# dp-promise: Differentially Private Diffusion Probabilistic Models for Image Synthesis

Haichen Wang[1], Shuchao Pang[1*], Zhigang Lu[2*],
Yihang Rao[1], Yongbin Zhou[1] & Minhui Xue[3]

[1]Nanjing University of Science and Technology, China
[2]James Cook University, Australia
[3]CSIRO's Data61, Australia

August 2024

# Outline

- **Background & Preliminaries**
- **Existing Work**
- **Our Method**
- **Experimental Evaluation**
- **Conclusion**

# Outline

- **Background & Preliminaries**
- Existing Works
- Our Method
- Experimental Evaluation
- Conclusion

# Background

- Large-scale data is crucial for DNN performance.
- Synthetic images produced by generative models can still lead to privacy leakage in sensitive domains.

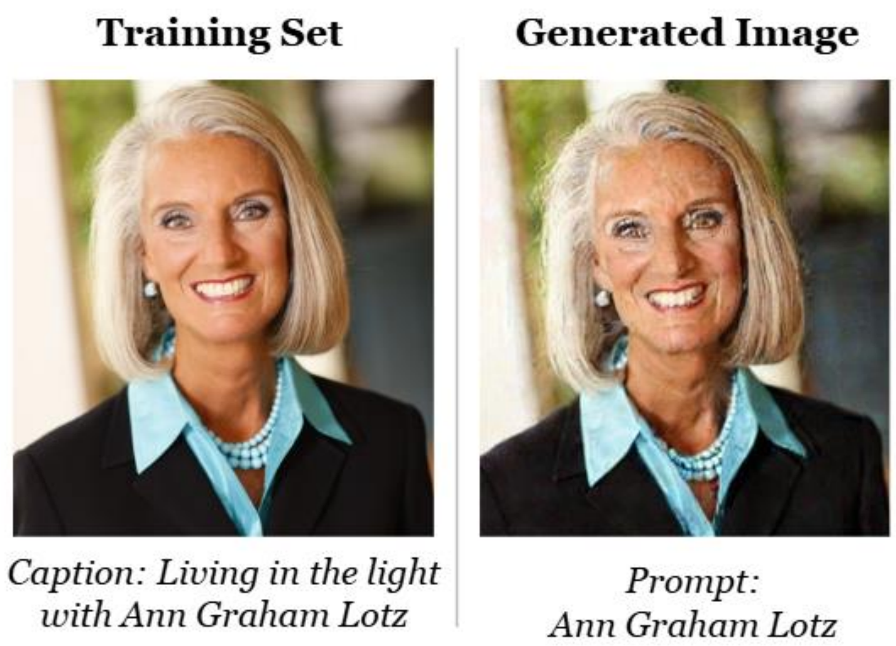


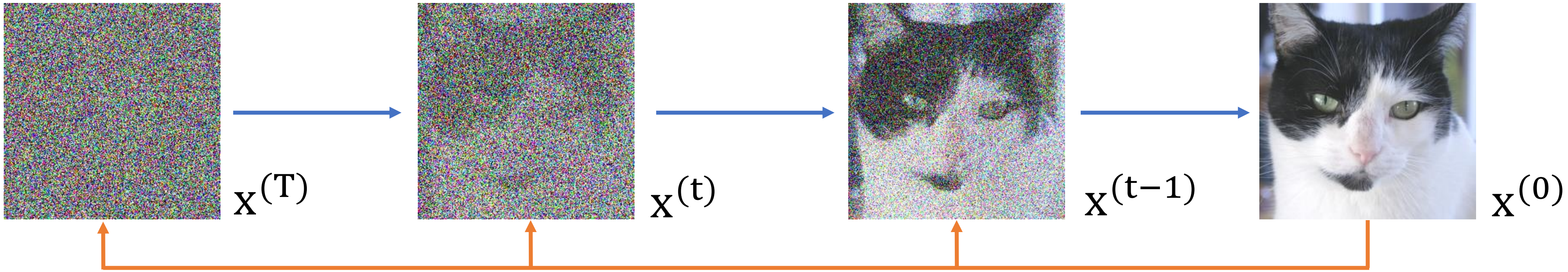

Fig: Left is an image from Stable Diffusion's training set. Right is a Stable Diffusion generation when prompted with "Ann Graham Lotz". [1]

[1] Carlini, Nicolas, et al. "Extracting training data from diffusion models." *32nd USENIX Security Symposium*. 2023.

