

Classifier-Free Guidance inside the Attraction Basin May Cause Memorization

Anubhav Jain^{1*}, Yuya Kobayashi², Takashi Shibuya², Yuhta Takida², Nasir Memon¹, Julian Togelius¹, Yuki Mitsufuji^{2,3}

¹New York University, ²Sony AI, ³Sony Group Corporation

{aj3281,memon,julian.togelius}@nyu.edu

{u.kobayashi,takashi.tak.shibuya,yuta.takida,yuhki.mitsufuji}@sony.com

Abstract

Diffusion models are prone to exactly reproduce images from the training data. This exact reproduction of the training data is concerning as it can lead to copyright infringement and/or leakage of privacy-sensitive information. In this paper, we present a novel way to understand the memorization phenomenon, and propose a simple yet effective approach to mitigate it. We argue that memorization occurs because of an attraction basin in the denoising process which steers the diffusion trajectory towards a memorized image. However, this can be mitigated by guiding the diffusion trajectory away from the attraction basin by not applying classifier-free guidance until an ideal transition point occurs from which classifier-free guidance is applied. This leads to the generation of non-memorized images that are high in image quality and well-aligned with the conditioning mechanism. To further improve on this, we present a new guidance technique, opposite guidance, that escapes the attraction basin sooner in the denoising process. We demonstrate the existence of attraction basins in various scenarios in which memorization occurs, and we show that our proposed approach successfully mitigates memorization. The code-base is publicly available here https://github.com/anubhav1997/mitigating_memorization.

1. Introduction

Recent advancements in image-generation models have enabled the generation of hyper-realistic images. Diffusion models [18, 22] have been at the forefront of this advancement with their ability to align image generation with different conditioning mechanisms such as text giving users more control over the output. However, a key challenge has emerged: these models can memorize and reproduce training examples verbatim (verbatim memorization) [3, 20, 21], leading to concerns over copyright infringement and leak-

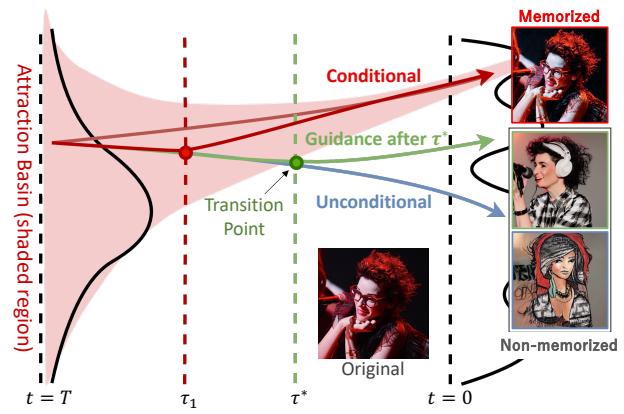
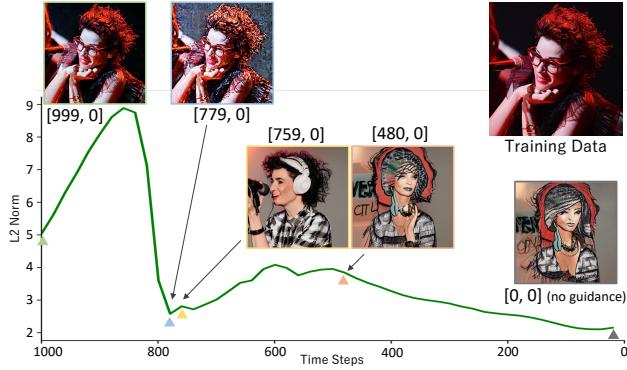


Figure 1. The diffusion trajectory contains an attraction basin (red region) which steers conditioned samples towards their memorized images. It can be avoided by applying zero classifier-free guidance when the trajectory is inside the attraction basin, such that there is an ideal transition point τ^* after which applying CFG leads to non-memorized output. Applying CFG at an earlier point such as τ_1 inside the attraction basin leads to the same memorized sample.

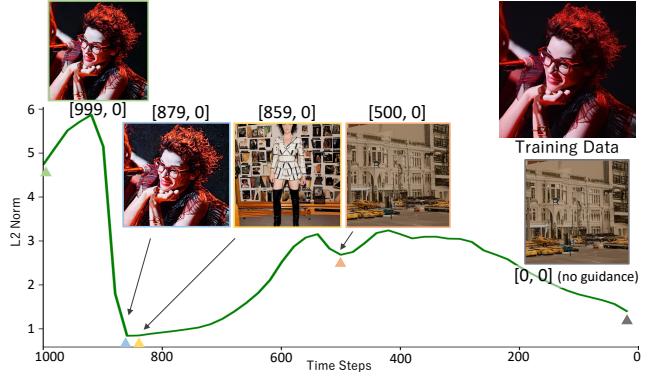
age of privacy sensitive information. Alternatively, the model can generate images copying the composition and structure present in one or more training images (template memorization)[17]. Researchers have identified factors contributing to memorization and reproduction, such as the duplication of training images and repeated captions [21]. However, not all causes are understood, as memorization can still occur even when these issues are avoided. Therefore, there is a growing need for effective mitigation strategies, in particular methods that can be applied during inference without requiring computationally intensive retraining of the core diffusion models.

In text-to-image (T2I) diffusion models, strong associations between a text prompt and an image in the training dataset lead to memorization. This can occur due to the presence of "trigger tokens" in overly specific text-prompts [17, 25], the model seeing multiple occurrences of these dur-

*Work done during an internship at Sony AI.



(a) Prompt: “Here’s What You Need to Know About St. Vincent’s Apple Music Radio Show”



(b) Prompt: “Here’s What You Need to Know About St. Vincent’s Apple Music Radio Show”

Figure 2. Plots showing magnitude of $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_0)$ when denoising without classifier-free guidance (CFG) at each time step. The figures show the generated image if you start applying CFG at that time step. We get non-memorized output if we apply CFG after the ideal transition point τ^* which coincides with the fall in the conditional noise prediction. This value is dependent on both the prompt and the initialization ((a) and (b) contain the same prompt with different initializations). More examples in the Appendix.

ing training (data duplication) [21, 25], or other causes such as fine-tuning on smaller datasets [21]. Previous approaches have focused on perturbing either the text prompt/embeddings [21, 25] or the cross-attention between the text and image embeddings [17]. While these approaches help weaken the associations between prompts and corresponding images, they are limited to specific conditions in which they have been tested and fail to generalize effectively. Additionally, since these techniques work at the prompt level, they are only applicable to text-conditional generation, and also assume the ability to update the prompt.

In this paper, we present a more generalizable understanding of the dynamics behind memorization in diffusion models. When memorization occurs, the text-conditioned noise prediction is uncharacteristically high [25], suggesting that there is a strong steering force towards the memorized output, where even different initializations lead to the same output. We show an attraction basin forms around this diffusion trajectory that steers the denoising process from any random initialization towards this memorized trajectory when applying classifier-free guidance (CFG). The attraction basin can be thought of as regions in the diffusion trajectory with very high text-conditioned noise prediction scores steering the trajectory in a specific direction. However interestingly, the attraction basin, as we show in this paper, gets narrower with every denoising time step. After a few denoising steps, the strong steering force is not present everywhere, and can only be experienced when the trajectory is inside the attraction basin. Such that if we apply CFG outside of the attraction basin it is unlikely to yield a memorized image. We can do so by initially applying no guidance until a transition point occurs where the disagreement between conditional and unconditional guidance sharply drops and applying CFG

henceforth leads to non-memorized images.

