

## 1 A Details on our Aug\_PE\_Image\_Voting algorithm

2 We present the pseudo-code for the Aug\_PE\_Image\_Voting algorithm. While it may appear that the  
 3 generated private text  $\mathcal{T}$  is unused in Algorithm 1, this is by design. Image voting serves solely as a  
 4 quality enhancement technique within our pipeline. In contrast, the original text voting mechanism  
 5 explicitly requires the generated private text. To maintain the completeness and integrity of the overall  
 6 pipeline, we retain the captioning process.

---

### Algorithm 1 Aug-PE with Image Voting (Aug\_PE\_Image\_Voting)

---

**Require:** Private image dataset  $\mathcal{D}$ , generated private text  $\mathcal{T}$ , population size  $N$ , number of iterations  $G$ ,  $\text{RANDOM\_API}(\cdot)$ ,  $\text{VARIATION\_API}(\cdot)$ , diffusion model  $\text{API Diffusion}(\cdot)$ , image encoder  $\text{EncodeImage}(\cdot)$ , DP noise multiplier  $\sigma, K, L$   
**Ensure:** Differentially private synthetic text data  $\mathcal{T}'$

```

1:  $\mathcal{C}_0 = \text{RANDOM\_API}(N)$ 
2:  $\mathcal{E}_{\text{priv}} = \text{EncodeImage}(\mathcal{D})$ 
3: for  $g = 0$  to  $G - 1$  do
4:   if  $K == 0$  then
5:      $\mathcal{I}_g = \text{Diffusion}(\mathcal{C}_g)$ 
6:      $\mathcal{E}_{\text{gen}} = \text{EncodeImage}(\mathcal{I}_g)$ 
7:   else if  $K > 0$  then
8:      $\mathcal{C}_g^k = \text{VARIATION\_API}(N), k = 1, 2, \dots, K$ 
9:      $\mathcal{I}_g^k = \text{Diffusion}(\mathcal{C}_g^k), k = 1, 2, \dots, K$ 
10:     $\mathcal{E}_{\text{gen}}^k = \text{EncodeImage}(\mathcal{I}_g^k), k = 1, 2, \dots, K$ 
11:     $\mathcal{E}_{\text{gen}} = \frac{1}{K} \sum_{k=1}^K \Phi(\mathcal{E}_{\text{gen}}^k)$ 
12:   end if
13:    $H = \mathbf{0}^N = [0, 0, \dots, 0]$ 
14:   for each  $e_{\text{priv}} \in \mathcal{E}_{\text{priv}}$  do
15:      $i \leftarrow \arg \min_j d(e_{\text{priv}}, \mathcal{E}_{\text{gen}}[j])$ 
16:      $H[i] \leftarrow H[i] + 1$ 
17:   end for
18:    $H \leftarrow H + \mathcal{N}(0, \sigma^2 \mathbf{I})$ 
19:    $P \leftarrow H / \sum H$ 
20:   if  $L == 1$  then
21:      $\mathcal{C}'_g \leftarrow$  draw  $N$  samples with replacement from  $\mathcal{C}_g$ 
22:      $\mathcal{C}_{g+1} \leftarrow \text{VARIATION\_API}(\mathcal{C}'_g, N)$ 
23:     save  $\mathcal{C}_{g+1}^{\text{syn}} \leftarrow \mathcal{C}'_g$ 
24:   else if  $L > 1$  then
25:      $\mathcal{C}'_g \leftarrow$  rank samples by probabilities  $P$  and draw top  $N$  samples
26:      $\mathcal{C}_{g+1}^k \leftarrow \text{VARIATION\_API}(\mathcal{C}'_g, N), k = 1, 2, \dots, K - 1$ 
27:      $\mathcal{C}_{g+1} = [\mathcal{C}'_g, \mathcal{C}_{g+1}^1, \dots, \mathcal{C}_{g+1}^{K-1}]$ 
28:     save  $\mathcal{C}_{g+1}^{\text{syn}} \leftarrow \mathcal{C}_{g+1}$ 
29:   end if
30: end for
31: return Final text candidates  $\mathcal{C}_G$  as synthetic data  $\mathcal{T}'$ 

