NEURAL INFORMATION PROCESSING SYSTEMS

Membership Inference on Text-to-Image Diffusion Models via Conditional Likelihood Discrepancy

Shengfang Zhai, Huanran Chen, Yinpeng Dong[✉], Jiajun Li,

Qingni Shen[✉], Yansong Gao, Hang Su, Yang Liu

Content

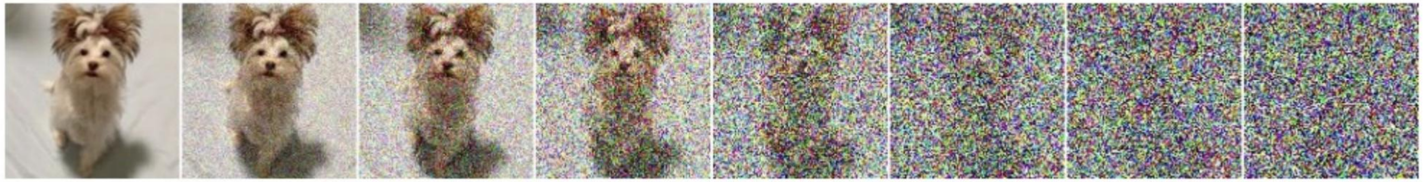
- Motivation: Why we need ***Membership Inference (MI)*** *for T2I models?*

+ Background

- Key intuition: Conditional Overfitting
- Methods
- Experiments
- Conclusion

Diffusion models

$$x_0 \rightarrow x_1 \rightarrow x_2 \longrightarrow \ldots \longrightarrow x_N \approx N(0, I)$$



➢ Training objective is simple (MSE loss, **E**vidence **L**ower **Bo**und)

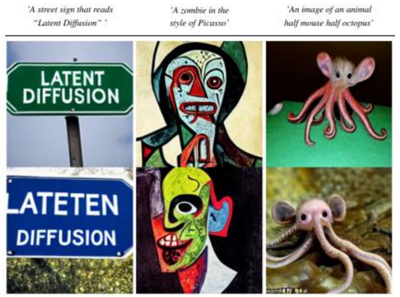
Unconditional Diffusion Models (DDPM):

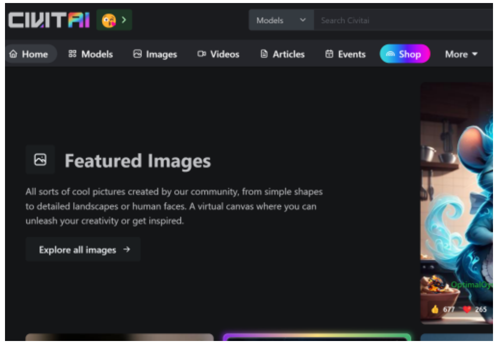
$$\log p_\theta(\mathbf{x}_0) \geq \mathbb{E}_{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}\left[\log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)}\right] = -\mathbb{E}_{\epsilon,t}\left[||\epsilon_\theta(\mathbf{x}_t, t) - \epsilon||^2\right] + C,$$

Conditional Diffusion Models (T2I Models):

$$\log p_\theta(\mathbf{x}_0|\mathbf{c}) \geq -\mathbb{E}_{\epsilon,t}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right] + C.$$

Simple but effective.

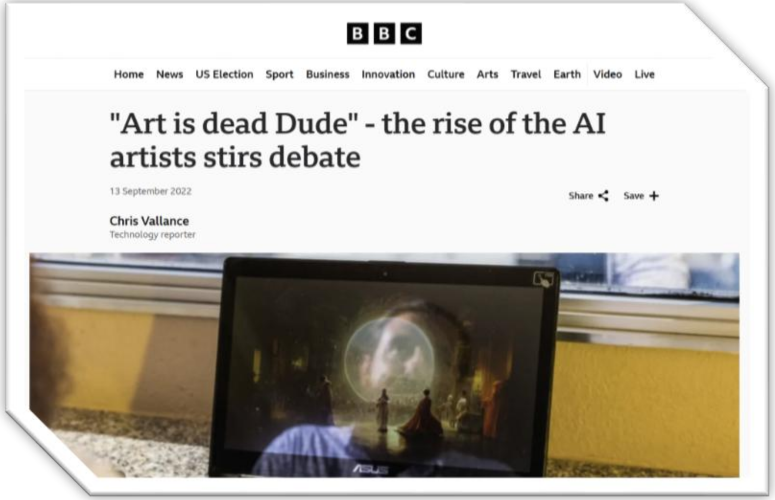




Why we need *Membership Inference* on Text-to-image diffusion models?