To summarize, we make the following contributions in this paper,

- We present a new way to understand memorization in conditional diffusion models by understanding the diffusion trajectory from a dynamical systems theory perspective. We show that an attraction basin forms in the denoising trajectory.
- We present approaches to avoiding the attraction basin in various scenarios that incur no additional computational costs and require neither access to the prompt nor modification of weights.
- We present a new guidance technique that we call Opposite Guidance (not to be confused with negative prompts [2]) to push the trajectory away from the attraction basin.
- We comprehensively show that previous approaches are not generalizable in mitigating memorization and only work in certain scenarios studied in those papers. On the contrary, our approach is able to mitigate memorization in all the explored scenarios.

2. Related Work

2.1. Understanding Memorization

In deep generative models, when memorization occurs, a generated sample can match verbatim with a single or set of training images (verbatim memorization) [3, 20, 21] or copy the same template as the training image (template memorization) [17] leading to privacy and/or copyright infringement issues. Researchers [3, 20, 21, 24] have identified that when images or prompts are duplicated in the training dataset, it can lead to memorization. Additionally, training on smaller datasets has also been shown to be a major factor in causing memorization [21]. However, as Somepalli et al. [21]

pointed out, even though data has been de-duplicated in the newer models [14] such as Stable Diffusion v2.1 (SDv2.1), the issue of memorization is yet to be resolved, and it does not explain much of the observed replication behavior. Factors such as conditioning and dataset complexity play a role even when the training datasets are large enough and are de-duplicated where "simpler" images and overly specific prompts have a higher chance of getting memorized [21]. Thus, there is a strong need for mitigation techniques beyond dataset manipulations.

2.2. Detecting Memorization

Similarity scores based on the SSCD (self-supervised copy detection) model [15] can be used for getting an estimate on the extent of memorization wherein scores > 0.5 can suggest that it is memorized [20]. However, this metric is prone to false positives when the textual similarity is high between a generated and real image [21]. Wen et al. [25] showed that for memorized prompts, the difference between the text-conditioned and the unconditioned score predictions is uncharacteristically high and can be informative in detecting memorization. They demonstrated high performance even in the first step of noise prediction. Ren et al. [17], on the other hand, showed that for memorized prompts, the cross-attention scores pertaining to certain "trigger tokens" in the memorized prompts are unusually high. Suggesting that they are given more importance, and thus lead to memorization. Other authors [3, 24] used a black box membership inference attack for finding memorized images. Daras et al. [7] showed that when an image is memorized, the model is able to fully reconstruct it from a noised version of it. This observation could be used for detection.

2.3. Mitigating Memorization

2.3.1. Training Time Mitigation

Wen et al. [25] proposed monitoring the text-conditioned noise prediction scores for each sample and excluding it from the current mini-batch if it surpasses a certain *pre-determined threshold*. On similar lines, Ren et al. [17] proposed removing samples from the mini-batch when their cross-attention entropy is above a particular *pre-determined threshold*. Somepalli et al. [21] proposed generating multiple BLIP captions per image during training. They also proposed perturbing the prompts by adding Gaussian noise or adding/replacing random words/numbers. Daras et al. [7] proposed training/finetuning on corrupted images can reduce instances of data replication by the model. Chavhan et al. [5] proposed pruning the UNet model weights assuming prior knowledge of memorized prompts.

Most training time mitigations require re-training the base diffusion model from scratch. Not only is this computationally expensive, it also requires control over the entire training process including access to the training datasets.

2.3.2. Inference Time Mitigation

Wen et al. [25] proposed perturbing the prompt embedding by minimizing the difference between the text-conditioned and unconditioned noise prediction scores such that it falls below a certain target loss. Ren et al. [17] proposed applying a mask and rescaling the attention scores to give less attention to "trigger tokens". However, this approach does not mitigate memorization when "trigger tokens" are not present. Somepalli et al. [21] proposed adding random tokens to the text prompt or repeating certain tokens during inference so as to add noise to the text embedding indirectly. Chen et al. [6] proposed different CFG weight schedulers to mitigate different causes of memorization. The approach presumes knowledge of not only whether memorization is present but also the type of memorization.

Most of these approaches work by updating text prompt/embeddings leads to a trade-off between the text alignment and memorization. Another drawback of these approaches is that they work better when a trigger token is present such that we can reduce the importance/weightage of that token and they do not generalize well to other types of memorization as we show in this paper. Lastly, since they work directly on the text prompt/embeddings, they cannot be applied to other conditioning mechanisms such as class labels.

3. Preliminaries

Diffusion models [22] such as Stable Diffusion (SD) [18] and Imagen [19] are trained with the objective of learning a model ϵ_θ to denoise a noisy input vector at different levels of noise characterized by a time-dependent noise scheduler. We can define the diffusion process as a stochastic differential equation (SDE) in the form,

$$dx_t = f(x_t, t)dt + g(t)dw_t.$$

where $x_t \in X$ is a sample at time t , $f(x_t, t)$ is the drift term, $g(t)$ is the diffusion coefficient, and w_t is a standard Wiener process.

During training, the forward diffusion process comprises a Markov chain with fixed time steps T . Given a data point $x_0 \sim q(x)$, we iteratively add Gaussian noise with variance β_t at each time step t to x_{t-1} to get x_t such that $x_T \sim \mathcal{N}(0, I)$. This process can be expressed as,

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t}x_{t-1}, \beta_t I), \quad \forall t \in \{1, \dots, T\}.$$

We can get a closed-form expression of x_t ,

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, \quad (1)$$

where $\bar{\alpha}_t = \prod_{i=1}^t (1 - \beta_i)$ and $\alpha_t = 1 - \beta_t$.

During training, we learn the reverse process through a network ϵ_θ to iteratively denoise x_t by estimating the noise ϵ_t at each time step t . The loss function is expressed as,

$$\mathcal{L} = \mathbb{E}_{t \in [1, T], \epsilon \sim \mathcal{N}(0, I)} [\|\epsilon_t - \epsilon_\theta(x_t)\|_2^2]. \quad (2)$$

Using the learned noise estimator network ϵ_θ , we can compute the previous sample x_{t-1} from x_t as follows,

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} (x_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(x_t)). \quad (3)$$

The learned noise estimator network ϵ_θ can be conditioned using a conditioning input such as text or class label, expressed as an embedding e_p . Ho et al. [9] proposed classifier-free guidance (CFG) as a mechanism to guide the diffusion trajectory towards generating outputs that align with the conditioning. The trajectory is directed towards the conditional score predictions and away from the unconditional score predictions, where s controls the degree of adjustment and e_\emptyset are empty prompt embeddings used for unconditional guidance.

$$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, e_\emptyset) + s \underbrace{(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset))}_{\text{conditional guidance}}. \quad (4)$$

4. Understanding Diffusion Trajectory during Memorization

In text-to-image diffusion models, different initializations generally lead to different outputs for the same text prompt. However, when the diffusion model has memorized a particular sample, the outputs are similar and closely resemble one or more training data samples regardless of initialized noise [25]. In such scenarios, the conditional guidance $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ used to predict the next sample x_{t-1} from the current sample x_t (Eq. 3) becomes uncharacteristically high, i.e. there is a strong force that steers the diffusion trajectory towards a specific sample (training sample). Interestingly, when observing this value $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ after a few denoising steps, the strong steering force is not always present, i.e. only subset of the sample space has uncharacteristically high values. Looking at the diffusion trajectory from the perspective of trajectories in dynamical systems, this observation suggests that in a certain region, there is an attraction basin [1, 13], and if the latent x_t at any time step t falls within this region, applying CFG will lead to the memorized output. However, if x_t is not in the attraction basin, applying CFG with any guidance weight s will not lead to the memorized output. Intuitively, one can imagine the attraction basin having a funnel shape, where it is broader at $t = T$ and becomes narrower as the reverse denoising process progresses. To define the attraction basin, we introduce the concept of a denoiser.