```

---

## 7 B More on our experiment results

8 Here we post more results of our experiments, and provide more experiment details.

9 **B.1 FID comparison between PE and SPTI**

10

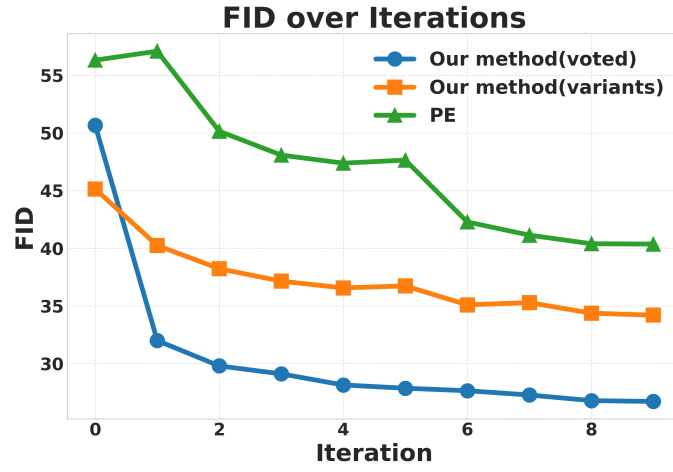


Figure 1: Experiment results on **LSUN Bedroom** dataset. ( $\epsilon = 1.0$ )

11

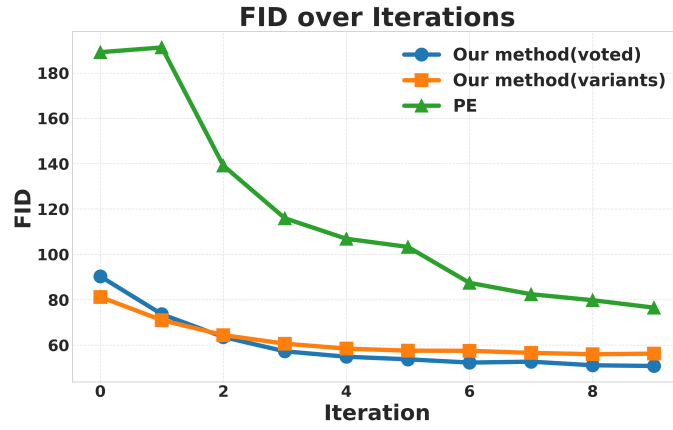


Figure 2: Experiment results on **European Art** dataset. ( $\epsilon = 1.0$ )

