JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

THE UNIVERSITY OF SYDNEY

UNSW CANBERRA

CSIRO

Private Image Generation with Dual-Purpose Auxiliary Classifier

Chen Chen, Daochang Liu, Siqi Ma, Surya Nepal, Chang Xu

Tag: THU-AM-369

Motivation

- **Privacy-preserving image generation** has been important for segments such as medical domains that have sensitive and limited data.

Research Gap

- The benefits of guaranteed privacy come at **substantial** costs of generated images' quality and utility due to the privacy budget constraints.

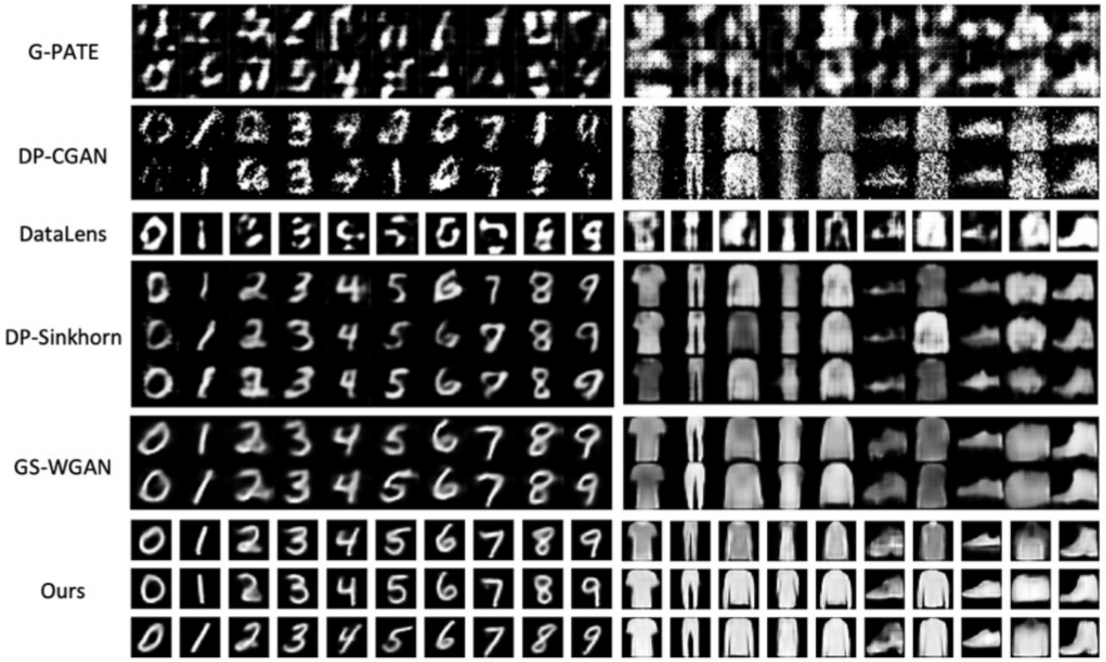


Figure 3. Image generated at privacy budget $\epsilon = 10$ for MNIST (Left) and F-MNIST (Right) by various methods.

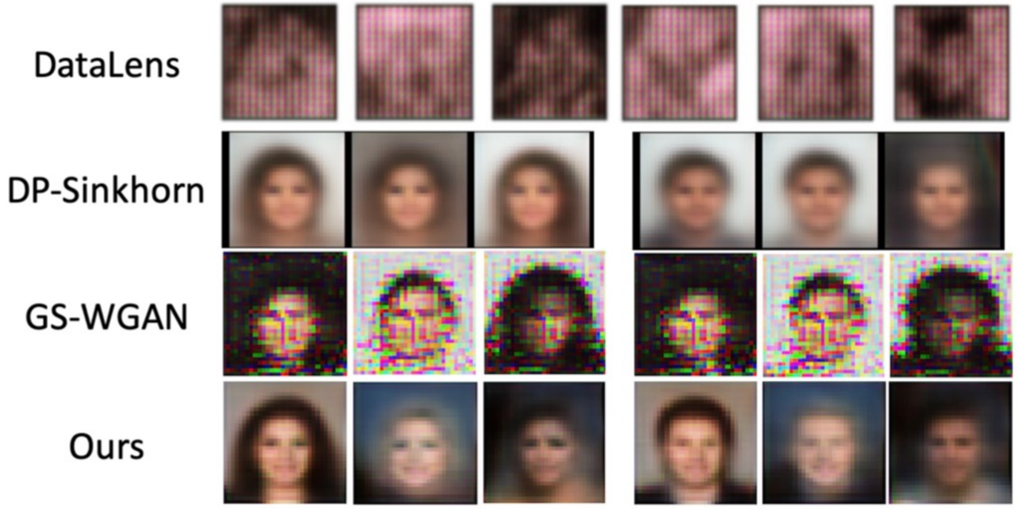


Figure 4. Image generated at privacy budget $\epsilon = 10$ for CelebA by various methods conditioned on gender. Left: Female. Right: Male.

