Detecting and Mitigating Memorization in Diffusion Models through Anisotropy of the Log-Probability

Rohan Asthana
rohan.asthana@fau.de

Vasileios Belagiannis
vasileios.belagiannis@fau.de

Friedrich-Alexander-Universität Erlangen-Nürnberg

Problem & Motivation

Setting.

- Large-scale text-to-image diffusion models (e.g. Stable Diffusion[1]) can unintentionally memorize training images, reproducing near-exact copies tied to specific prompts and seeds.
- This raises privacy, copyright, and safety concerns in real-world deployments.

[1] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

Limitations of Existing Metrics

Current denoising-free detection.

- Current methods[2,3] utilize norm of the difference between conditional and unconditional scores $\|s_\theta^\Delta\| := \|\tilde{s}_\theta - s_\theta\| \approx \|\nabla_{\boldsymbol{x}_t} \log p_t(c \mid \boldsymbol{x}_t)\|$.
- This measures the overall sharpness of the log-probability.

[2] Wen, Yuxin, et al. "Detecting, explaining, and mitigating memorization in diffusion models." The Twelfth International Conference on Learning Representations. 2024.

[3] Jeon, Dongjae, Dueun Kim, and Albert No. "Understanding and Mitigating Memorization in Generative Models via Sharpness of Probability Landscapes." International Conference on Machine Learning. PMLR, 2025.

Limitations of Existing Metrics

- **High/medium noise**.
  - Log-probability $\rightarrow$ isotropic.
  - Curvature $\approx$ same for all directions.
  - Norm-based metrics ✓
- **Low noise**.
  - Log-probability $\rightarrow$ *anisotropic*.
  - Curvature different across different directions.
  - norm-based metrics ✘

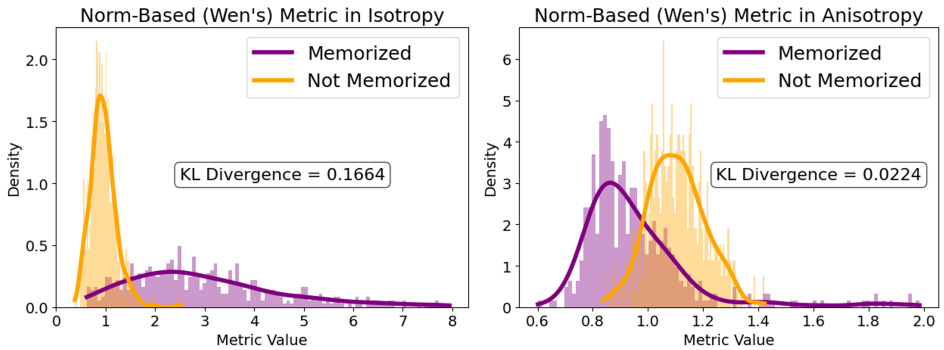


Figure: Failure of denoising-free norm-based methods in Anisotropy.

Angular Alignment in Anisotropy

Core geometric insight.

- Memorized prompts $\rightarrow$ strong alignment between the unconditional score s_θ and the guidance vector s_θ^Δ.
- This is because when the modes of $\log p_t(c \mid \boldsymbol{x}_t)$ and $\log p_t(\boldsymbol{x}_t)$ are close, the cosine similarity has a lower bound controlled by differences in covariances[4].