Definition 1 (Denoiser). *Let X denote the entire sample space of the diffusion trajectory, and let E denote the embedding space. A denoiser is a function $\varphi : X \times (0, T] \times E \rightarrow X$ that outputs the result of the reverse diffusion process based on Eq. (3) with CFG. Specifically, $\varphi(x, t, e)$ represents the resulting sample at time $t = 0$, starting from a state (x, t) with embeddings $e \in E$, where $x \in X$ and $t \in (0, T]$.*

For an attractor x^a from the training dataset, we then define the attraction basin as follows.

Definition 2 (Attractor and attraction basin). *Suppose a prompt is given with embeddings denoted as e_p . A sample $x^a \in X$ is considered an (ϵ -)attractor if the diffusion trajectories denoised with CFG tend to converge to samples within $B_D(x^a, \epsilon)$ at $t = 0$, regardless of the initial noise, i.e., $\varphi(x, T, e_p) \in B_D(x^a, \epsilon)$ for any $x \in X$, where $B_D(\cdot, \cdot)$ represents a perceptual ball given by a certain perceptual distance $D : X \times X \rightarrow \mathbb{R}_{\geq 0}$ as*

$$B_D(x^a, \epsilon) = \{x' \in X \mid D(x^a, x') \leq \epsilon\}.$$

An attraction basin is a region in the state space within which samples will converge to around x^a at $t = 0$ under CFG inference. More formally, the $((\epsilon, \delta)$ -)attraction basin defined using x^a is expressed as

$$X^b(x^a, \epsilon) = \{(x, t) \mid \mathbb{P}(\varphi(x, t, e) \in B_D(x^a, \epsilon)) > 1 - \delta\},$$

representing all points in the state space from which generated samples using CFG will reach samples perceptually similar to the attractor at $t = 0$.

A simple way to observe the attraction basin is by applying zero CFG ($s = 0$ in Eq. 4), such that you are not considering the text-prompt at all and guiding the denoising process towards an approximation of the training distribution i.e. $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, e_\emptyset)$. Interestingly, as shown in Figure 2 the conditional noise prediction $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ remains high in the initial time steps and then steeply falls. We refer to the region of high conditional noise prediction as the attraction basin, and the fall in conditional noise prediction is associated with the denoising trajectory leaving the attraction basin. The time step at which the denoising trajectory leaves the attraction basin is referred to as the transition point from getting a memorized output to a non-memorized output. In Figure 2(a), if you were to start applying CFG at any point on or before $t = 779$, you get the same memorized output. However, surprisingly, applying CFG at exactly one denoising step after this $t = 759$ leads to a non-memorized output (step size is 20 when denoising with 50 inference steps). We define the transition point as follows.

Definition 3 (Transition point). *Given prompt embeddings, let $x^a \in X$ be an attractor. We define a transition point as (x_τ, τ) for $x_\tau \in X$ and $\tau > 0$ such that $(x_\tau, \tau + 1)$ lies within its attraction basin, while (x_τ, τ) does not. More formally, an $((\epsilon, \delta)$ -)transition point satisfies*

$$\begin{aligned} \mathbb{P}(\varphi(x_\tau, \tau + 1) \in B_D(x^a, \epsilon)) &> 1 - \delta \quad \text{and} \\ \mathbb{P}(\varphi(x_\tau, \tau) \notin B_D(x^a, \epsilon)) &< 1 - \delta, \end{aligned}$$

where B_D is a perceptual ball defined in Definition 2.

This phenomenon occurs only for memorized samples as illustrated in Figure 6, the conditional guidance $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ remains low and almost flat across different time steps for non-memorized samples. We further analyze the diffusion trajectory during memorization, demonstrating the presence of an attraction basin in Appendix 8.

5. Simple Mitigation Strategy

Based on the intuition from the previous section, we can simply mitigate memorization by finding an ideal transition point to switch from zero CFG to applying CFG. Based on our preliminary experiments, we found that transition points can either be static or dynamic.

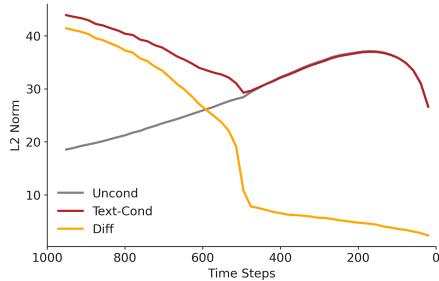


Figure 3. Average magnitude of the text-conditioned ($\epsilon_\theta(x_t, e_p)$) and unconditional noise predictions ($\epsilon_\theta(x_t, e_\emptyset)$) and their difference ($\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$) when applying zero CFG. We see a static transition point ($t = 500$) appear when SDv2.1 is finetuned on the LAION-10k dataset [21].

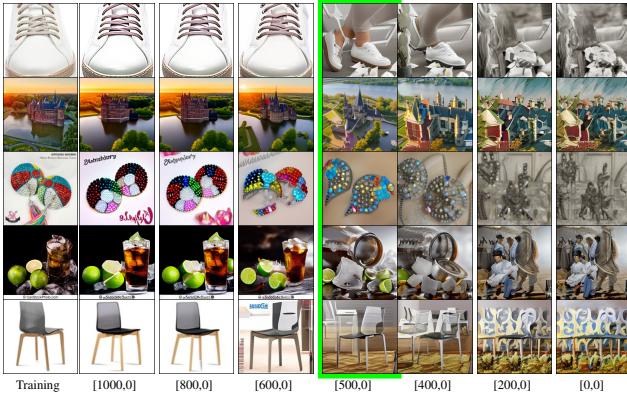


Figure 4. Applying CFG before the static transition point ($T=500$) leads to memorized outputs while applying CFG after the fixed transition point leads to non-memorized outputs. Applying CFG too late results in poor quality images that resemble the unconditional generations.

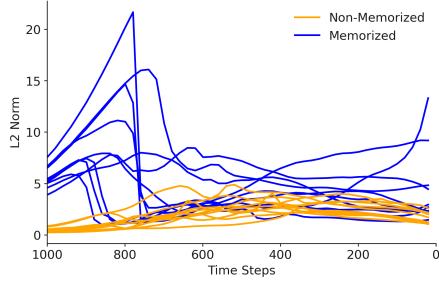
Static Transition Point. In some trained diffusion models, we observed a universal trend in the mean conditional $\epsilon_\theta(x_t, e_p)$ and unconditional noise prediction $\epsilon_\theta(x_t, e_\emptyset)$ across different samples. As shown in Figure 3, the mean



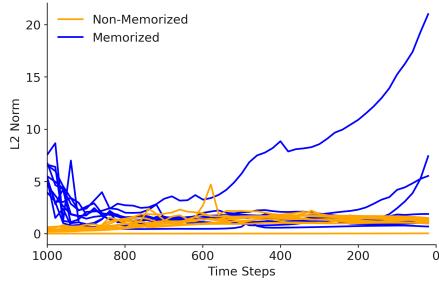
Figure 5. In some models, transition points can occur at a different time step for each sample, as seen for pre-trained SDv1.4. In row 1 the transition point is approximately $t = 800$ while for row 2 it is $t = 650$.

conditional noise prediction $\epsilon_\theta(x_t, e_p)$ is higher than unconditional noise prediction $\epsilon_\theta(x_t, e_\emptyset)$ in the initial time steps ($t \geq 500$). At a static time step ($t = 500$), the two scores align in magnitude until the end of the denoising process, and at the same time, the L_2 distance between them drastically drops. This suggests that the zero CFG trajectory leaves the attraction basin at that fixed time step ($t = 500$). This can be further validated by applying CFG from any time step before the fixed transition point ($t = 500$), as illustrated in Figure 4. Applying CFG from an early time step leads to the same memorized image, but the output immediately switches at the transition point ($t = 500$) such that we generate text-aligned non-memorized images. On the other hand, if we start applying CFG very late in the denoising process, the outputs are not well aligned with the text prompts and resemble their unconditional generations.