12

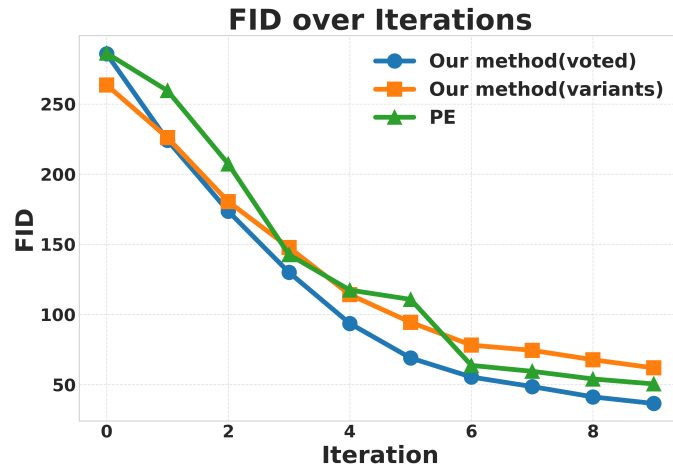
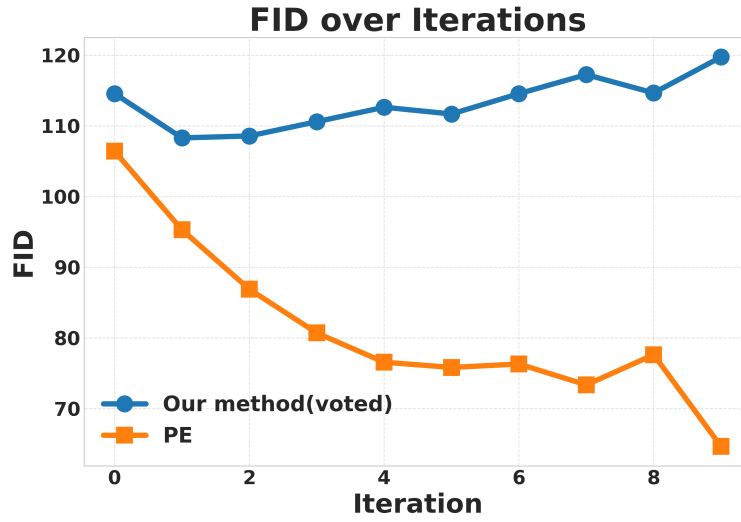
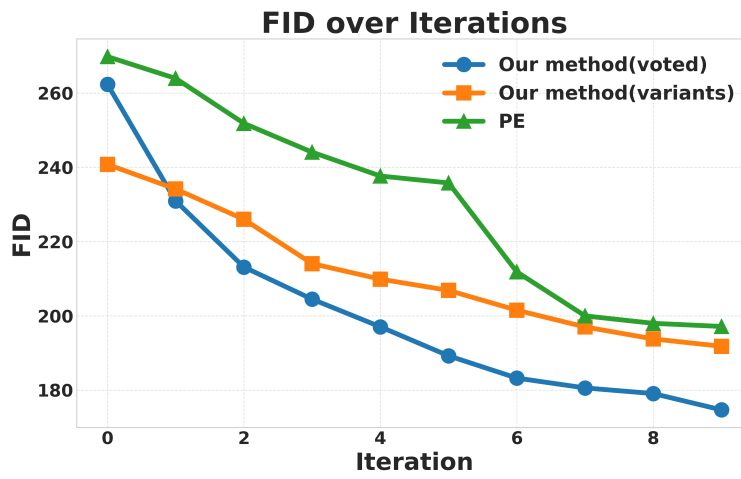


Figure 3: Experiment results on **Wave-ui-25k** dataset. ( $\epsilon = 1.0$ )

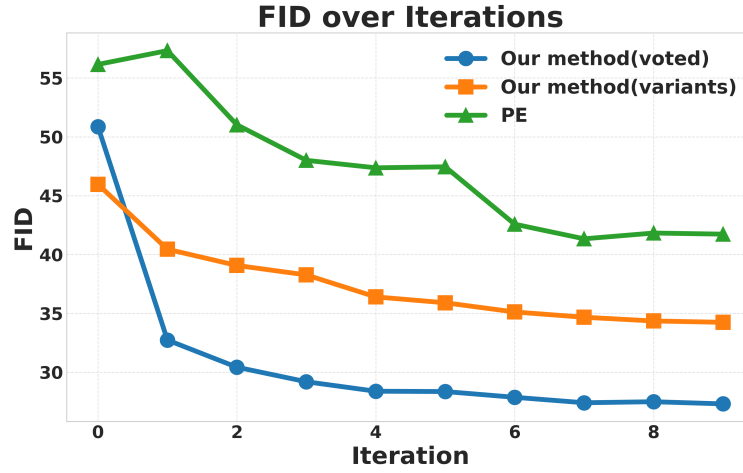
13

Figure 4: Experiment results on **Cat** dataset. ( $\epsilon = 1.0$ )

