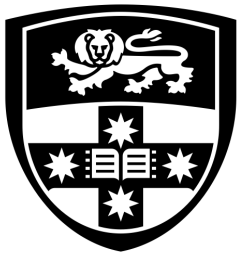
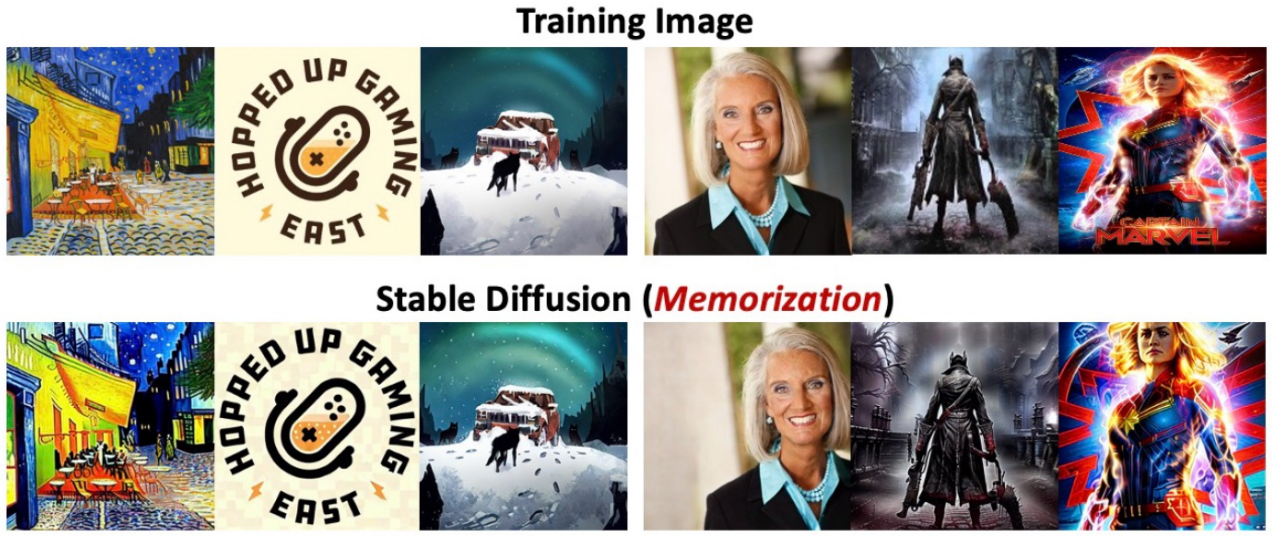


THE UNIVERSITY OF SYDNEY

Towards Memorization-Free Diffusion Models

Chen Chen, Daochang Liu, Chang Xu

Motivation



- Pretrained diffusion models can memorize and regurgitate training data during inference without informing data owners and model users.
- This has exposed the owners and users to potential violation of copyright laws and introduction of ethical dilemmas.
- Two factors has heightened such litigation risks: (1) the widespread use and deployment of open-source state-of-the-art diffusion models, and (2) the extensive size of training sets, which impedes detailed human review.

Research Gap

Previous research often necessitates the followings:

- Reducing memorization at a cost of output quality and utility (text-alignment).
- Model re-training.
- Not automated by nature:
  - Extensive manual intervention.
  - Indiscriminative applications of mitigation strategy to both memorized and un-memorized prompts.

Background: Metrics

- **Pixel-level similarity**: negative normalized L2-norm distance

$$\sigma_t = -\frac{\ell_2(\hat{x}_0, n_0)}{\alpha \cdot \frac{1}{k} \sum_{z_0 \in S_{\hat{x}_0}} \ell_2(\hat{x}_0, z_0)}$$

- **Object-level similarity**: dot product of Self-supervised Copy Detection (SSCD) embeddings

