Memorization In Stable Diffusion Is Unexpectedly Driven by CLIP Embeddings

Supplementary Material

A. Additional Experimental Results

A.1. Applying Our Method To Stable Diffusion v1.4

We apply our mitigation method, consisting of `<pad>` replacement and $\mathbf{v}^{\mathbf{eot}}$ masking. As shown in Figure 9, our method significantly reduces memorization without degrading image quality.

Figure 9. **<pad> replacement and $\mathbf{v}^{\mathbf{eot}}$ masking.** The first column (Original) shows the memorized image consistently reproduced from the original embedding regardless of seeds. The remaining five columns (Ours) are generated using our mitigation method with five different random seeds.

A.2. Comparison with Prior Mitigation Methods

Comparison with Baselines (1) Ren et al. [24], which rescales cross-attention. (2) Wen et al. [36], which minimize the magnitude of text-conditional noise prediction. (3) Random Token Addition (RTA) [31], which inserts randomly sampled tokens. (4) Random Number Addition (RNA) [31], which inserts a random number between $[0, 10^6]$. We use their optimal configurations `optim_target_loss = 3` for Wen et al. [36] and `rescale_attention = 1.25` for Ren et al. [24]. Results are shown in Figure 10.

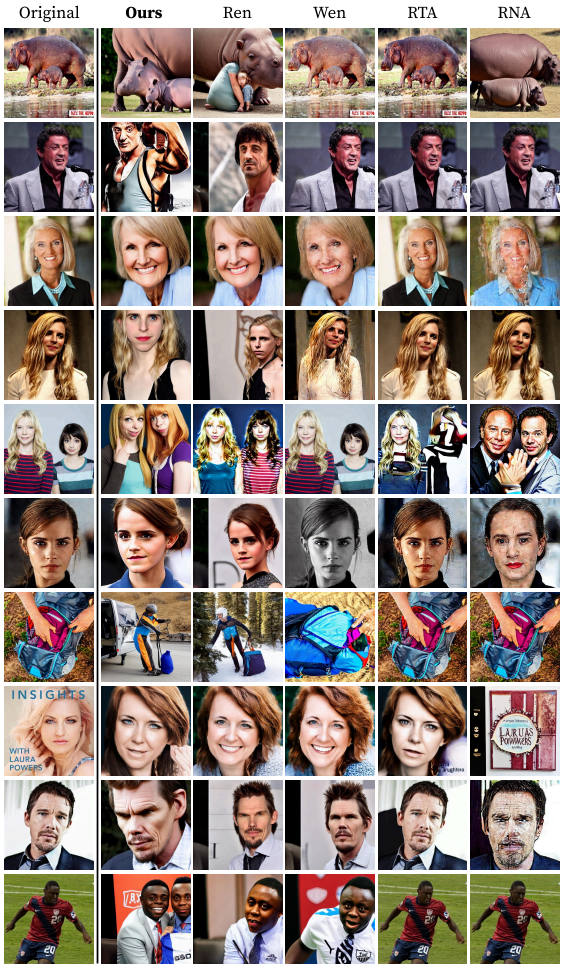


Figure 10. Each row shows images generated from the same prompt using different mitigation methods. Visually and quantitatively, our method produces a more favorable quality-preserving trade-off.

Inference Time Comparison. As shown in Table 6, our mitigation method achieves the fastest inference among all evaluated approaches. Since it does not require any additional optimization or computation during inference, its runtime closely matches that of standard text-to-image generation without mitigation. In contrast, other methods incur additional computational overhead, resulting in slower inference.

Method	Inference Time (s)
Ours	3.17 ± 0.01
Ren et al. [24]	3.37 ± 0.02
Wen et al. [36]	3.55 ± 0.16
RNA [31]	3.26 ± 0.02
RTA [31]	3.31 ± 0.02

Table 6. Average inference time per image for each method.

Evaluation on Webster 500 Prompts. Many prior memorization studies [4, 18, 24, 36] report results directly on all 500 candidate prompts released by Webster [35]. In the main paper, we use a stricter curated benchmark of 458 MV prompts, constructed from Webster [35] and Membench [13], for primary analysis, as the full 500-prompt set includes unstable and degraded cases that can confound mechanistic interpretation. At the same time, to facilitate direct comparison with this widely used evaluation setting, we additionally evaluate all mitigation methods on the full Webster 500 set.