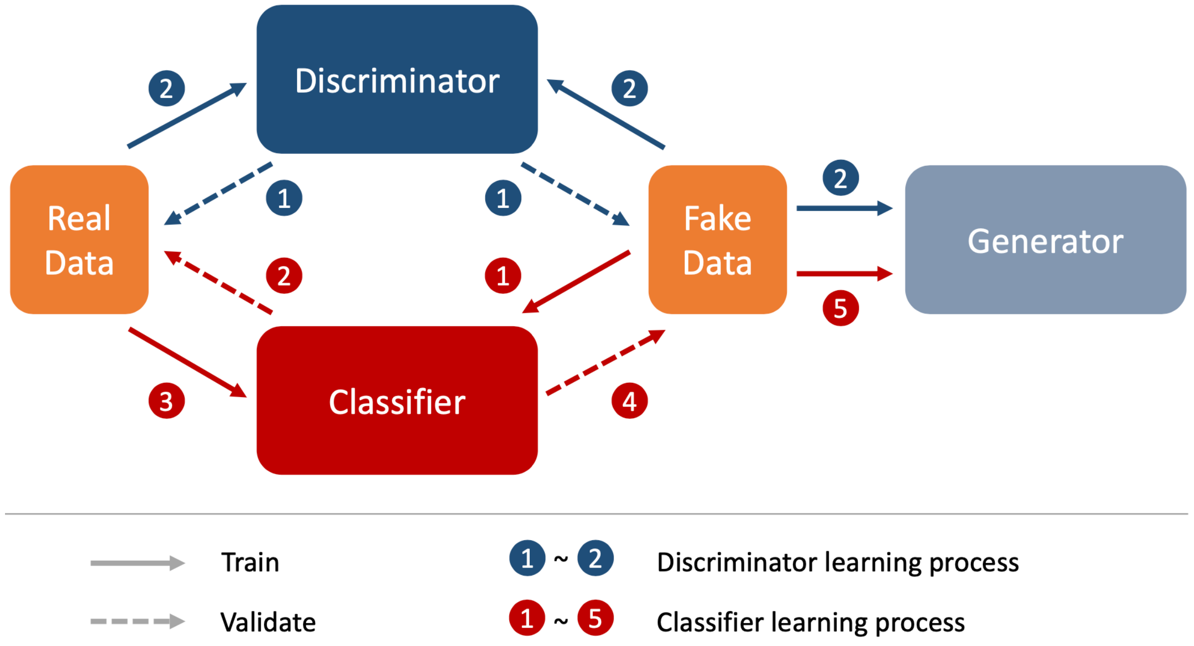
Research Gap

- The commonly used utility metric is the Standard Utility: gen2real accuracy (g2r%), while the Reversed Utility: real2gen accuracy (r2g%) is neglected.

- No work so far has investigated whether incorporating utility measures in the model design would result in better utility performance under the given privacy budget.

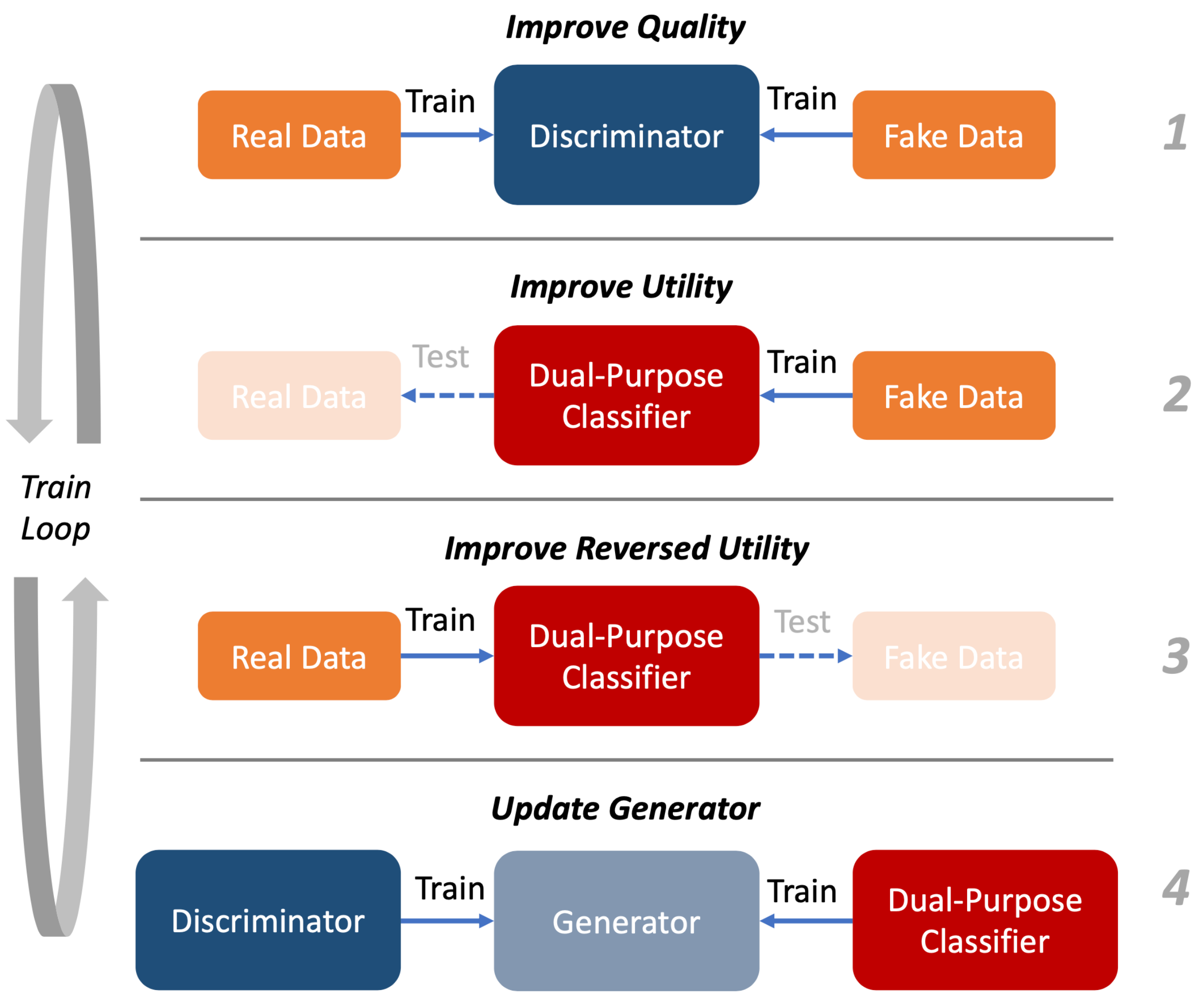
Method: DP-GAN-DPAC

- Incorporate a dual-purpose auxiliary classifier (DPAC) into the training of differentially private GAN (DP-GAN), making it a 3-player game.



