Improved Diffusion-based Generative Model with Better Adversarial Robustness

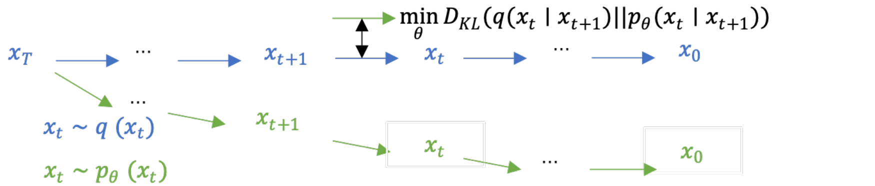
Zekun Wang, Mingyang Yi, Shuchen Xue, Zhenguo Li, Ming Liu, Bing Qin, Zhi-Ming Ma

Motivation: *Distribution Mismatch in Diffusion-based Models*

- **Exposure Bias:** Mismatch in training and inference:

$$\sum_{t=0}^{T-1}\underbrace{D_{KL}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))}_{L_t} \text{ vs } p_{\boldsymbol{\theta}}(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}) = \mathcal{N}(\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1}, t+1), \sigma_{t+1}^2 \boldsymbol{I}),$$

- Training: condition $x_{t+1} \sim q(x_{t+1})$ Inference: condition $x_{t+1} \sim p_\theta(x_{t+1})$



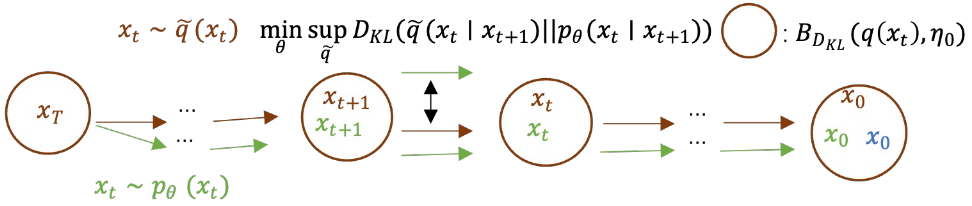
- The mismatch error is severe with **fewer sampling steps** and **is accumulated during inference process!**

- Exposure Bias also in Consistency Distillation Models

$$\min_{\boldsymbol{\theta}} \mathcal{L}_{CD}(\boldsymbol{\theta}) = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_{t+1} \sim q(\boldsymbol{x}_{t+1})} \left[d\left(f_{\boldsymbol{\theta}}(\Phi_t(\boldsymbol{x}_{t+1}), t), f_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1}, t+1)\right)\right],$$

- $\boldsymbol{\Phi_t(x_{t+1}) \neq \hat{\Phi}_t(x_{t+1}, \epsilon_\phi)}$

- Intuition: Improving **Distributional Robustness** can alleviate mismatch
  - **Distributional Robustness (DRO)**: Condition $x_{t+1} \sim \tilde{q}\ (x_{t+1})$,
- $\tilde{q}(x_{t+1})$ is in a ball over the ground truth $q(x_{t+1})$, which covers $p_\theta(x_{t+1})$
  - Counting Biases in Training: Minimization conducted on a ball over $x_{t+1} \sim q\ (x_{t+1})$



Method

- Objectives:
  - Diffusion Models

$$\min_{\boldsymbol{\theta}} \sum_{t=0}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_0)}\left[\mathbb{E}_{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}\left[\sup_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\| \leq \eta} \left\| \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t + \boldsymbol{\delta}) - \boldsymbol{\epsilon}_t - \frac{\boldsymbol{\delta}}{\sqrt{1-\bar{\alpha}_t}} \right\|^2\right]\right]$$

  - Consistency Distillation

$$\hat{\mathcal{L}}_{CD}^{Adv}(\boldsymbol{\theta}) = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{x}_{t+1}}\left[\sup_{\|\boldsymbol{\delta}\| \leq \eta} d\left(f_{\boldsymbol{\theta}}(\hat{\Phi}_t(\boldsymbol{x}_{t+1}, \boldsymbol{\epsilon}_\phi) + \boldsymbol{\delta}, t), f_{\boldsymbol{\theta}}(\boldsymbol{x}_{t+1}, t+1)\right)\right]$$

- Efficient Implementation
  - **Fine-tuning** Models with **Free AT**!

