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 were 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 examples of transition points coinciding with a fall in conditional guidance in Figure 13.
- More visual results on different scenarios, comparing with baselines in Section 15.

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 [5, 13, 26] 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. [20] 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 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 [28] and [24] lead to poorer similarity scores.

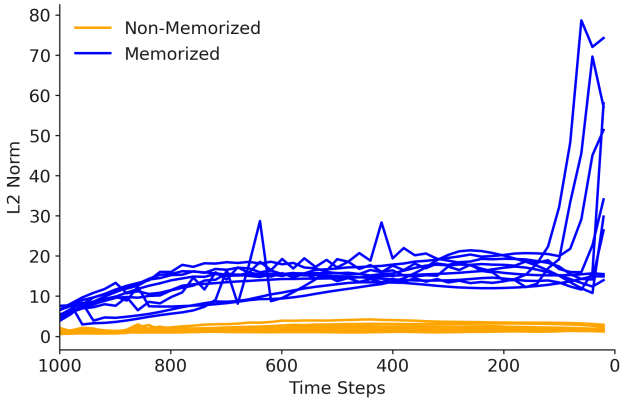


Figure 9. When you apply CFG from the beginning, the magnitude of $\epsilon_\theta(x_t, e_p) - \epsilon_\theta(x_t, e_\emptyset)$ remains high throughout.

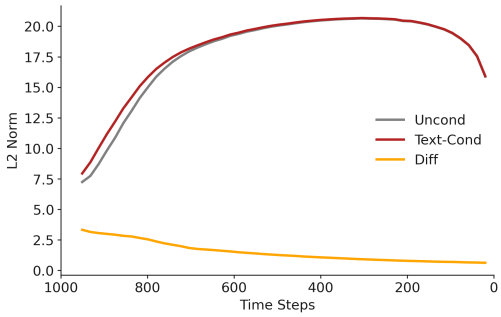


Figure 10. Mean conditional and unconditional noise predictions when SDv2.1 is finetuned on the Imagenette dataset.

Table 4. Results on Scenario 4. Ren et al. [20] studied this scenario.

	Similarity		CLIP Score	FID
	95pc	Mean ± Std		
No mitigation	0.9262	0.5508 ± 0.31	0.3144	155.35
Add rand word [24]	0.9049	0.3779 ± 0.29	0.3046	**143.10**
Add rand numb [24]	0.8934	0.3740 ± 0.28	0.3068	146.01
Wen et al. [28] (l_{target} =3)	0.9155	0.4913 ± 0.31	**0.3134**	154.70
Wen et al. [28] (l_{target} =1)	0.9011	0.3287 ± 0.27	0.3088	147.74
Ren et al. [20]	**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

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

We observed that finetuning SDv2.1 on the LAION-10k dataset [24] leads to exact memorization when using the

Training | Add rand numb [24] | Add rand word [24] | Wen et al. [28] | Ren et al. [20] | 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 Beauty and the Beast

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.

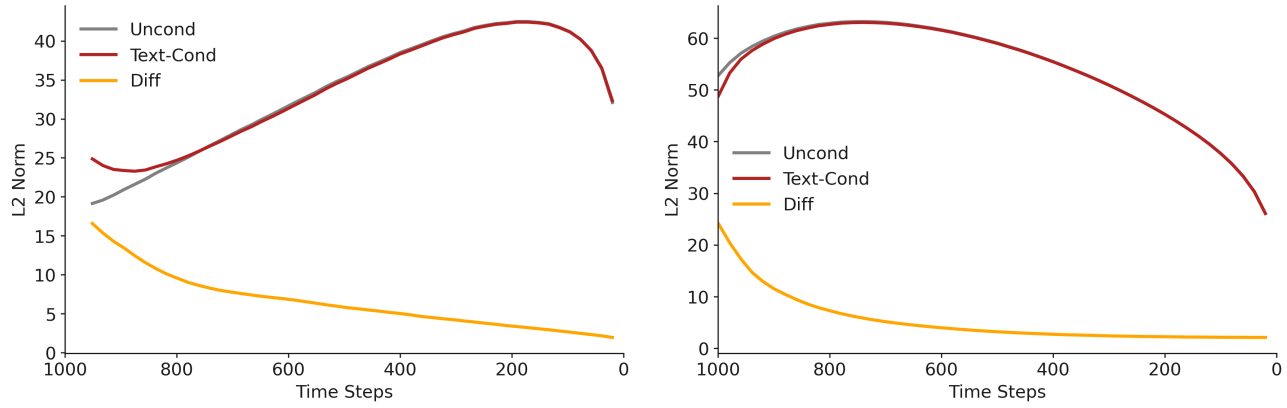
	Similarity (95pc)	**CLIP Score**	**FID**
No mitigation	0.3952	0.314	11.46
Ours (STP)	**0.2861**	0.309	16.06

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

[1] https://drive.google.com/drive/folders/1TT1x1yT2B-mZNXuQPg7gqAhxN_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. [24]. We followed the same inference protocol with 50 inference steps using the DPM Multi-step solver [14].

