MAGIC: Diffusion Model Memorization Auditing via Generative Image Compression

Gunjan Dhanuka[1] **Sumukh Aithal**[1] **Avi Schwarzschild**[1] **Zhili Feng**[1] **J Zico Kolter**[1] **Zachary Lipton**[1] **Pratyush Maini**[1 2]

Abstract

Diffusion models have revolutionized generative modeling by producing high-fidelity images. However, concerns about *memorization*—where models reproduce specific training images—pose ethical and legal challenges, especially regarding copyrighted content. In this paper, we critically analyze current memorization criteria, highlighting their brittleness due to reliance on specific caption-image pairs and vulnerability to common prompt modifications standard in industry, at both training and inference time. We propose a novel method for **M**emorization **A**uditing via **G**enerative **I**mage **C**ompression (`MAGIC`) that reframes memorization detection as an image compression problem. Specifically, we investigate whether the model can regenerate a particular image, independent of textual prompts. By compressing an image into a short learned conditioning (embedding), we directly measure how faithfully a diffusion model can reconstruct it. Experimentally, `MAGIC` significantly improves robustness and accuracy (by over 20%) in detecting memorized content compared to existing approaches. `MAGIC` thus enhances our understanding of memorization and provides practical tools for developing safer generative systems.

1. Introduction

Diffusion models have rapidly emerged as a transformative class of generative models, capable of synthesizing images with striking fidelity and diversity (Ho et al., 2020; Song & Ermon, 2020). Despite their impressive capabilites, these models' tendency to *memorize* and exactly reproduce specific training images raises significant ethical and legal concerns. Recent lawsuits brought by artists and content creators against generative AI companies highlight the urgency of addressing concerns regarding intellectual property rights and privacy (Sustainable Technology Partners, 2023). For example, popular text-to-image diffusion models (such as Stable Diffusion, Ideogram, Dall-E 3), trained on massive web-sourced datasets, have been accused of reproducing copyrighted artwork without consent, fueling heated debates around intellectual property infringement and responsible AI development (Somepalli et al., 2023a; Carlini et al., 2023).

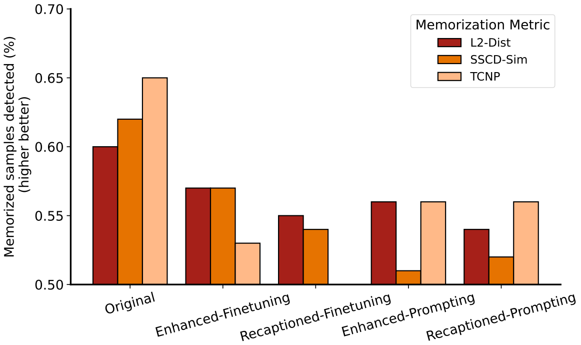

Figure 1: Memorization detection performance by existing metrics after applying simple train-time and inference-time modifications. Existing metrics heavily rely on the exact text prompt that triggers the image. Hence, they suffer a drop in detection performance under benign modifications to prompts during training and/or inference. Such modifications are commonplace in industry today.

Existing memorization metrics rely heavily on precise text-image pairs or consistent reproduction across multiple generations (Carlini et al., 2023; Wen et al., 2024). In practice, these definitions are fragile, easily circumvented by minor modifications to textual prompts during either training or inference. Real-world practices, such as refining or entirely re-captioning training images (Betker et al., 2023; Nguyen

[1]Carnegie Mellon University, Pittsburgh, PA, USA. [2]DatologyAI. Correspondence to: Gunjan Dhanuka <gdhanuka@cs.cmu.edu>, Pratyush Maini <pratyushmaini@cmu.edu>.

Published at ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Vancouver, Canada. PMLR 267, 2025.

et al., 2023), routinely evade detection. Similarly, at inference time, the use of prompt enhancement to improve user prompts has become standard practice. These practices highlight critical limitations of current auditing methods (Section 2).

Motivated by these insights, we propose MAGIC (**M**emorization **A**uditing via **G**enerative **I**mage **C**ompression), a novel approach that reframes the problem of memorization detection as an image compression task (Section 3). MAGIC employs *soft embedding optimization* to compress target images into short, learned embeddings that, when used as conditioning, can regenerate the original image through the diffusion model. Intuitively, if a diffusion model has memorized an image, it should readily compress and reconstruct that image from a compact embedding with few tokens. Conversely, images not memorized by the model either cannot be reconstructed faithfully or require substantially larger number of tokens, effectively distinguishing memorization from generalized learning.

Empirically, MAGIC significantly enhances robustness and accuracy in detecting memorized content, maintaining high detection rates even when traditional metrics degrade drastically (nearly 15%, see Figure 1) under realistic textual interventions. Our evaluations demonstrate that MAGIC reliably identifies memorization overlooked by prior methods.