- Unauthorized data usage auditing:

Issues about copyright infringement [1,2,3…]

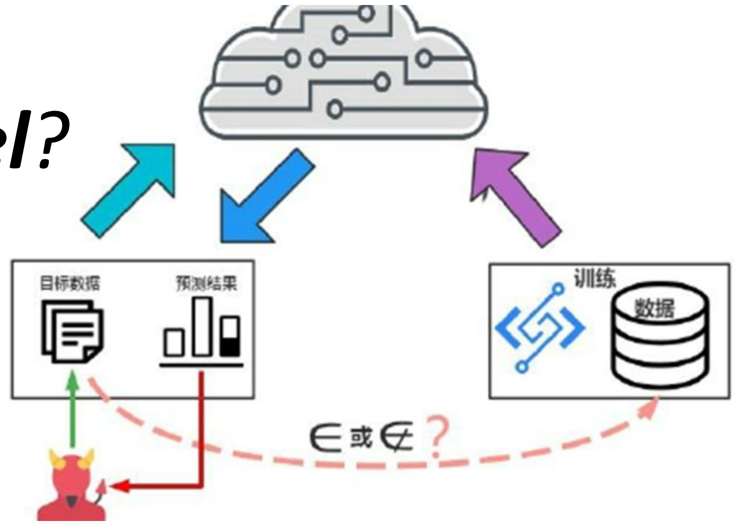


- Exploring memorization in T2I models:

[1] BBC. "Art is dead Dude" - the rise of the AI artists stirs debate. 2022. URL https://www.bbc.com/news/technology-62788725.
[2] CNN. AI won an art contest, and artists are furious. 2022. URL https://www.cnn.com/2022/09/03/ tech/ai-art-fair-winner-controversy/index.html.
[3] Reuters. Lawsuits accuse AI content creators of misusing copyrighted work. 2023. URL https://www.reuters.com/legal/transactional/ lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/.
[4] WashingtonPost. He made a children's book using AI. Then came the rage. 2022. URL https://www.washingtonpost.com/technology/2023/01/19/ ai-childrens-book-controversy-chatgpt-midjourney/.

Membership Inference:

*Is this **data point** used to train the **target model**?*



In **traditional tasks**:

For a given data point **x**:

$$\mathcal{M}(\mathbf{x}, f_\theta) = \mathbb{1}\left[\mathcal{M}'(\mathbf{x}, f_\theta) > \tau\right]$$

In **text-to-image synthesis**:

For a given data pair (image, text) → **(x, c)**

$$\mathcal{M}(\mathbf{x}, \mathbf{c}, f_\theta) = \mathbb{1}\left[\mathcal{M}'(\mathbf{x}, \mathbf{c}, f_\theta) > \tau\right]$$

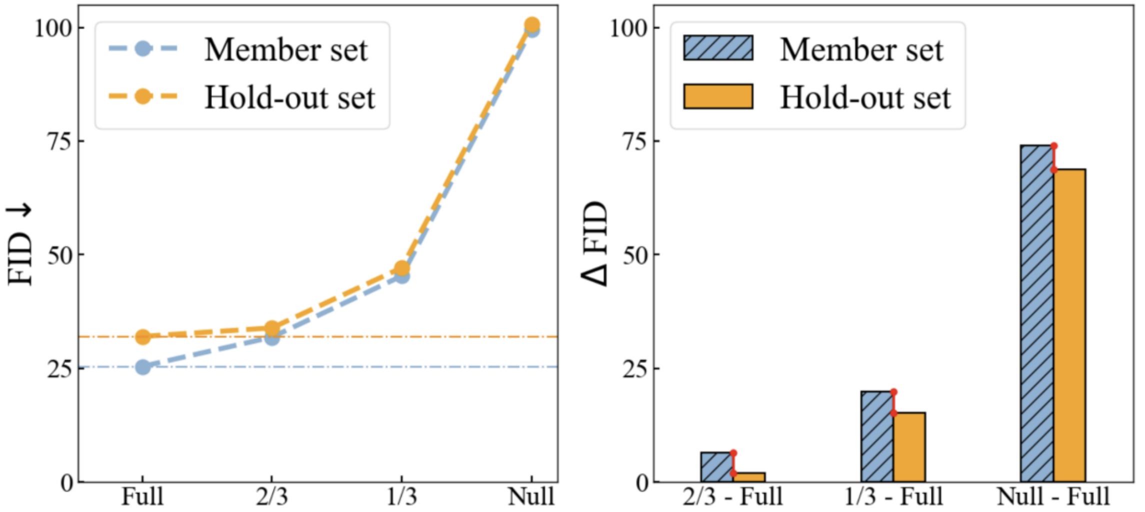
Are existing works good enough?