# Diffusion Models

- Forward process
- Reverse process

$$P_\theta(\mathrm{x}^{(t-1)}|\mathrm{x}^{(t)})$$




$$q\left(\mathrm{x}^{(t)}\middle|\mathrm{x}^{(0)}\right) = \mathrm{N}(\mathrm{x}^{(t)}; \sqrt{\alpha_t}x^{(0)}, (1-\alpha_t)I)$$

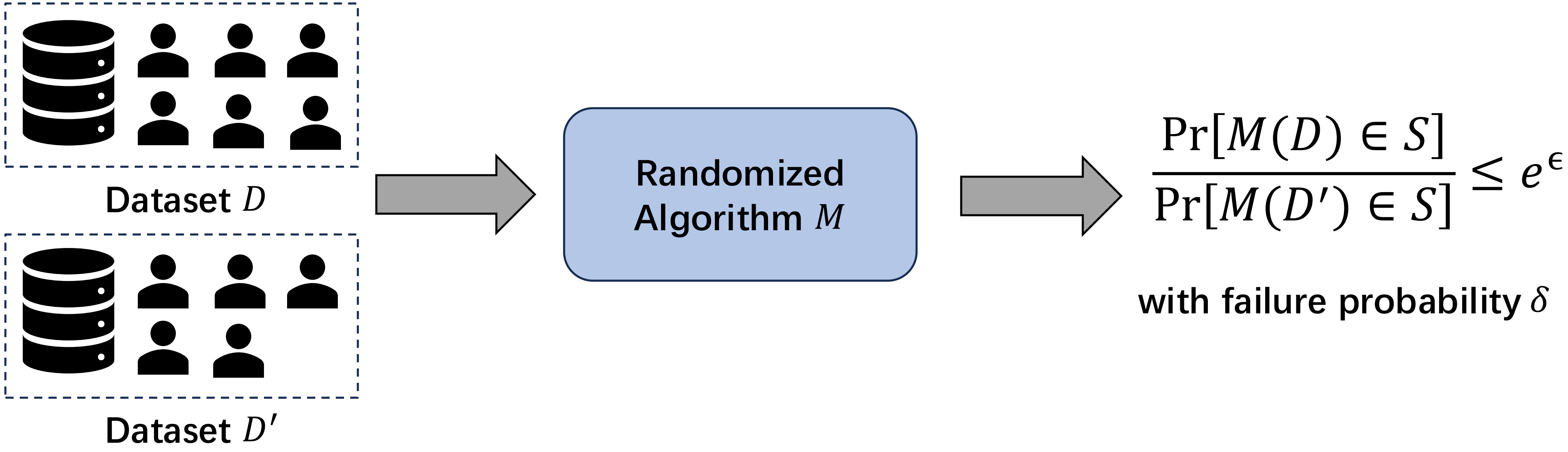
$\alpha_t$: noise scale

# Differential Privacy

- A randomized algorithm $M$ is $(\epsilon, \delta)$-DP if and only if

$$\Pr[M(D) \in S] \leq e^{\epsilon} \Pr[M(D') \in S] + \delta$$

where $D$ and $D'$ are two neighboring datasets.