Dynamic Transition Point. In other trained diffusion models, we saw that every prompt and initialization pair leads to a different transition point. As shown in Figures 6(a) and 5, different trajectories leave the attraction basin at different time steps, characterized by their fall in the conditional guidance $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$. We found that it is dependent on both the conditional embedding e_p and the initialization x_T . We show an example in Figures 2(a) and 2(b), where the transition point occurs at different time steps even for the same prompt. Thus, we propose a method to find this dynamic transition point (DTP). As shown in Figure 2, it is the point after the first local minima in the graph. Thus to mitigate memorization, this difference is tracked until the first local minimum occurs, and CFG is applied henceforth. This process requires no additional computations as the text-conditioned and unconditional noise predictions are computed by default when denoising using CFG. We present the pseudo-code for this approach in Algorithm 1.



(a) Applying zero CFG



(b) Applying opposite guidance

Figure 6. Magnitude of $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ when denoising (a) without CFG and (b) with opposite guidance. Transition point occurs sooner with opposite guidance.

Algorithm 1 Reverse Diffusion with Dynamic Transition Point Method and Opposite Guidance

Require: $x_T \sim \mathcal{N}(0, I_d)$, $\lambda > 0$, $v_{OG} \in \{0, 1\}$

- 1: $d_{T+2} = -\infty$; $d_{T+1} = -\infty$
- 2: $s = -\lambda * v_{OG}$
- 3: **for** $t = T$ **to** 1 **do**
- 4: $d_t = \|\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)\|_2^2$
- 5: **if** $d_{t+2} > d_{t+1}$ **and** $d_{t+1} < d_t$ **then**
- 6: $s = \lambda$
- 7: **end if**
- 8: $\hat{\epsilon} = \epsilon_\theta(x_t, e_\emptyset) + s(\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset))$
- 9: $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\hat{\epsilon})$
- 10: **end for**
- 11: **return** x_0

5.1. Opposite Guidance

An issue with this method is that when the transition point occurs very late ($t \leq 500$) in the denoising process, CFG is applied for fewer time steps, and this can impact the image quality.

To ensure that the transition point occurs earlier, we introduce a new concept of opposite guidance (OG). Unlike negative prompting, we apply negative or opposite CFG, to push the diffusion trajectory in the opposite direction to that of traditional CFG and thus, push the trajectory away from the attraction basin of the traditional CFG trajectory sooner in the denoising process, as also shown in Figure 6(b). Opposite guidance can be expressed as follows for $s > 0$,

$$\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, e_\emptyset) - s(\underbrace{\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)}_{\text{conditional guidance}}). \quad (5)$$

Similar to unconditional denoising, when denoising using opposite guidance, there is an ideal time step where the magnitude of conditional guidance falls. At this time step, switching from opposite guidance to traditional positive guidance leads to non-memorized samples with high image quality and text alignment. Thus, it can easily be integrated along with the previous method to find a static or dynamic transition point. We present the pseudo-code for this approach in conjunction with the DTP method in Algorithm 1 with the parameter $v_{OG} = 1$.

6. Memorization under Different Scenarios

We study memorization under different scenarios and show how this simple approach of finding an ideal transition point to avoid the attraction basin can be applied to effectively mitigate memorization.

- Scenario 1: Fine-tuned SDv2.1 on 10,000 LAION datapoints (followed by [21]). Memorization occurs due to overfitting on the small dataset.
- Scenario 2: Fine-tuned SDv2.1 on the Imagenette dataset [10] containing 10 ImageNet classes (followed by [21]). Exact memorization is not always observed but similarity with the training dataset increases.
- Scenario 3: Fine-tuned SDv1.4 on 200 prompts duplicated 200 times with an additional 120,000 prompts that were not duplicated (followed by [25]). Memorization occurs due to data duplication.
- Scenario 4: 500 memorized prompts for pre-trained SDv1.4 found using membership inference attack [3, 24] (followed by [17]). The exact cause of memorization for each prompt is unknown, but the strong presence of "trigger tokens" is noticeable. These are in the form of nouns such as celebrity and movie names with little to no additional descriptions. Results are in Appendix 9.

Notably, different researchers have independently studied these different scenarios, but, as we show later, their approaches do not generalize to other scenarios as memorization occurs due to different causes in each scenario. Refer to Appendix 11 for other implementation details.

Evaluation Metrics We utilize the similarity metric based on the SSDC [15] embeddings to judge the level of memorization [21]; the CLIP score [16] to assess how well the generated images align with the text prompts; and the Fréchet Inception Distance (FID) [8] for image quality.

6.1. Scenario 1

We observed that in this scenario, a static transition point at $t = 500$ occurs as shown in Figure 3. We present qualitative

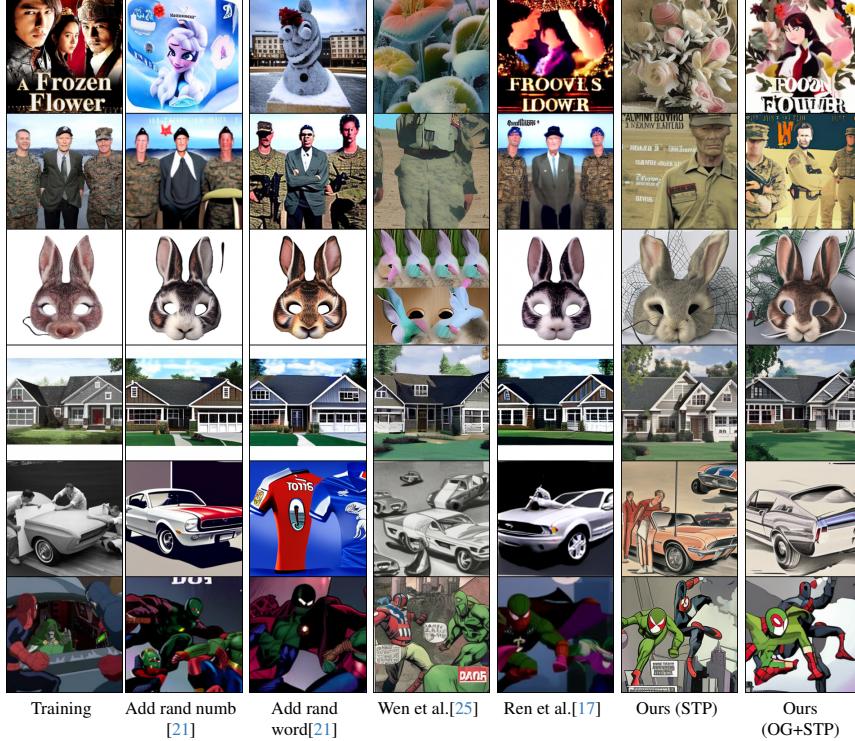


Figure 7. Qualitative results comparing the proposed approach with the baselines in Scenario 1. Ren et al. [17] is unable to mitigate memorization in most cases while Wen et al. [25] has poor image quality and text-alignment. More enlarged examples (Fig. 14) and prompts for this figure are in Appendix 12.

Table 1. Comparison of results on Scenario 1. None of the other approaches perform well. The only comparable method [25] qualitatively results in poorer quality images (see visual examples in Fig. 7).

	Similarity (95pc)	CLIP Score	FID
No Mitigation	0.6504	0.3027	16.8373
Add rand word [21]	0.5254	0.2941	17.4142
Add rand numb [21]	0.5416	0.2993	17.0245
Wen et al. [25]	0.3853	0.2895	16.7176
Ren et al. [17]	0.6028	0.2959	20.2931
Ours (STP)	0.2857	0.2976	19.8494
Ours (OG + STP)	<u>0.3811</u>	0.3020	15.6679

6.2. Scenario 2

Even in this scenario, as illustrated in Figure 10, a static transition point occurs. We applied CFG from $t = 700$ and $t = 600$ to show that we can reduce the similarity with the training dataset at no cost to the image quality, as shown in Table 2. We would like to point out that previous approaches are specific to text-to-image models and thus cannot be extrapolated to other conditioning mechanisms. This shows a further advantage of our CFG-based mitigation technique.

