerank	**35.02** (68.7/43.3/29.2/20.1)	27.67 (57.9/33.2/21.4/14.3)	**34.33** (63.1/40.1/27.8/19.8)	37.49/18.68/**34.82**
Diffusion Process				
D3PM	27.61 (65.4/37.7/22.8/14.2)	22.94 (54.9/28.8/16.9/10.4)	27.84 (59.8/34.9/22.1/14.5)	33.92/14.96/31.72
DiffuSeq	28.78 (- / - / - / -)	15.37 (- / - / - / -)	25.45 (- / - / - / -)	31.17/12.23/29.24
SeqDiffuSeq	30.03 (- / - / - / -)	17.14 (- / - / - / -)	26.17 (- / - / - / -)	31.90/12.36/29.22
Difformer	31.58 (68.6/41.4/26.7/17.5)	24.80 (58.7/32.0/19.7/12.5)	30.08 (64.4/39.5/26.5/18.2)	35.47/15.17/32.82
SEDD	31.87 (68.7/41.8/27.2/18.0)	24.98 (59.2/32.4/20.1/12.9)	29.38 (62.2/38.0/24.9/16.9)	34.33/15.22/32.06
Dinoiser	31.91 (67.1/40.9/26.7/17.7)	24.77 (57.2/31.0/19.0/12.0)	31.49 (62.8/38.4/25.5/17.3)	35.17/15.63/32.53
Ours	**33.42** (68.0/42.0/27.7/18.6)	**26.69** (57.7/32.3/20.4/13.4)	**33.15** (63.4/39.9/27.4/19.2)	**36.44/16.09/33.56**

Boundary Conditional Diffusion Processes

- Discrete Image Generation



(A) Bit Diffusion *repro* (FID 10.37) (B) DDIM (FID 4.04) (C) Ours (FID 3.86)

Figure 3: Generated images of Bit Diffusion *repro*, DDIM, and Ours on CIFAR-10.

Table 3: FID scores on CIFAR-10.

Models	CIFAR-10 (FID ⇓) 200K	500K	Final
Continuous Pixels			
DDPM	-	-	3.17
DDIM	-	-	4.04
Discrete Ordinal Pixels			
D3PM GAUSS	-	-	7.34
τ**LDR**-0	-	-	8.10
τ**LDR**-10	-	-	3.74
BINARY CODING (UINT8):			
Bit Diffusion	-	-	3.48
Bit Diffusion *repro*	22.12	13.23	10.37
Ours	8.17	5.03	**3.86**
FIXED EMBEDDING:			
Bit Diffusion *repro*	19.69	16.61	12.96
Ours	12.32	10.09	**9.15**
Categorical Pixels			
D3PM UNIFORM	-	-	51.27
D3PM ABSORBING	-	-	30.97
VECTOR QUANTIZATION:			
D3PM-VQ	-	-	16.47
τ**LDR**-VQ	-	-	40.06
SDDM-VQ	-	-	12.23
TRAINABLE EMBEDDING:			
Bit Diffusion *repro*	33.09	27.21	19.26
Ours	21.17	15.32	**10.99**

Conclusion

- Observe the challenges in continuous diffusion models when applied to discrete modeling.
- Attempt to add the discreteness as a prior into the diffusion processes