- Only targeting at small-scale diffusion model [1] (NOT text-to-image)
- Unrealistic evaluation setting → **Hallucination of success!**
  1. Over-training
  2. Distribution shift

Methods	Evaluation (Fine-tuning)	Evaluation (Pretraining)
SecMI [2]	~ 60 Epochs (**Over-training**)	LAION / COCO as mem/hold-out set (**Different distribution**)
PIA [3]	N/A	LAION / COCO as mem/hold-out set (**Different distribution**)
PFAMI [4]	~ 60 Epochs (**Over-training**)	LAION / COCO as mem/hold-out set (**Different distribution**)

[1] Nicolas Carlini et al. Extracting training data from diffusion models. In 32nd USENIX Security Symposium (USENIX Security 23)
[2] Jinhao Duan et al. Are diffusion models vulnerable to membership inference attacks? In International Conference on Machine Learning, 2023.
[3] Fei Kong et al. An efficient membership inference attack for the diffusion model by proximal initialization. In The Twelfth International Conference on Learning Representations, 2024.
[4] Wenjie Fu et al. A probabilistic fluctuation based membership inference attack for generative models. arXiv preprint arXiv:2308.12143, 2023

Conditional Overfitting



(a) Text Condition (b) Discrepancy of Text Conditions

➢Observation: T2I training process involves ***Conditional Overfitting.***

Training overfitting:

$$D(q_{\text{mem}}(\mathbf{x}), p(\mathbf{x})) \leq D(q_{\text{out}}(\mathbf{x}), p(\mathbf{x}))$$

Conditional overfitting:

FID *(Fréchet Inception Distance)*

$$\underbrace{\mathbb{E}_{\mathbf{c}}[D(q_{out}(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c})) - D(q_{mem}(\mathbf{x}|\mathbf{c}), p(\mathbf{x}|\mathbf{c}))]}_{\textit{overfitting to conditional distribution}} \geq \underbrace{D(q_{out}(\mathbf{x}), p(\mathbf{x})) - D(q_{mem}(\mathbf{x}), p(\mathbf{x}))}_{\textit{overfitting to marginal distribution}}$$

CLiD (**C**onditional **Li**kelihood **D**iscrepancy)

Using Kullback-Leibler (KL) divergence as the distance metric, we can get (Proof in Appendix B):

$$\mathbb{E}_{q_{mem}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x})] \geq \mathbb{E}_{q_{out}(\mathbf{x},\mathbf{c})}[\log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x})] + \delta_H$$

Ignoring δ_H, we have the indicator ***CLiD***:

$$\mathbb{I}(\mathbf{x},\mathbf{c}) = \log p(\mathbf{x}|\mathbf{c}) - \log p(\mathbf{x})$$

Using ELBOs to approximate likelihood:

$$\mathbb{I}(\mathbf{x},\mathbf{c}) = \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{null}}) - \epsilon||^2\right] - \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right]$$

To simplify computation, we directly estimate likelihood difference by Monte Carlo Sampling [1]:

$$\mathbb{I}(\mathbf{x},\mathbf{c}) = \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_{\text{null}}) - \epsilon||^2 - ||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right]$$

[1] Li, Alexander C., et al. "Your diffusion model is secretly a zero-shot classifier." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

CLiD-MI

$$\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_i^*} = \mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}_i^*) - \epsilon||^2 - ||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right], \ \mathbb{C} = \{\mathbf{c}_1^*, \mathbf{c}_2^* ..., \mathbf{c}_k^*\}$$

$$\mathcal{L}_{\mathbf{x},\mathbf{c}} = -\mathbb{E}_{t,\epsilon}\left[||\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c}) - \epsilon||^2\right]$$