Method	SSCD	CLIPScore	Aesthetic
Ours	0.39 ± 0.20	0.29 ± 0.03	5.33 ± 0.26
Wen	0.65 ± 0.22	0.31 ± 0.03	5.35 ± 0.23
Ren	0.20 ± 0.15	0.26 ± 0.03	4.90 ± 0.31
RTA	0.70 ± 0.23	0.31 ± 0.03	5.36 ± 0.23
RNA	0.70 ± 0.23	0.32 ± 0.03	5.33 ± 0.25

Table 7. Comparison of mitigation methods on all 500 candidate prompts from Webster [35].

As shown in Table 7, our method (<pad> $\rightarrow$! & $\mathbf{v}^{\mathbf{eot}} \rightarrow \mathbf{0}$) continues to provide a favorable trade-off between memorization reduction and generation quality on this broader benchmark. In particular, compared to Wen et al. [36], RNA [31], and RTA [31], our method achieves substantially lower SSCD while maintaining comparable CLIPScore and Aesthetic Score. Although Ren et al. [24] attains lower SSCD, it does so at a noticeably larger cost in prompt alignment and visual quality. These results are consistent with our main-paper findings: even on the noisier Webster 500 set, our method remains a competitive mitigation method with a strong quality-preserving trade-off.

Evaluation on Membench. To ensure the robustness of our findings, we conduct extensive evaluations across multiple datasets. As shown in Table 8, on the Membench [13], which consists of 3,000 prompts, our method shows consistently comparable or better results against existing baselines. We do not report RNA or RTA results, as their performance is not comparable to these three representative baselines.

Method	SSCD	CLIPScore	Aesthetic
Ours	0.16 ± 0.23	0.30 ± 0.07	5.14 ± 0.53
Ren	0.15 ± 0.19	0.29 ± 0.06	5.17 ± 0.51
Wen	0.79 ± 0.37	0.31 ± 0.06	5.22 ± 0.48

Table 8. Comparison of mitigation methods on Membench [13].