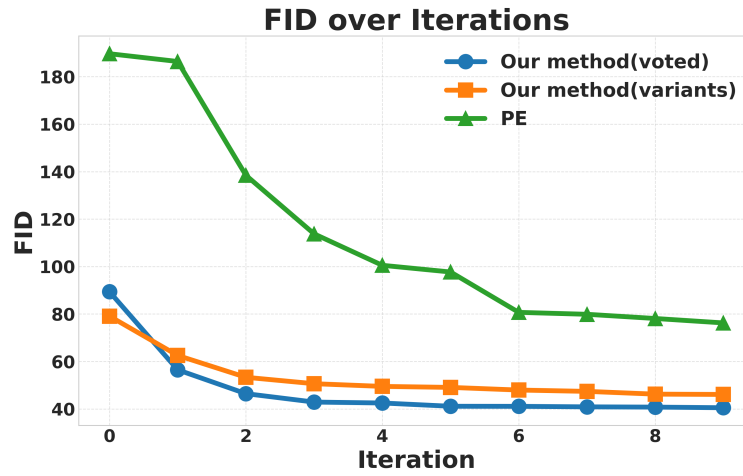
14

Figure 5: Experiment results on **Sprite Fright** dataset. ( $\epsilon = 1.0$ )

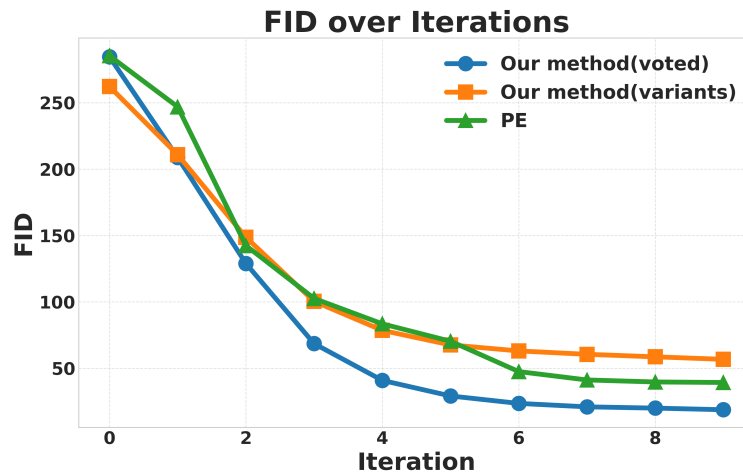
15

Figure 6: Experiment results on **LSUN Bedroom** dataset. ( $\epsilon = 10.0$ )

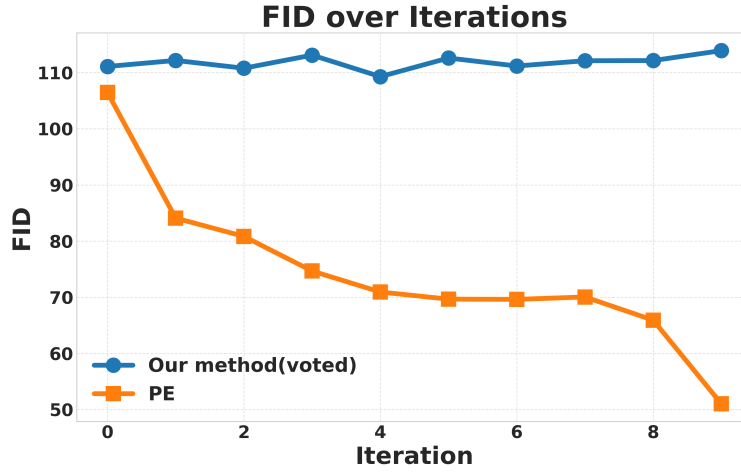
16

Figure 7: Experiment results on **European Art** dataset. ( $\epsilon = 10.0$ )

