# Memorization is Localized within a Small Subspace in Diffusion Models

**Anonymous Authors**[1]

## 1. Introduction

Large-scale text-to-image diffusion models excel at producing high-quality images from text but often memorize and replicate their training data, leading to copyright and privacy concerns (Verge, 2023; Newspaper, 2024). Recent research on diffusion models (Wen et al., 2024; Ren et al., 2024; Yoon et al., 2023; Gu et al., 2024; Chen et al., 2024; Somepalli et al., 2023) links memorization to data duplication or specific prompts that trigger the generation of training data. Proposed mitigation strategies include inference-time techniques like detecting and perturbing trigger tokens, as well as training-based methods like filtering and perturbing training data. However, these approaches often disrupt either the training or inference processes, posing challenges for real-world deployment.

In this preliminary study, we reveal a surprising finding: *memorization in pre-trained diffusion models tends to localize within a specific subset of neurons*. Unlike previous research focusing on individual memorized prompts, we identify critical neurons within models that show heightened responses to a small set of memorized prompts compared to non-memorized ones. These specialized neurons, termed "memorized neurons," exhibit significant overlap across different sets of memorized prompts, indicating the presence of a distinct memorization subspace in the model. Leveraging this discovery, we propose a novel, one-time, training-free strategy to mitigate memorization that involves selectively pruning weight regions associated with these memorized neurons. Unlike traditional methods, our technique offers ease and speed, requiring no modifications to the model's training or inference processes.

## 2. Methodology

Recent studies on Large Language Models (LLMs) have uncovered specialized neurons (Zhang et al., 2023; Suau et al., 2020) crucial for safety (Wei et al., 2024), by leveraging pruning techniques (Sun et al., 2024). Inspired by this research, we propose a method to identify and prune specific neurons in Diffusion Models (DMs) to mitigate memorization, building upon the Wanda pruning approach (Sun et al., 2024).

**Candidate neurons to prune in DMs:** We consider the Feed-Forward Network (FFN) in the UNets within diffusion model pipelines. At time step $t$ of the denoising process and layer $l$, we denote the input to the FFN for text prompt $p$ by $z^{t,l}(p) \in \mathbb{R}^{d\times m}$ and output of the FFN by $z^{t,l+1}(p) \in \mathbb{R}^{d\times m}$. Here $m$ is the number of latent tokens. FFN in Stable Diffusion consists of GEGLU activation (Shazeer, 2020), which operates as shown in Equation 1:

$$h^{t,l}(p) = \text{GEGLU}(\text{Linear}(z^{t,l}(p)) \qquad (1)$$
$$z^{t,l+1}(p) = \mathbf{W}^l \cdot h^{t,l}(p),$$

$\mathbf{W}^l \in \mathbb{R}^{d\times d'}$ is the weight matrix in the *second* linear layer.

**Layer-wise Wanda score for memorized prompts at time $t$:** We randomly sample $n$ memorized prompts from a total of 500 prompts identified by the training data extraction attack (Webster, 2023), denoted as $P = \{p_1, p_2, ...p_n\}$.

We collect neuron activations corresponding to the set of known memorized prompts $P$, flatten each of them, and arrange them in a matrix denoted by $\mathbf{H}^{t,l}(P) = [h^{t,l}(p_1), h^{t,l}(p_2), ..., h^{t,l}(p_n)]$ such that $\mathbf{H}^{t,l}(P) \in \mathbf{R}^{d'\times n*m}$. Then, we calculate the importance score for FFN weights $\mathbf{W}^l$ using input neurons for memorized prompts similar to Wanda (Sun et al., 2024) in Equation 2.

$$\mathbf{S}^{t,l}(P)_{(i,j)} = \left|\mathbf{W}^l\right|_{(i,j)} \cdot \left\|\mathbf{H}^{t,l}(P)_{(j,:)}\right\|_2 \qquad (2)$$

where $|\cdot|$ computes the absolute value, and $\|\cdot\|_2$ denotes the $l_2$-norm. For details on Wanda, we direct the readers to (Sun et al., 2024). Similarly, we calculate the importance score for the null prompt $p_\emptyset$ as $\mathbf{S}^{t,l}(P_\emptyset)$, where $P_\emptyset$ is formulated by stacking $n$ repetitions of $h^{t,l}(p_\emptyset)$. Note that this process only requires one forward pass per prompt.