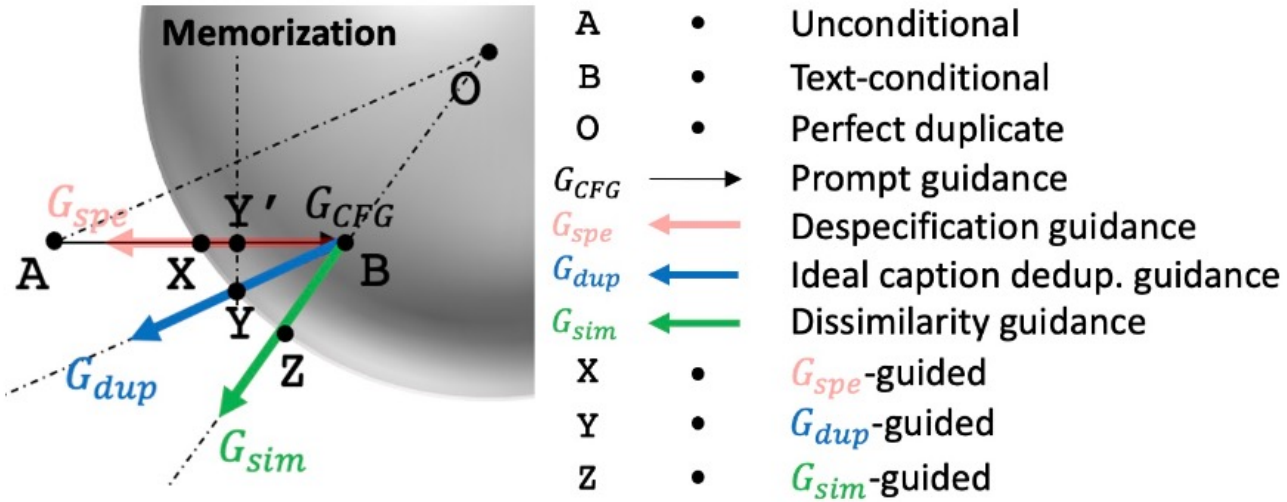
$$\sigma_t = E(\hat{x}_0)^T \cdot E(n_0)$$

Background: Causes of Memorization

- **Overly specific user prompts** act as a “key” to the pretrained model’s memory.
- **Duplicated training images** are more inclined to be memorized by diffusion models.
- **Duplicated captions across those duplicated images** can exacerbate the memorization issue by overfitting the text-image pairs to text-conditional diffusion models.

Method: AMG

- **AMG** is a unified framework comprises three guidance strategies, G_{spe}, G_{dup}, and G_{sim}, each meticulously crafted to address one of the identified causes of memorization.



$$\hat{\epsilon} \leftarrow \hat{\epsilon} + 1_{\{\sigma_t > \lambda_t\}} \cdot (G_{spe} + G_{dup} + G_{sim}) \quad (10)$$

$$x_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1-\bar{\alpha}_{t-1}}\hat{\epsilon} \quad (11)$$

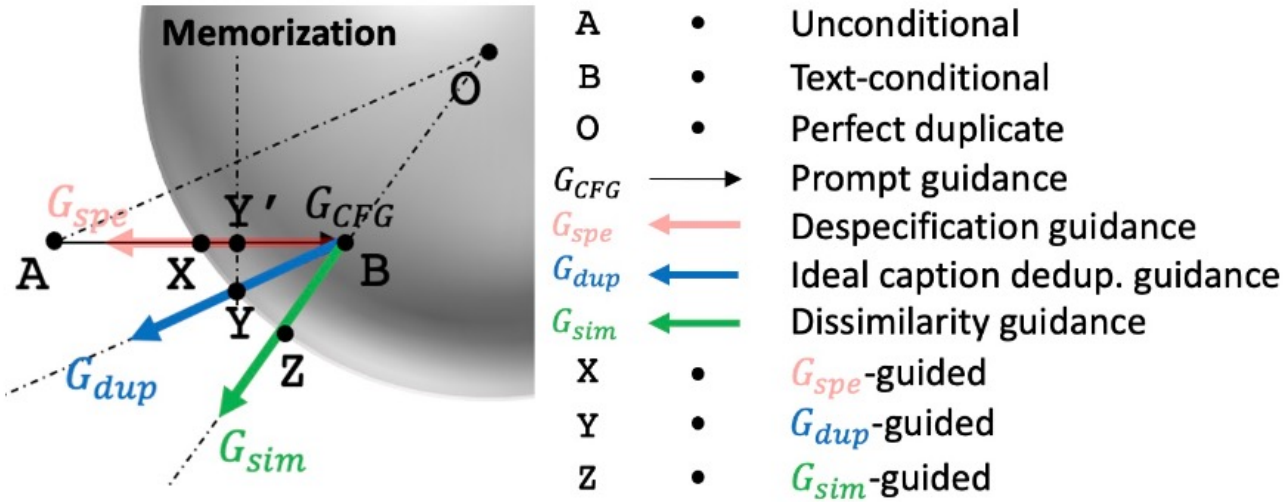
- All three guidance methods can steer the generation away from memorization.
- Each method solely requires updating the epsilon prediction.
- An indicator function is used to activate our guidance only during high similarity cases to maximally preserve output quality and utility.