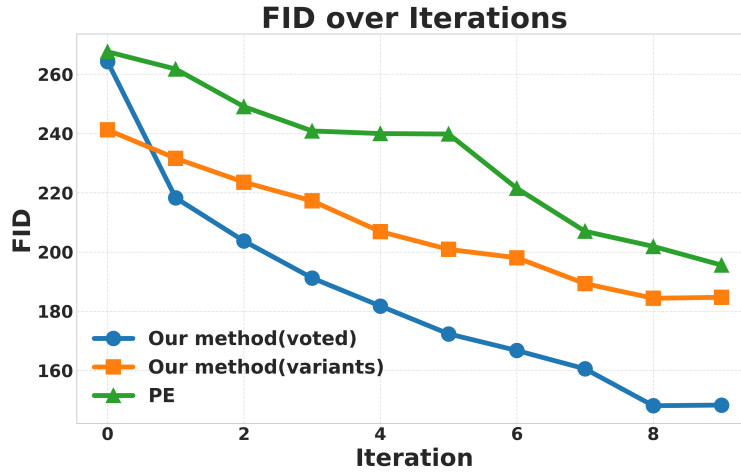
17

Figure 8: Experiment results on **Wave-ui-25k** dataset. ( $\epsilon = 10.0$ )

18

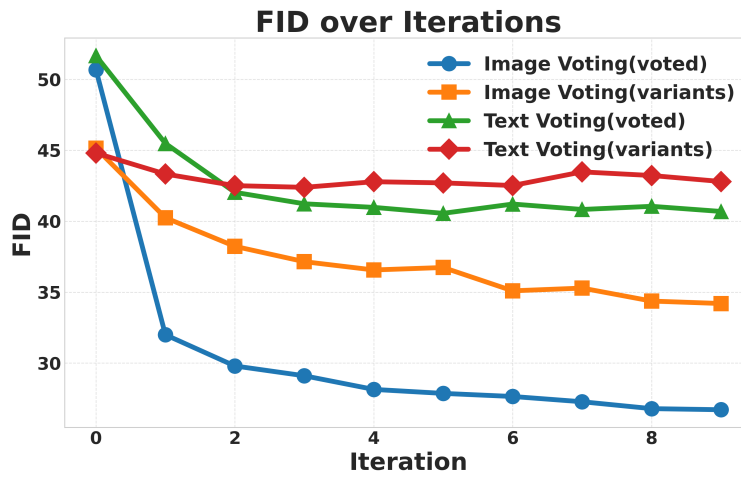
Figure 9: Experiment results on **Cat** dataset. ( $\epsilon = 10.0$ )

19

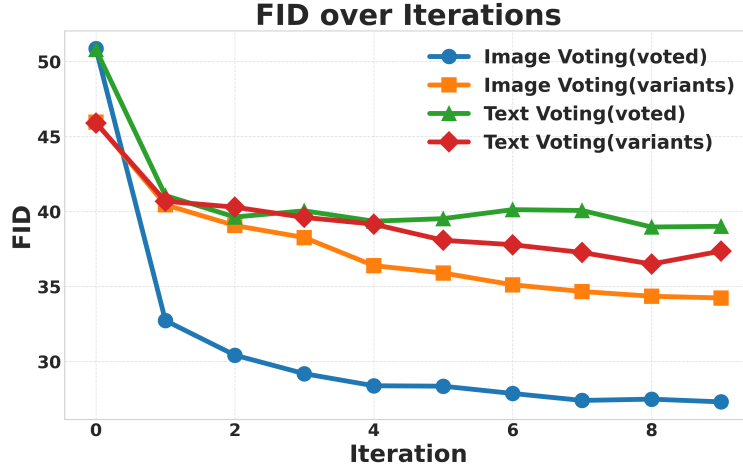
Figure 10: Experiment results on **Sprite Fright** dataset. ( $\epsilon = 10.0$ )

