Towards Memorization-Free Diffusion Models

Supplementary Material

Outline. In Sec. 1, we analyze additional insights into the synergistic use of Anti-Memorization Guidance (AMG) with our deliberately designed conditional guidance strategy and parabolic scheduling. Sec. 2 introduces KDE plots as an alternative evaluative method for memorization, showcasing the distribution of the top 1 similarity score across generated images, complementing the main paper's quantitative results. Sec. 3 demonstrates AMG's adaptability in switching samplers or similarity metrics within its guidance to meet specific user needs, maintaining effectiveness in memorization eradication. Sec. 4 delves into the implementation details. Finally, Sec. 5 offers additional qualitative results.

1. Additional Analysis

The power of AMG can be amplified when paired with our deliberately designed conditional guidance strategy that incorporates a parabolic scheduling threshold for determining the activation of AMG at each inference step.

Firstly, the efficacy of G_{sim}, our dissimilarity guidance component, is clear in Fig. 1 to Fig. 5, with guided versions consistently registering lower similarity scores compared to their ablated counterparts, indicating successful deviation from memorized training images:

$$G_{sim} = c_3\sqrt{1-\bar{\alpha}_t} \cdot \nabla_{x_t}\sigma_t \tag{1}$$

where c_3 further enables users to control the guidance intensity, thus balancing privacy and utility. This is also capable of providing a guarantee of memorization-free generations by simply setting a large c_3 at the cost of output quality. **Pairing G_{sim} with our conditional guidance strategy can optimize such trade-off**.

Specifically, our conditional guidance uses a parabolic schedule (Eq. (2)) that is closely aligned with the characteristics of the denoising stages as can be observed from the blue lines in Fig. 1 to Fig. 5, where early-stage predictions have lower similarity scores, which is then increased exponentially in the following mathematical form:

$$\lambda_t = a + (b-a)e^{-ct} \tag{2}$$

This distinctive pattern of similarity scores during the denoising stages can be attributed to the initial noised and imprecise predictions, which register exceedingly low similarity scores according to nL2 (Eq. (3)). As t decreases, the denoising process yields more distinct predictions $\hat{x}_t$, resulting in elevated similarity scores that can potentially reveal cases of memorization.

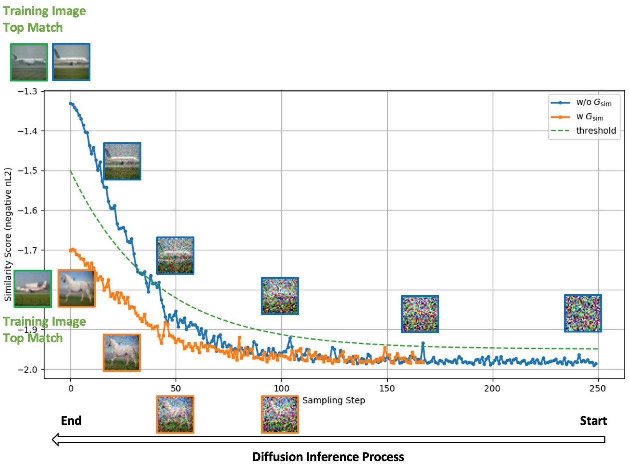


Figure 1. Example illustrating the detection of potential memorization at an **early** stage of reverse diffusion process, enabling AMG to influence the coarser structures in generated outputs.

$$\sigma_t = -\frac{\ell_2(\hat{x}_t, n_t)}{\alpha \cdot \frac{1}{k}\sum_{z_t \in S_{\hat{x}_t}} \ell_2(\hat{x}_t, z_t)} \tag{3}$$

Our parabolic schedule is tailored to this trend, serving as the threshold for AMG's activation. This design facilitates the early detection and intervention of potential memorization instances, as demonstrated in Fig. 1, Fig. 2, and Fig. 3. Specifically, in Eq. (2), the parameter a represents the asymptotic threshold value that the parabolic schedule approaches as t increases towards infinity, which we have set to -1.95. The parameter b denotes the value of the parabolic schedule at $t = 0$, and we have set this to -1.5, intentionally lower than the threshold defining memorization, which is -1.4. c controls the shape of the parabolic schedule, which we set to -0.025 to produce the green dashed lines in Fig. 1 to Fig. 5.