Method: DP-GAN-DPAC

- The DPAC alternates between learning from real and fake data **sequentially**, incorporates both g2r% and r2g% in the model design and accelerates Generator convergence.



Results

- Quality (measured by IS and FID)

Method	MNIST		F-MNIST		CelebA	
	IS	FID	IS	FID	IS	FID
PATE-GAN [25]	1.46	253.55	2.35	229.25	-	-
DP-CGAN [36]	-	179.20	-	243.80	-	-
G-PATE [30]	5.16	150.62	4.33	171.90	1.37	350.92
DataLens [37]	5.78	137.50	4.58	167.70	1.42	320.84
DP-MERF [22]	-	121.40	-	110.40	-	-
GS-WGAN [8]	9.23	61.34	5.32	131.34	1.85	297.35
DPSinkhorn [6]	-	55.56	-	129.40	-	168.40
Ours	**9.71**	**54.06**	**6.60**	**90.77**	**1.90**	**139.99**

Table 1. Comparing IS $\uparrow$ and FID $\downarrow$ on various datasets.

Results

- Utility (measured by downstream classification accuracy: gen2real & real2gen)

Method	MNIST MLP	MNIST CNN	F-MNIST MLP	F-MNIST CNN	CelebA MLP	CelebA CNN
DP-CGAN [36]	0.60	0.63	0.50	0.46	-	-
G-PATE [30]	-	0.81	-	0.69	-	0.71
DataLens [37]	-	0.81	-	0.71	-	0.73
DP-MERF [22]	0.81	0.82	0.71	**0.73**	-	-
GS-WGAN [8]	0.79	0.80	0.65	0.65	0.68	0.66
DPSinkhorn [6]	0.80	0.83	0.73	0.71	0.76	0.76
Ours	**0.85**	**0.88**	**0.75**	**0.73**	**0.80**	**0.85**

Table 2. Comparing gen2real accuracy ↑ on various datasets.

Method ↑	MNIST MLP	MNIST CNN	F-MNIST MLP	F-MNIST CNN	CelebA MLP	CelebA CNN
GS-WGAN [8]	0.99	0.99	0.85	0.85	0.66	0.60
Ours	**1.00**	**1.00**	**0.97**	**0.98**	**0.99**	**0.98**

Table 3. Comparing real2gen accuracy ↑ on various datasets.

Motivation

- Machine learning applications have achieved success in many domains.
- However, this might not be the case for domains whose real data is too **rare** or contains **sensitive** information.
- Generative Adversarial Networks (GANs) have been a successful data augmenter and privacy protector since their ability to generate synthetic images that can be difficult to tell from the real ones.
- However, GANs are subject to model inversion attacks and membership inference attacks in both white-box and black-box settings, thus may leak sensitive information about input data.

Motivation

- Recent work: DPGANs have integrated the state-of-the-art (SOTA) privacy protection framework called differential privacy (DP) into GAN training, to provide GAN methods with rigorous privacy guarantee.

- However, the benefits of guaranteed privacy comes with substantial costs of generated images' quality and utility.

Background: DP

- Differential Privacy (DP) is a strong technique for privacy guarantees.
- We define datasets $\mathcal{D}$ and $\mathcal{D}'$ that only differ in one entry as adjacent datasets.
- For a general training algorithm $f(\cdot)$, its $\mathcal{L}_2$ sensitivity on adjacent datasets $\mathcal{D}$ and $\mathcal{D}'$ is
$$\Delta_2 f = max_{\mathcal{D},\mathcal{D}'}||f(\mathcal{D}) - f(\mathcal{D}')||_2$$
- Gaussian sanitization mechanism $\mathcal{M}(\cdot)$ with range $\mathcal{R}$ simply adds Gaussian noise to $f(\cdot)$ based on its sensitivity:
$$\mathcal{M}(\mathcal{D}) = f(\mathcal{D}) + \mathcal{N}(0, (\sigma\Delta_2 f)^2)$$
- To adopt DP in GAN training, we clip the gradient norm of the generator to bound its sensitivity, then correspondingly adding Gaussian noise to be differentially private.

Background: DP

- This allows the mechanism to be (ε, δ)-DP, where the following equation would hold for any subsets of the mechanism's output $\mathcal{S} \subseteq \mathcal{R}$ with δ probability of failing the DP and privacy budget ε.

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

- Privacy accountant computes the privacy cost at each access to the training data and accumulates this cost as the training progresses, acting as a stopping criteria.