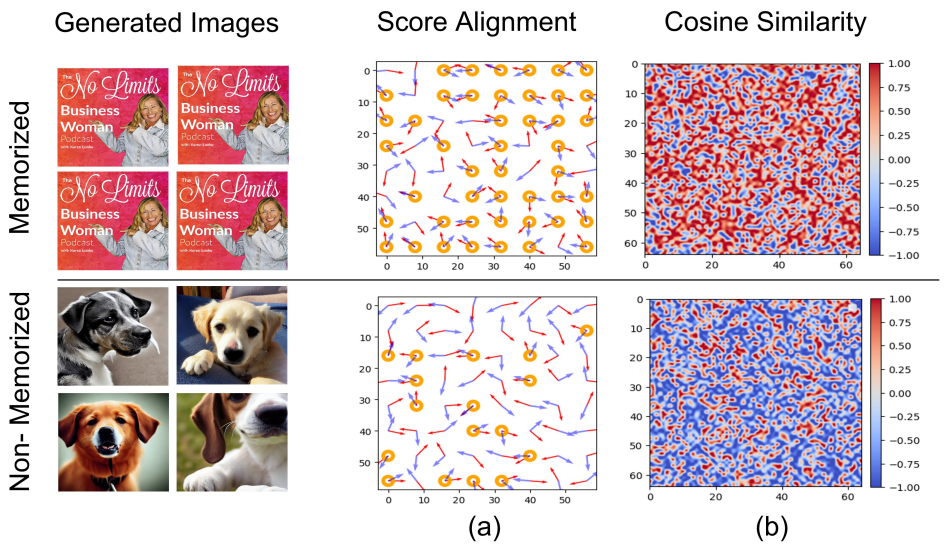


Figure: Strong alignment between the unconditional score and the guidance vector

[4] Theorem and proof provided in the paper.

Proposed Denoising-Free Metric

Denoising-free detection from pure noise.

- Utilizes only two forward passes[5] of the model on initial noise $\boldsymbol{x}_T$.
- Combines:
  - cosine similarity in the anisotropic low-noise regime,
  - norm of score difference $\|s_\theta^\Delta\|$ in the isotropic high-noise regime.

Definition.

$$\mathcal{M}(\mathbf{x}_T, c) = \gamma_1 \underbrace{\left\{ \frac{\left\langle s_\theta^\Delta(\mathbf{x}_T, t \approx 0, c),\ s_\theta(\mathbf{x}_T, t \approx 0) \right\rangle}{\left\| s_\theta^\Delta(\mathbf{x}_T, t \approx 0, c) \right\| \left\| s_\theta(\mathbf{x}_T, t \approx 0) \right\|} \right\}}_{\text{cosine similarity in anisotropy}} + \gamma_2 \underbrace{\left\| s_\theta^\Delta(\mathbf{x}_T, t \approx T, c) \right\|}_{\text{norm of score difference in isotropy}}$$

[5] One near $t \approx 0$, one near $t \approx T$.

Denoising-Free Detection Results

Experimental setup.

- Memorization detection on Stable Diffusion v1.4 and v2.0 comparing AUC and TPR@1%FPR.

Method	SD v1.4			SD v2.0		
	AUC ↑	TPR@1%FPR ↑	Time (sec.) ↓	AUC ↑	TPR@1%FPR ↑	Time (sec.) ↓
$n = 1$						
Ren et al. (2024)	0.846	0.116	0.05	0.848	0	0.07
Wen et al. (2024)	0.976	0.896	0.40	0.948	0.739	0.80
Jeon et al. (2025)	0.987	0.908	5.40	**0.959**	0.740	14.60
$\mathcal{M}(\mathbf{x_T}, \mathbf{c})$ **(ours)**	**0.994 ±0.001**	**0.935±0.002**	1.10	0.953±0.016	**0.791±0.015**	2.20
$n = 4$						
Ren et al. (2024)	0.839	0.130	0.05	0.853	0	0.07
Wen et al. (2024)	0.992	0.944	1.20	0.980	0.876	2.70
Jeon et al. (2025)	0.998	0.982	19.40	**0.991**	**0.895**	56.40
$\mathcal{M}(\mathbf{x_T}, \mathbf{c})$ **(ours)**	**0.999±0.001**	**0.984±0.002**	3.40	0.981±0.003	0.890 ± 0.009	7.30

Detection performance.

- AUC up to 0.999 on SD v1.4.
- TPR@1%FPR up to 0.984.
- 5–8× faster than the previous best denoising-free approach.

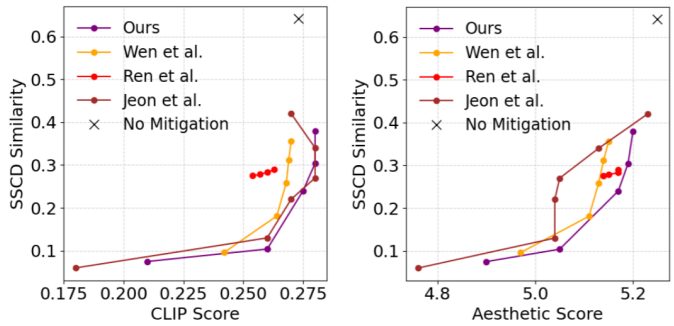
Inference-Time Mitigation

Using the metric for mitigation.