Algorithm 1 Adversarial Training for Diffusion Model

Input: dataset $\mathcal{D}$, model parameter $\boldsymbol{\theta}$, learning rate κ, loss weighting $\lambda(\cdot)$, adversarial steps K, adversarial learning rate α
while do not converge **do**
  Sample $\boldsymbol{x} \sim \mathcal{D}$ and $t \sim \mathcal{U}[1, T]$
  Sample $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
  $\boldsymbol{\delta} \leftarrow \boldsymbol{0}$
  for $i = 1, 2, \ldots, K$ **do**
    $\mathcal{L} \leftarrow \left\| \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} + \boldsymbol{\delta}) - \boldsymbol{\epsilon} - \frac{\boldsymbol{\delta}}{\sqrt{1-\bar{\alpha}_t}} \right\|^2$ in (14)
    $\boldsymbol{\delta} \leftarrow \boldsymbol{\delta} + \alpha \cdot \frac{\nabla_{\boldsymbol{\delta}}\mathcal{L}}{\|\nabla_{\boldsymbol{\delta}}\mathcal{L}\|}$ ▷ maximize perturbation
    $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \kappa \cdot \nabla_{\boldsymbol{\theta}}\mathcal{L}$ ▷ update model
  end for
end while

Experiment on Consistency Models

- Results of LCM on MS-COCO 2014 at 512x512

Methods	FID ↓				CLIP Score ↑			
	1 step	2 step	4 step	8 step	1 step	2 step	4 step	8 step
LCM	25.43	12.61	11.61	12.62	29.25	30.24	30.40	30.47
LCM-AT (Ours)	**23.34**	**11.28**	**10.31**	**10.68**	**29.63**	**30.43**	**30.49**	**30.53**

Experiment on Diffusion Models

- Results on CIFAR-10 32x32

(a) IDDPM

Methods \ NFEs	5	8	10	20	50
ADM (original)	37.99	26.75	22.62	10.52	4.55
ADM (finetune)	**36.91**	26.06	21.94	10.58	4.34
ADM-IP	47.57	26.91	20.09	7.81	3.42
ADM-AT (Ours)	37.15	**23.59**	**15.88**	**6.60**	**3.34**

(b) DDIM

Methods \ NFEs	5	8	10	20	50
ADM (original)	34.28	14.34	11.66	7.00	4.68
ADM (finetune)	29.30	15.08	12.06	6.80	4.15
ADM-IP	43.15	15.72	10.47	4.58	4.89
ADM-AT (Ours)	**26.38**	**12.98**	**9.30**	**4.40**	**3.07**

(c) ES

Methods \ NFEs	5	8	10	20	50
ADM (original)	82.18	29.28	17.73	5.11	2.70
ADM (finetune)	63.46	24.80	17.03	5.19	2.52
ADM-IP	91.10	31.44	18.72	5.19	2.89
ADM-AT (Ours)	**41.07**	**21.62**	**14.68**	**4.36**	**2.48**

(d) DPM-Solver

Methods \ NFEs	5	8	10	20	50
ADM (original)	23.95	8.00	5.46	3.46	3.14
ADM (finetune)	22.98	7.61	5.29	3.41	3.12
ADM-IP	43.83	6.70	6.80	9.78	10.91
ADM-AT (Ours)	**18.40**	**5.84**	**4.81**	**3.28**	**3.01**

- Results on ImageNet 64x64

(a) IDDPM

Methods \ NFEs	5	8	10	20	50
ADM (original)	76.92	33.74	27.63	12.85	5.30
ADM (finetune)	78.87	33.99	27.82	12.80	5.26
ADM-IP	67.12	29.96	22.60	8.66	**3.83**
ADM-AT (Ours)	**45.65**	**23.79**	**19.18**	**8.28**	4.01

(b) DDIM

Methods \ NFEs	5	8	10	20	50
ADM (original)	60.07	20.10	14.97	8.41	5.65
ADM (finetune)	60.32	20.26	15.04	8.32	5.48
ADM-IP	76.51	26.25	18.05	8.40	6.94
ADM-AT (Ours)	**43.04**	**16.08**	**12.15**	**6.20**	**4.67**

(c) ES

Methods \ NFEs	5	8	10	20	50
ADM (original)	71.31	28.97	21.10	8.23	3.76
ADM (finetune)	72.30	29.24	21.58	8.25	3.64
ADM-IP	88.37	33.91	23.32	7.80	3.54
ADM-AT (Ours)	**43.95**	**19.57**	**14.12**	**6.16**	**3.45**

(d) DPM-Solver

Methods \ NFEs	5	8	10	20	50
ADM (original)	27.72	10.06	7.21	4.69	4.24
ADM (finetune)	27.82	9.97	7.22	4.64	**4.15**
ADM-IP	32.43	9.94	8.87	9.16	9.68
ADM-AT (Ours)	**17.36**	**6.55**	**5.78**	**4.56**	4.34

Visualization

Baseline Ours


A photo of beautiful mountain with realistic sunset and blue lake, highly detailed, masterpiece


Astronaut in a jungle, cold color palette, muted colors, detailed, 8k.