We further leverage MAGIC to conduct a comprehensive audit of contemporary diffusion models, focusing on high-stakes copyrighted content (Appendix D). Notably, we uncover that prominent models memorize over 50% of top images from major intellectual property holders, underscoring the practical importance and immediate applicability of our auditing framework. By shifting the discourse from mere output consistency to the deeper question: *can the model regenerate specific training content under plausible conditions?*, MAGIC provides a first of its kind tool for auditing image memorization by diffusion models at scale.

2. Brittleness of Existing Memorization Definitions

In this section, we investigate the robustness of existing memorization detection methods for diffusion models, specifically highlighting how they can be easily circumvented through common, practical modifications to textual prompts. We systematically analyze the vulnerability of established metrics under both training-time and inference-time interventions, clearly demonstrating their fragility and motivating the need for a more robust, text-independent detection approach.

2.1. Preliminaries: Existing Memorization Metrics

We first define three widely-used memorization metrics in the context of diffusion models.

L2 Distance. The L2-Dist metric considers an image memorized if the pixel-level Euclidean distance between a generated image and a corresponding training image is below a predefined threshold (Carlini et al., 2023). This method is extremely sensitive to minor pixel-level differences and augmentations, like generating a mirror-image of the target image.

Self-Supervised Copy Detection. The SSCD metric was proposed by Pizzi et al. (2022) as a measure of semantic similarity between two images. This improves upon L2-Dist by evaluating perceptual similarity using features extracted from self-supervised models, and has been used in measuring memorization in past works (Somepalli et al., 2023b).

Text-conditioned Noise prediction. Wen et al. (2024) examine the magnitude of text-conditional noise predictions, observing that for memorized prompts, the text condition consistently guides the generation toward the memorized image regardless of initializations. Their method achieves high detection accuracy (AUC of 0.960) even at the first generation step with a single generation per prompt, making it significantly more efficient than previous approaches that require multiple generations or querying large training datasets.

2.2. Training-Time Interventions

Training-time prompt modifications are common in modern generative modeling practices. For instance, state-of-the-art models like Stable Diffusion-2 and DALL-E 3 frequently employ synthetic or refined captions to enhance training data quality (Betker et al., 2023).

Training on Enhanced Prompts (Enh-FT). We use an LLM (GPT-4o-mini) to generate semantically richer prompts from the original training captions, simulating practical enhancements aimed at clearer, more descriptive training data. By aligning training data with prompt-engineering techniques—such as appending keywords like *high-quality, 4k, ultra-resolution* or imitating specific artistic styles (e.g., camera types or film aesthetics), the model is adapted to diverse user prompts.

Training on Recaptioned Prompts (Recap-FT). We simulate a more comprehensive data labeling overhaul by completely recaptioning images with entirely new, descriptive text generated independently via an multimodal LLM. This practice has shown great promise in improving model training (Nguyen et al., 2023; Li et al., 2024). To achieve this, we

Algorithm 1 Memorization Auditing via Generative Image Compression (MAGIC)

```
Input: Target image $I_0$, diffusion model $\epsilon_\theta$, VAE encoder $\mathcal{E}_{\text{img}}$, initial embedding $e$, learning rate $\eta$, iterations $N$
Encode image: $x_0 \leftarrow \mathcal{E}_{\text{img}}(I_0)$
for $i = 1$ to $N$ do
    Sample timestep $t \sim \text{Uniform}(1, \ldots, T)$
    Sample noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
    Create noised latent: $x_t \leftarrow \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$
    Predict noise: $\hat{\epsilon} \leftarrow \epsilon_\theta(x_t, t, e)$
    Compute loss: $L \leftarrow \frac{1}{2\sigma_t^2}\|\hat{\epsilon} - \epsilon\|_2^2$
    Update embedding: $e \leftarrow e - \eta \nabla_e L$
end for
return Optimized embedding $e$
```

utilize an off-the-shelf VLM, Qwen-2.5 ((Wang et al., 2024; Bai et al., 2025)), to generate $n = 50$ unique captions per image. The captions are created by posing diverse questions about various aspects of the image, such as its attributes, context, or artistic style.

2.3. Inference-Time Interventions

We further demonstrate vulnerabilities by evaluating inference-time prompt modifications, which represent typical end-user interactions.

Inference using Enhanced Prompts (Enh-Prompt). Platforms like Ideogram and Midjourney have "Prompt Enhancers" that improve the user's input prompt by appending keywords and detailed explanations to elicit high-quality generations. To simulate this kind of honest inference-time intervention, we use an LLM (GPT-4o-mini) to enhance the original prompt, and pass the enhanced prompt to the Diffusion Model for image generation.

Inference using Recaptioned Prompts (Recap-Prompt). Since artists can't recover the exact training prompt, they can use a vision–language model to auto-caption their image and feed that into existing memorization metrics. To simulate this scenario, we use the Qwen-2.5-VL model to generate a descriptive caption of the image, and use this description as input to the Diffusion Model.

2.4. Experiments.

To ensure a fair comparison, we reproduced the fine-tuning setup and dataset released by (Somepalli et al., 2023a). We used Stable Diffusion 2.1 (SD-2.1) (Rombach et al., 2022) as the pre-trained diffusion model and fine-tuned it for 100,000 steps across all experimental setups. To construct the memorized subset, we select the 100 samples with the highest SSCD similarity scores between the generated images and their corresponding originals. Visual inspection of these samples confirms strong resemblance, suggesting likely memorization. In contrast, the non-memorized subset comprises the 100 samples with the lowest similarity scores.