Fig. 4 and Fig. 5, serving as two counterexamples, underscores the importance of timely intervention. Utilizing a constant guidance schedule results in late memorization detection, necessitating larger late-stage adjustments. These adjustments often involve last-ditch efforts like adding noise to the image background, noticeable upon close inspection in Fig. 4 and Fig. 5, to reduce the similarity score. Such measures can diminish image quality and might still fail to effectively prevent memorization.

By identifying and intervening memorization early, AMG primarily affects the coarser structures, preserving

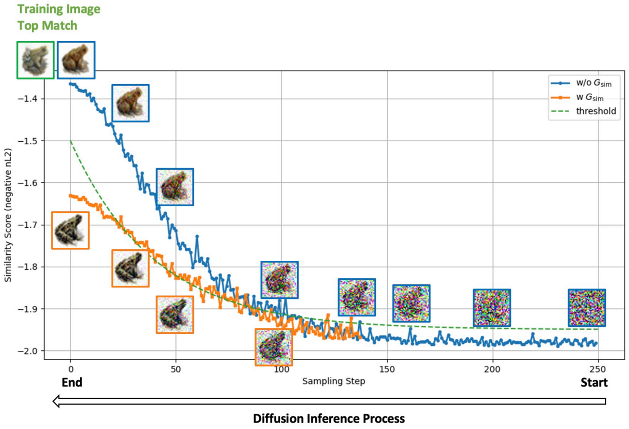


Figure 2. Example illustrating the early detection of potential memorization at an **intermediate** stage of reverse diffusion process, enabling AMG to influence the generation's structures that are between coarse and fine detail.

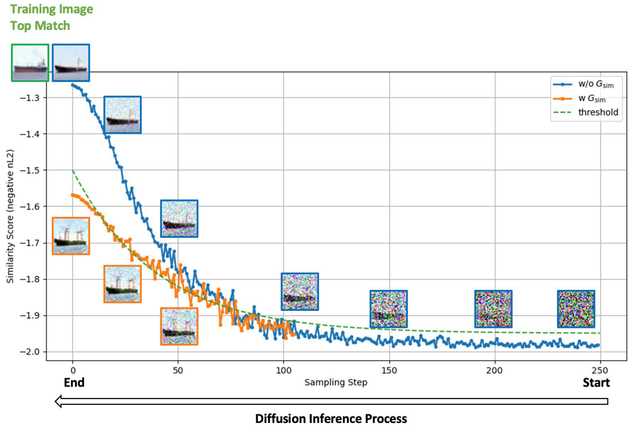


Figure 3. Another example illustrating the early detection of potential memorization at an **intermediate** stage of reverse diffusion process, enabling AMG to influence the generation's structures that are between coarse and fine detail.

the finer details and overall aesthetic of the generated images. This approach avoids resorting to last-minute measures to lower the similarity score at the cost of quality. Thus, this optimizes the privacy-utility trade-off, results in more visually appealing outputs that are distinct from the training data, as demonstrated in Fig. 1 to Fig. 3.

2. Alternative Evaluative Method

In the main paper, we first evaluate memorization using two key quantitative metrics in accordance with the established standards in the literature: (1) the 95th percentile of the top 1 similarity scores of all generated images, as per [6], and (2) the proportion of images exceeding certain similarity thresholds, indicating memorization, following [2]. Additionally, we assess the maximum similarity score to understand the worst-case scenario. Relying solely on the 95th percentile might not fully capture the distribution, particularly if there's a significant upper tail beyond this percentile.