**Localizing and pruning memorized neurons:** For a given sparsity level $s\%$, we now define the top-$s\%$ *important neurons* in the $i$-th row of $\mathbf{W}^l$ as $\mathbf{A}^{t,l}(P) = \{(i,j)|$ if $\mathbf{S}^{t,l}(P)_{(i,j)}$ in top-s% $(\mathbf{S}^{t,l}(P)_{(i,:)})\}$.

A weight neuron is defined as a *memorized* neuron if it ranks among the top $s$-% of important neurons and its Wanda score exceeds that of the null string. We define the set of memorized neurons denoted by $\mathbf{V}^{t,l}(P, P_\emptyset)$ which is formulated as

$$\mathbf{V}^{t,l}(P, P_\emptyset) = \{(i,j)| \text{ if } \mathbf{S}^{t,l}(P)_{(i,j)} > \mathbf{S}^{t,l}(P_\emptyset)_{(i,j)} \quad \forall (i,j) \in \mathbf{A}^{t,l}(P)\} \qquad (3)$$

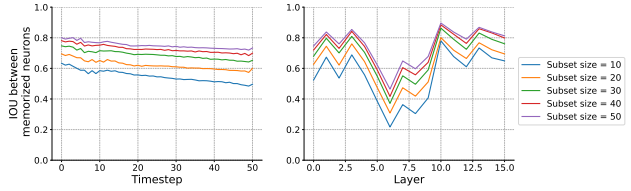


Figure 1: Average Pairwise IOU averaged over timestep (left) and layer (right) for $N = 10$ and varying subset size.

To prune the memorized neurons, we first take a union of indices in $\mathbf{V}^{t,l}(P, P_\emptyset)$ for the first $\tau$ time steps and set the weights in $\mathbf{W}^l$ corresponding to this union zero. Subsequently, we will use the pruned $\mathbf{W}^l$ for image sampling. Empirically, we found that aggregating $\tau = 10$ time steps is enough for memorization mitigation and quality image generation.

## 3. Experiments and Results

**Experimental setup:** We use 500 memorized prompts identified for Stable Diffusion v1 (Webster, 2023) and denote this dataset by $\mathcal{D}$. We select $N$ different subsets of prompts from $\mathcal{D}$, each containing $k$ memorized prompts. We denote the *collection* of these *subsets* by $\mathbb{P}^{N,k} = \{P^i\}\ \forall i \in [1, N]$, such that $|P^i| = k$.

**Memorized Neurons can be Localized within a Small Subspace:** We proceed to compute the *Average Pairwise* intersection-over-union (IOU) for time step $t$ and layer $l$ between the memorized neurons activated by two distinct subsets within $\mathbb{P}^{N,k}$, given by $\text{IOU}^{t,l}(\mathbb{P}^{N,k}) = \frac{1}{n(n-1)} \sum_{i \neq j}^{N} \text{iou}(\mathbf{V}^{t,l}(P^i), \mathbf{V}^{t,l}(P^j))$. We then average $\text{IOU}^{t,l}(\mathbb{P}^{N,k})$ over time steps and layers and report these quantities in Figure 1. Figure 1 shows that, for fixed $N$ and $k$, the average IOU remains high across all denoising iterations and most FFN layers. This consistency indicates that memorized neurons are similar across different subsets, suggesting that memorization is localized within a compact subspace in the pre-trained model.