A.3. Cross-Prompt Swapping Experiment

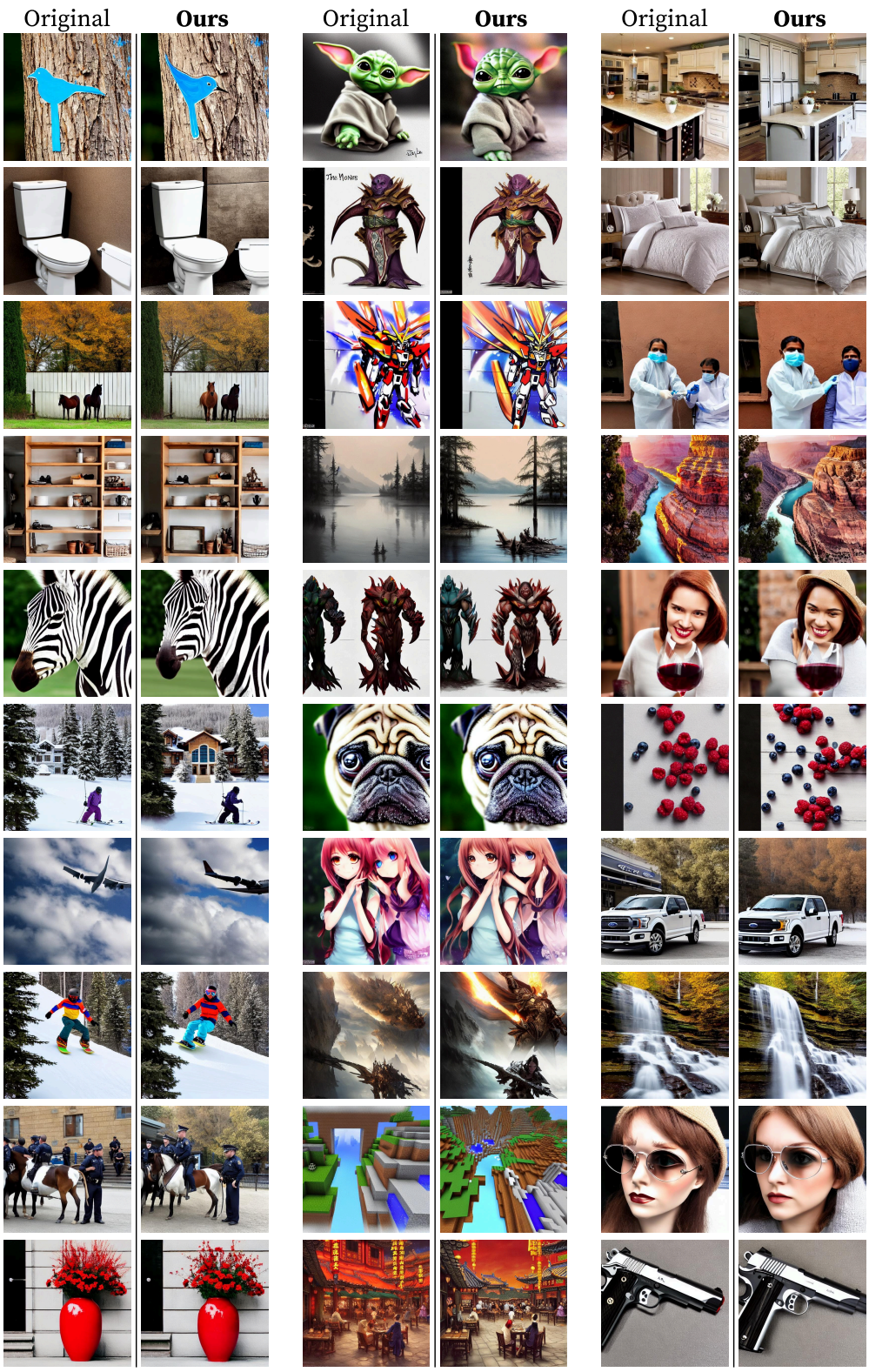
To further investigate the role of $\mathbf{v}^{\mathbf{eot}}$, we conduct an experiment swapping $\mathbf{v}^{\mathbf{eot}}$ and $\mathbf{v}_i^{\mathbf{pad}}$ between different prompts. Our experiment reveals that swapping only $\mathbf{v}^{\mathbf{eot}}$ has no discernible effect on the output, whereas swapping both $\mathbf{v}^{\mathbf{eot}}$ and $\mathbf{v}_i^{\mathbf{pad}}$ alters the generated image to match the swapped prompt (Figure 11). This suggests that $\mathbf{v}_i^{\mathbf{pad}}$, much like $\mathbf{v}^{\mathbf{eot}}$, encode critical semantic information rather than serving as mere placeholders.

(a) **(b)**

Figure 11. (a) Original generation without swapping. (b) Generation after swapping both $\mathbf{v}^{\mathbf{eot}}$ and $\mathbf{v}_i^{\mathbf{pad}}$, showing that the output changes to reflect the swapped prompt.

A.4. Robustness of our mitigation method on non-memorized prompts

We utilize non-memorized prompts to evaluate whether our mitigation method impacts general generation quality. Fol-



(a) MS COCO **(b)** Lexica Art **(c)** LAION

Figure 12. "Original" is generated using the default embeddings, while "Ours" applies our mitigation strategy (<pad> $\rightarrow$! & $\mathbf{v}^{\mathbf{eot}} \rightarrow \mathbf{0}$).

lowing the prior studies [4, 36], we sample 500 prompts each from MS COCO, Lexica Art, and LAION. To ensure a fair and rigorous comparison, we apply additional filtering so that the selected prompts resemble memorized prompts in structure and complexity: each prompt must consist only of ASCII characters; exclude URLs (e.g., "http", "www"); and have a tokenized length between 5 and 40.

As shown in Figure 12, our mitigation method preserves visual quality and prompt alignment across a diverse range of non-memorized prompts, demonstrating robustness and deployability without degradation in general generation.

B. Detailed Experimental Settings

Model We perform all experiments using Stable Diffusion, specifically v1.4 for the SD 1 series and v2.1 for the SD 2 series. Image generation is carried out with the DDIM scheduler [32], using a guidance scale of 7.5 and 50 inference steps, following prior memorization works [31, 36]. Detailed software libraries and configurations are provided in the Supplementary Material.

Computational Resources All experiments were conducted on a single NVIDIA L40S GPU (48 GB VRAM) and an AMD EPYC 9754 processor (192 CPU cores). Because our study focuses exclusively on inference rather than training, the overall computational requirements were modest, and the full experimental pipeline ran efficiently on a single GPU.