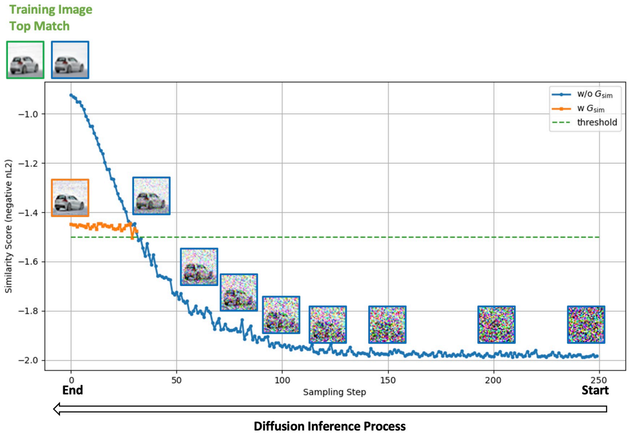


Figure 4. An **ablated counterexample** highlighting the risk of incomplete memorization prevention and the consequence of altering only the **finest** details and **injecting noises to image background** to lower the similarity score when employing a constant guidance schedule, as opposed to the more effective parabolic scheduling in our conditional guidance approach.

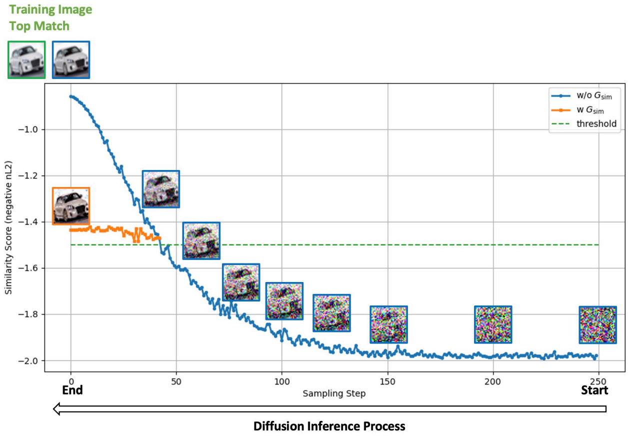


Figure 5. Another **ablated counterexample** highlighting the risk of incomplete memorization prevention and the consequence of altering only the **finest** details and **injecting noises to image background** to lower the similarity score when employing a constant guidance schedule, as opposed to the more effective parabolic scheduling in our conditional guidance approach.

However, these metrics still don't fully represent the distribution of the top 1 similarity scores. To address this, we introduce qualitative KDE (Kernel Density Estimation) plots as a complementary qualitative evaluation method for memorization, showcased in Fig. 6, Fig. 7, and Fig. 8, providing a more comprehensive view of the data distribution.

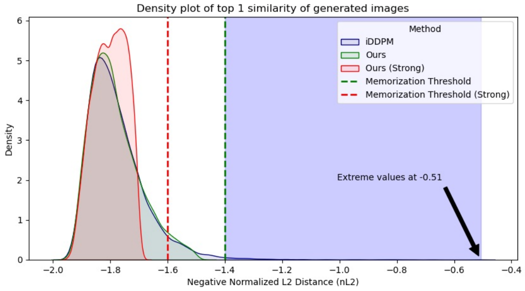


Figure 6. KDE plot depicting the top 1 similarity scores for **unconditional CIFAR-10 generation**. AMG effectively shifts all outputs to the left, maintaining similarity scores below the established memorization threshold.

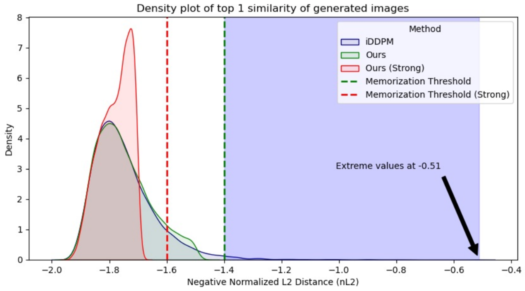


Figure 7. KDE plot depicting the top 1 similarity scores for **class-conditional CIFAR-10 generation**. AMG effectively shifts all outputs to the left, maintaining similarity scores below the established memorization threshold.