Figure 1 shows that simple interventions to the prompts dramatically reduce memorization detection accuracy across all metrics. It is noteworthy that the interventions are not strictly adversarial in nature, and can be made by an honest model developer to improve the performance of their Diffusion Models.

3. MAGIC: Memorization Auditing by Generative Image Compression

Given the shortcomings of existing memorization metrics (Section 2), a robust approach to auditing memorization must satisfy two critical properties: (1) independence from the original text-image pair, requiring only the image itself, and (2) direct probing of the model's learned visual representation rather than its entire generative pipeline. Motivated by these insights, we propose MAGIC (**M**emorization **A**uditing via **G**enerative **I**mage **C**ompression), which frames memorization detection as an image compression task. Intuitively, if a model has memorized an image, it should have *compressed* information about it in the visual encoder, and hence should be able to *reconstruct* it using minimal conditioning information.

MAGIC finds an optimal continuous embedding (conditioning vector) for each target image by minimizing the model's reconstruction error on that image. This embedding can be thought of as a "compressed representation" for the image within the model's learned space. We then use properties of this embedding (such as its size) as a measure of memorization. We also show that this procedure can be used to generate the compressed reconstruction, providing a visual confirmation of memorization.

3.1. Soft Embedding Optimization for Image Compression

The core idea behind MAGIC is optimizing a short, continuous conditioning embedding e to reconstruct a given target image I_0. Specifically, we find e by minimizing reconstruction loss over the diffusion process. Let $x_0 = \mathcal{E}_{\text{img}}(I_0)$ represent the VAE-encoded image latent. Starting from a generic embedding (e.g., encoding "an image"), we iteratively update e to minimize the standard diffusion training loss:

$$L(e) = \mathbb{E}_{t,\epsilon}\left[\frac{1}{2\sigma_t^2}\|\hat{\epsilon}_\theta(x_t, t, e) - \epsilon\|_2^2\right], \qquad (1)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $t \sim \text{Uniform}(1, T)$, and $x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$. The full procedure is detailed in Algorithm 1.

Table 1: **Memorization Detection Results Across Metrics and Interventions.** We compare the performance of `MAGIC` against existing memorization detection methods, including Density$_{\ell_2}$ (Carlini et al., 2023), Density$_{\mathrm{SSCD}}$ (Somepalli et al., 2023b), and TCNP (Wen et al., 2024). Each column corresponds to a variant of the diffusion model (original or subjected to training-time or inference-time prompt modifications). Each row reports the accuracy at a 5% false positive rate (FPR) for a specific method. All baseline methods rely on fixed prompts, whereas `MAGIC` uses image-only compression. Darker green indicates higher detection performance (closer to 1.0), while red/yellow indicates performance close to random guessing.

Method	Original	Enh-FT	Recap-FT	Enh-Prompt	Recap-Prompt
Density$_{\mathtt{SSCD}}$@$\mathtt{SSCD} = 0.5, n = 4$	0.62	0.57	0.54	0.51	0.52
	Accuracy (↑) at 5% FPR Threshold on Original				
Density$_{\ell_2}, n = 4$	0.60	0.57	0.55	0.56	0.54
`TCNP`, $n = 4$	0.65	0.53	0.50	0.56	0.56
`MAGIC` (w/ L2 Norm)	0.73	0.72	0.52	0.73	0.73
`MAGIC` (w/ Token Length)	0.79	0.80	0.84	0.79	0.79
`MAGIC` (w/ Recon. Sim.)	0.81	0.79	0.73	0.81	0.81
`MAGIC` (All)	0.84	0.81	0.85	0.84	0.84

3.2. Variants of MAGIC

We test three variants of our method, each leveraging a distinct property of the optimized embedding:

Embedding Norm (L2 Norm). This variant quantifies memorization by measuring the Frobenius norm $|e|_F$ of the optimized embedding, normalized by JPEG compression size. Intuitively, a lower embedding norm indicates the model requires less information to reconstruct the image, suggesting stronger memorization. To ensure that the embedding norm is minimized during the optimization process, we apply L2 weight decay regularization during embedding optimization.

Token Length (Compression Factor). Each embedding e in Stable Diffusion has dimensions $512 \times N_{\text{tokens}}$ (e.g., 77 tokens). In this variant, we search for the minimum number of tokens needed to achieve faithful image reconstruction (SSCD score > 0.7). A smaller required token count indicates stronger memorization.

Reconstruction Similarity (SSCD). This metric represents the maximum achievable similarity (as measured by SSCD) between the original and the reconstructed image, given the full embedding dimension. A high SSCD score indicates successful reconstruction and is therefore a direct indicator of memorization capability.

Combined Metric (All). We also consider a combined approach (`MAGIC` *All*), where we jointly leverage embedding norm, token length, and reconstruction similarity through a logistic regression model trained on a validation set to predict memorization.

