Additional Experiments for ICML Submission 4722: Trustworthy Machine Learning through Data-Specific Indistinguishability

Experiment 1: Copyright/Contribution Control in Finetuning Diffusion Models with DSI Framework

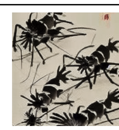
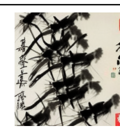
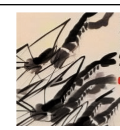
	Original Artwork	Before Tuning	Regular Tuning ($\epsilon = +\infty$)	$\epsilon = 1$	$\epsilon = 2$	$\epsilon = 4$	$\epsilon = 6$	$\epsilon = 8$
Qi.								
Xu.								
Remb.								

Table 1: Finetuing a Stable Diffusion (v1-4) on 425 paintings collected from 10 artists with/out DSI noise on **per-sample contribution** with epoch number selected to be **10**. The original artwork, and the generated images before tuning, after tuning without noise ($\epsilon = \infty$), and after provable-trust tuning with DSI noises and various indistinguishability budget in ϵ, ($\delta = 0.002$) of three selected artists, Baishi Qi (Qi.), Beihong Xu (Xu.) and Rembrandt Harmenszoon van Rijn (Remb.) are presented.

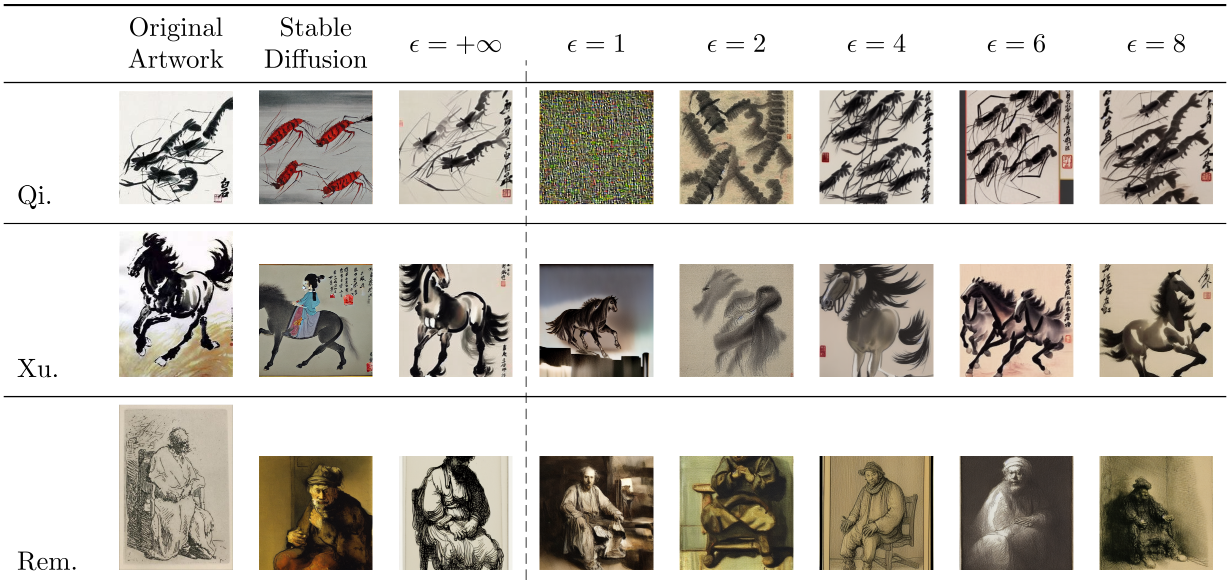


Table 2: Finetuning a Stable Diffusion (v1-4) on 425 paintings collected from 10 artists with/out DSI noise on **per-sample contribution** with epoch number selected to be **20**. The original artwork, and the generated images before tuning, after tuning without noise ($\epsilon = \infty$), and after provable-trust tuning with DSI noises and various indistinguishability budget in ϵ, ($\delta = 0.002$) of three selected artists, Baishi Qi (Qi.), Beihong Xu (Xu.) and Rembrandt Harmenszoon van Rijn (Remb.) are presented.