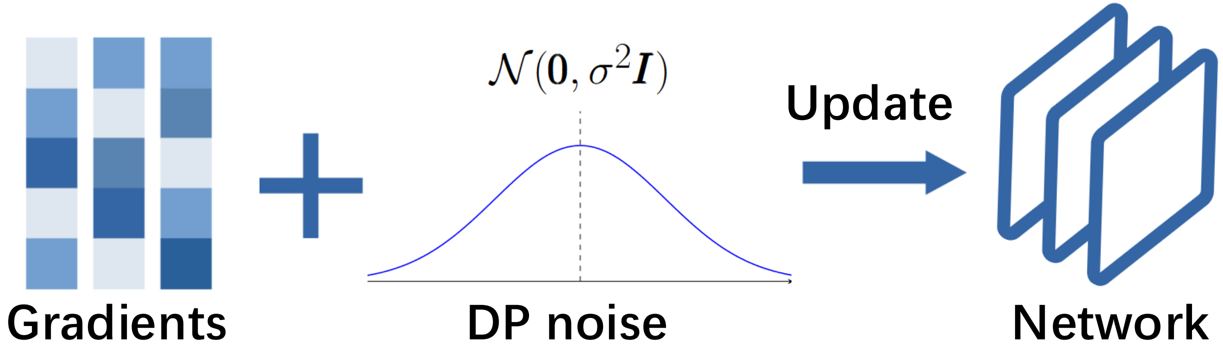


# Outline

- Background & Preliminaries
- **Existing Works**
- Our Method
- Experimental Evaluation
- Conclusion

# Existing Approaches

- Based on **GANs** [DP-GAN'18] [GS-WGAN'20] [G-PATE'21]
- Based on **Feature Matching** [DP-MERF'21] [DP-MEPF'23]
- Based on **Diffusion Models** [DPDM'23] [DP-Diffusion'23]




- DM-based methods overlook inherent "**privacy features**".

# Outline

- Background & Preliminaries
- Existing Works
- **dp-promise**
- Experimental Evaluation
- Conclusion

# Threat Model

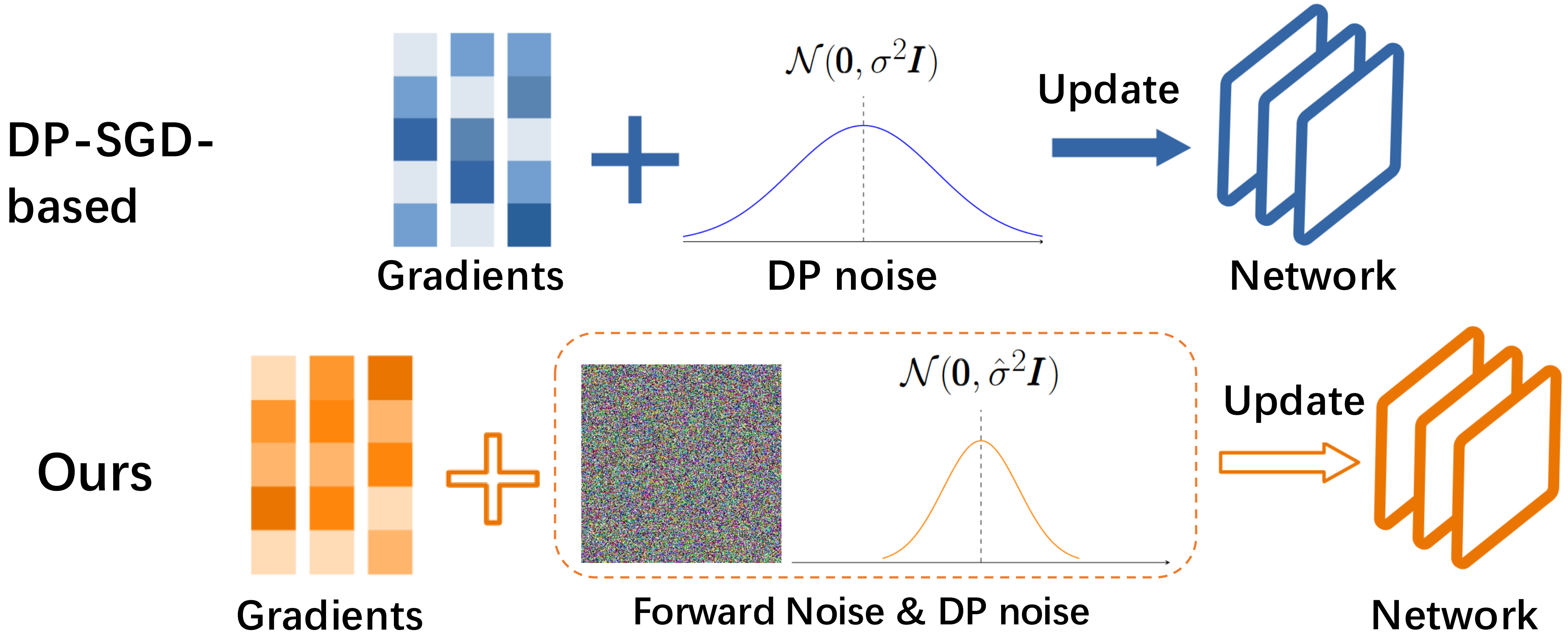
- White-box adversaries against DMs
  - Given access to the images generated by DMs and the model parameters of the trained DMs.
  - Infer the existence of a particular image or reconstruct a set of images belonging to the DMs training data,