3.3. Experiments and Results

We present a thorough evaluation of `MAGIC` and comparison against state-of-the-art memorization detection methods:

Setup. The evaluation setup for this section extends from the discussion about the limitations of existing memorization metrics. Following established experimental setups from recent literature (Carlini et al., 2023; Somepalli et al., 2023b; Wen et al., 2024), we fine-tune Stable Diffusion 2.1 on 10,000 image-caption pairs from LAION. We then choose the 100 most memorized images based on SSCD score of the original and generated image. Since we use SSCD to find the examples, we use a fixed threshold of 0.5 to report its accuracy, as opposed to finding the threshold at 5% FPR. We adapt all methods to use the SD v2.1 model.

Results in Table 1 indicate robustness to Interventions. `MAGIC` maintains high accuracy and robustness against training-time and inference-time textual interventions, significantly outperforming prior caption-dependent methods. `MAGIC` fundamentally shifts memorization auditing from a reliance on fixed prompts towards directly probing model capabilities. This offers (i) **Robustness to prompt variability**: Independent of textual conditioning, making it resilient to common real-world modifications; (ii) **Practical scalability**: No requirement to access original training captions or large datasets—only the diffusion model and target image.

4. Conclusion

We introduced `MAGIC`, a novel method for auditing memorization in diffusion models by framing it as a generative image compression task. Unlike prior metrics that relied on prompts and failed under textual variations, `MAGIC` optimized soft embeddings to test whether a model could reconstruct an image from learned representations alone.

Empirically, `MAGIC` improved robustness and accuracy, maintaining strong performance where existing methods degraded. It successfully identified memorized images missed by other metrics and reduced false positives, providing a scalable and reliable tool for auditing memorization.

References

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3): 8, 2023.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 267–284. USENIX Association, 2019. URL https://www.usenix.org/conference/usenixsecurity19/presentation/carlini.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chavhan, R., Li, D., and Hospedales, T. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*, 2024.

Chen, Z., Liu, Z., Lin, T., Jiang, Y., and Zhang, Q. Towards memorization-free diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jiang, Y., Lin, H., Bai, Y., Peng, B., Liu, Z., Lyu, Y., Yang, Y., Xingzheng, and Dong, J. Image-level memorization detection via inversion-based inference perturbation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vwOq7twk7L.

Li, X., Tu, H., Hui, M., Wang, Z., Zhao, B., Xiao, J., Ren, S., Mei, J., Liu, Q., Zheng, H., et al. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*, 2024.

Lu, Y., Li, Y., Wang, Z., Li, Q., Zhang, H., and Yang, H. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ma, Z., Li, Q., Zhang, X., Du, T., Lin, R., Wang, Z., Ji, S., and Chen, W. An inversion-based measure of memorization for diffusion models. *arXiv preprint arXiv:2405.05846*, 2024.

Nguyen, T., Gadre, S. Y., Ilharco, G., Oh, S., and Schmidt, L. Improving multimodal datasets with image captioning. *arXiv preprint arXiv:2307.10350*, 2023.

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

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2017. doi: 10.1109/SP.2017.41. URL https://arxiv.org/abs/1610.05820.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Song, Y. and Ermon, S. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Sustainable Technology Partners. Ai lawsuit timeline. https://sustainabletechpartner.com/topics/ai/generative-ai-lawsuit-timeline/, 2023. Accessed: 2023-09-25.

Vyas, V. and Abbe, E. On provable copyright protection for generative models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., and Lin, J. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Webster, R., Rabin, J., Simon, L., and Jurie, F. Detecting gan-generated imagery using color cues. In *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4234–4238. IEEE, 2019.

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.

Zhang, W., Iwamura, M., Zeng, M., Shi, Y., and Kiya, H. Copyright protection in generative ai: A technical perspective. *arXiv preprint arXiv:2302.02333*, 2023.

A. Related Work

Memorization in generative models, especially diffusion models, has become a critical area of research due to its implications for privacy, security, and intellectual property rights. This section reviews prior work on analyzing, detecting, and mitigating memorization in diffusion models, as well as technical perspectives on copyright protection in generative AI.

Memorization in Generative Models. Memorization has studied across classes of generative models such as GANs and LLMs. Webster et al. (2019) and others showed that GANs can sometimes overfit and reproduce training examples, proposing metrics to detect such behavior. In NLP, Carlini et al. (2019) was one of the first to demonstrated that large language models can emit training data verbatim, raising privacy concerns. However, diffusion models differ substantially in generation process (iterative noising/denoising) and conditioning mechanisms, rendering many existing techniques inapplicable or ineffective. Early evidence of memorization in diffusion models was reported by Somepalli et al. (2023a), who noted that models trained on extremely limited or duplicated data tend to regurgitate specific images. Subsequent work by Carlini et al. (2023) provided a large-scale confirmation: they extracted dozens of exact training images from Stable Diffusion by cleverly searching the latent space and prompt space. These studies established that diffusion models *do* memorize some training images, especially those that are repeated in the data or have unique features.