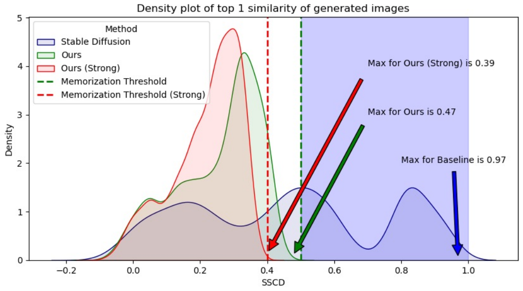


Figure 8. KDE plot depicting the top 1 similarity scores for **text-conditional LAION generation**. AMG effectively shifts all outputs to the left, maintaining similarity scores below the established memorization threshold.

	Memorization Metrics by SSCD ↓			
	Top5%	Top1	%>0.90	FID↓
iDDPM [4]	0.83	0.96	0.42	**7.44**
Ours	**0.79**	**0.85**	**0.00**	7.58

Table 1. Comparisons on unconditional generation of CIFAR-10 based on SSCD similarity. AMG effectively eliminates memorization without affecting image quality.

Text-conditional image generation presents a greater memorization challenge, as depicted in Fig. 8, where a larger portion of the blue curves (representing the top 1 similarity scores of pretrained models) cross the memorization threshold compared to unconditional and class-conditional generations in Fig. 6 and Fig. 7. Irrespective of the generation task, AMG reliably shifts these distributions leftward, reducing similarity to training data. The main AMG version ensures all top 1 similarity scores fall below the conventional thresholds (-1.40 for CIFAR-10 and 0.50 for LAION), and the strong AMG version meets even stricter thresholds (-1.60 for CIFAR-10 and 0.40 for LAION), affirming its efficacy.

3. AMG's Wide Adaptability

This section illustrates AMG's flexibility as a unified framework adaptable to different similarity measures for guidance signals and evaluation, as well as to different sampling methods including DDPM and accelerated techniques like DDIM [7], while retaining its efficacy.

3.1. Switching Similarity Measures

Self-supervised Copy Detection (SSCD) has emerged as the preferred method for detecting memorization in the LAION dataset [3, 5, 6], outperforming other metrics like the normalized L2 distance (nL2) and CLIP. Although nL2 has been the metric of choice for CIFAR-10 in prior studies [2], no work has verified its superiority to other measures such as SSCD. Thus, to complement, we also adopt SSCD as the guiding and evaluation metric within AMG, as an alternative to nL2. This change allows SSCD to influence guidance activation and scale, and also to direct updates in the prediction process as described in Eq. (1). For CIFAR-10, we find a threshold of 0.50 for top 1 SSCD similarity does not accurately indicate memorization, often resulting in false positives. Therefore, we have adjusted the threshold, considering instances with top 1 SSCD similarity scores above 0.90 as true memorization cases to improve the test's precision. Tab. 1 shows that AMG retains its effectiveness in eliminating memorization even when transitioning to SSCD, confirming the framework's flexibility and robustness.

3.2. Switching Samplers

In the main paper, we detail the implementation of our Anti-Memorization Guidance (AMG) framework utilizing the DDIM sampler. For tasks with lower computational demands such as unconditional and class-conditional generation on the CIFAR-10 dataset, employing the DDPM sampler is a practical alternative. With the DDPM sampler, the dissimilarity guidance G_{sim} in AMG framework simplifies to the following expressions:

$$G_{sim} = c \cdot \nabla_{x_t} \sigma_t \tag{4}$$

$$x_{t-1} \leftarrow \text{sample from } \mathcal{N}(\mu - 1_{\{\sigma_t > \lambda_t\}} \cdot \Sigma G_{sim}, \Sigma) \tag{5}$$

where μ and Σ represent the mean and variance of the model's predicted distributions, respectively. These adjustments maintain the integrity of AMG's guidance while accommodating the operational characteristics of the DDPM sampler.

4. Implementational Details

This section provides additional implementational details, please refer to the code for even more comprehensive details, which is also included in the supplementary material.