- Threshold-based CLiD_{th}:

$$\mathcal{M}_{\text{CLiD}_{th}}(\mathbf{x}, \mathbf{c}) = \mathbb{1}\left[\alpha \cdot \mathcal{S}(\frac{1}{k}\sum_i^k \mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_i^*}) + (1-\alpha) \cdot \mathcal{S}(\mathcal{L}_{\mathbf{x},\mathbf{c}}) > \tau\right]$$

- Vector-based CLiD_{vec}:

$$\mathbf{V} = \left(\mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_1^*}, \mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_2^*} \cdots \mathcal{D}_{\mathbf{x},\mathbf{c},\mathbf{c}_k^*}, \mathcal{L}_{\mathbf{x},\mathbf{c}}\right)$$

$$\mathcal{M}_{\text{CLiD}_{vec}}(\mathbf{x}, \mathbf{c}) = \mathbb{1}\left[\mathcal{F}_{\mathcal{M}}(\mathbf{V}) > \tau\right]$$

Main Experiments

- Settings
  1. Fine-tuning (Over-training): consistent with exiting works

     Data (member/hold-out set Size): Pokemon (~400), COCO (2500), Flickr (2500);

     Training steps: 15,000, 150,000, 150,000

     No augmentation.

  2. Fine-tuning (Real-world training): following Huggingface scripts [1]

     Data (member/hold-out set Size): Pokemon (~400), COCO (2500), Flickr (100,000);

     Training steps: 7,500, 50,000, 200,000

     Default augmentation.

  3. Pretraining (Ensuring the distribution consistency).
- Metrics

  ASR, AUC, TPR@1%FPR

[1] Huggingface. The training script of stable-diffusion, 2024. URL https://huggingface.co/docs/diffusers/training/text2image#launch-the-script. Accessed: May 22, 2024.

Table 1: Results under *Over-training* setting. We mark the best and second-best results for each metric in **bold** and underline, respectively. Additionally, the best results from baselines are marked in blue for comparison.

Method	MS-COCO			Flickr			Pokemon			Query
	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	
Loss	81.92	89.98	32.28	81.90	90.34	40.80	83.76	91.79	25.77	1
PIA	68.56	75.12	5.08	68.56	75.12	5.08	83.37	90.95	13.31	2
M. C.	82.04	89.77	36.04	83.32	91.37	41.20	79.35	86.78	23.74	3
SecMI	83.00	90.81	50.64	62.96†	89.29	48.52	80.49	90.64	9.36	12
PFAMI	94.48	98.60	78.00	90.64	96.78	50.96	89.86	95.70	65.35	20
CLiD_{th}	99.08	**99.94**	**99.12**	91.42	97.39	**74.00**	**97.96**	99.28	**97.84**	15
CLiD_{vec}	**99.74**	99.31	95.20	**91.78**	**97.52**	73.88	97.36	**99.46**	96.88	15

† When conducting SecMI [15], we observe that the thresholds obtained on the shadow model sometimes do not transfer well to the target model.

Over-training:

1. No obvious effectiveness difference of MI methods (Query 1 vs Query 12)
2. Excessive and unrealistic overfitting.

Fail to adequately reflect the effectiveness differences among various methods !

Table 2: Results under *Real-world training* setting. We also highlight key results according to Tab. 1.

Method	MS-COCO			Flickr			Pokemon			Query
	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	
Loss	56.28	61.89	1.92	54.91	56.60	1.83	61.03	65.96	2.82	1
PIA	54.10	55.52	1.76	51.96	52.73	1.28	58.34	59.95	2.64	2
M. C.	57.98	61.97	2.64	54.92	56.78	2.16	61.10	66.48	3.84	3
SecMI	60.94	65.40	3.92	55.60	63.85	2.76	61.28	65.56	0.84	12
PFAMI	57.36	60.39	2.72	54.68	56.13	1.80	58.94	63.53	5.76	20
CLiD$_{th}$	<u>88.88</u>	<u>96.13</u>	**67.52**	<u>87.12</u>	<u>94.74</u>	<u>53.56</u>	**86.79**	**93.28**	**61.39**	15
CLiD$_{vec}$	**89.52**	**96.30**	<u>66.36</u>	**88.86**	**95.33**	**53.92**	<u>85.47</u>	<u>92.61</u>	<u>59.95</u>	15