Definitions and Detection of Memorization in Diffusion Models. Due to the continuous nature of image generation, defining what constitutes memorization is non-trivial. A straightforward definition is that a model has memorized an image if it can reproduce it with sufficiently high fidelity (e.g., low perceptual distance) when given the right prompt or latent representation. This definition was operationalized by Carlini et al. (2023) by searching for prompts that yield images nearly identical to known training examples. A key challenge is setting a similarity threshold — too strict a threshold misses cases of near-memorization, while too lenient a threshold may flag merely similar outputs as memorized. Somepalli et al. (2023b) and Wen et al. (2024) observed that certain rare *prompt phrases* or *trigger words* can consistently cause a model to output the same image, effectively acting as keys to memorized content. Building on this, Wen et al. (2024) proposed an automated test: if multiple generations with different random seeds for a given prompt yield nearly identical images, then the model has memorized that content. Another related field is that of *membership inference*: given an image, determine if it was in the training set (Shokri et al., 2017; Carlini et al., 2022). For diffusion models, the task is harder because the model does not explicitly output training samples unless specifically prompted. Our compression-based approach can be seen as a type of membership inference attack specialized for diffusion generative models.

Recently, (Jiang et al., 2025) introduced introduced Inversion-based Inference Perturbation (IIP), a framework for detecting image-level memorization in diffusion models without relying on prompt information. InvMM (Ma et al., 2024) is another inversion-based metric that quantifies image-level memorization in diffusion models by estimating the KL divergence between sensitive latent noise distributions and a standard Gaussian prior. Both IIP and InvMM are computationally intensive, making it less practical for large-scale applications.

Mitigating Memorization and Content Removal. Several works address how to prevent or limit memorization in generative models. Chen et al. (2024) proposed strategies to train diffusion models that are less prone to memorizing training data. Data pruning and de-duplication prior to training have been suggested to reduce overfitting on near-duplicate images (Somepalli et al., 2023a; Carlini et al., 2023). Recently, techniques for *concept erasure* in diffusion models have emerged. Lu et al. (2024) introduced `MACE`, a finetuning framework that uses targeted LoRA updates to remove the ability to generate specific concepts from a model. Similarly, Chavhan et al. (2024) explored pruning or editing model weights associated with memorized content, effectively ”forgetting” that content without retraining from scratch. These approaches are complementary to ours: while they aim to scrub memorized data from models, our focus is on *detecting and characterizing* memorization.

Ethical and Legal Perspectives. The ability of generative models to reproduce training images verbatim has direct implications for copyright and privacy. Zhang et al. (2023) provided a technical perspective on copyright protection in generative AI, discussing how models might infringe on intellectual property rights. Vyas & Abbe (2023) explored provable copyright protection mechanisms for generative models. Our work falls under the broader umbrella of AI model auditing and transparency, sometimes termed the ”blue team” approach in AI security: developing tools to ensure models behave responsibly. We note that distinguishing memorization from mere style imitation is an open question: generative models often learn to mimic artistic styles or compositions from training data without copying any single image exactly. This gray area—between permissible inspiration and impermissible plagiarism—remains an area for future work and likely policy intervention.

B. Preliminaries on Diffusion Models

B.1. Diffusion Model Background and Notation

We briefly review the diffusion model setup. Given an image latent x_0, the forward diffusion process progressively corrupts it into a noisy latent x_t:

$$q(x_t \mid x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t} x_0, (1 - \bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$ with a predefined variance schedule β_s. The reverse (denoising) process employs a U-Net parameterized by θ to predict noise ϵ given latent x_t and text conditioning e:

$$\hat{\epsilon}_\theta(x_t, t, e) \approx \epsilon, \quad \text{where } x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon. \tag{3}$$

In text-to-image setups (e.g., Stable Diffusion), e typically originates from a CLIP-based text encoder. Critically, e need not correspond to actual textual prompts—it can be directly optimized as a continuous representation, providing the basis for our soft embedding approach.

C. More Details about MAGIC

Normalization via JPEG Compression. To account for the intrinsic complexity of images (i.e., some images inherently compress better due to their simplicity), we normalize all our embedding metrics relative to JPEG compression size, a “zero knowledge” baseline as introduced by Somepalli et al. (2023b). This normalization helps control for the inherent compressibility of an image, thus isolating the memorization signal more accurately.

The Importance of Initialization. In order to converge to the solution fast, we notice that initializing the text guidance with the actual prompt of the image can be extremely beneficial. In particular, by assuming access to either the (i) original text associated with caption; or (ii) a VLM generated caption of an image, we can significantly speed up the optimization process. This observation both points to a practical strategy to yield the best results out of MAGIC, but also an important limitation of the optimization process itself, which would be of interest for future work.