Method: AMG

- **Despecification guidance** (G_{spe}) aims to reduce the specificity of text prompt.

$$s_1 = \max(\min(c_1\sigma_t, s_0 - 1), 0) \quad (13)$$

$$G_{spe} = -s_1(\epsilon_\theta(x_t, y) - \epsilon_\theta(x_t)) \quad (14)$$



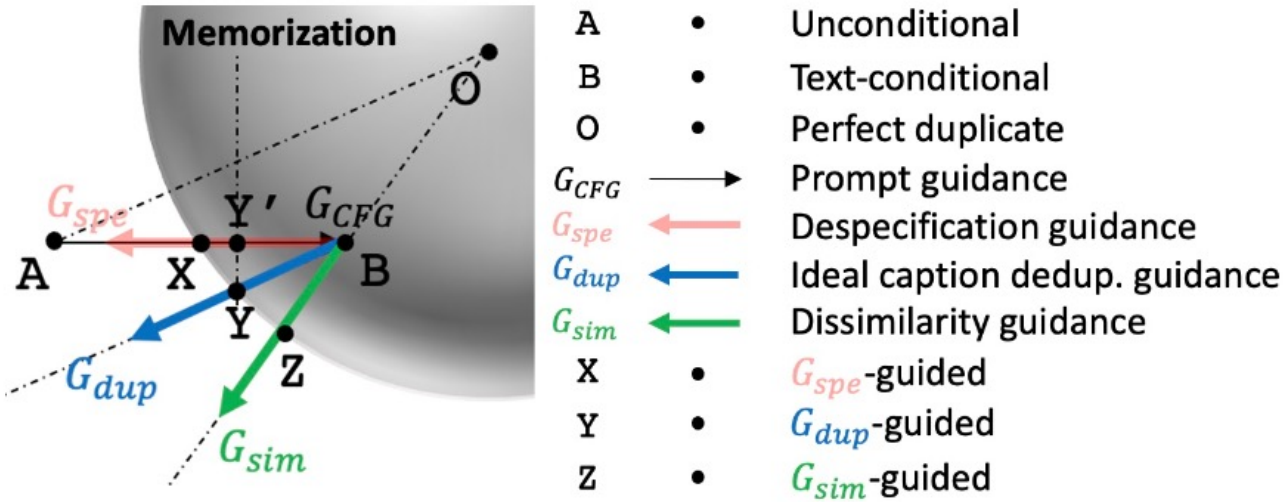
- This method aligns with the principles of CFG but pursues the inverse goal: linearly adjust the epsilon prediction to be less aligned with prompt-conditional prediction.
- The guidance scale is proportional to the similarity score during each inference step.

Method: AMG

- **Caption deduplication guidance** (G_{dup}) aims to instruct the generations away from those duplicated captions' generations.