## 20 B.2 FID comparison between Image Voting and Text Voting

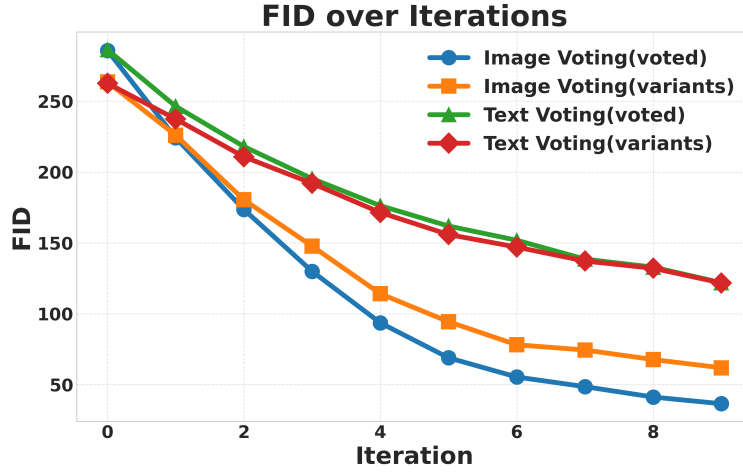
21

Figure 11: Experiment results on **LSUN Bedroom** dataset. ( $\epsilon = 1.0$ )

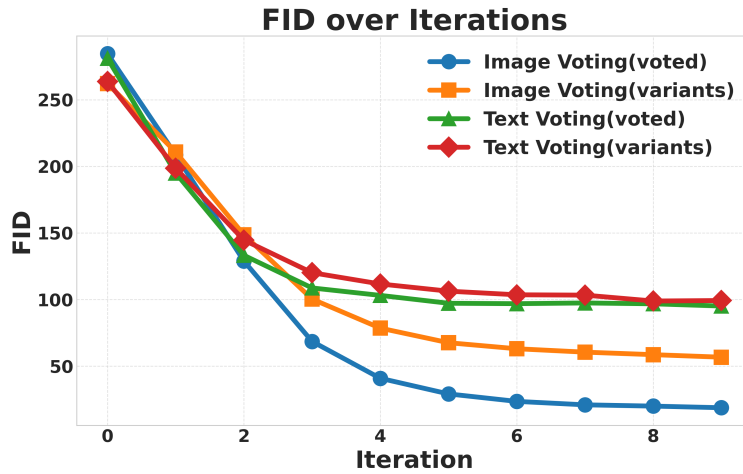
22

Figure 12: Experiment results on **LSUN Bedroom** dataset. ( $\epsilon = 10.0$ )

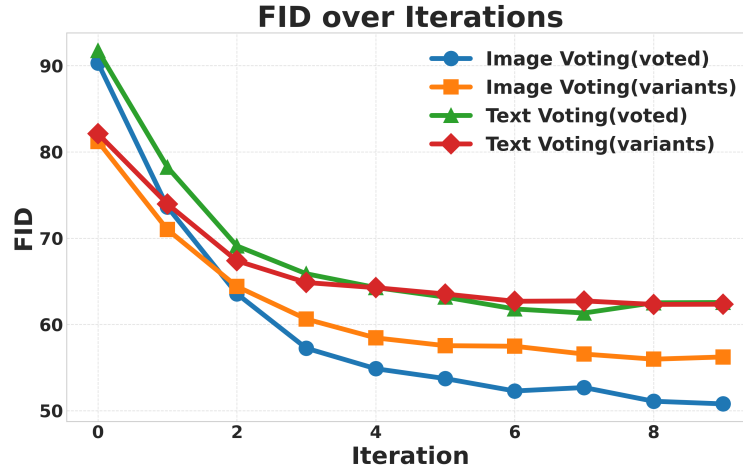
23

Figure 13: Experiment results on **Wave-ui-25k** dataset. ( $\epsilon = 1.0$ )