Method	LAION			Query
	ASR	AUC	TPR@1%FPR	
Loss	51.78	50.90	1.75	1
PIA	52.13	52.42	1.25	2
M. C.	53.18	53.96	1.25	3
SecMI	57.43	58.59	2.45	12
PFAMI	59.08	61.11	1.45	20
CLiD$_{th}$	**64.53**	**67.82**	**5.01**	15

Table 3: The performance of membership inference methods on Stable Diffusion v1-5 [47] in pretraining setting. We utilize the processed LAION dataset to ensure the distribution consistency between member / hold-out sets [13, 16]. The best results are highlighted in **bold**.

Real-world training & Pretraining setting:
Outperforming the baselines across all three metrics

Other Experiments

1. Effectiveness trajectory

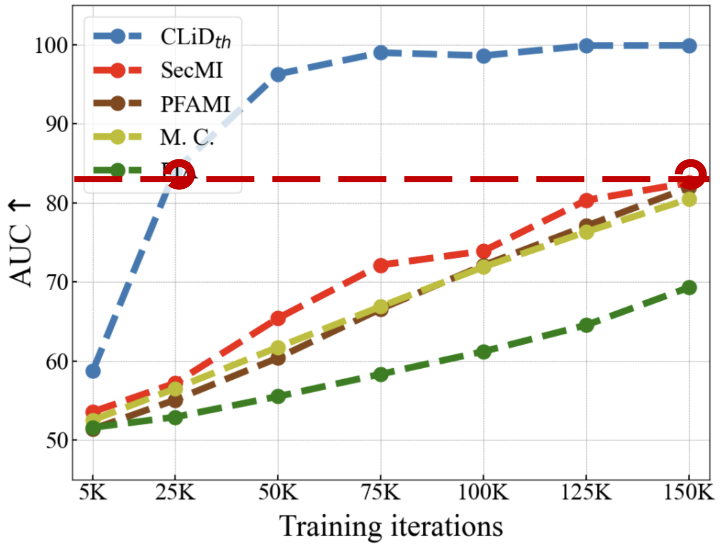


Figure 2: Effectiveness trajectory on various training steps.

2. Ablation Study

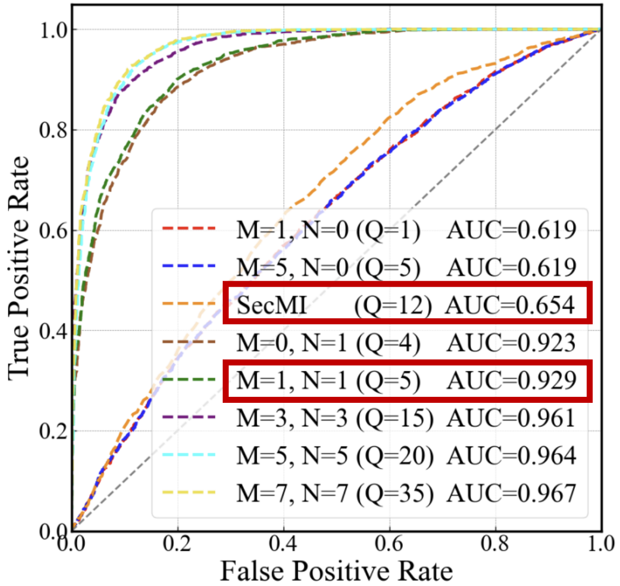


Figure 3: Performance of CLiD$_{th}$ and SecMI under various Monte Carlo sampling numbers (i.e., query count). The legend labels are sorted in ascending order by AUC values.

Other Experiments

3. Resistance to Defense

Table 4: The performance of different methods under no augmentation and default augmentation.