Dataset and Evaluation To build the dataset used in our experiments, we apply filtering criteria to a subset of Membench [13] prompts, using images generated with 10 fixed random seeds: 0, 1, 10, 42, 100, 441, 515, 1000, 2025, 10000 (see Section 3), and average the results over them to ensure stability and comparability.

B.1. Analysis of Memorized Prompt Datasets from Prior Works

Webster [35] released 500 candidate memorized prompts for Stable Diffusion v1.4 (SD1). However, as we discussed earlier, a considerable portion of these prompts are not truly *memorized* cases. As shown in Figure 13, many of them produce degraded or entirely unrelated images, and in some cases, the generated image corresponds to a completely different prompt. Therefore, directly using all 500 prompts without verification would be inappropriate.

Webster [35] further categorized 345 prompts for SD1 into 86 *Matching Verbatim* (MV), 30 *Retrieval Verbatim* (RV), and 219 *Template Verbatim* (TV) cases. For Stable Diffusion v2.1 (SD2), the same study includes 219 prompts, with only 4 labeled as MV and 215 as TV.

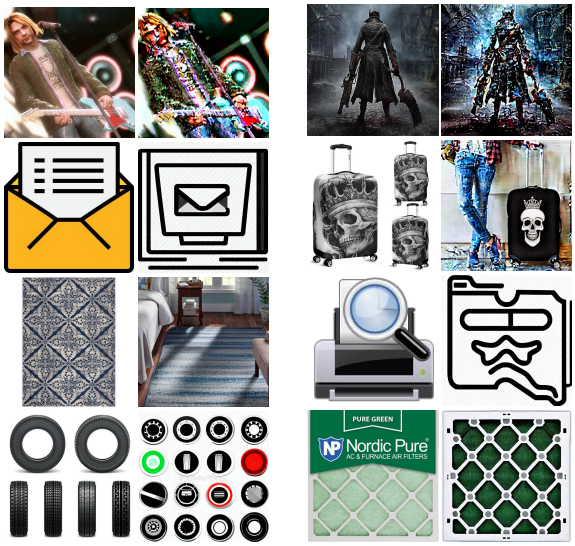


Figure 13. Each left column shows the reference image, and each right column shows the corresponding model-generated image.

Hong et al. [13] provided 3000 memorized prompts for SD1 without further categorization. Prompts corresponding to SD2 are not publicly available.

To better understand the characteristics of memorized prompts, we analyze the token length statistics of prompts used in prior memorization studies. Table 9 summarizes the statistics for SD1 and SD2 prompts reported in prior memorization studies.

B.2. Prompts used for Figures

To enhance reproducibility and transparency, we list the exact prompts used for generating the example images shown in our figures. The memorized prompts are drawn from publicly released datasets provided by Webster [35] and Hong et al. [13], while the non-memorized prompts are sampled from MS COCO [19], Lexica [28] and LAION [29].

Figure 1
- Living in the Light with Ann Graham Lotz
- Mothers influence on her young hippo
- Here's Who Ian McShane May Be Playing in Game of Thrones Season Six

Figure 3
- Living in the Light with Ann Graham Lotz
- Talks on the Precepts and Buddhist Ethics

Figure 4
- Mothers influence on her young hippo
- The No Limits Business Woman Podcast

Metric	SD1		SD2
	Webster [35]	Hong et al. [13]	Webster [35]
Number of Prompts	345	3000	219
Mean Length	19.74	15.31	21.51
Max / Min Length	38 / 4	38 / 5	35 / 6

Table 9. Prompt Token Length Statistics for Stable Diffusion v1.4 (SD1) and Stable Diffusion v2.1 (SD2)

Figure 6
- Björk Explains Decision To Pull Vulnicura From Spotify
- ”As Punisher Joins Daredevil Season Two, Who Will the New Villain Be?”
- Talks on the Precepts and Buddhist Ethics

Figure 7
- Emma Watson to play Belle in Disney’s Beauty and the Beast
- ”The Long Dark Gets First Trailer, Steam Early Access”
- ”Long-Lost F. Scott Fitzgerald Story Rediscovered and Published, 76 Years Later”