11.2. Scenario 2

We finetuned Stable Diffusion v2.1 on the Imagenette dataset [12] 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 [24]. 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 [14] for 50 inference steps.

11.3. Scenario 3

We used the pre-trained model weights from Wen et al. [28] available publicly[2]. 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 [14] for 50 inference steps.

11.4. Scenario 4

Webster et al. [27] 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 [14] 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.

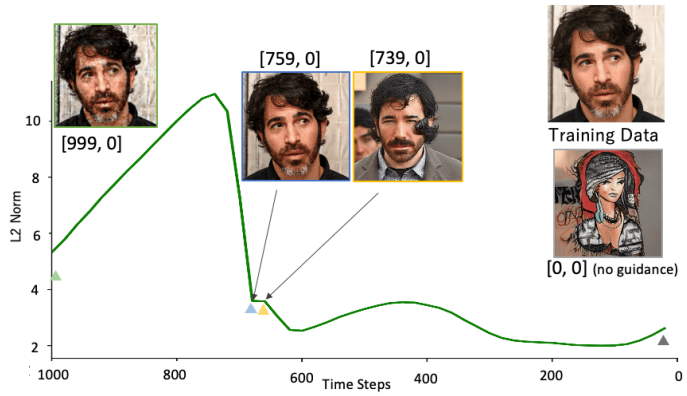
12. Prompts for Figure 7

- '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'

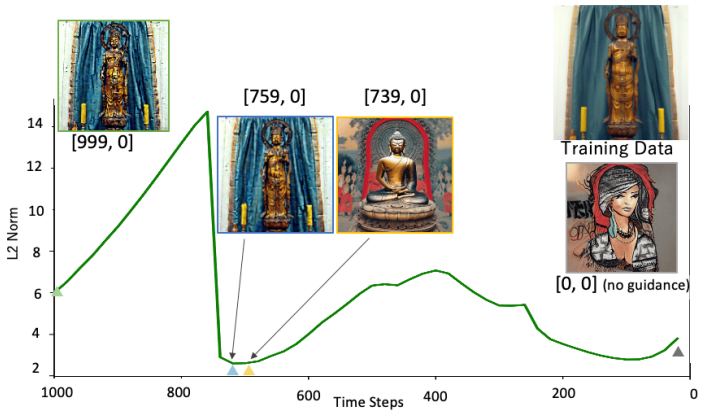
13. Prompts for Figure 8

- 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)"

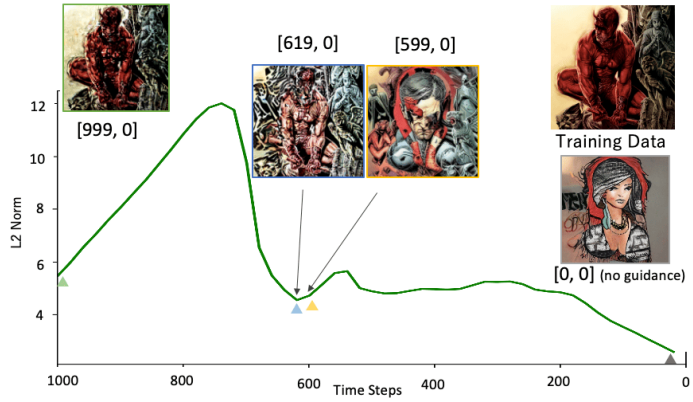
[2] https://drive.google.com/drive/folders/1XiYtYySpTUmS_9OwojNo4rsPbkfCQKBl



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



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



(c) Prompt: As Punisher Joins Daredevil 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 Examples of Transition Points

We provide additional visualizations of transition points coinciding with a fall in conditional guidance in Figure 13.

15. More Visual Examples

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.

Original | Add rand numb [24] | Add rand word [24] | Wen et al. [28] | Ren et al. [20] | Ours (STP) | Ours (OG+STP)

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 Thicken 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 following 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.

Training | Add rand numb [24] | Add rand word [24] | Wen et al. [28] | Ren et al. [20] | Ours (DTP) | Ours (OG+DTP)

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 (Columns 2-11) based on the similarity metric. Since exact memorization is not observed for class conditional models, we look at the most similar real images for qualitative results. We did not observe any verbatim memorization.