Applying AMG on Latent Diffusion Models (LDMs). As discussed in the main paper, to compute similarity scores, we need to obtain the model's prediction of $\hat{x}_0$ using the following Diffusion Kernel:

$$\hat{x}_0 = \frac{x_t - \sqrt{1 - \bar{\alpha}_t} \cdot \hat{\epsilon}}{\sqrt{\bar{\alpha}_t}} \tag{6}$$

This presumes the representation x_t is in the pixel space across the diffusion steps from $t = T$ to 0, so that we can then search for its nearest neighbor n_0 in the training set, which is also in the pixel space. However, in the context of LDMs, $\hat{x}_0$ would be in latent space, thus it necessitates an additional conversion step using the decoder $\mathcal{D}$ from LDM's pre-trained autoencoder to obtain the pixel-space representation: $\hat{x}_0 \leftarrow \mathcal{D}(\hat{x}_0)$, before searching for its nearest training image in this updated pixel-space representation.

Scope of memorization eradication. Our research ambitiously targets the most practically significant scope of eliminating memorization in diffusion models, specifically ensuring the generation of images with a similarity score below a commonly accepted threshold (*e.g.*, SSCD < 0.50) across the extensive LAION5B dataset.

Initially daunting due to the sheer volume of comparisons required, the task was made feasible by utilizing the official clip-retrieval tool [1] provided by the LAION5B dataset creator, which includes a pre-trained k-nearest neighbor index for the entire LAION5B dataset. Specifically, its ClipClient allows remote querying of a clip-retrieval backend via python for very efficient retrieval of nearest neighbors based on the dot product of CLIP embeddings.

A challenge arises from the discrepancy between the CLIP-based retrieval provided by ClipClient and the SSCD embedding we use for object-level similarity. To bridge this gap, for each generated image, we retrieve the closest 1000 neighbors according to CLIP embeddings and then compute SSCD similarities to pinpoint the closest match. While this introduces a small risk of missing the global SSCD minimum, it's manageable by adjusting the search breadth at additional or less computational costs.

Furthermore, our approach allows for targeted scrutiny of specific images, where the user can directly supply URLs of the images of interest, circumventing broader searches to focus on preventing memorization of chosen images. Implementationally, our method necessitates only the image URLs as additional input. The algorithm automates the process thereafter, loading the URLs as images, computing their SSCD embeddings, and ensuring these specified images are included in the SSCD similarity computation, regardless of their retrieval status via ClipClient.

This personalized approach can drastically cut computational demands and refine the search to user-defined priorities. It also encompasses the narrower scope adopted in baseline methods like those in [6], treating them as a subset scenario where our methodology is adapted to their set of 10,000 selected LAION5B images, sidelining the broader ClipClient search.

5. Additional Qualitative Results

Finally, we present additional qualitative results in Fig. 9, Fig. 10, and Fig. 11 to demonstrate AMG's effectiveness in guiding pretrained diffusion models to produce memorization-free outputs.

References

[1] Romain Beaumont. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval, 2022. 4

[2] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security Symposium*, pages 5253–5270, 2023. 2, 3

[3] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023. 3

[4] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 3

[5] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *CVPR*, pages 6048–6058, 2023. 3