Table 2. Results on Scenario 2. The value in the brackets in the time steps between which we apply CFG. Applying CFG later from $t = 700$ (row 2) and $t = 600$ (row 3) reduces similarity with comparable or improved FID score.

	Similarity (95pc)	FID
No mitigation	0.3702	43.35
CFG=7.5 [700,0]	0.3018	38.86
CFG=7.5 [600,0]	0.2756	45.73

6.3. Scenario 3

In this scenario, we observed that a dynamic transition point occurs. We experimentally validate our approach of DTP method and opposite guidance qualitatively and quantita-

and quantitative results in Figure 7 and Table 1, respectively. Our approach outperformed all previous methods, and we observed the previously proposed techniques that focused solely on the presence of "trigger tokens" [17, 25] did not work well. This shows that when memorization occurs due to factors other than data duplication or the presence of trigger tokens, these methods may fail. We present further results when finetuning on the LAION-100k dataset [21] in Appendix 10.

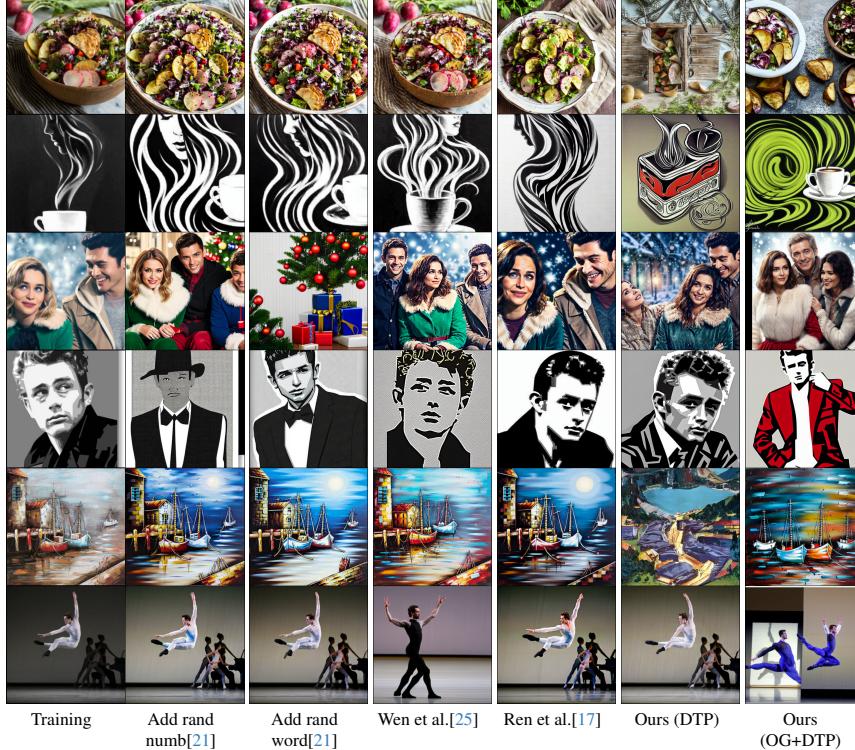


Figure 8. Qualitative results comparing the proposed approach with the baselines in Scenario 3. Prompts used to generate these figures are given in the Appendix. Rows 1, 2, 5 and 6 show examples, where for most baselines the output images remain closely related to their memorized versions. More enlarged examples (Fig. 15) and prompts for this figure are in Appendix 13.

Table 3. Table showcasing results in Scenario 3. Wen et al. [25] studied this scenario and is the only other method which performs well but does not generalize to other scenarios (see Table 1).

	Similarity 95pc	Mean ± Std	CLIP Score	FID
No Mitigation	0.7977	0.5513 ± 0.16	0.3105	106.49
Add rand word [21]	0.7480	0.4312 ± 0.21	0.3071	116.92
Add rand numb [21]	0.7366	0.3850 ± 0.21	0.3027	126.54
Wen et al. [25] ($l_{target}=3$)	0.7747	0.5147 ± 0.17	0.3100	109.28
Wen et al. [25] ($l_{target}=1$)	0.6038	0.2808 ± 0.16	0.3050	136.34
Ren et al. [17]	0.6881	0.4036 ± 0.16	0.3066	124.38
Ours (DTP)	0.5885	0.2866 ± 0.16	0.3020	138.92
Ours (OG + DTP)	0.6915	0.2844 ± 0.20	0.2910	140.05

tively in Figure 8 and Table 3. We saw that our method performs at par with [25] that studied this specific scenario and even outperforming it in terms of memorization in the top 95 percentile. Approaches by Ren et al. [17] and Somepalli et al.[21] that were proposed for other scenarios do not work well showcasing their lack of generalizability.

Inference Time. One of the major advantages of our approach is its simplicity and lack of computational time overhead. Image generation using our approach as well as standard CFG-based image generation using SDv2.1 on one A100 GPU takes 1.26 seconds, while in comparison the mitigation technique proposed by Wen et al. [25] took 2.86

seconds and that of Ren et al. [17] took 2.08 seconds. These are the average values across 10,000 generations.

Limitations. As transition points do not exist for non-memorized samples, we require detecting memorization before we can apply our approach. This can either be done by denoising twice, once with traditional CFG to look at the trends in the magnitude of conditional guidance and the second time to apply our mitigation technique if the prompt is memorized. Alternatively, as [17, 25] showed, we can do detection at $t = 0$ by simply looking at the magnitude of conditional guidance $\|\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_0)\|_2^2$. If this magnitude is above a certain threshold it indicates that the sample is memorized with high accuracy and we can apply our guidance accordingly.

7. Conclusion

We present a novel understanding of the dynamics behind memorization in conditional diffusion models. We showcase the presence of an attraction basin that steers randomly initialized latent vectors towards a memorized output. We propose an approach to steer away from the attraction basin and detect the point at which the trajectory leaves the attraction basin referred to as the transition point. We show that applying CFG after the transition point leads to non-

memorized outputs. We presented results in various scenarios such as the presence of data replication, fine-tuning on smaller datasets, and memorized examples found in existing pre-trained models to showcase its efficiency.

To summarize, our method is simple, works well across all tested scenarios both in terms of image quality metrics and visual inspection, requires less computational time, and does not require altering the prompt.

References

- [1] Joseph Auslander, Nam P Bhatia, and Peter Seibert. Attractors in dynamical systems. Technical report, 1964. [4](#)
- [2] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? *arXiv preprint arXiv:2406.02965*, 2024. [2](#)
- [3] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. [1, 2, 3, 6](#)
- [4] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. [1](#)
- [5] Ruchika Chavhan, Ondrej Bohdal, Yongshuo Zong, Da Li, and Timothy Hospedales. Memorized images in diffusion models share a subspace that can be located and deleted. *arXiv preprint arXiv:2406.18566*, 2024. [3](#)
- [6] Chen Chen, Daochang Liu, and Chang Xu. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8425–8434, 2024. [3](#)
- [7] Giannis Daras, Alexandros G Dimakis, and Constantinos Daskalakis. Consistent diffusion meets tweedie: Training exact ambient diffusion models with noisy data. *arXiv preprint arXiv:2404.10177*, 2024. [3](#)
- [8] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. [6](#)
- [9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. [4](#)
- [10] Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet, 2019. [6, 3](#)
- [11] Tuomas Kynkänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. *arXiv preprint arXiv:2404.07724*, 2024. [1](#)
- [12] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022. [3](#)
- [13] John W. Milnor. Attractor. *Scholarpedia*, 1(11):1815, 2006. [4](#)
- [14] Alex Nichol. Dall-e 2 pre-training mitigations. <https://openai.com/index/dall-e-2-pre-training-mitigations/>, 2022. [Accessed 10-10-2024]. [3](#)
- [15] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14532–14542, 2022. [3, 6](#)
- [16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [6](#)
- [17] Jie Ren, Yixin Li, Shenglai Zen, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. *arXiv preprint arXiv:2403.11052*, 2024. [1, 2, 3, 6, 7, 8](#)
- [18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. [1, 3](#)
- [19] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. [3](#)
- [20] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023. [1, 2, 3](#)
- [21] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023. [1, 2, 3, 5, 6, 7, 8](#)
- [22] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. [1, 3](#)
- [23] Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *arXiv preprint arXiv:2404.13040*, 2024. [1](#)
- [24] Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023. [2, 3, 6](#)
- [25] Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. [1, 2, 3, 4, 6, 7, 8](#)