**Definition 4** (White-box membership inference attacks). *Let $\mathcal{A}$ be a white-box adversary, $\mathcal{D}$ be data distribution, A be training algorithm, and $\mathcal{G}$ be a diffusion model with a neural network $\mathbf{z}_\theta$. The white-box membership inference attack is*

0. *$\mathcal{A}$ has full access to $\mathcal{G}$ and $\mathbf{z}_\theta$.*
1. *Select a private dataset $D_{priv} \in \mathcal{D}$.*
2. *Train $\mathcal{G}$ on $D_{priv}$ with algorithm A as $\mathcal{G}_{A,D_{priv}} = A(\mathcal{G}, D_{priv})$.*
3. *Flip a coin to decide whether $b = 0$ or $b = 1$.*
4. *Sample $\mathbf{x} \in D_{priv}$ if $b = 0$, $\mathbf{x} \in \mathcal{D}$ if $b = 1$.*
5. *Attack is successful if $\mathcal{A}(\mathbf{x}, \mathcal{G}_{A,D_{priv}}, \mathcal{D}) = b$, and fails otherwise.*

# Overview

- How to leverage forward process noise? Existing vs Ours

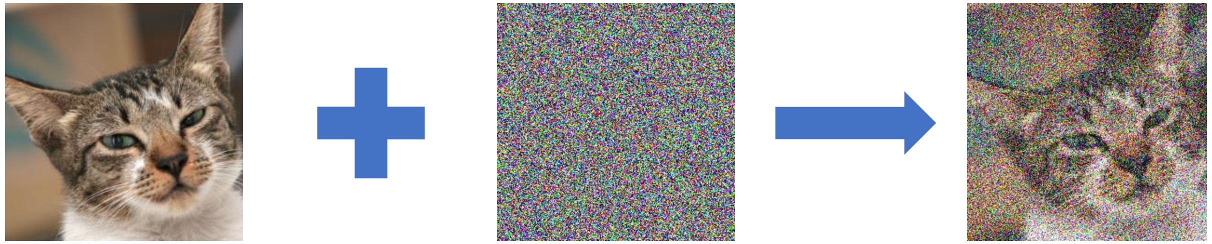


# dp-promise

- Recall forward process

$$x^{(t)} = \sqrt{\alpha_t} x^{(0)} + \sqrt{1 - \alpha_t} z, z \sim N(0, I)$$

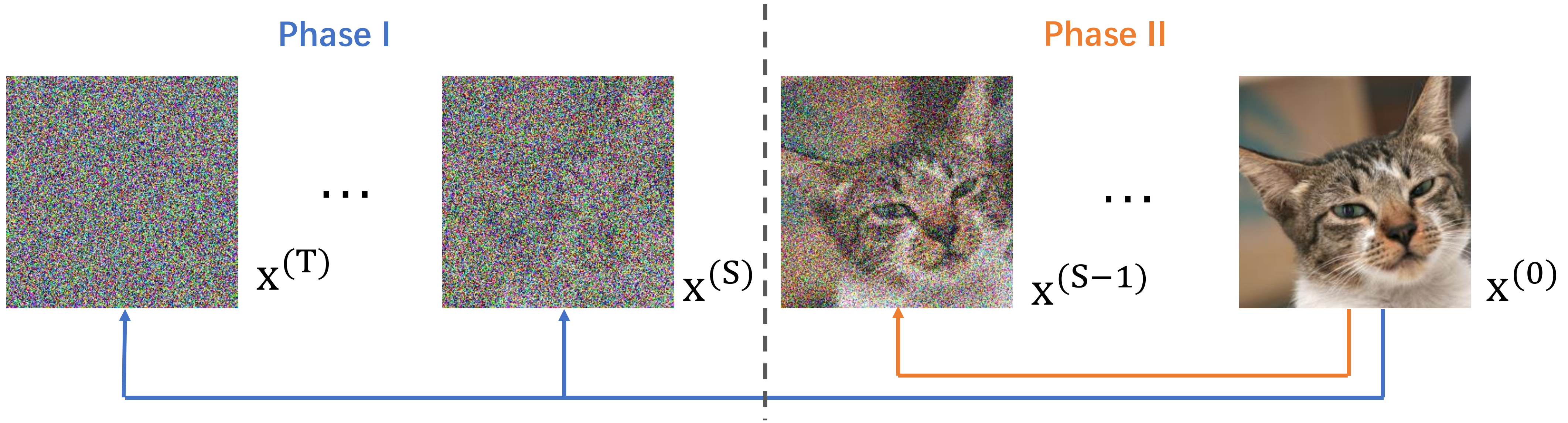
- Forward process is differentially private



**Original image** $x^{(0)}$ **Noise** $z$ **Noisy image** $x^{(t)}$

# dp-promise

- **Phase I**: Non-private Training
- **Phase II**: Private Training