Figure 9
- Brit Marling-Zal Batmanglij Drama Series The OA Gets Picked Up By Netflix
- Sarah Silverman Will Star in HBO Pilot from Secret Diary of a Call Girl Creator
- Ethan Hawke to Star as Jazz Great Chet Baker in New Biopic
- ”George R.R. Martin to Focus on Writing Next Book, World Rejoices”
- Hawkgirl Cast in Arrow/Flash Spinoff Series For The CW
- South Park: The Stick of Truth Review (Multi-Platform)
- Ava DuVernay Won’t Direct Black Panther After All
- Freddy Adu Signs For Yet Another Club You Probably Don’t Know
- ”””””””””””Listen to The Dead Weather’s New Song, ”””””””””””Buzzkill(er)”””””””””””””””””””””””””
- Emma Watson Set to Star Alongside Tom Hanks in Film Adaptation of Dave Eggers’ The Circle

Figure 10
- Mothers influence on her young hippo
- Rambo 5 und Rocky Spin-Off - Sylvester Stallone gibt Updates
- Living in the Light with Ann Graham Lotz
- Brit Marling-Zal Batmanglij Drama Series The OA Gets Picked Up By Netflix
- Watch the First Episode of Garfunkel and Oates
- Emma Watson Set to Star Alongside Tom Hanks in Film Adaptation of Dave Eggers’ The Circle
- Full body U-Zip main opening - Full body U-Zip main opening on front of bag for easy unloading when you get to camp
- Insights with Laura Powers
- Ethan Hawke to Star as Jazz Great Chet Baker in New Biopic
- Freddy Adu Signs For Yet Another Club You Probably Don’t Know

Figure 12

MS COCO
- A painted sign of a blue bird in a tree in the woods.
- a small white toilet is in a tiny bathroom
- Two horses standing behind fence with grassy leaves
- A lot of wooden shelves filled with lots of clutter.
- A zebra standing on the grass holding its head near the ground.
- a skier and a snowboarder in front of a large house
- Large airplane flying below the clouds from underneath
- A snowboarder in a colorful jacket racing down a slope.
- There are police men standing and some are sitting on horses
- A large red vase sitting in front of a building.

Lexica Art
- Baby Yoda by Jeszika Le Vye, trending on artstation, hyperdetalied,
- the dnd monster pung_ concept art
- graffiti on a wall of anime gundam, speed painting, trending on artstation
- A beautiful concept art painting of a gloomy lake with a body floating slightly below the surface by
- three different views of a predator set for dota 2, concept art by senior character artist, trending on artstation, artstation hd, full body
- closeup of a pug with moss growing out of its face folds, macro photography, overgrown pug, high resolution photo, trending on artstation
- two beautiful idols standing face to face, trending art-

station, pixiv, detailed anime art
- guild wars 2, cinematic battlefield, Hyperrealistic CGI Photorealistic cinematic volume lighting fanart on ArtStation full sun shine day concept art, digital art, high detail by Daniel Dociu
- minecraft style concept art of a blue portal to a fantasy world
- fantasy art of a bustling tavern in china, at night, by justin gerard, highly detailed digital art, trending on artstation

LAION
- kitchen ideas with islands simply home designs home design ideas 3
- 78 most terrific supple duvet sets with matching let curtains covers curtain duvets cover definition define snazzy aequorea plus thresh expensive quilts
- India begins next phase of COVID vaccination from Monday:
- Gary Hart Photography: Bright Angel Lightning, Grand Canyon
- Beautiful happy lady drinking glass of red wine.
- berries_lead
- 2018 Ford F-150
- Alpine Falls in Loyalsock State Forest
- Retro Classic Full Metal Tear Drop Clear Lens Aviator Glasses C936
- Well M1911 A1 Vollmetall Springer 6mm BB Two-Tone

Figure 13
- 33 Screenshots of Musicians in Videogames
- “communication, email, mail, message, online, open icon”
- Anzell Blue/Gray Area Rug by Andover Mills
- Vector Tyre Icons isolated on white background Illustration
- Bloodborne Video: Sony Explains the Game’s Procedurally Generated Dungeons
- 3D Black & White Skull King Design Luggage Covers 007
- “document,print,preview icon”
- 12x12x1 Pure Green AC Furnace Air Filters Qty 6 - Nordic Pure

C. Difference of embeddings between CLIP and OpenCLIP

In the CLIPText encoder used in Stable Diffusion v1.4, text embeddings are grouped into three clusters: $\mathbf{v}^{\mathbf{sot}}$, $\mathbf{v}_i^{\mathbf{pr}}$, and a merged cluster of $\mathbf{v}^{\mathbf{eot}}$ and $\mathbf{v}_i^{\mathbf{pad}}$. This is because the tokenizer duplicates the `<eot>` for padding, making $\mathbf{v}_i^{\mathbf{pad}}$ nearly identical to $\mathbf{v}^{\mathbf{eot}}$ after encoding. In contrast, the CLIPText encoder of Stable Diffusion v2.1 yields a separate cluster for $\mathbf{v}^{\mathbf{eot}}$ and $\mathbf{v}_i^{\mathbf{pad}}$, indicating a semantic distinction between the two.