$$s_2 = \max(\min(c_2\sigma_t, s_0 - s_1 - 1), 0) \quad (15)$$

$$G_{dup} = -s_2(\epsilon_\theta(x_t, y_N) - \epsilon_\theta(x_t)) \quad (16)$$

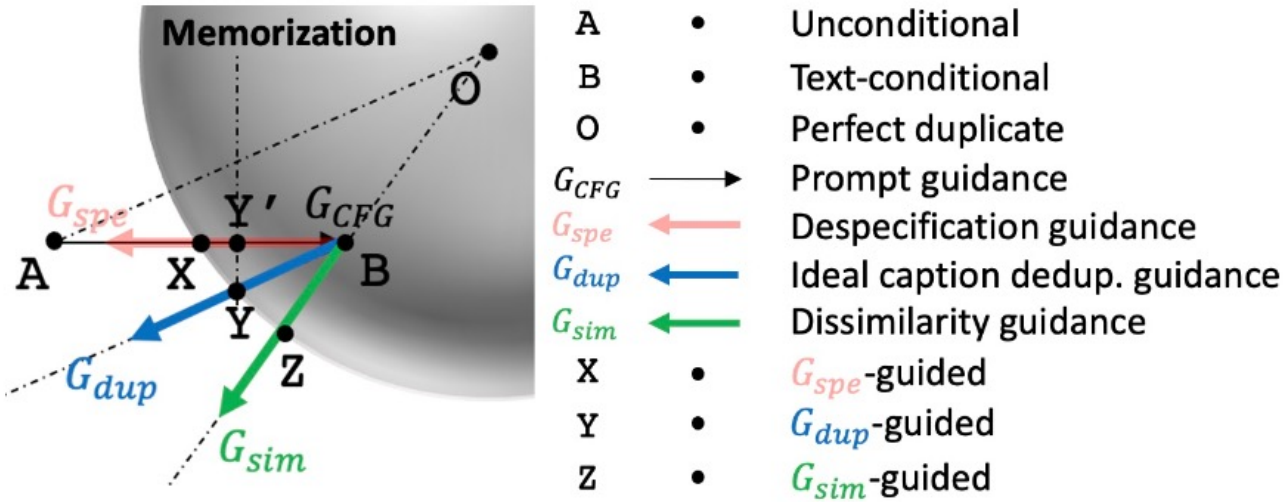


- This method intentionally uses duplicated captions as prompts to generate memorized images, then guides the inference away from these images.
- The guidance scale is proportional to the similarity score during each inference step.

Method: AMG

- **Dissimilarity guidance** (G_{sim}) extends the discrete class label in classifier guidance to a continuous embedding represented by the similarity score.

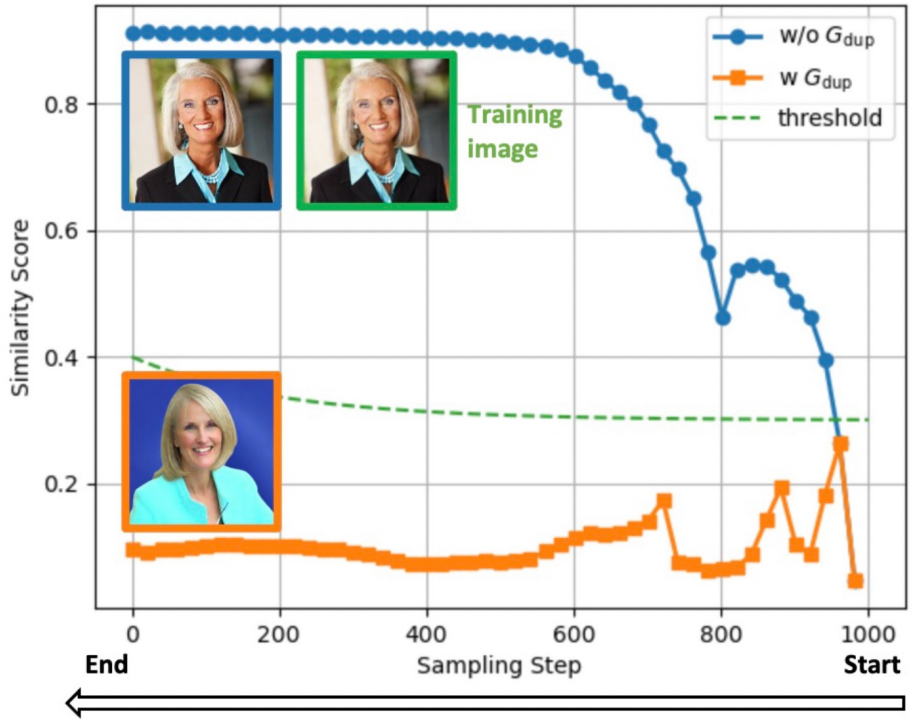
$$G_{sim} = c_3\sqrt{1-\bar{\alpha}_t} \cdot \nabla_{x_t}\sigma_t \qquad (17)$$