Classifier-Free Guidance inside the Attraction Basin May Cause Memorization

Supplementary Material

We present the following contents in the Appendix:

- Additional analysis on the attraction basin in Section 8.
- Experimental results on Scenario 4 that was discussed in the main text in Section 9.
- Additional analysis of Scenario 1, where we provide experimental results when SDv2.1 is finetuned on LAION-100K dataset in Section 10.
- Figure on Scenario 2, showing the occurrence of a static transition point in Figure 10.
- Prompts used in Figure 7 and Figure 8 in Sections 12 and 13 respectively.
- Additional example of the transition point coinciding with the fall in conditional guidance in Figure 13.
- More visual results on different scenarios, comparing with baselines in section 14.

8. Additional Analysis of the Attraction Basin

In the paper, we discuss observing the attraction basin when applying zero CFG in the denoising process by observing the magnitude of $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$. Now the question arises, what is the trend when denoising with CFG? Does the value still drop after a particular time step? We show in Figure 9 that the magnitude of $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ remains high throughout the denoising process when we denoise using CFG. Further validating our observation, as in this case, the sample will be inside the attraction basin throughout the denoising process.

Interestingly, our observations also align with previous studies on improving diversity and fidelity in diffusion models where they showed the negative impacts of high CFG in the initial time steps [4, 11, 23] by showing that monotonically increasing CFG weight schedulers lead to improved performance. These studies, however, are not in the context of memorization.

9. Experimental Results on Scenario 4

For Scenario 4, where memorization occurs due to the presence of trigger words, we observed a dynamic transition point. We apply the same approach as present in Alg. 1.

Experimental Results: We compare our approach with previous baselines in Figure 11 and Table 4, and show that our simple approach is able to mitigate memorization in this scenario as well. [17] had initially studied this scenario and we report comparable similarity results while still being generalizable to other scenarios. Our opposite guidance and dynamic transition point method yield a 0.2611 similarity score as compared to theirs of 0.2544. Additionally, we do

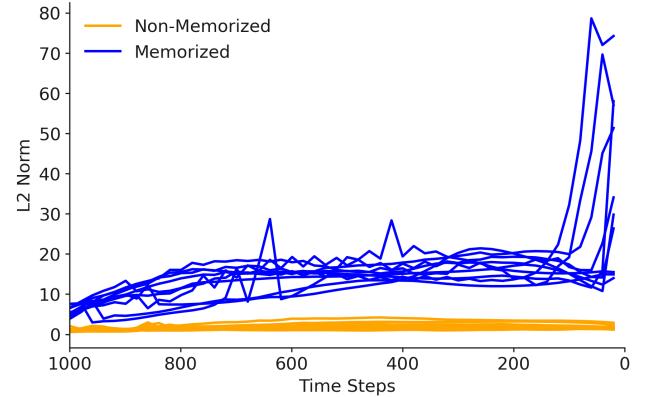


Figure 9. When you apply CFG from the beginning, the magnitude of $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ remains high throughout.

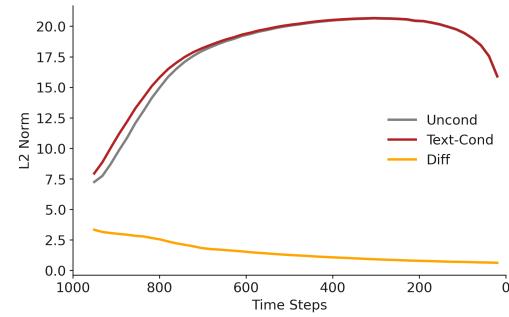


Figure 10. Mean conditional and unconditional noise predictions when SDv2.1 is finetuned on the Imagenette dataset.

Table 4. Results on Scenario 4. Ren et al. [17] studied this scenario.

	Similarity 95pc	Mean \pm Std	CLIP Score	FID
No mitigation	0.9262	0.5508 ± 0.31	0.3144	155.35
Add rand word [21]	0.9049	0.3779 ± 0.29	0.3046	143.10
Add rand numb [21]	0.8934	0.3740 ± 0.28	0.3068	146.01
Wen et al. [25] ($l_{target} = 3$)	0.9155	0.4913 ± 0.31	0.3134	154.70
Wen et al. [25] ($l_{target} = 1$)	0.9011	0.3287 ± 0.27	0.3088	147.74
Ren et al. [17]	0.6718	0.2544 ± 0.18	0.3110	148.93
Ours (DTP)	0.8722	0.3001 ± 0.25	0.2897	169.67
Ours (OG + DTP)	0.8680	0.2611 ± 0.24	0.2873	155.41

not deteriorate the image quality as observed by the FID score of 155 which is the same as the FID score without any mitigation strategy. Other approaches such as [25] and [21] lead to poorer similarity scores.

10. Results when Finetuning SDv2.1 on LAION-100k

We observed that finetuning SDv2.1 on the LAION-10k dataset [21] leads to exact memorization when using the



Training

Add rand
numb[21]Add rand
word[21]

Wen et al.[25]

Ren et al.[17]

Ours (DTP)

Ours (OG +
DTP)

Figure 11. Qualitative results comparing the proposed approach with the baselines in Scenario 4. Prompts - (a) "Listen to The Dead Weather's New Song, ""Buzzkill(er)"""; (b) 2020 Honda FourTrax Foreman Rubicon 4x4 Automatic DCT in New Haven, Connecticut - Photo 1; (c) "Listen to Ricky Gervais Perform ""Slough"" as David Brent"; (d) Gabriel García Márquez's Collection Is Going to Austin; (e) Emma Watson to play Belle in Disney's <i>Beauty and the Beast</i>"

same text prompts. However, on the LAION-100k dataset, there are stylistic similarities in the output, but verbatim memorization is not always present. In the former scenario, a major proportion of prompts lead to memorized outputs that are extremely similar to the original training images (similarity score > 0.5). Thus, we choose to focus on mitigating memorization when finetuning on 10,000 samples. However, for both these dataset sizes, the disparity in the text-conditioned and unconditional scores appears in the initial time steps as visualized in Figure 3(a) for LAION-10k and 12(a) for LAION-100k.

Applying our approach to a SDv2.1 finetuned on the LAION-100k dataset shows similar improvements in similarity scores. We summarize the results in Table 5.

Table 5. Results when fine-tuning SDv2.1 on LAION-100k.