C.1. Visualization of the Soft Prompt Optimization

In Figures 2 and 3, we show how the Soft Embedding Optimization method proceeds for different Token Length values, on one example from the memorized and non-memorized set each. For the memorized example in Figure 2, we see that token length of 8 onwards start producing a similar looking final image as the target image. It also shows that increasing the number of tokens beyond a certain point does not help much in improving the quality of the final reconstruction. However, in Figure 3, we see that for all different token lengths, the final image is still quite different from the target image in both details and style. This implies that no such soft embedding was found by the optimization algorithm that would elicit generation of this target image easily.

D. How much copyrighted content do frontier models memorize?

D.1. Audited Characters from Major Copyright Holders

We systematically selected the top ten most popular characters or images from each copyright holder using Google image search queries. For reproducibility, the exact characters used in our audit are listed below.

Marvel Characters

Thanos, Deadpool, Doctor Strange, Wolverine, Black Panther, Hulk, Thor, Iron Man, Captain America, Spiderman

Pokémon Characters

Pikachu, Charizard, Mewtwo, Eevee, Snorlax, Jigglypuff, Gengar, Bulbasaur, Squirtle, Charmander

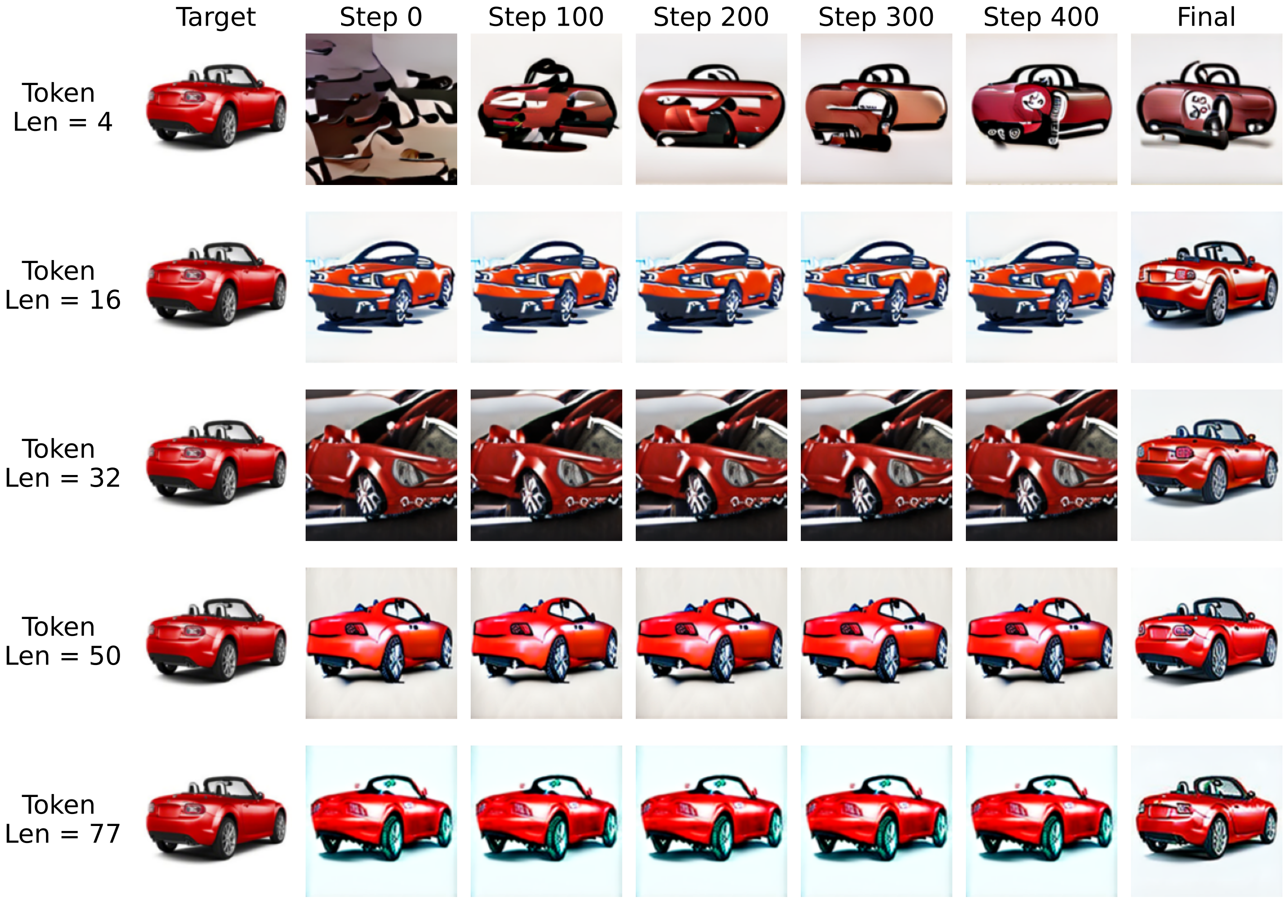

Figure 2: Soft Prompt Optimization Process for varying Token Lengths for an image from the Memorized Set.