# Privacy Analysis

- dp-promise asymptotically satisfies $\left(\epsilon, \delta(\epsilon)\right)$-DP, where

**Lemma 4.** *Given a time-step boundary S for splitting Phase I and Phase II, a batch size $m_1$, the size of the private dataset n, the data dimensions d, the pre-defined diffusion noise scale $\alpha_S$, and the number of iterations $N_1$, Phase I in Algorithm 1 asymptotically satisfies $\mu_1$-GDP, where*

$$\mu_1 = \frac{m_1}{n}\sqrt{N_1(\exp(4d\alpha_S/(1-\alpha_S)-1)}. \tag{15}$$

**Lemma 5.** *Given a DP-SGD noise scale σ, a batch size $m_2$, the size of the private dataset n, and the number of iterations $N_2$, Phase II in Algorithm 1 satisfies $\mu_2$-GDP, where*

$$\mu_2 = \frac{m_2}{n}\sqrt{N_2(\exp(1/\sigma^2)-1)}. \tag{16}$$

**Theorem 2** (Differential privacy for dp-promise)**.** *Algorithm 1 asymptotically satisfies $(\varepsilon, \delta(\varepsilon))$-DP, it holds that*

$$\delta(\varepsilon) = \Phi(-\frac{\varepsilon}{\mu}+\frac{\mu}{2}) - \exp(\varepsilon)\Phi(-\frac{\varepsilon}{\mu}-\frac{\mu}{2}), \tag{17}$$

$$\mu = \sqrt{\mu_1^2+\mu_2^2}, \tag{18}$$

*where $\mu_1$ is defined in Equation (15) and $\mu_2$ is defined in Equation (16).*

# Outline

- Background & Preliminaries
- Existing Works
- Our Proposal
- **Experimental Evaluation**
- Conclusion