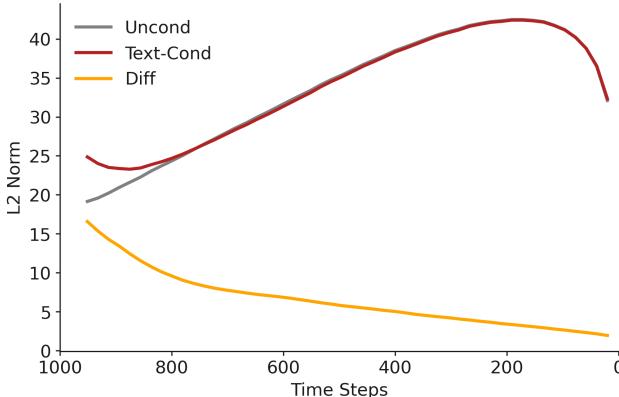
	Similarity (95pc)	CLIP Score	FID
No mitigation	0.3952	0.314	11.46
Ours (STP)	0.2861	0.309	16.06

11. Detailed Experiment Settings

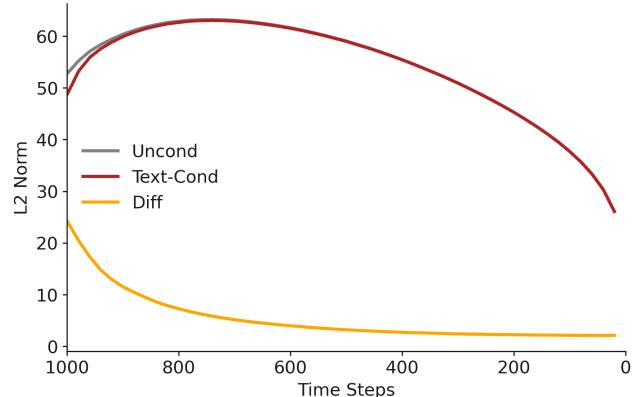
11.1. Scenario 1

We finetune Stable Diffusion v2.1 on 10,000 examples from the LAION dataset, publicly available here ¹. The model was finetuned with image sizes 256x256 for 100,000 steps to allow it to memorize the small dataset entirely. During inference, it was observed that the same prompts as the

¹https://drive.google.com/drive/folders/1TT1x1yT2BmZNXuQPg7gqAhxN_fWCD_



(a) SDv2.1 fine-tuned on LAION-100k



(b) SDv2.1 without fine-tuning

Figure 12. Plots depicting the trends in the L_2 norms of the text-conditioned noise predictions, unconditional noise predictions, and their difference when applying zero CFG during the denoising steps. We see universal transition points appear when SDv2.1 is finetuned on larger datasets as well such as LAION-100k, but this is not visible in the pretrained model.

training dataset lead to similar outputs. This memorization scenario was initially studied by Somepalli et al. [21]. We followed the same inference protocol with 50 inference steps using the DPM Multi-step solver [12].

11.2. Scenario 2

We finetuned Stable Diffusion v2.1 on the Imagenette dataset [10] comprising 10 classes of the full ImageNet dataset. These classes are - bench, English springer, cassette player, chain saw, church, French horn, garbage truck, gas pump, golf ball, and parachute. Similar to Scenario 1, this was initially studied by [21]. We finetuned the model for 40,000 steps which led to the best memorization vs image quality trade-off. All images were given the same prompt template, *An image of {Object}*, where *Object* is an ImageNet class. During sampling, we used the DPM multi-step solver [12] for 50 inference steps.

11.3. Scenario 3

We used the pre-trained model weights from Wen et al. [25] available publicly². The model was finetuned using 200 samples that were duplicated 200 times with an additional 120,000 samples from the LAION dataset. We only expect the 200 duplicated samples to be memorized. During image generation, we used the DPM multi-step solver [12] for 50 inference steps.

11.4. Scenario 4

Webster et al. [24] had found 500 prompts memorized by the pre-trained Stable Diffusion v1.4. We used this prompt dataset directly. Similar to other scenarios for sampling images, we used DPM multi-step solver [12] for 50 inference steps. Since the memorized samples were found in a

pre-trained model, the exact cause of memorization is unknown. We observed a large number of template memorization samples as well in addition to verbatim memorization. We observed a strong presence of trigger tokens in these memorized prompts, where the outputs were closely related to either celebrity, movie, or book titles.

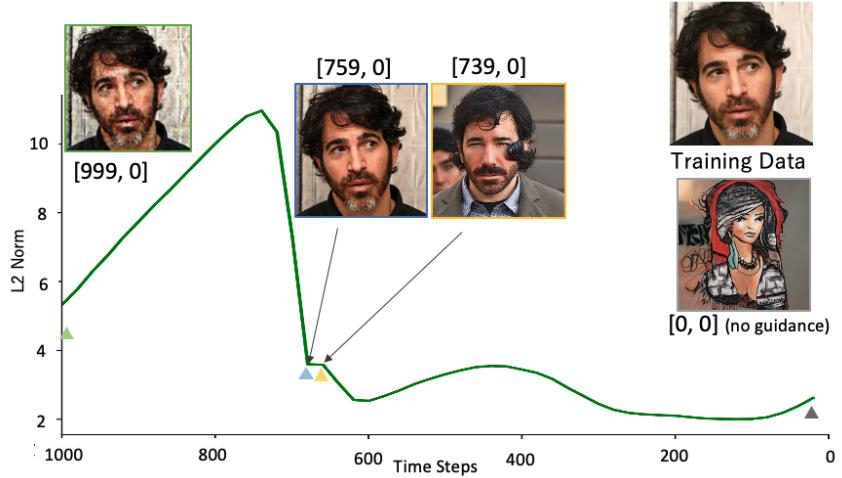
12. Prompts for Figure 7

- 'Frozen Flower A'
- 'Clint Eastwood - Camp Pendleton'
- 'Adult Kids Half Face Rabbit Bunny Mask for Halloween/Easter/Masquerade/Carnival/Party-Luckyfine'
- 'Eplans Craftsman House Plan Open Layout With Flex Space'
- 'How the Mustang got its clothes'
- 'Spidey and Cap team up against Doctor Doom'

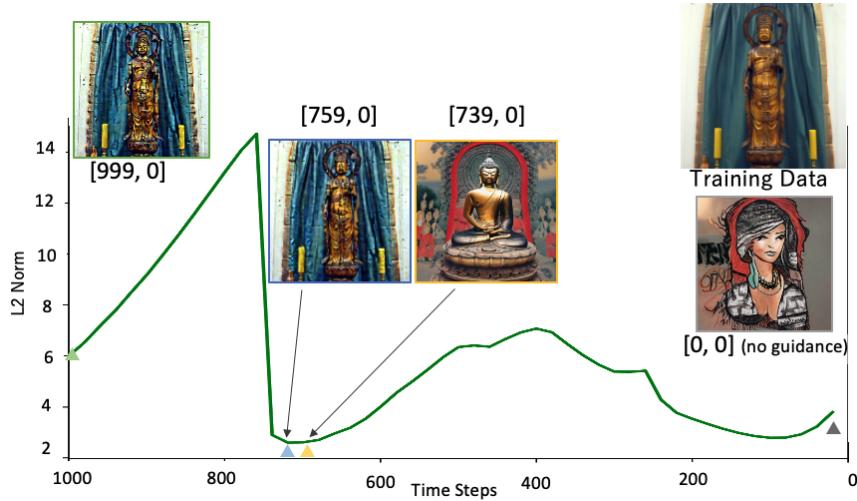
13. Prompts for Figure 8

- Fattoush Salad with Roasted Potatoes
- illusion art step by step ; Illusion Kunst, Illusion Art, Illusion Paintings, Coffee Drawing, Coffee Art, Coffee Time, Coffee Shop, Coffee Cups, Pencil Art Drawings
- Christmas Comes Early to U.K. Weekly Home Entertainment Chart
- James Dean In Black And White Greeting Card by Douglas Simonson
- 3D Metal Cornish Harbour Painting
- "In this undated photo provided by the New York City Ballet, Robert Fairchild performs in ""In Creases"" by choreographer Justin Peck which is being performed by the New York City Ballet in New York. (AP Photo/New York City Ballet, Paul Kolnik)"