Disney Characters

Mickey Mouse, Minnie Mouse, Goofy, Donald Duck, Simba, Elsa (Frozen), Woody (Toy Story), Ariel (The Little Mermaid), Buzz Lightyear, Belle

New Characters

Envy, Anxiety, Ennui, Embarassment (from Inside Out 2), Pecharunt (Pokemon), Ironheart Riri, Maystorm (Marvel), Grape (Nintendo)

All image selections were performed in May 2025, capturing current search-engine popularity, and the first image appearing on Google image results was consistently chosen.

Having redefined memorization detection through generative image compression (Section 3), we now leverage our approach to empirically audit memorization in frontier diffusion models, focusing on popular copyrighted characters. Unlike prior methods, which depend heavily on static image-caption pairs and fail to reliably detect memorization under minor textual perturbations (Section 2), MAGIC robustly identifies memorization purely based on image content. Using the thresholds established in Section 3.3 (corresponding to a 5% FPR), we perform a rigorous evaluation on a systematically curated set of iconic images representing major intellectual properties.

Specifically, we evaluate ten popular and recognizable characters or images from each of the following major copyright holders: **Pokémon, Disney**, and **Nintendo**. Images are selected using standard Google searches (the first image result for each character), simulating plausible and common scenarios of memorization in publicly-trained models.

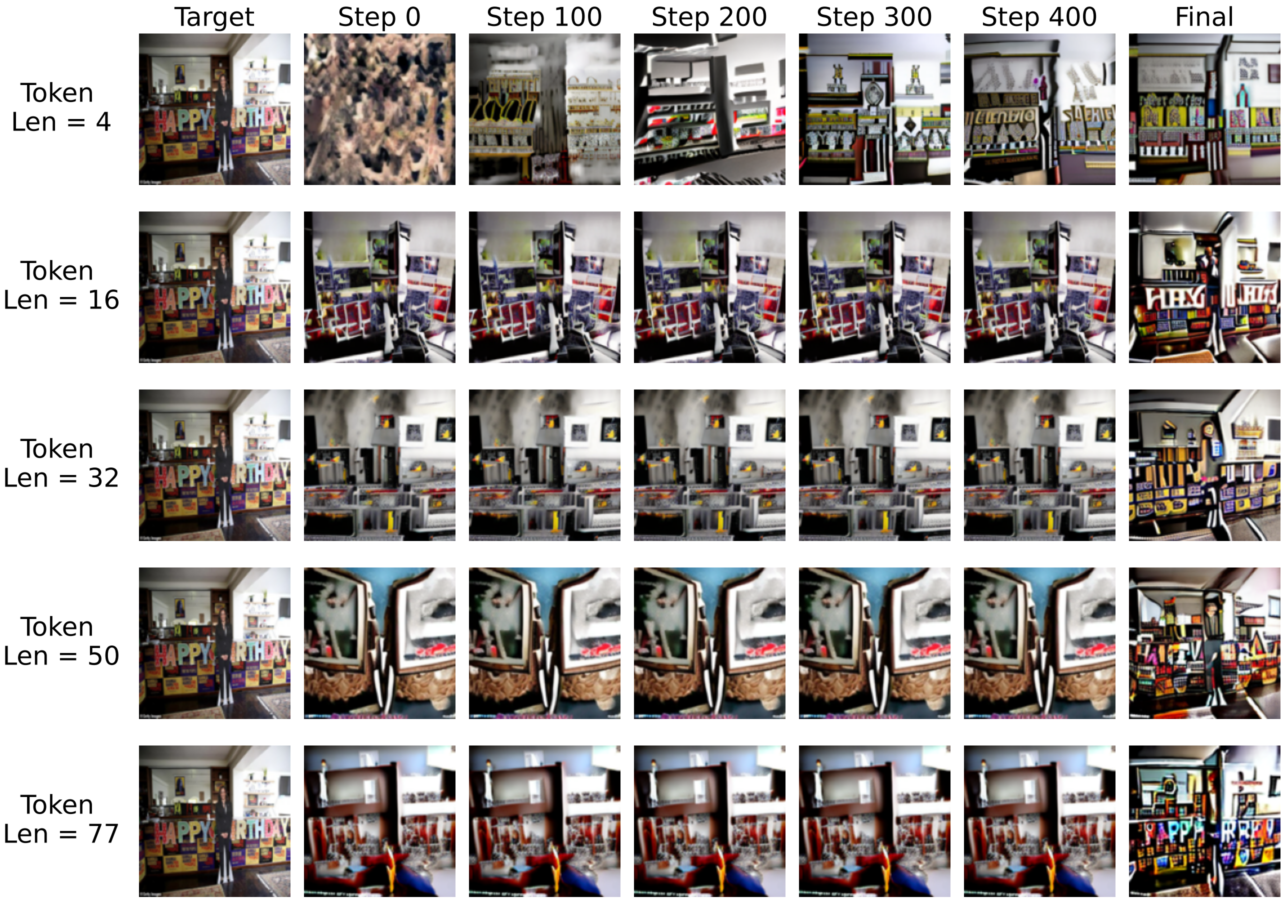

Figure 3: Soft Prompt Optimization Process for varying Token Lengths for an image from the Non-Memorized Set.