# Experimental Setup

- **Datasets:**
  - MNIST, Fashion-MNIST, CelebA, and CIFAR-10
- **Metrics:**
  - Sample quality (FID, IS)
  - Downstream utility (Classification accuracy)
- **Baselines:**
  - Feature Matching: DP-MERF, DP-MEPF
  - Diffusion Model: DPDM, DP-Diffusion

# Gray-scale Datasets (metrics)

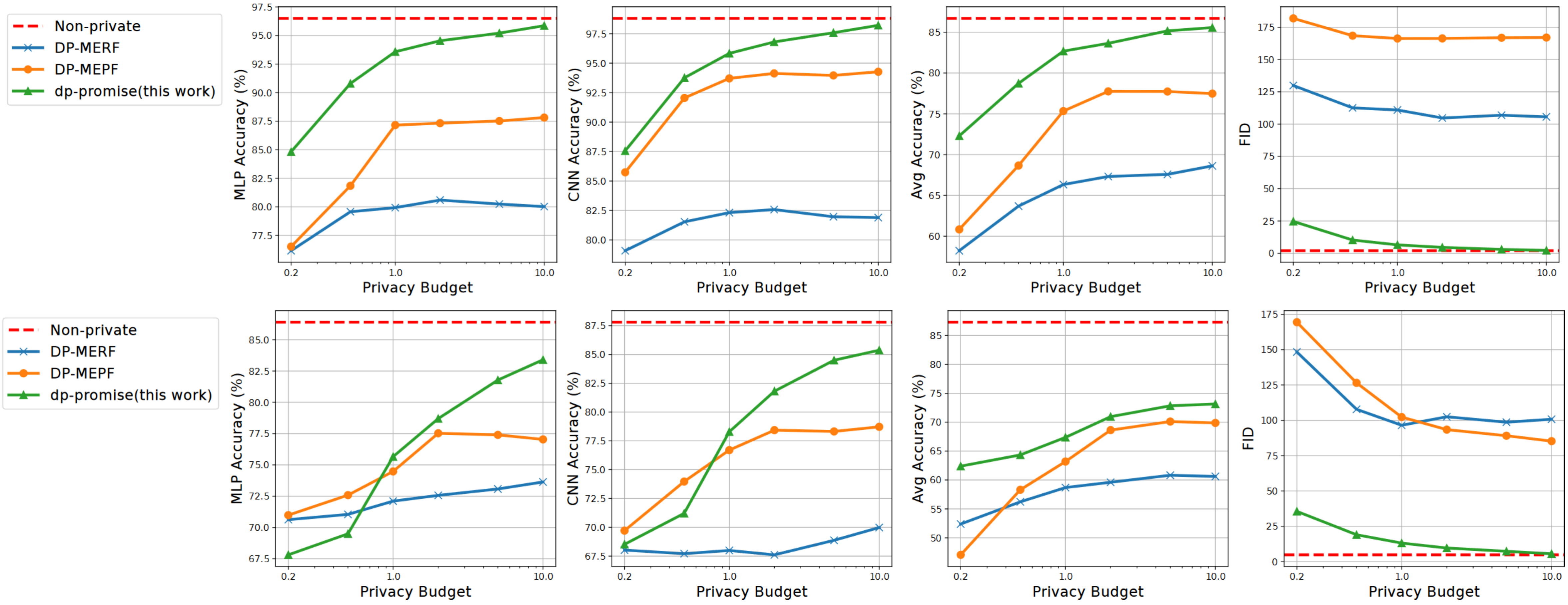
| MNIST | $D_{pub}$ | $\varepsilon = \infty$ (Non-private) | | | | $\varepsilon = 10$ | | | | $\varepsilon = 1$ | | | | $\varepsilon = 0.2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ |
| DP-MERF [18] | ✗ | 80.4 | 83.5 | 70.5 | 104.4 | 80.0 | 83.5 | 68.6 | 105.6 | 80.0 | 82.3 | 66.3 | 110.9 | 76.2 | 79.0 | 58.2 | 133.3 |
| DPDM (FID) [11] | ✗ | 95.7 | 98.6 | 85.7 | 2.0 | 94.5 | 97.8 | 85.4 | 4.4 | 87.7 | 92.7 | 77.8 | 22.4 | 66.4 | 71.2 | 54.1 | 60.8 |
| DPDM (Acc) [11] | ✗ | **96.6** | **98.9** | **86.4** | 1.9 | 95.2 | 98.0 | **85.8** | 5.9 | 91.5 | 95.1 | 82.1 | 34.1 | 78.0 | 84.6 | 71.6 | 101.9 |
| DP-MEPF [19] | ✓ | 87.6 | 94.3 | 77.9 | 167.2 | 87.8 | 94.3 | 77.5 | 167.0 | 87.2 | 93.7 | 75.3 | 166.3 | 76.5 | 85.7 | 58.3 | 180.2 |
| DP-SGD DM | ✓ | 96.4 | 98.6 | 86.2 | 1.7 | 94.5 | 97.6 | 85.1 | 3.0 | 90.8 | 94.1 | 75.5 | 8.6 | 56.8 | 65.3 | 42.8 | 28.3 |
| DPDM (Pub) | ✓ | 96.5 | 98.8 | **86.4** | 1.9 | 95.3 | 97.8 | 85.6 | 3.9 | 92.3 | 95.6 | 82.2 | 9.0 | 81.3 | 86.2 | **73.3** | 26.5 |
| dp-promise (this work) | ✓ | 96.4 | 98.7 | 86.1 | **1.6** | **95.9** | **98.2** | 85.6 | **2.3** | **93.6** | **95.8** | **83.0** | **6.6** | **84.8** | **87.6** | 72.3 | **23.1** |