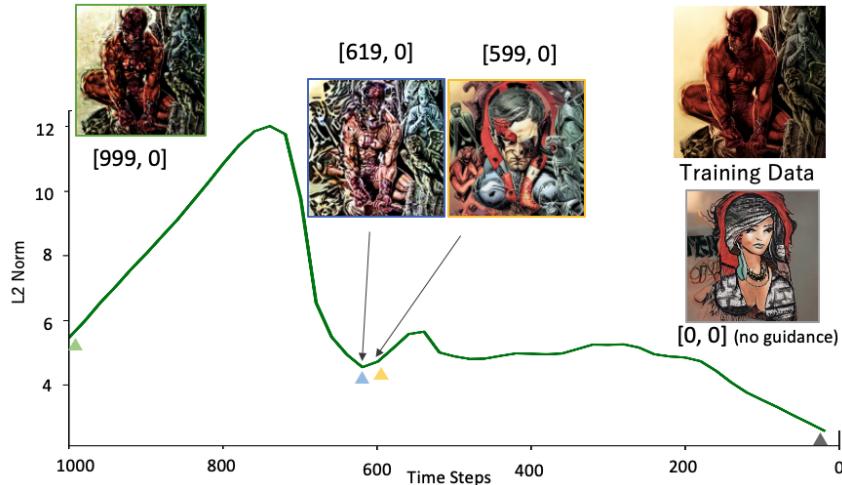
²https://drive.google.com/drive/folders/1XiYtYySpTUmS_9OwojNo4rsPbfCQK



(a) Prompt: Chris Messina In Talks to Star Alongside Ben Affleck in <i>Live By Night</i>



(b) Prompt: Talks on the Precepts and Buddhist Ethics



(c) Prompt: As Punisher Joins <i>Daredevil</i> Season Two, Who Will the New Villain Be?

Figure 13. Examples show the transition of images from memorized to non-memorized if we apply CFG starting from an ideal transition point.

14. More Visual Examples

In this section, we provide more visual examples comparing our approach with baselines to showcase the effectiveness of our approach in mitigating memorization. We provide examples for Scenario 1 in Figure 14, Scenario 2 in Figure 17, Scenario 3 in Figure 15 and Scenario 4 in Figure 16.



Figure 14. Qualitative results comparing the proposed approach with the baselines in Scenario 1. The following prompts have been used to generate these images: (a) 'Fila Disruptor Animal WMN Zebra/ Black'; (b) 'Hooded Coat Chicken Fluffy Faux Fur Jacket'; (c) 'Cerebro (Brain) Canvas Art Print'; (d) 'couple, lavender brown, and half-blood prince image'; (e) 'Baby Paintings - Two Mares and a Foal by George Stubbs'; (f) 'Hydrate Condition (For Dry Colour-Treated Hair)'; (g) 'Portable USB Electric Juicer'

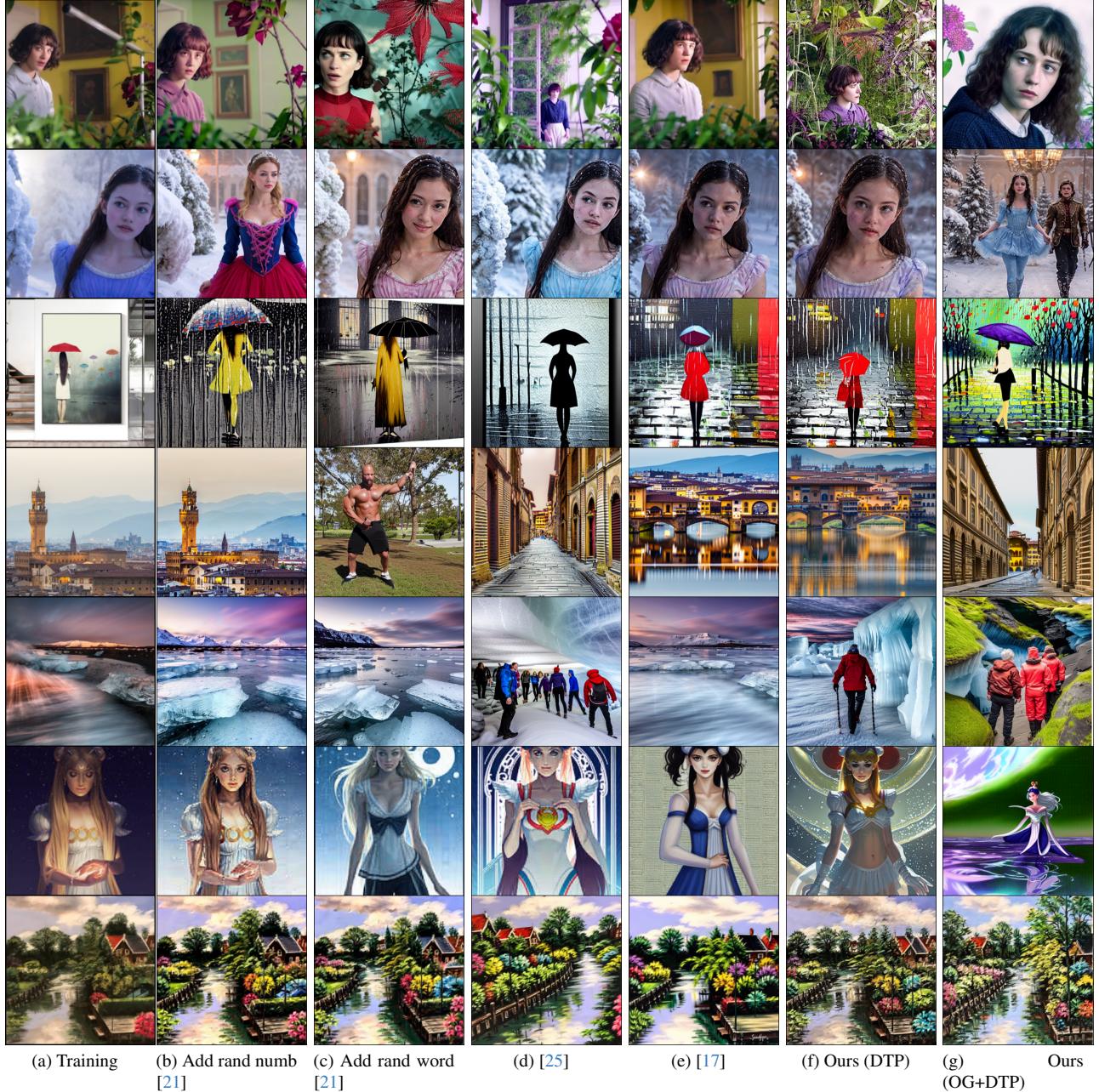


Figure 15. Qualitative results comparing the proposed approach with the baselines in Scenario 3. The follows prompts have been used (a) This Beautiful Fantastic (2016); (b) 91079f7c6f6 Disney s The Nutcracker and the Four Realms Movie Review - Theresa s Reviews; (c) Woman with Umbrella In The Rain Painting Printed on Canvas 1; (d) View of Florence during the day Stock Photo - 22581191; (e) Ice Cave Day Tour with Flights from Reykjavik; (f) Sailor Moon by Charlie-Bowater; (g) Pastel artwork of a canal in Edam, Netherlands, by Susan Marino.



Figure 16. Qualitative results comparing the proposed approach with the baselines in Scenario 4. The prompts used for generating these images are: (a) Read a Previously Unpublished F. Scott Fitzgerald Story; (b) Ethan Hawke to Star as Jazz Great Chet Baker in New Biopic; (c) Father Christmas Red Wall Tapestry Wall Tapestry; (d) Ricky Gervais Promises More David Brent Gigs; (e) Skull 5 barely there iPhone 6 case; (f) Pug Mom iPhone 5 Case; (g) Keep Calm and focus on Alissa iPhone 3 Tough Cover.



Figure 17. Qualitative results on scenario 2 showing the closest match of the generated sample (column 1) with real samples (column 2-11) based on the similarity metric. Since for class conditional models, exact memorization is not observed we look at the most similar real images for qualitative results. We did not observe any verbatim memorization.