Research Gap

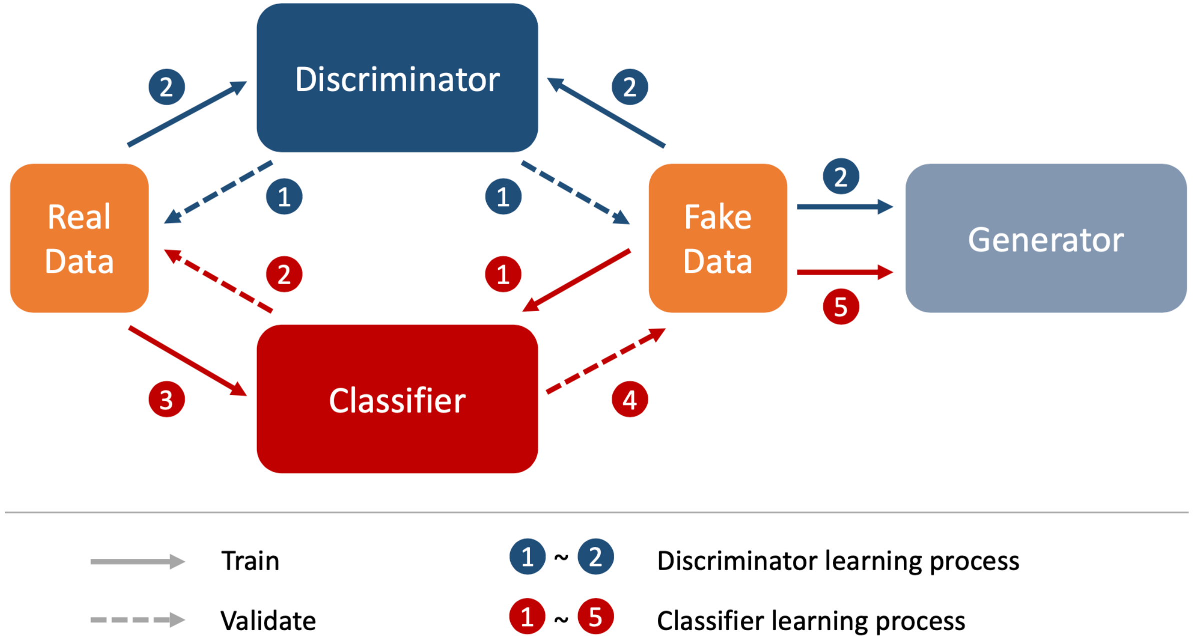
- No work so far has investigated whether incorporating utility measure in the model design would result in better utility performance under the given privacy budget.
  - Can we incorporate utility measure in the model design? Yes!
  - How? By adding an auxiliary classifier network to common GAN architecture (i.e., changing GAN from a 2-player game to a 3-player game).

- The commonly used utility metric is the Standard Utility: gen2real accuracy (g2r%), while the Reversed Utility: real2gen accuracy (r2g%) is neglected.
  - Why r2g% matters? It evaluates the outputs' generalizability.
  - Can we incorporate it in the model design as well? Yes!

- The gained privacy largely sacrifices output quality and utility.
  - Can we do better? Yes, use sequential training strategy for faster convergence!

Method: DP-GAN-DPAC

- Adds a dual-purpose auxiliary classifier (DPAC) into the training of differentially private GAN (DP-GAN), making it a 3-player game.



- When using fake data to train C:

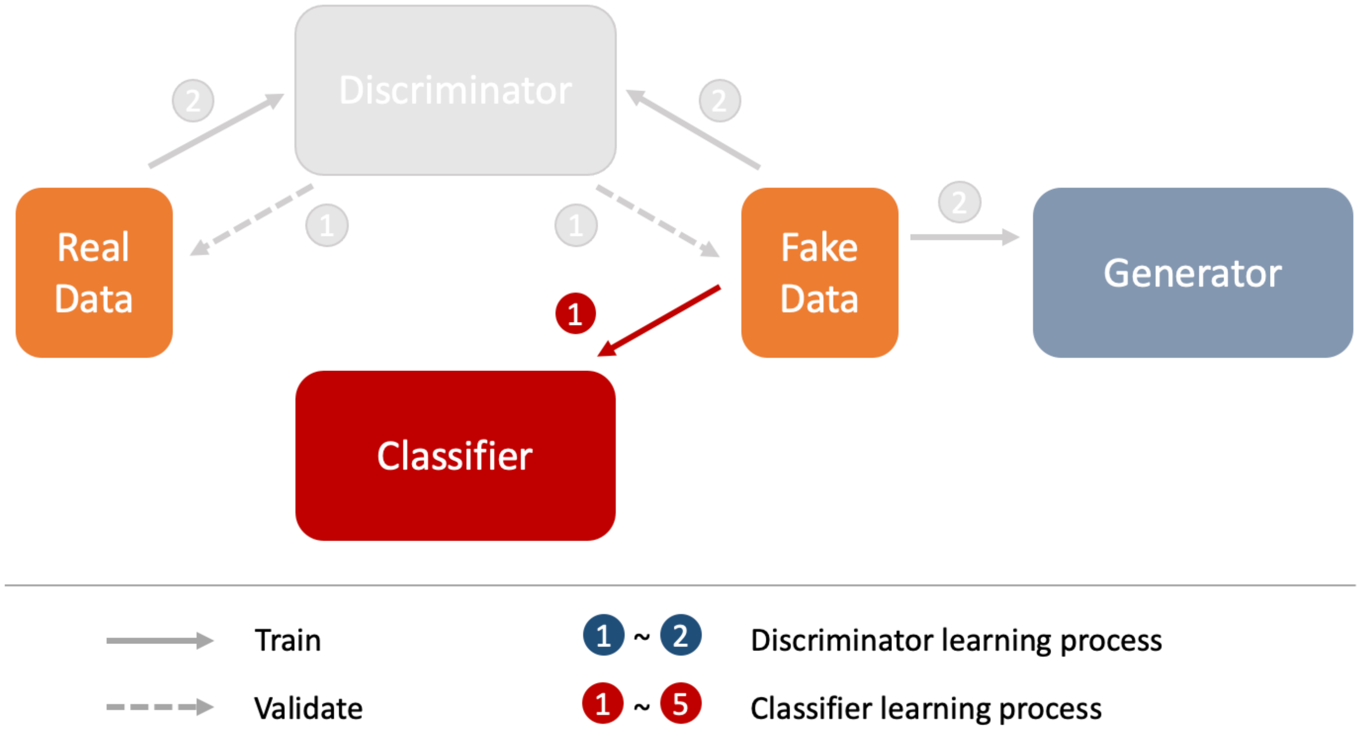
$$\min_G \max_D \min_C V(G, D, C) = -\beta \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{P}_{\tilde{\boldsymbol{x}}}}[D(G(\boldsymbol{z}, \boldsymbol{y}))] + \beta \mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}_{\boldsymbol{x}}}[D(\boldsymbol{x})] - (1-\beta)\mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{P}_{\tilde{\boldsymbol{x}}}}[C(G(\boldsymbol{z}, \boldsymbol{y}), \boldsymbol{y})] \quad (4)$$