| Fashion-MNIST | $D_{pub}$ | $\varepsilon = \infty$ (Non-private) | | | | $\varepsilon = 10$ | | | | $\varepsilon = 1$ | | | | $\varepsilon = 0.2$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ |
| DP-MERF [18] | ✗ | 73.8 | 63.4 | 63.2 | 103.3 | 72.6 | 70.0 | 60.6 | 100.7 | 75.1 | 64.0 | 58.7 | 96.5 | 70.6 | 69.0 | 52.4 | 149.8 |
| DPDM (FID) [11] | ✗ | 84.8 | 87.3 | **74.1** | 8.0 | 82.6 | 85.3 | 72.1 | 17.9 | 74.4 | 77.1 | 66.7 | 45.1 | 55.3 | 55.5 | 45.6 | 76.7 |
| DPDM (Acc) [11] | ✗ | 86.4 | 87.7 | 73.3 | 7.0 | 83.1 | 85.4 | 72.6 | 18.1 | 76.1 | 78.6 | 68.8 | 50.3 | 69.2 | 72.7 | 65.5 | 126.5 |
| DP-MEPF [19] | ✓ | 74.9 | 79.4 | 69.7 | 86.7 | 74.0 | 78.7 | 66.0 | 89.1 | 74.5 | 76.7 | 63.2 | 102.3 | **71.0** | 69.7 | 47.1 | 167.5 |
| DP-SGD DM | ✓ | 85.8 | 87.6 | 73.8 | 5.7 | 82.3 | 84.6 | 71.1 | 6.4 | 65.7 | 69.7 | 53.9 | 16.5 | 44.2 | 50.8 | 41.7 | 38.4 |
| DPDM (Pub) | ✓ | **86.5** | **87.9** | 73.9 | 5.2 | 82.0 | 85.0 | 71.2 | 10.4 | 76.5 | 80.2 | **69.8** | 20.9 | 70.4 | **73.8** | **68.3** | 40.2 |
| dp-promise (this work) | ✓ | 85.7 | 87.4 | 73.5 | **4.8** | **83.4** | **85.5** | **73.1** | **6.3** | **78.4** | **81.6** | 69.2 | **13.6** | 67.8 | 68.5 | 62.4 | **34.8** |

# Gray-scale Datasets (images)


Figure 3: The synthetic data generated by DP-MERF, DPDM, DP-MEPF, DP-SGD DM, and dp-promise under $\varepsilon = 10$ and $\delta = 10^{-5}$ on MNIST and Fashion-MNIST. The original data is presented in the last row.