[6] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *NeurIPS*, 2023. 2, 3, 4

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3

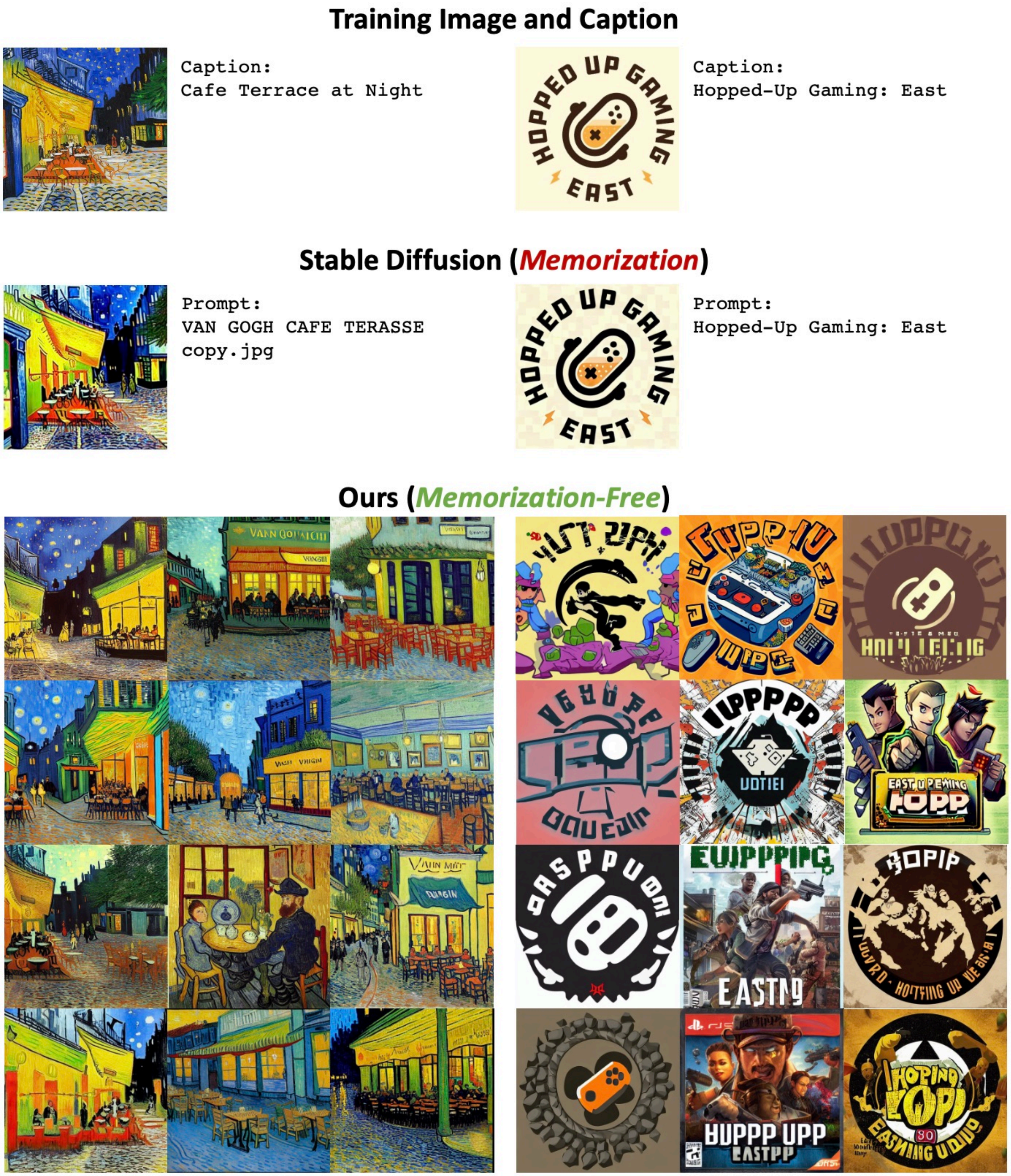


Figure 9. Additional qualitative comparisons showcase AMG's effectiveness in guiding pretrained diffusion models to produce memorization-free outputs.