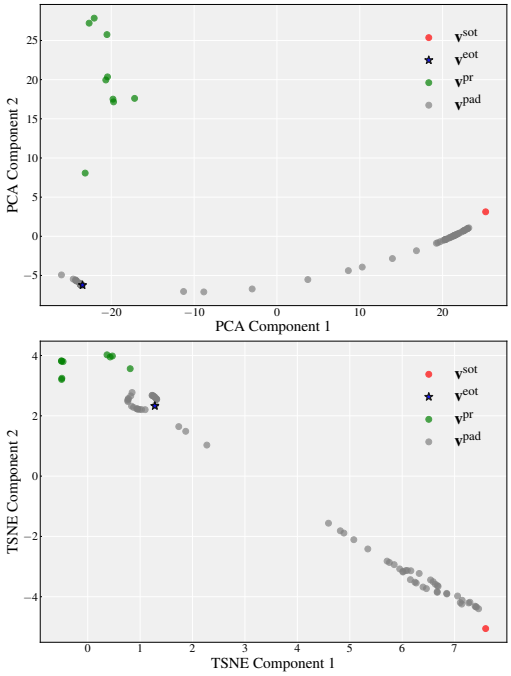
Stable Diffusion v1.4 employs the original CLIP text encoder, where the tokenizer uses `<eot>` as the default `<pad>`. As a result, $\mathbf{v}_i^{\mathbf{pad}}$ are near duplicates of $\mathbf{v}^{\mathbf{eot}}$, which amplifies the semantic influence of $\mathbf{v}^{\mathbf{eot}}$ and contributes significantly to memorization.

Stable Diffusion v2.1 employs an OpenCLIP text encoder. Importantly, OpenCLIP maintains the same overall architecture and embedding mechanism as the original CLIP, with only minor differences in training scale and data. The critical difference relevant to our analysis is that the OpenCLIP tokenizer uses a neutral `!` token for `<pad>` instead of duplicating `<eot>`. As a result, $\mathbf{v}_i^{\mathbf{pad}}$ are semantically distinct from the $\mathbf{v}^{\mathbf{eot}}$, eliminating the structural duplication that contributed to memorization in Stable Diffusion v1.4.

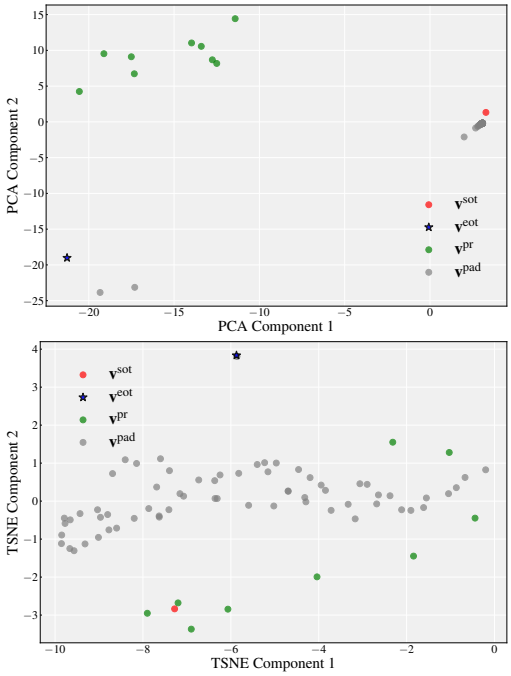
The PCA and t-SNE visualizations in Figure 14 further illustrate this difference.