- When using real data to train C:

$$\min_G \max_D \min_C V(G, D, C) = -\beta \mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{P}_{\tilde{\boldsymbol{x}}}}[D(G(\boldsymbol{z}, \boldsymbol{y}))] + \beta \mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}_{\boldsymbol{x}}}[D(\boldsymbol{x})] - (1-\beta)\mathbb{E}_{\tilde{\boldsymbol{x}} \sim \mathcal{P}_{\tilde{\boldsymbol{x}}}}[C'(G(\boldsymbol{z}, \boldsymbol{y}), \boldsymbol{y})] - \mathbb{E}_{\boldsymbol{x} \sim \mathcal{P}_{\boldsymbol{x}}}[C(\boldsymbol{x}, \boldsymbol{y})] \quad (10)$$

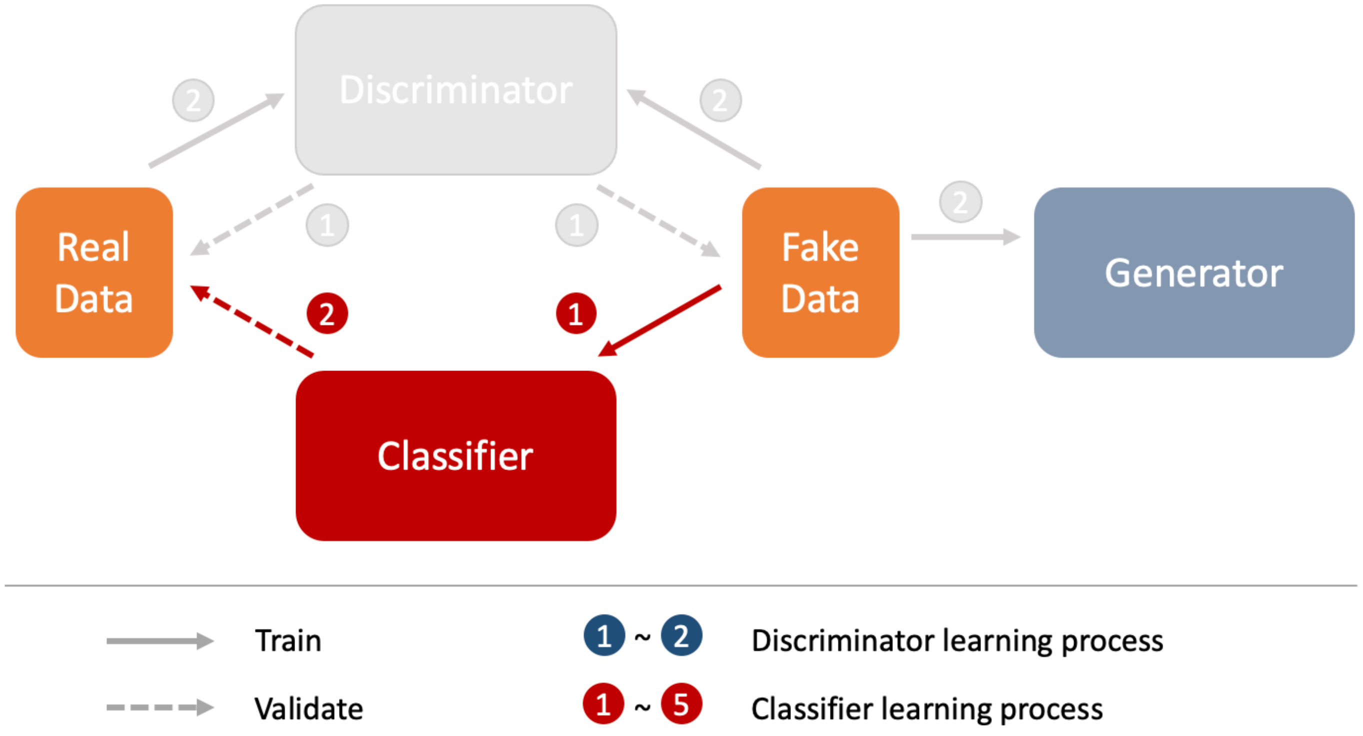
Method: DP-GAN-DPAC

- Step 1:



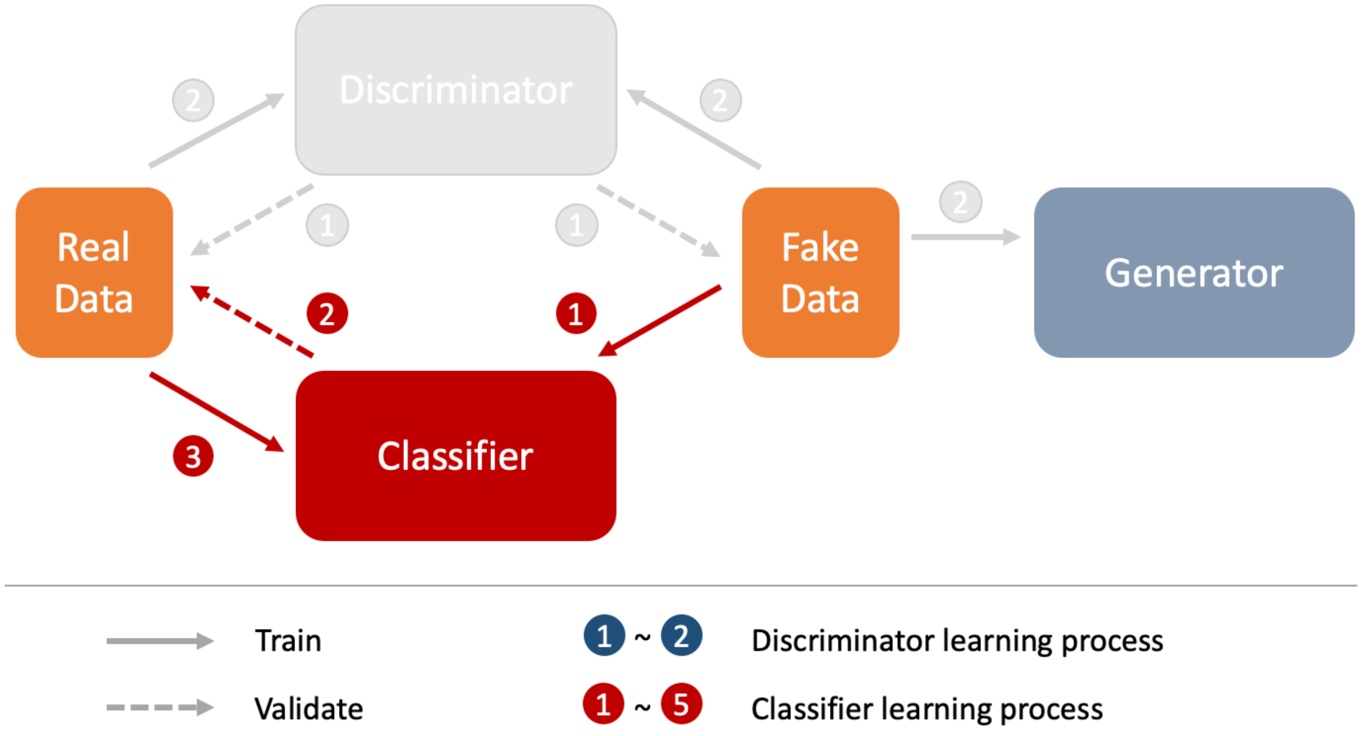
Method: DP-GAN-DPAC

- Step 2:



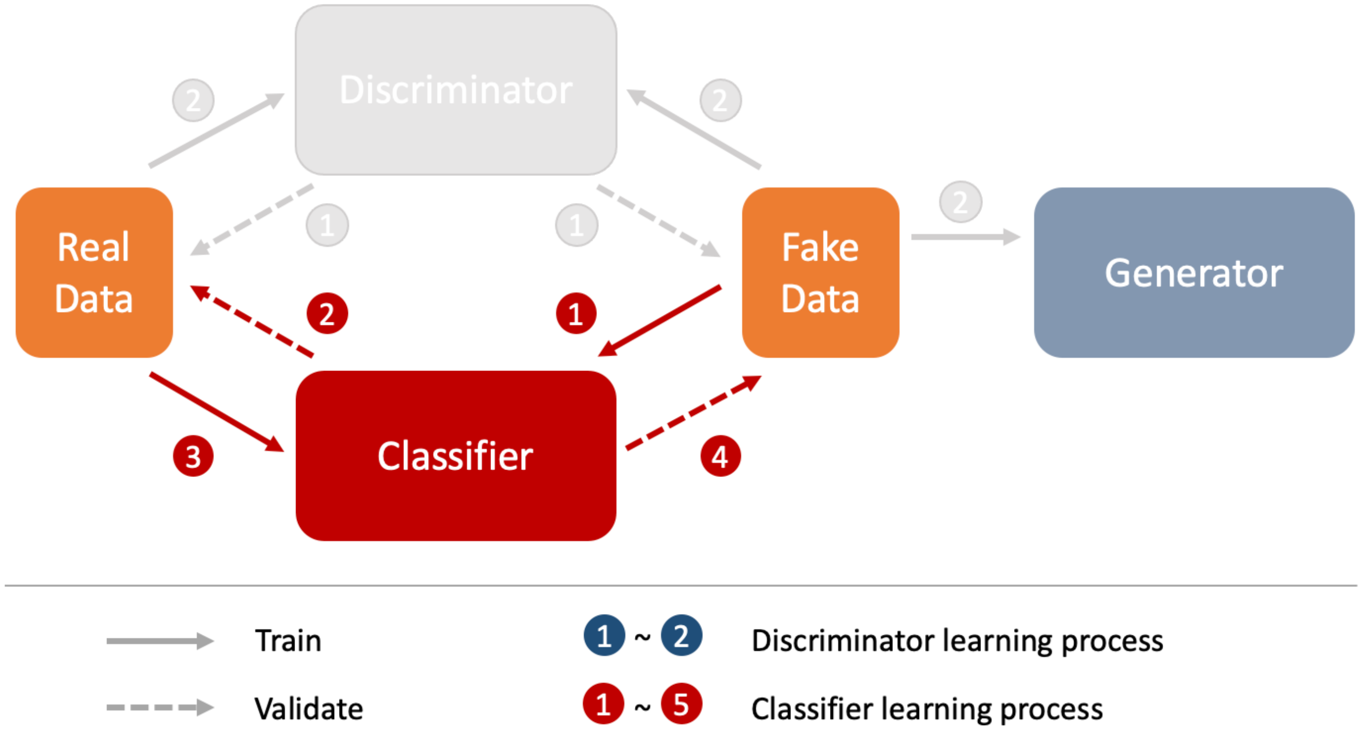
Method: DP-GAN-DPAC

- Step 3:



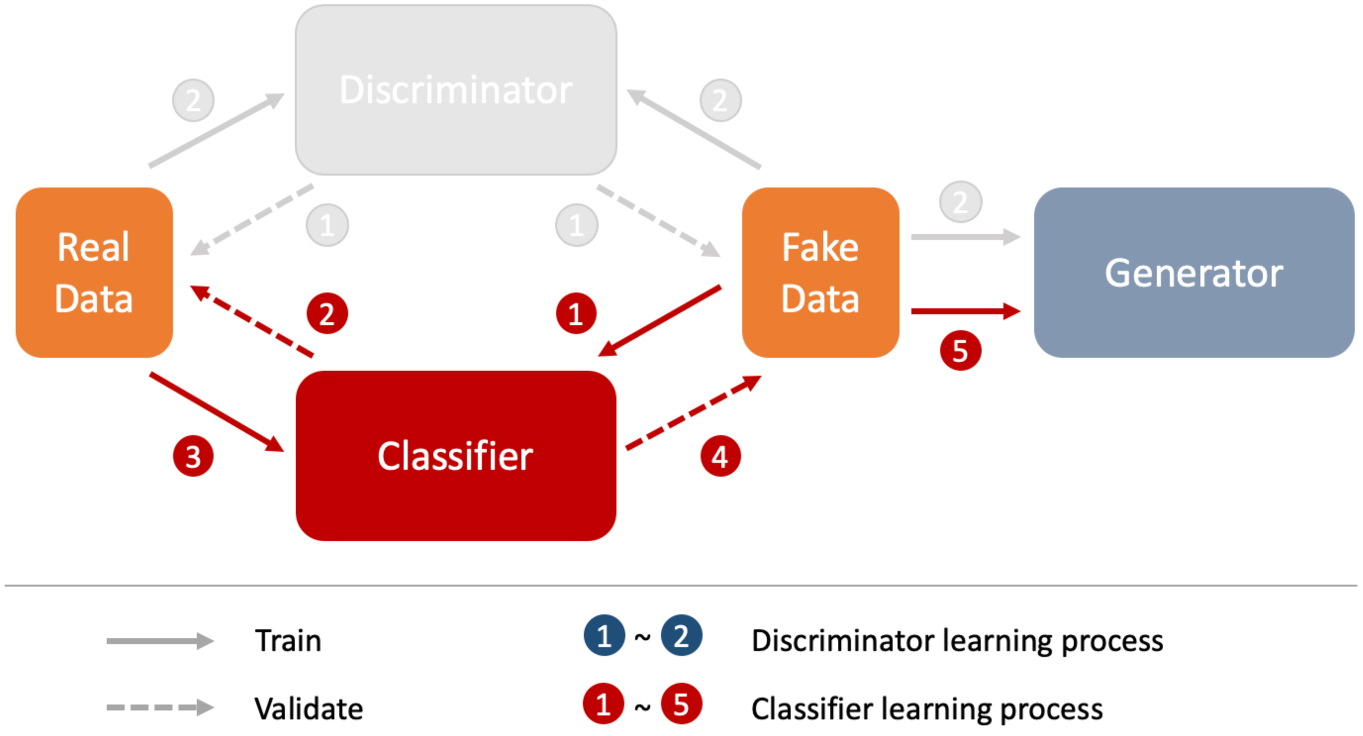
Method: DP-GAN-DPAC

- Step 4:



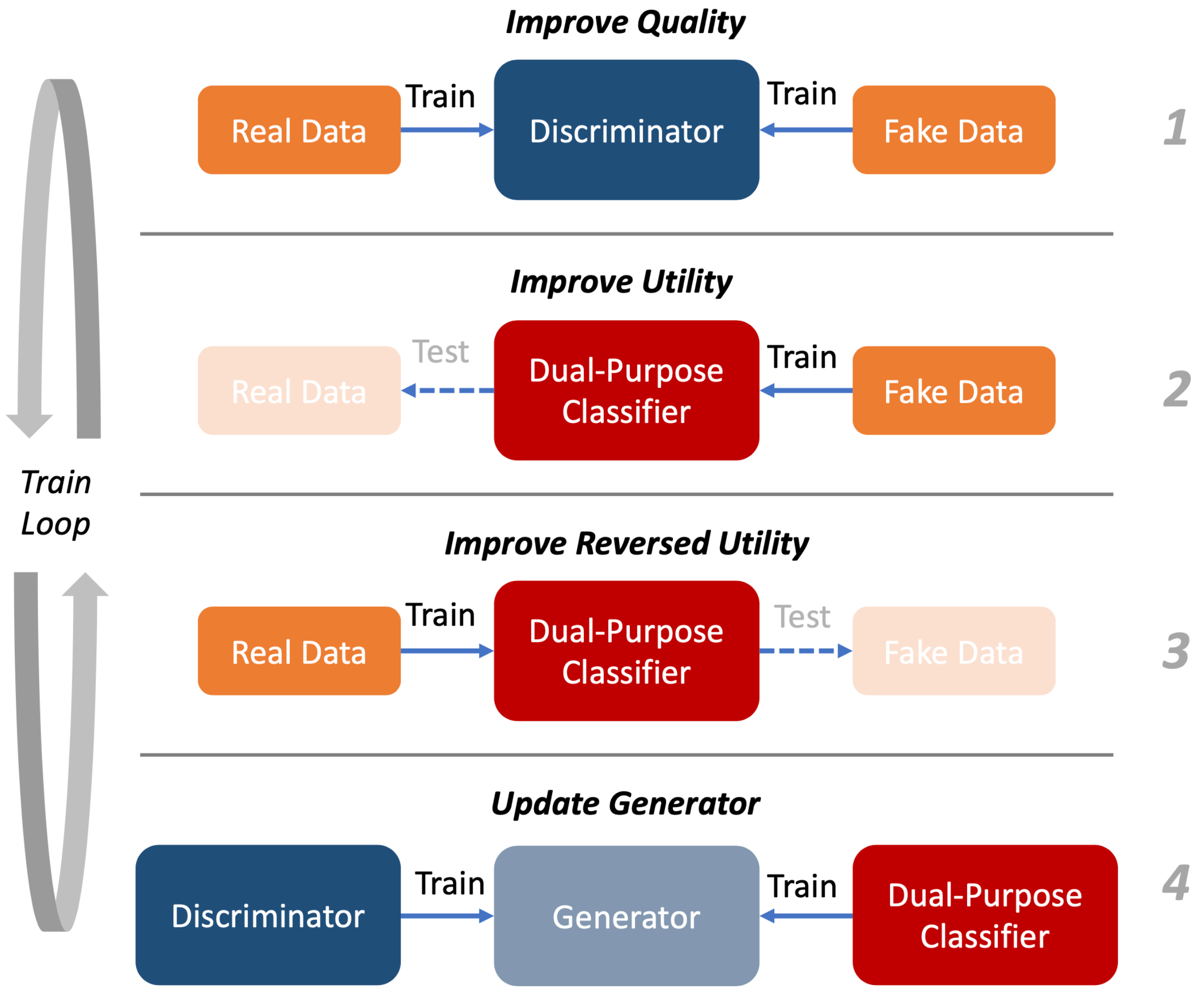
Method: DP-GAN-DPAC

- Step 5:



Method: DP-GAN-DPAC

- The DPAC alternates between learning from real and fake data **sequentially**, incorporates both g2r% and r2g% in the model design and accelerates Generator convergence.



Results

- Quality (measured by IS and FID)

Method	MNIST		F-MNIST		CelebA	
	IS	FID	IS	FID	IS	FID
PATE-GAN [25]	1.46	253.55	2.35	229.25	-	-
DP-CGAN [36]	-	179.20	-	243.80	-	-
G-PATE [30]	5.16	150.62	4.33	171.90	1.37	350.92
DataLens [37]	5.78	137.50	4.58	167.70	1.42	320.84
DP-MERF [22]	-	121.40	-	110.40	-	-
GS-WGAN [8]	9.23	61.34	5.32	131.34	1.85	297.35
DPSinkhorn [6]	-	55.56	-	129.40	-	168.40
Ours	**9.71**	**54.06**	**6.60**	**90.77**	**1.90**	**139.99**

Table 1. Comparing IS $\uparrow$ and FID $\downarrow$ on various datasets.

Results

- Utility (measured by downstream classification accuracy: gen2real & real2gen)

Method	MNIST MLP	MNIST CNN	F-MNIST MLP	F-MNIST CNN	CelebA MLP	CelebA CNN
DP-CGAN [36]	0.60	0.63	0.50	0.46	-	-
G-PATE [30]	-	0.81	-	0.69	-	0.71
DataLens [37]	-	0.81	-	0.71	-	0.73
DP-MERF [22]	0.81	0.82	0.71	**0.73**	-	-
GS-WGAN [8]	0.79	0.80	0.65	0.65	0.68	0.66
DPSinkhorn [6]	0.80	0.83	0.73	0.71	0.76	0.76
Ours	**0.85**	**0.88**	**0.75**	**0.73**	**0.80**	**0.85**

Table 2. Comparing gen2real accuracy ↑ on various datasets.

Method ↑	MNIST MLP	MNIST CNN	F-MNIST MLP	F-MNIST CNN	CelebA MLP	CelebA CNN
GS-WGAN [8]	0.99	0.99	0.85	0.85	0.66	0.60
Ours	**1.00**	**1.00**	**0.97**	**0.98**	**0.99**	**0.98**

Table 3. Comparing real2gen accuracy ↑ on various datasets.

Results

- Ablation studies

Method	IS ↑	FID ↓	gen2real ↑ MLP	gen2real ↑ CNN	real2gen ↑ MLP	real2gen ↑ CNN
Baseline	5.32	131.24	0.65	0.65	0.85	0.85
w/o g2r	6.33	88.17	0.73	0.68	0.94	0.95
w/o r2g	6.47	**86.91**	0.74	0.71	0.92	0.92
w/o seq	4.91	128.25	0.65	0.64	0.88	0.77
w/o init	6.56	101.69	0.72	0.65	**0.97**	0.95
Full	**6.60**	90.77	**0.75**	**0.73**	**0.97**	**0.98**

Table 4. Ablation studies.

Conclusions

- The "reversed utility" is identified as a beneficial part of an improved design of private GANs.
- A dual-purpose auxiliary classifier is developed in alignment with both the standard and reversed utility.
- The classifier is trained with strategies like sequentialization to accelerate the convergence of generator.