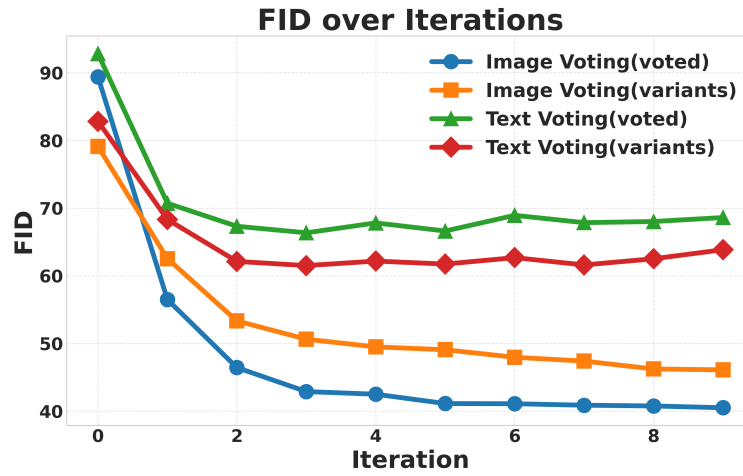
24

Figure 14: Experiment results on **Wave-ui-25k** dataset. ( $\epsilon = 10.0$ )

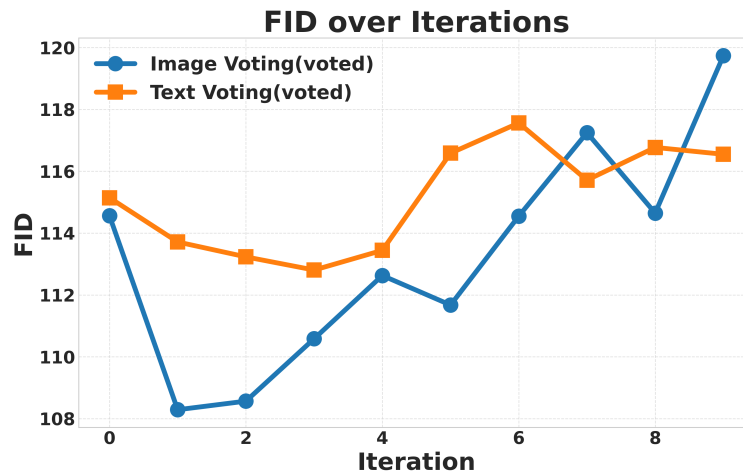
25

Figure 15: Experiment results on **European art** dataset. ( $\epsilon = 1.0$ )

26

Figure 16: Experiment results on **European art** dataset. ( $\epsilon = 10.0$ )

27

Figure 17: Experiment results on **Cat** dataset. ( $\epsilon = 1.0$ )

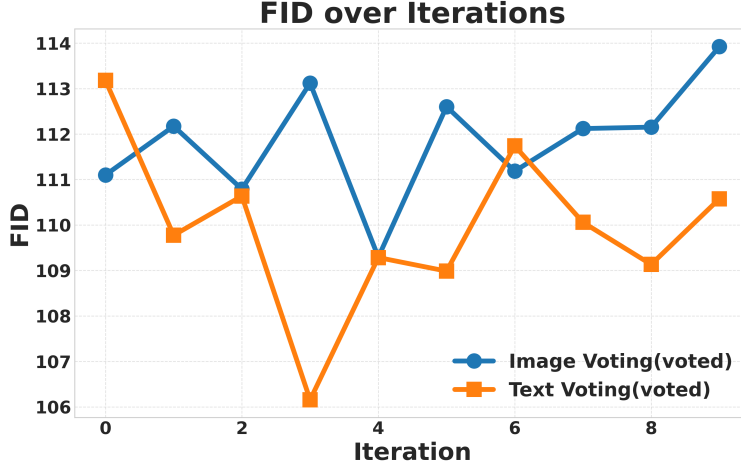


Figure 18: Experiment results on **Cat** dataset. ( $\epsilon = 10.0$ )

## C Extended Ablation Study

### C.1 Large Language Model and Diffusion Model APIs

Along with the effect of our image voting trick, we also tested our method with different LLM and diffusion model APIs along with other hyperparameters. For large language models, we choose **qwen-plus** as comparison to our original **Meta-Llama-3-8B-Instruct** model. For diffusion model APIs, we choose **stable-diffusion-xl-base-1.0** to compare with our original model.