D. Cross-Attention Maps by Embedding

Following Chen et al. [4], we visualize the cross-attention maps from the first two 64-pixel downsampling layers of the U-Net at the last denoising step. To ensure consistent brightness across all maps, we normalize the intensity by setting the minimum and maximum brightness levels equal across all cross-attention maps. As shown in Figure 15 and 16, not only does the attention map corresponding to $\mathbf{v}^{\mathbf{eot}}$ appear bright, but the adjacent $\mathbf{v}_i^{\mathbf{pad}}$ also exhibit high attention scores.



(a) CLIP: PCA (top) and t-SNE (bottom)



(b) OpenCLIP: PCA (top) and t-SNE (bottom)

Figure 14. Text embeddings of "Living in the Light with Ann Graham Lotz" visualized using PCA and t-SNE for (a) CLIP and (b) OpenCLIP models. The two projection methods reveal distinct clustering behaviors.

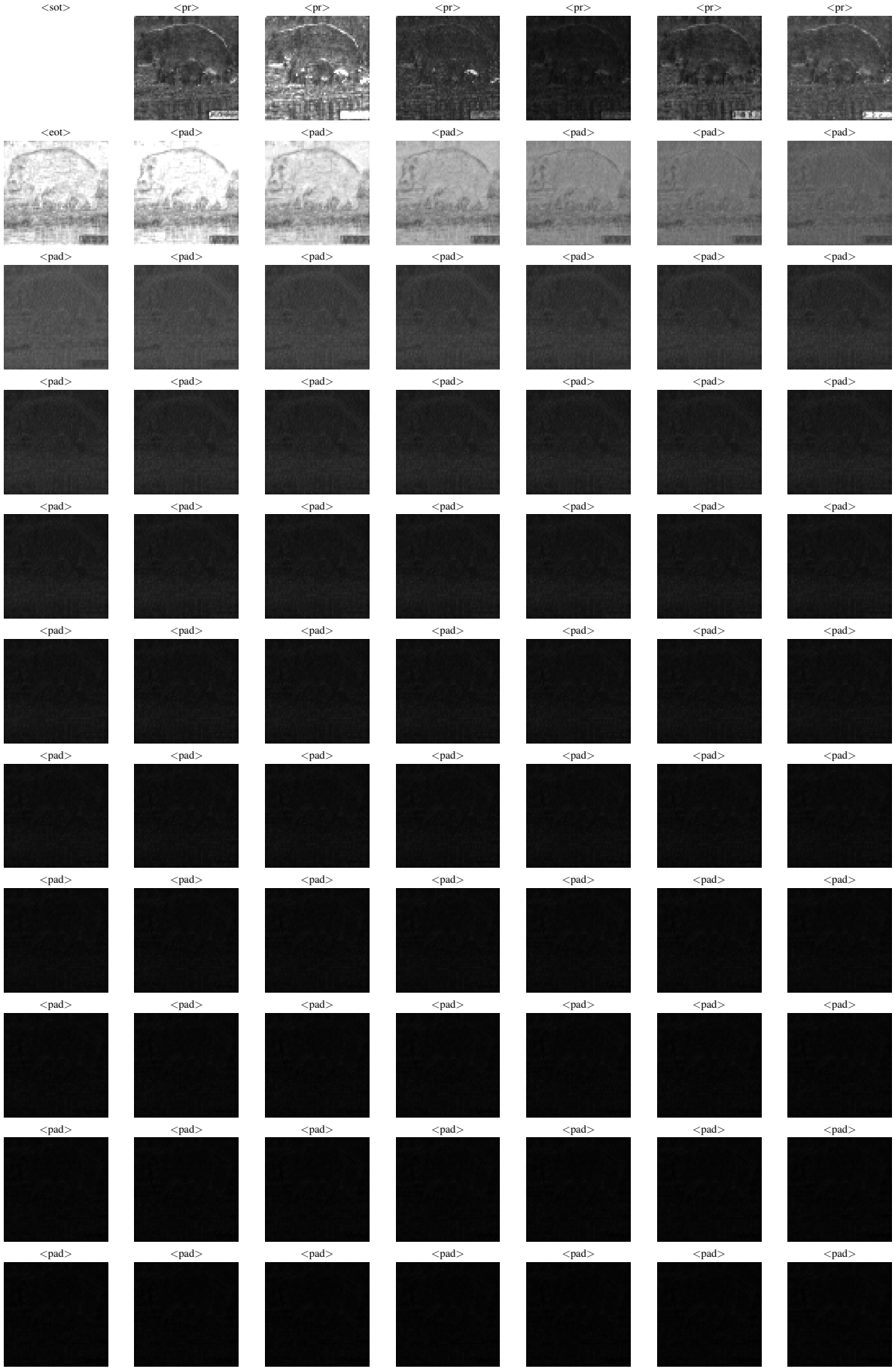


Figure 15. Cross Attention maps of “Mothers influence on her young hippo”