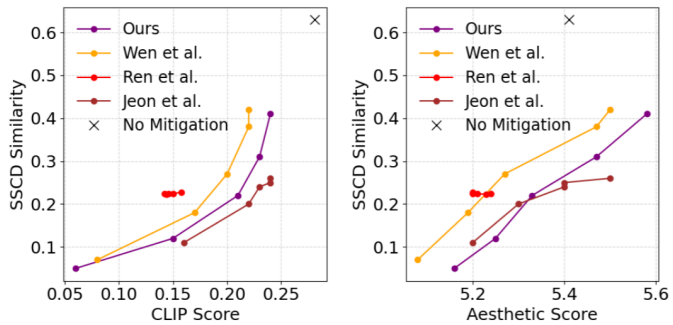
- We integrate $M(\boldsymbol{x}_T, c)$ into inference-time sampling to suppress memorized generations.
- Mitigation evaluated on MemBench[6].

Mitigation Performance.

- Low SSCD similarity to memorized images while maintaining CLIP and aesthetic scores.
- Provides a more favorable quality–safety-efficiency trade-off than prior methods.



(a) SD v1.4



(b) SD v2.0

[6] Hong, C., Oh, T.-H., & Sung, M. (2025). MemBench: Memorized Image Trigger Prompt Dataset for Diffusion Models. Transactions on Machine Learning Research.

Qualitative Results

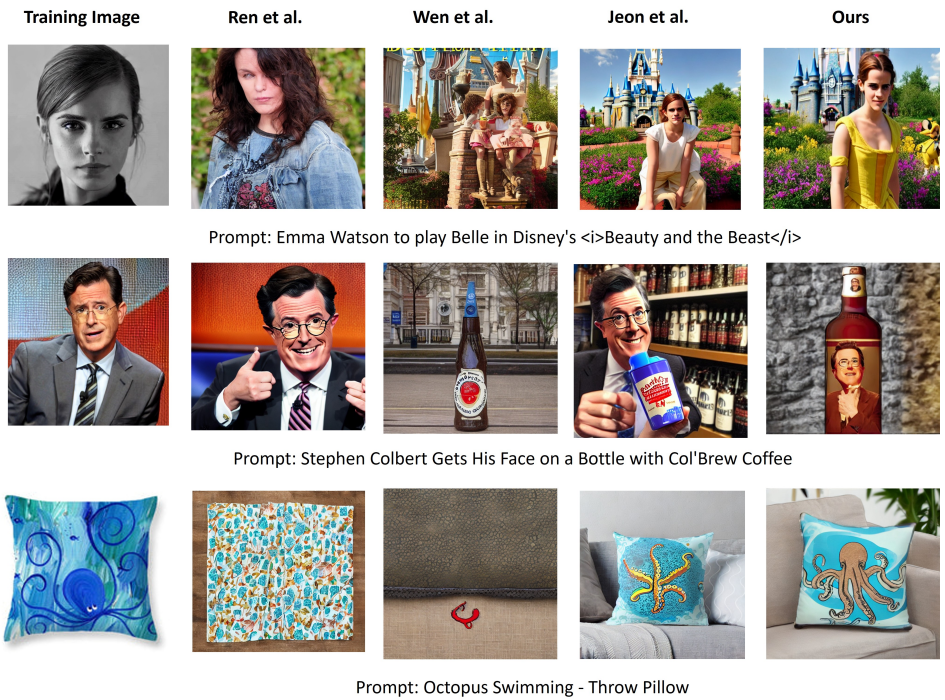


Figure: Our method enables quick memorization mitigation while preserving the text-image alignment.

Conclusion

- Norm-based denoising-free memorization metrics implicitly assume isotropy and fail in the anisotropic low-noise regime.
- Angular alignment between unconditional and conditional scores is a robust anisotropy-aware signal.
- Our denoising-free metric:
  - detects memorization from pure noise using just two forward passes,
  - achieves SOTA detection,
  - is 5–8$\times$ faster than the previous best method.
- Fast memorization mitigation while preserving semantic and aesthetic quality.

References

- Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
- Wen, Yuxin, et al. "Detecting, explaining, and mitigating memorization in diffusion models." The Twelfth International Conference on Learning Representations. 2024.
- Jeon, Dongjae, Dueun Kim, and Albert No. "Understanding and Mitigating Memorization in Generative Models via Sharpness of Probability Landscapes." International Conference on Machine Learning. PMLR, 2025.
- Hong, C., Oh, T.-H., & Sung, M. (2025). MemBench: Memorized Image Trigger Prompt Dataset for Diffusion Models. Transactions on Machine Learning Research.

Thank you!

Code & paper available online