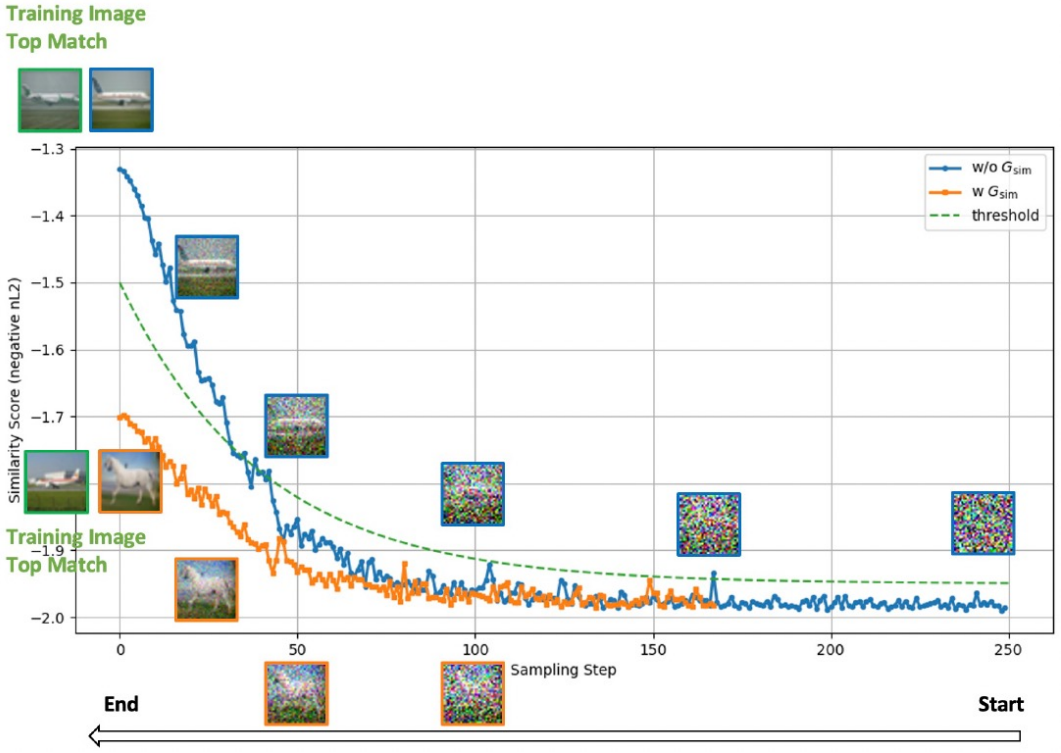


- This method assures generated images to be actively directed towards reducing their similarity score.

Analysis

- AMG uses similarity scores for early identifications of potential memorization, then employ guidance methods to instruct those generations away from memorized training image.





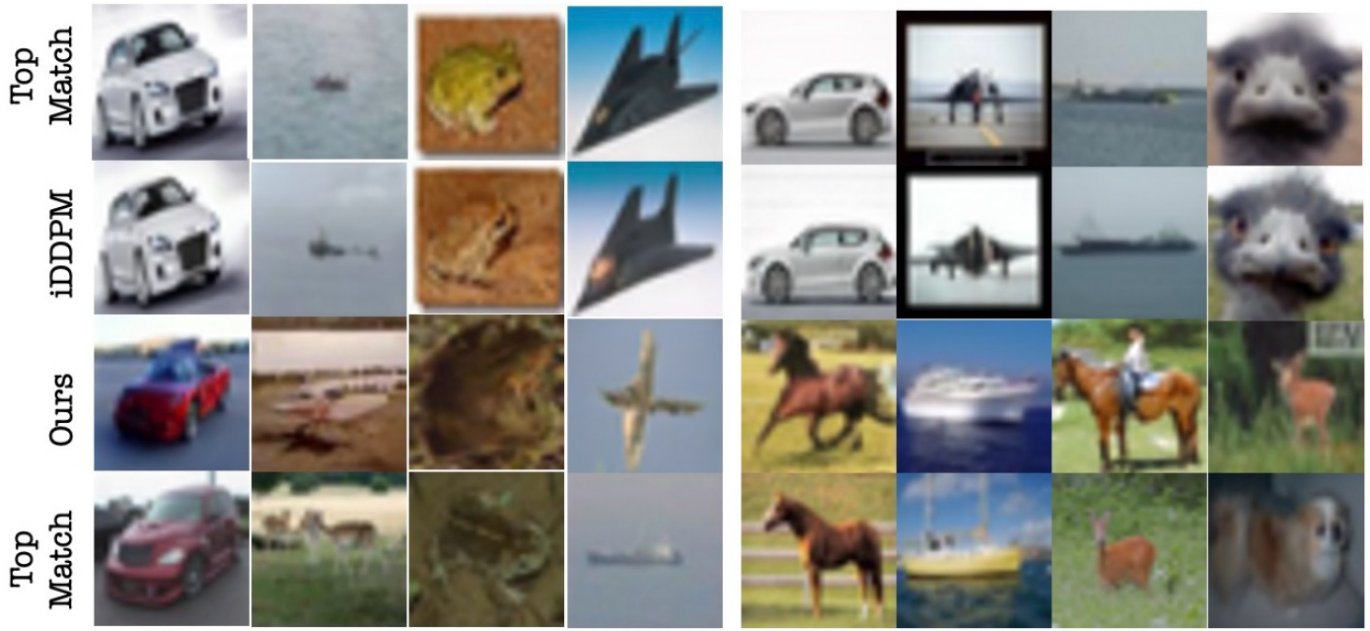
The denoising trajectories with and without AMG's guidance strategies.