**Memorization Can be Edited via Pruning:** We start by pruning a pre-trained Stable Diffusion model according to Equation 3 for each subset $P^k \in \mathbb{P}^{N,k}$ resulting in a set of $N$ models. We then evaluate the set of pruned models on the dataset of 500 memorized prompts $\mathcal{D}$ released by Webster (2023). To assess the extent of memorization exhibited by pruned models, we report SSCD (Pizzi et al., 2022) and CLIP similarity between generated images and input prompt. Figure 2 shows that the pruned models achieve decreased SSCD compared to no-mitigation baseline and comparable scores to inference-time mitigation (Wen et al., 2024), indicating effective mitigation of memorization. Moreover, pruned models derived from different subsets exhibit similar scores in mitigating memorization. This observation is

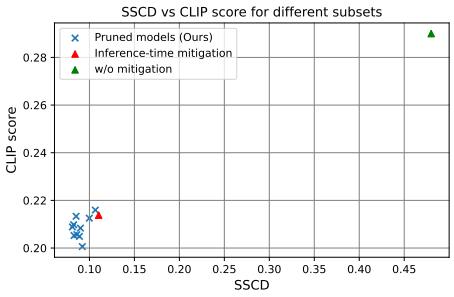


Figure 2: Quality (CLIP similarity score, ↑) vs Memorization (SSCD, ↓) for 10 different pruned models compared with inference-time mitigation in Wen et al. (2024).


Figure 3: Despite sharing the same seed, different pruned models yield semantically similar images.

corroborated by Figure 3 which shows that different pruned models generate semantically similar images for a fixed initialization. We conducted the extraction attack from (Carlini et al., 2023) on Stable Diffusion v1.5 fine-tuned on Imagenette (Howard, 2019), duplicating 100 images 50 times to identify memorized images. Applying our pruning method to the fine-tuned model, the extraction attack initially identified 9 potentially memorized examples. After pruning, we reduced the memorization rate to 0%, demonstrating our method's effectiveness. Evaluating the pruned models on the COCO dataset revealed an FID of 14.35, closely matching the pre-trained model's 14.49, indicating that the pruned models retain their general image-generation capabilities while effectively mitigating memorization.

## 4. Conclusions

We leverage existing pruning heuristics and present an intriguing discovery - *Memorization resides within a potentially unique compact subspace in pre-trained models*. We demonstrate that simply pruning these subspaces leads to effective mitigation of memorization. This marks the first instance of mitigating memorization in diffusion models without additional training, paving the way for more trustworthy text-to-image generation.

## References


Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. *arXiv*, 2023.

Chen, C., Liu, D., and Xu, C. Towards memorization-free diffusion models. *arXiv*, 2024.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv*, 2024.

Howard, J. Imagenette: A smaller subset of 10 easily classified classes from imagenet, March 2019. URL https://github.com/fastai/imagenette.

Newspaper, T. A. Deviantart and midjourney deny wrongdoing in copyright infringement lawsuit over in ai image generators, 2024.

Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., and Douze, M. A self-supervised descriptor for image copy detection. *arXiv*, 2022.

Ren, J., Li, Y., Zen, S., Xu, H., Lyu, L., Xing, Y., and Tang, J. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. *arXiv preprint arXiv:2403.11052*, 2024.

Shazeer, N. Glu variants improve transformer. *arXiv*, 2020.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *arXiv*, 2023.

Suau, X., Zappella, L., and Apostoloff, N. Finding experts in transformer models. *arXiv preprint arXiv:2005.07647*, 2020.

Sun, M., Liu, Z., Bair, A., and Kolter, J. Z. A simple and effective pruning approach for large language models. *arXiv*, 2024.

Verge, T. Getty images sues ai art generator stable diffusion in the us for copyright infringement, 2023.

Webster, R. A reproducible extraction of training images from diffusion models. *arXiv*, 2023.

Wei, B., Huang, K., Huang, Y., Xie, T., Qi, X., Xia, M., Mittal, P., Wang, M., and Henderson, P. Assessing the brittleness of safety alignment via pruning and low-rank modifications. *arXiv preprint arXiv:2402.05162*, 2024.

Wen, Y., Liu, Y., Chen, C., and Lyu, L. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=84n3UwkH7b.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023. URL https://openreview.net/forum?id=shciCbSk9h.

Zhang, Z., Zeng, Z., Lin, Y., Xiao, C., Wang, X., Han, X., Liu, Z., Xie, R., Sun, M., and Zhou, J. Emergent modularity in pre-trained transformers. *arXiv preprint arXiv:2305.18390*, 2023.