# Gray-scale Datasets (privacy-utility trade-off)



# Color Datasets (metrics)

| CelebA | $D_{pub}$ | $\varepsilon = 10$ | | $\varepsilon = 5$ | | $\varepsilon = 1$ | |
|---|---|---|---|---|---|---|---|
| | | FID↓ | IS↑ | FID↓ | IS↑ | FID↓ | IS↑ |
| DPDM (FID) [11] | ✗ | 20.9 | 2.0 | 45.8 | 2.1 | 72.5 | 2.1 |
| DP-MEPF [19] | ✓ | 18.0 | **2.5** | 18.9 | 2.4 | 19.7 | **2.6** |
| DPDM (Pub) | ✓ | 8.6 | **2.5** | 8.8 | 2.4 | 10.4 | 2.4 |
| DP-Diffusion [17] | ✓ | 8.5 | 2.4 | 9.5 | **2.6** | 12.2 | 2.6 |
| dp-promise (this work) | ✓ | **6.0** | **2.5** | **6.5** | 2.5 | **9.0** | **2.6** |

| CIFAR-10 | $D_{pub}$ | $\varepsilon = 10$ | | $\varepsilon = 5$ | | $\varepsilon = 1$ | |
|---|---|---|---|---|---|---|---|
| | | FID↓ | IS↑ | FID↓ | IS↑ | FID↓ | IS↑ |
| DPDM (FID) [11] | ✗ | 92.8 | 3.7 | 106.5 | 3.5 | 128.4 | 3.4 |
| DP-MEPF [19] | ✓ | 32.6 | 7.3 | 38.8 | 6.5 | 43.2 | 6.1 |
| DPDM (Pub) | ✓ | 20.9 | 8.4 | 22.7 | 8.3 | 27.6 | 8.2 |
| DP-Diffusion [17] | ✓ | 19.8 | 8.2 | 23.5 | 8.1 | 26.5 | 8.5 |
| dp-promise (this work) | ✓ | **17.9** | **8.6** | **18.9** | **8.7** | **21.8** | **9.1** |

# Higher Resolution & Ablation Studies

| Methods | $\varepsilon = 10$ | | $\varepsilon = 5$ | | $\varepsilon = 1$ | |
|---|---|---|---|---|---|---|
| | FID↓ | IS↑ | FID↓ | IS↑ | FID↓ | IS↑ |
| DPDM (Pub) | 46.5 | 2.0 | 50.2 | 2.1 | 58.3 | 2.5 |
| dp-promise (this work) | **25.3** | **2.5** | **26.2** | **2.6** | **29.1** | **2.7** |

| Methods | MNIST | | | | Fashion-MNIST | | | |
|---|---|---|---|---|---|---|---|---|
| | MLP | CNN | Avg | FID↓ | MLP | CNN | Avg | FID↓ |
| without Phase I | 95.7 | 97.8 | 84.4 | 2.5 | 82.2 | 83.5 | **72.7** | 6.8 |
| with Phase I | **95.8** | **98.1** | **84.8** | **2.3** | **82.4** | **84.9** | 72.5 | **6.5** |

# Outline

- Background & Preliminaries
- Existing Works
- Our Method
- Experimental Evaluation
- **Conclusion**

# Conclusion

- We introduce dp-promise, a new framework for training differentially private diffusion models.
- Our method first leverages the noise in the forward process to reduce information loss in private training.
- We provide DP theoretical analysis.
- Experimental results show dp-promise's effectiveness under practical privacy budgets.

# Thank you for your time!

**Contact**: wanghch@njust.edu.cn

**Code**: https://github.com/deabfc/dp-promise