	SDXL-Turbo	SDXL-base-1.0
<b>Meta-Llama-3-8B-Instruct</b>	26.71	25.42
<b>qwen-plus</b>	27.82	25.29

Table 1: **FID** (lower is better) on **LSUN Bedroom** with  $\epsilon = 1.0$ . Tested on different LLM and diffusion APIs. Other experiment settings are the same as default.



Figure 19: Images generated by DALL-E from DP-synthetic text of dataset **LSUN Bedroom** ( $\epsilon = 1.0$ ). FID value is 29.96



Figure 20: Images generated by DALL-E from DP-synthetic text of dataset **Wave-ui-25k** ( $\epsilon = 1.0$ ). FID value is 150.68



## 36 D Details of Experimental Configurations

37 To facilitate reproducibility, we detail the hyperparameter settings used for all experiments in Section  
 38 4. For hyperparameters not mentioned, the default values are used. For experiment settings of ablation  
 39 studies between image voting and text voting, we use the same hyperparameters in Table 2 below.

Dataset	Configurations		<i>SPTI(ours)</i>	PE image
LSUN bedroom	PE.run()	num_samples_schedule	2000	2000
		iterations	10	10
	ImageVotingNN() NNhistogram()	lookahead_degree	0	-
		lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1
Cat	PE.run()	num_samples_schedule	200	200
		iterations	10	10
	ImageVotingNN() NearestNeighbors()	lookahead_degree	8	-
		lookahead_degree	-	8
	PEPopulation()	initial_variation_api_fold	0	0
		next_variation_api_fold	1	1
wave-ui	PE.run()	num_samples_schedule	2000	2000
		iterations	10	10
	ImageVotingNN() NNhistogram()	lookahead_degree	0	-
		lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1
Europeart	PE.run()	num_samples_schedule	2000	2000
		iterations	10	10
	ImageVotingNN() NNhistogram()	lookahead_degree	0	-
		lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1
Sprite Fright	PE.run()	num_samples_schedule	1000	1000
		iterations	10	10
	ImageVotingNN() NNhistogram()	lookahead_degree	0	-
		lookahead_degree	-	4
	PEPopulation()	initial_variation_api_fold	6	0
		next_variation_api_fold	6	1

Table 2: Experiment Settings for **FID** evaluation.

Configurations		<i>SPTI(ours)</i>	PE image
PE.run()	num_samples_schedule	2000	2000
	iterations	10	10
ImageVotingNN()	lookahead_degree	0	-
NNhistogram()	lookahead_degree	-	4
PEPopulation()	initial_variation_api_fold	6	0
	next_variation_api_fold	6	1

Table 3: Experiment Settings of *SPTI* and PE for **FID** evaluation of **MM-Celeba-HQ** dataset

Table 4: Hyperparameters for fine-tuning diffusion models with DP constraints  $\epsilon = 10, 1$  and  $\delta = 10^{-5}$  on text-conditioned CelebAHQ.

	$\epsilon = 10$	$\epsilon = 1$
batch size	256	256
base learning rate	$1 \times 10^{-7}$	$1 \times 10^{-7}$
learning rate	$2.6 \times 10^{-5}$	$2.6 \times 10^{-5}$
epochs	10	10
clipping norm	0.01	0.01
noise scale	0.55	1.46
ablation	-1	-1
num of params	280M	280M
use_spatial_transformer	True	True
cond_stage_key	caption	caption
context_dim	1280	1280
conditioning_key	crossattn	crossattn
transformer depth	1	1