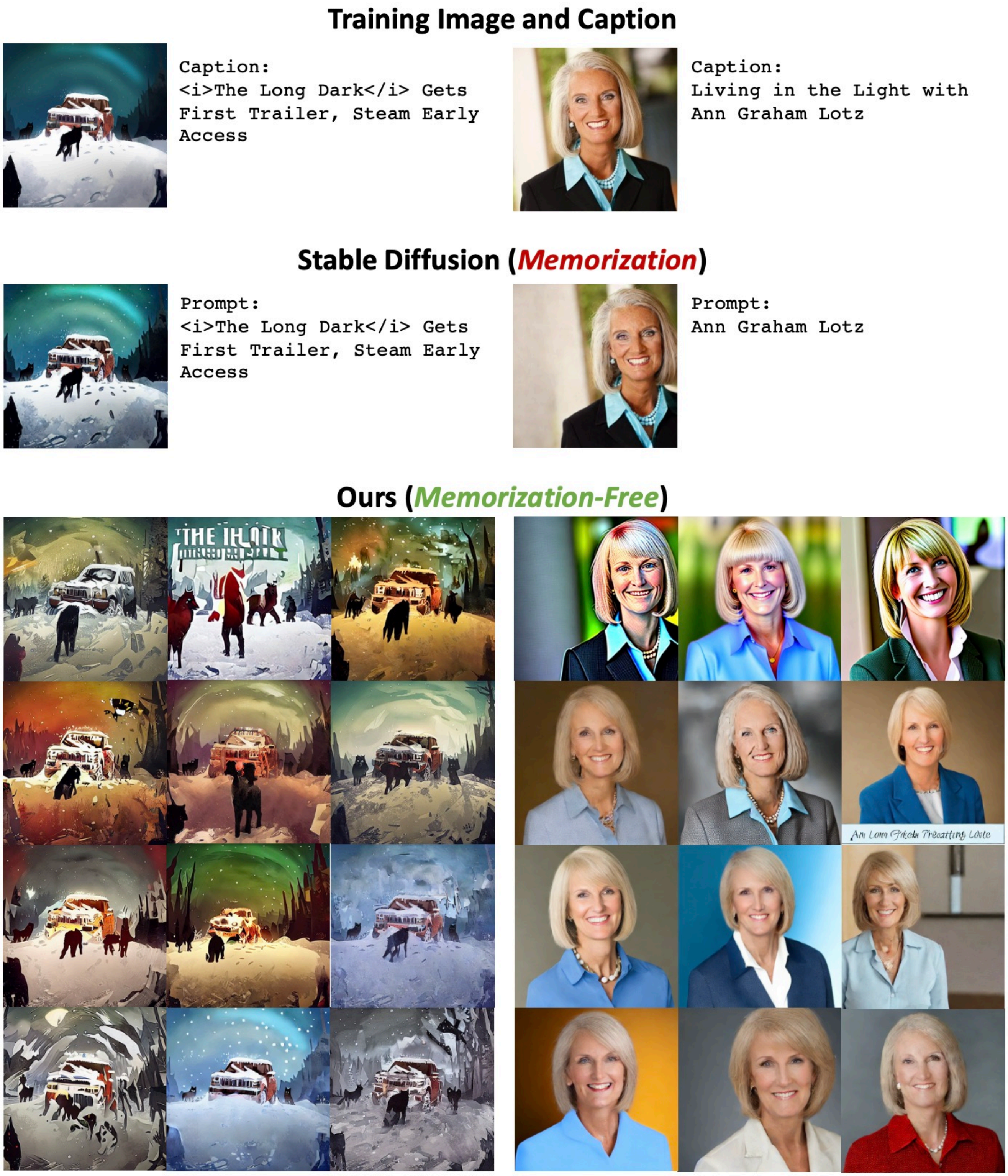
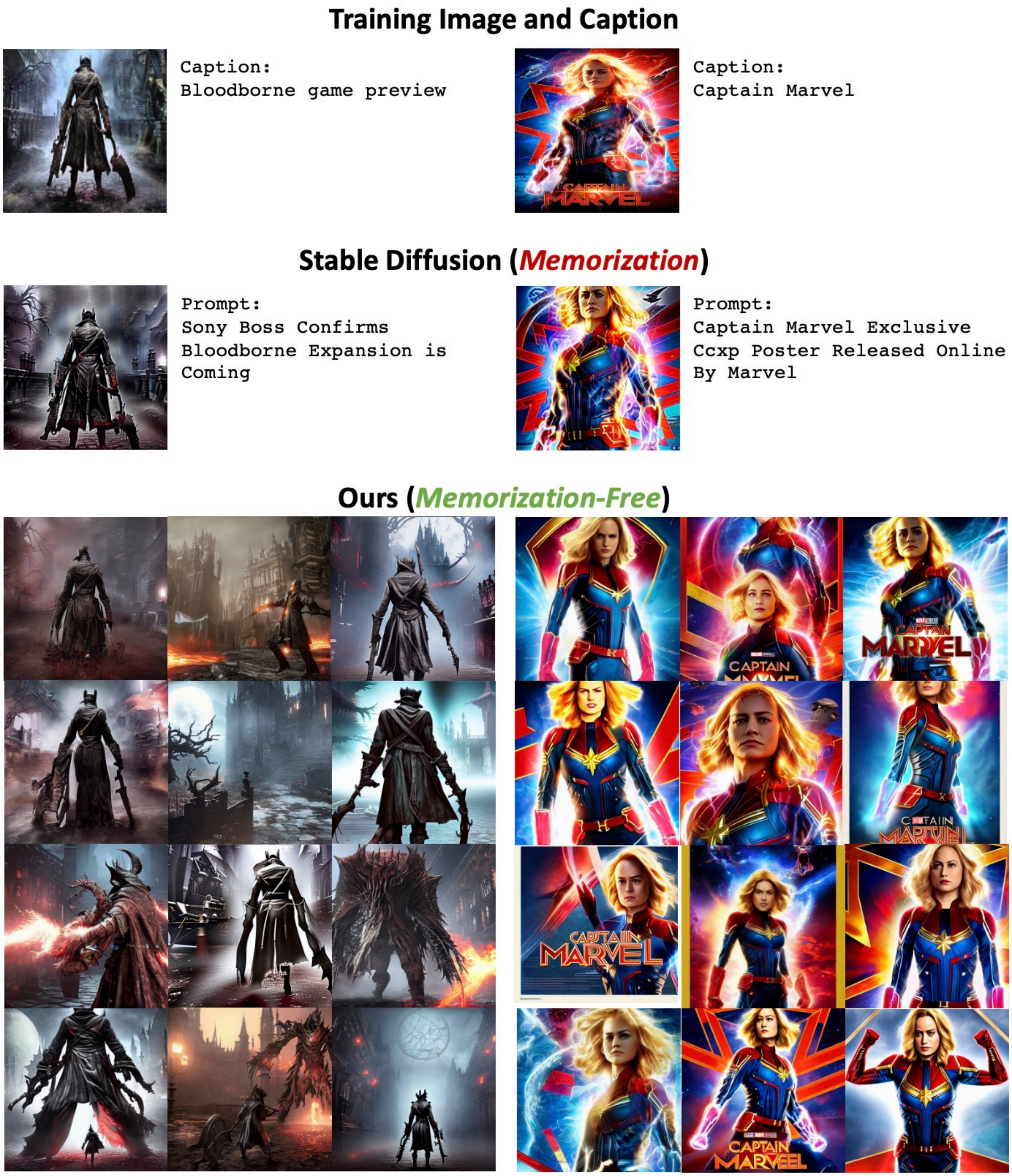

Training Image and Caption
Caption:
<i>The Long Dark</i> Gets First Trailer, Steam Early Access
Caption:
Living in the Light with Ann Graham Lotz
Stable Diffusion (Memorization)
Prompt:
<i>The Long Dark</i> Gets First Trailer, Steam Early Access
Prompt:
Ann Graham Lotz
Ours (Memorization-Free)

Figure 10. Additional qualitative comparisons showcase AMG's effectiveness in guiding pretrained diffusion models to produce memorization-free outputs.



Training Image and Caption

Caption:
Bloodborne game preview

Caption:
Captain Marvel

Stable Diffusion (*Memorization*)

Prompt:
Sony Boss Confirms Bloodborne Expansion is Coming

Prompt:
Captain Marvel Exclusive Ccxp Poster Released Online By Marvel

Ours (*Memorization-Free*)

Figure 11. Additional qualitative comparisons showcase AMG's effectiveness in guiding pretrained diffusion models to produce memorization-free outputs.