Results

- AMG successfully guides Stable Diffusion pretrained on LAION to produce memorization-free outputs. Left: pixel-level memorizations. Right: object-level memorizations.

Results

- AMG successfully guides iDDPM pretrained on CIFAR-10 to produce memorization-free outputs. Left: Class-conditional generation. Right: Unconditional generation.



Results

- Comparisons on text-conditional generation of LAION5B based on SSCD similarity. AMG successfully eliminates memorization with minimal impact on quality and text-alignment.

	Memorization Metrics by SSCD ↓					
	Top5%	Top1	%>0.5	%>0.4	FID ↓	CLIP ↑
SD [30]	0.91	0.93	44.85	59.23	106.41	**28.04**
Ablation [19]	-	-	0.30*	-	-	-
GNI [36]	0.91	0.94	42.75	58.18	97.81	27.79
RT [36]	0.61	0.84	15.07	26.75	101.69	22.63
CWR [36]	0.79	0.85	26.45	40.93	**96.25**	25.96
RNA [36]	0.75	0.82	17.78	29.05	99.68	23.37
Ours(Main)	0.41	0.47	**0.00**	7.07	99.12	26.98
Ours(Strong)	**0.34**	**0.39**	**0.00**	**0.00**	100.45	26.72

Results

- Comparisons on unconditional generation of CIFAR-10 based on nL2 similarity. AMG effectively eliminates memorization without affecting image quality.

	Memorization Metrics by nL2				
	Top5%↑	Top1↑	%<1.4↓	%<1.6↓	FID↓
iDDPM [23]	1.58	0.51	0.93	5.78	7.44
Ours(Main)	1.61	1.47	**0.00**	4.34	7.25
Ours(Strong)	**1.71**	**1.68**	**0.00**	**0.00**	**6.98**

- Comparisons on class-conditional generation of CIFAR-10 based on nL2 similarity. AMG effectively eliminates memorization without affecting image quality..

	Memorization Metrics by nL2				
	Top5%↑	Top1↑	%<1.4↓	%<1.6↓	FID↓
iDDPM [23]	1.53	0.51	1.53	9.77	11.81
Ours(Main)	1.56	1.46	**0.00**	8.70	11.54
Ours(Strong)	**1.71**	**1.68**	**0.00**	**0.00**	**11.44**

Results

- Ablation studies on text-conditional generation based on SSCD. Grey-colored font denotes areas of sacrifice.

	Mem. by SSCD ↓		FID ↓	CLIP ↑
	Top5%	%>0.5		
Baseline [30]	0.9133	44.85	106.41	**28.04**
$G_{sim} + G_{spe}$	0.4072	**0.00**	119.13	26.67
$G_{sim} + G_{dup}$	0.4073	**0.00**	120.48	26.17
$G_{spe} + G_{dup}$	0.7396	31.62	**87.10**	27.18
Full	**0.4066**	**0.00**	99.12	26.98

Conclusion and Limitations

- We introduce AMG, a unified framework featuring three specialized guidance strategies, each addressing a specific cause of memorization in diffusion models.

- Theoretical analysis and empirical results, including ablation studies, confirm the essential role of each strategy in achieving an optimal privacy-utility trade-off.

- However, AMG relies on the availability of training images to compute similarity scores during inference. Additionally, although it is a training-free method, the nearest neighbor search largely increases its computational cost compared to the original inference process. We look forward to future work proposing new detection and mitigation methods that can achieve a good privacy-utility trade-off without the use of training data.