Method	No Augmentation			Defaut Augmentation		
	ASR	AUC	TPR@1%FPR	ASR (Δ)	AUC (Δ)	TPR@1%FPR (Δ)
Loss	66.54	72.73	7.72	56.28 (-10.26)	61.89 (-10.84)	1.92 (-5.80)
PIA†	56.56	59.28	2.00	54.10 (-2.46)	55.52 (-3.76)	1.76 (-0.24)
SecMI	72.02	81.07	13.72	60.94 (-11.08)	65.40 (-15.08)	3.92 (-9.80)
PFAMI	79.20	87.05	18.44	57.36 (-21.84)	60.39 (-26.66)	2.72 (-15.72)
CLiD$_{th}$	96.76	99.47	91.72	88.88 **(-7.88)**	96.13 **(-3.34)**	67.52 (-24.20)‡

†We omit the discussion of PIA as it shows no effectiveness at this training steps, with the metrics consistently approximating random guessing.

‡The TPR@1%FPR value changes significantly here because its ROC curve is very sharp when FPR close to 0.

Stronger resistance to data augmentation

Table 5: Effectiveness of CLiD$_{th}$ in adaptive defense. We calculate the FID [20] with $10,000$ unseen MS-COCO samples to assess the model utility.

Defense	CLiD$_{th}$ on MS-COCO			FID ↓ / Δ
	ASR	AUC	TPR@1%FPR	
None	88.88	96.13	67.52	13.17
Reph	85.32	93.83	55.67	13.58 / +0.41
Del-1	86.40	93.59	59.52	13.18 / -0.01
Del-3	83.91	91.52	52.03	12.92 / -0.25
Shuffle	65.89	67.37	0.15	18.26 / +5.09†

†Compared to other methods, the increase in FID caused by shuffling is unacceptable for generative models.

Resistance to adaptive defense

Other Experiments

Images	Text	
	**Groundtruth Text (MS-COCO):**	*A big burly grizzly bear is show with grass in the background.*
	Generated by BLIP:	*A brown bear sitting on the grass.*
	Generated by GPT4o-mini:	*A close-up of a large brown bear with thick fur, sitting in a grassy area.*
	**Groundtruth Text (MS-COCO):**	*A large white bowl of many green apples.*
	Generated by BLIP:	*A bowl of green apples.*
	Generated by GPT4o-mini:	*A bowl filled with fresh, shiny green apples stacked on top of each other.*
	**Groundtruth Text (MS-COCO):**	*A little girl holds up a big blue umbrella.*
	Generated by BLIP:	*A young girl holding an umbrella.*
	Generated by GPT4o-mini:	*A young girl holds a blue umbrella while wearing a pink jacket and jeans.*

4. Weaker Assumption:

What if we don't have groundtruth text?

→ Use Image-Caption model (BLIP) to generate Pseudo-Text.

Table 6: Results without access to the corresponding text under *Over-training* setting and *Real-world training* setting. We fine-tune MS-COCO on SDv1-4. Key results are highlighted as Tab. 1.

Method	*Over-training* (Pseudo-Text)			*Real-world training* (Pseudo-Text)			Query
	ASR	AUC	TPR@1%FPR	ASR	AUC	TPR@1%FPR	
Loss	73.80	81.01	9.71	56.08	58.47	1.60	1
PIA	61.40	65.75	1.20	53.44	54.38	1.52	2
M. C.	74.36	81.55	11.28	56.68	60.00	1.28	3
SecMI	82.04	88.97	40.80	60.48	64.04	3.28	12
PFAMI	91.56	95.16	68.16	58.12	59.77	2.64	20
CLiD$_{th}$	92.84	95.43	**72.36**	76.16	83.27	**19.76**	15
CLiD$_{vec}$	**93.26**	**96.59**	71.73	**77.76**	**84.48**	18.06	15

Conclusion

1. Identifying ***Conditional Overfitting***, i.e., T2I diffusion models overfit more to conditional distribution p(x, y) than to marginal distribution p(x)
2. Revealing the ***hallucination success*** of existing membership inference methods and providing a more reasonable evaluation setting
3. Proposing to conduct membership inference via Conditional Likelihood Discrepancy (CLiD). CLiD-MI significantly **outperforms baselines across various data distributions and scales**.

Limitation

Superiority of CLiD-MI over the baselines in the pretraining setting ***is not as evident*** compared to fine-tuning setting.

→ We emphasize our experiments under pretraining setting (Tab. 3) reveal the hallucination success of existing works and ***encourage future research to focus on this more challenging and practical scenario***.

Thanks!

Paper

Repository

shengfang.zhai@gmail.com