Figure 4 summarizes our findings. We observe substantial memorization across all tested copyright holders. Notably, Marvel exhibits the highest number of memorized instances (9 out of 10), closely followed by Pokémon and Disney, each with 8 instances memorized.

Importantly, our evaluation reveals that certain images, like iconic depictions of characters (e.g., Batman or Pikachu), have been memorized multiple times under various different captions. This insight highlights a crucial limitation of prior caption-dependent metrics (Wen et al., 2024; Carlini et al., 2023), which cannot effectively handle such real-world memorization scenarios.

This analysis underscores the pressing need for robust memorization auditing tools, particularly in light of ongoing lawsuits, such as the well-known litigation against Stability AI by artists and corporations alleging unauthorized reproduction of protected content (Sustainable Technology Partners, 2023). By quantifying memorization concretely and independently from textual conditioning, MAGIC provides valuable transparency for regulatory scrutiny, legal contexts, and ethical development of generative AI technologies.

E. Future Work and Limitations

Several promising directions arise from our work. First, extending MAGIC to other generative modeling paradigms, such as video (like SORA) or 3D generative models, is an immediate and compelling avenue. Given the increasing prominence and potential privacy implications of these modalities, adapting our compression-based auditing framework could significantly advance the understanding and control of memorization across diverse media types. Second, our method currently focuses strictly on exact or near-exact image reconstructions. This specificity may miss nuanced forms of memorization, such as semantic or stylistic memorization, which may also carry significant ethical implications. Examining partial or semantic-level

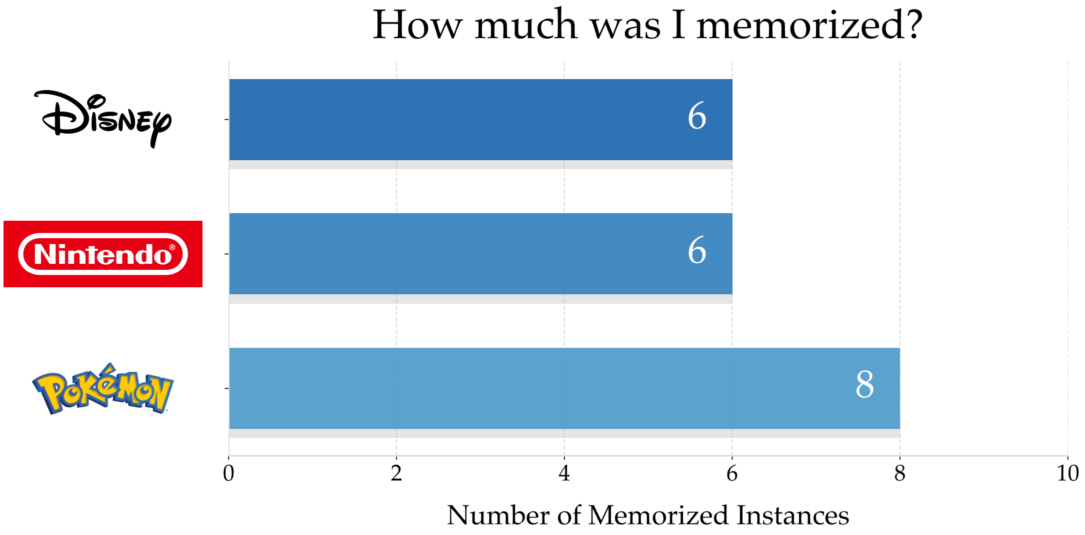

Figure 4: How much copyrighted content does SD-2.1 memorize based on the copyright holder? We take the top 10 most famous copyrighted characters for each of the above companies, and then use MAGIC on those images to test whether they were memorized by the diffusion model. We find that more than 50% of copyrighted characters were memorized by the model.

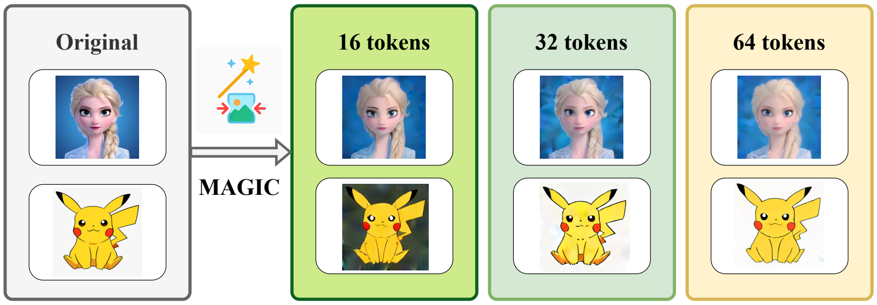

Figure 5: **Examples of Memorized Instances Detected by MAGIC.** One representative memorized image per company as detected by our proposed method. The images shown were faithfully regenerated by frontier diffusion models, clearly illustrating the practical importance of auditing memorization.

memorization remains a critical next step. Third, MAGIC still requires iterative optimization, especially if initialized poorly or without prior textual guidance. Future work optimizing this process could make the optimization process more reliable and fast.