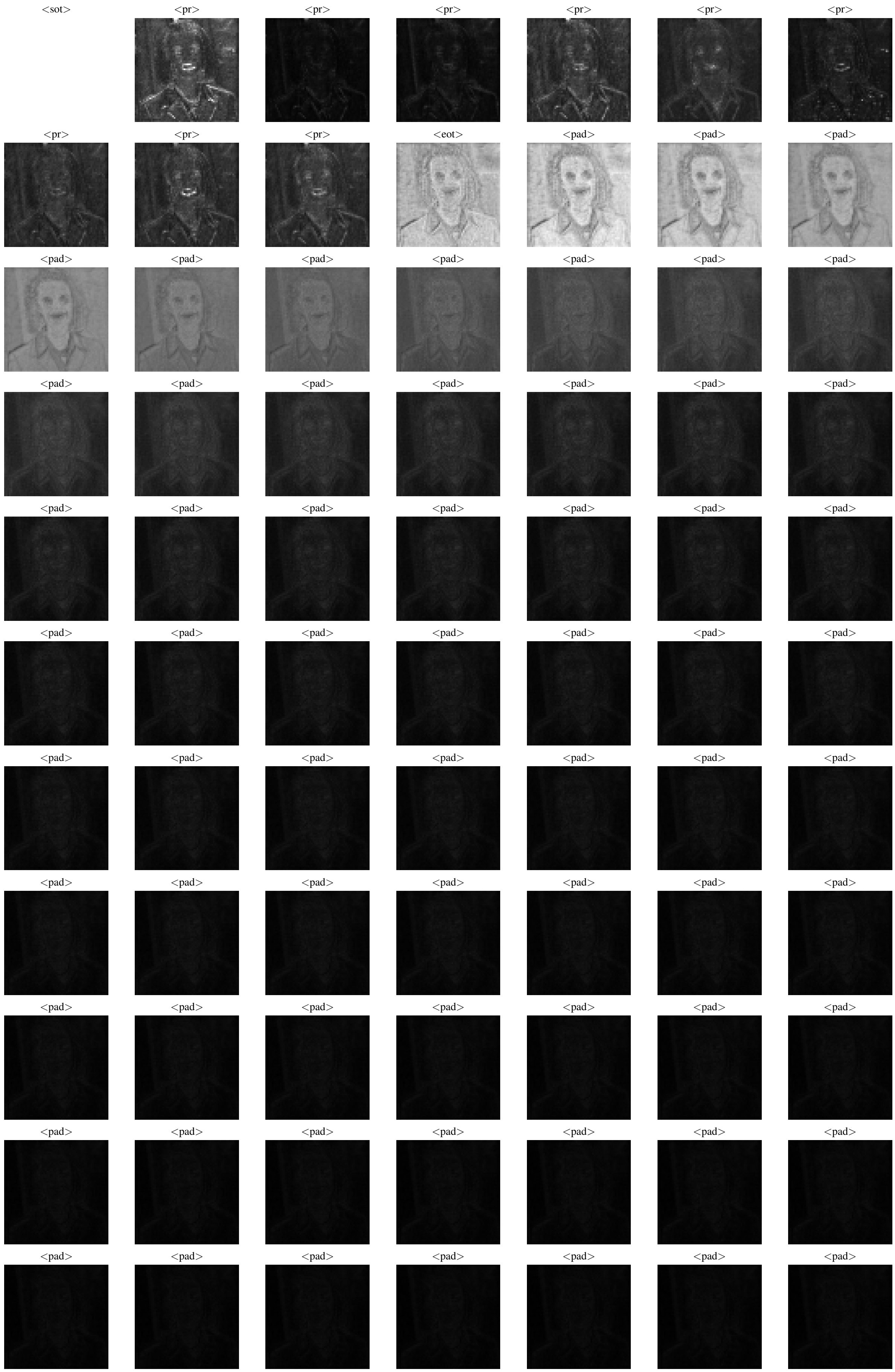


Figure 16. Cross Attention maps of “Living in the Light with Ann Graham”