Experiment 2: Comparison between DSI Local SGD with Indistinguishability Control Methods from DP-SGD

Model	Method \ ϵ	∞	1	2	3	4	5	6	7	8
ResNet-20	[XXWD23]	91.7	55.3	63.1	67.6	72.4	73.7	74.3	75.8	76.0
	[YZCL21]		/	59.7	/	/	70.1	/	/	74.9
	DSI-Local-SGD		57.4	71.2	73.8	78.7	79.8	83.1	83.6	83.9
WideResNet-16	[DBH+22]	94.6	56.8	64.9	69.2	71.9	74.1	77.0	78.8	79.5
	[BPBB23]		57.2	64.6	/	70.5	/	/	/	79.8
	DSI-Local-SGD		63.7	76.1	80.5	82.4	84.4	86.6	86.8	87.1

Table 3: **Test Accuracy** (%) Comparison between standard distinguishability control through per-sample gradient clipping and isotropic noise in DP-SGD [XXWD23, DBH+22] and the augmented versions with additional public data – [YZCL21] projects per-sample gradients into a 2000-rank subspace estimated by public gradients and [BPBB23] conducts mixup between every datapoint and synthetic data – and DSI-Local-SGD with $\mathcal{O}$ being 20-local-GD-iteration with R_i as a leaving-one subset of CIFAR-10 training data U from scratch across different ϵ selections with $\delta = 10^{-5}$.

Experiment 3: Comparison between DSI Noise and Isotropic Noise in Defending Backdoor Attacks

Table 4: Comparison on indistinguishability control and defense efficiency in Adversarial Success Rate (ASR) against **Low-Frequency Attacks** [ZPMJ21] between **DSI Noise** and **Isotropic Noise** with **fixed** test accuracy on clean data.

(a): $m = 10$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
75.6	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	**13.9**
	$(\epsilon = 2019, \delta = 10^{-5})$ (Iso.)	25.6
65.5	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	**2.0**
	$(\epsilon = 575, \delta = 10^{-5})$ (Iso.)	20.2

(b): $m = 20$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
79.1	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	**8.4**
	$(\epsilon = 1661, \delta = 10^{-5})$ (Iso.)	16.2
70.9	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	**5.3**
	$(\epsilon = 479, \delta = 10^{-5})$ (Iso.)	11.5

(c): $m = 40$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
78.4	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	6.4
	$(\epsilon = 627, \delta = 10^{-5})$ (Iso.)	**7.3**
73.6	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	**6.6**
	$(\epsilon = 312, \delta = 10^{-5})$ (Iso.)	7.2

(d): $m = 80$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
79.5	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	8.9
	$(\epsilon = 350, \delta = 10^{-5})$ (Iso.)	**8.2**
74.1	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	**6.1**
	$(\epsilon = 114, \delta = 10^{-5})$ (Iso.)	7.7

Table 5: Comparison on indistinguishability control and defense efficiency in Adversarial Success Rate (ASR) against **Blended Attacks** [CLL+17] between **DSI Noise** and **Isotropic Noise** with **fixed** test accuracy on clean data.

(a): $m = 10$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
73.9	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	15.7
	$(\epsilon = 2019, \delta = 10^{-5})$ (Iso.)	**9.9**
57.5	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	**0.1**
	$(\epsilon = 350, \delta = 10^{-5})$ (Iso.)	9.8

(b): $m = 20$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
77.9	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	**7.7**
	$(\epsilon = 1661, \delta = 10^{-5})$ (Iso.)	8.8
67.1	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	**1.1**
	$(\epsilon = 241, \delta = 10^{-5})$ (Iso.)	2.4

(c): $m = 40$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
78.8	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	7.1
	$(\epsilon = 627, \delta = 10^{-5})$ (Iso.)	**4.1**
73.5	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	12.3
	$(\epsilon = 312, \delta = 10^{-5})$ (Iso.)	**5.1**

(d): $m = 80$ Sources

Test ACC (%)	Ind. Guarantee	ASR (%)
79.3	$(\epsilon = 8, \delta = 10^{-5})$ (DSI)	**3.4**
	$(\epsilon = 350, \delta = 10^{-5})$ (Iso.)	3.9
73.7	$(\epsilon = 4, \delta = 10^{-5})$ (DSI)	4.7
	$(\epsilon = 179, \delta = 10^{-5})$ (Iso.)	**3.6**

References

[BPBB23] Wenxuan Bao, Francesco Pittaluga, Vijay Kumar BG, and Vincent Bindschaedler. Dp-mix: mixup-based data augmentation for differentially private learning. *Advances in Neural Information Processing Systems*, 36:12154–12170, 2023.

[CLL+17] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[DBH+22] Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

[XXWD23] Hanshen Xiao, Zihang Xiang, Di Wang, and Srinivas Devadas. A theory to instruct differentially-private learning via clipping bias reduction. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2170–2189. IEEE, 2023.

[YZCL21] Da Yu, Huishuai Zhang, Wei Chen, and Tie-Yan Liu. Do not let privacy overbill utility: Gradient embedding perturbation for private learning. In *International Conference on Learning Representations*, 2021.

[ZPMJ21] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16473–16481, 2021.