STYLE MEMORIZATION IN TEXT-TO-IMAGE GENERATIVE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

While the intended use of generative models is to create novel content, they can generate expressive content on a large scale that closely resembles training data, particularly memorized styles of specific artists, competing with human creators in existing markets. Past research overlooks this type of subtle memorization by focusing on exact/near-exact copies and by failing to account for the inherit difficulty of reproducing atypical semantic properties, like artistic styles. To distinguish expected from unexpected/undesired capabilities, this work proposes (τ, λ)-memorization that defines an extractable artistic style as memorized if it is sufficiently atypical at the image-level (λ) and caption-level (τ), as these *outlier-properties* are inherently difficult to reproduce. Experiments show that the popular general-purpose model Stable Diffusion memorized 34% of artistic styles, producing fake artworks based on non-overly descriptive prompts that enable artist re-identification with visual search, highlighting potential market impact. With adequate thresholds τ and λ to define caption and style atypicality, style memorization is detected with over 20% TPR at 1% FPR, unlike prior work that only achieves TPR as high as FPR. This work shows that atypical artistic styles are more likely to be memorized, particularly underrepresented ones, exposing an underestimation of memorization risks of current frequency-based indicators, thus informing data owners of risks of style memorization and the need for appropriate mitigation strategies.

1 INTRODUCTION

Advances in generative models (GMs) enable the generation of high-quality images based on natural prompts that can be used for commercial art and graphic design purposes (Somepalli et al., 2023a) and the significant efficiency improvement brought by Stable Diffusion (Rombach et al., 2022) increased accessibility, leading to widespread use (Rombach et al., 2022). Alongside remarkable advances in image synthesis, concerns have emerged from memorization of training data and its implications for privacy (Carlini et al., 2023a; 2019; 2021) and intellectual property preservation (Somepalli et al., 2023a;b), as GMs exceed their expected capabilities by generating exact/near-exact replicas of memorized training samples instead of novel images. Particularly, since training datasets often contain copyrighted works, artists fight against AI companies for repurposing their work (Associated Press, 2023), claiming that synthetic images very similar to their own distinctive style can be generated by these models when prompted with their name. This threat is even more acute due to widespread use of AI systems, such as the viral Ghibli-style imagery case (Deckker and Sumanasekara, 2025), motivating discussions on the copyright implications of using copyrighted works to train GMs (U.S. Copyright Office, 2024; 2025a;b). Solutions to this problem include blocking requests to generate images in the style of living artists OpenAI (2023). However, users can use other trigger-tokens that co-occur with the artist's name, such as "starry night" that co-occurs with "Van Gogh" (Naseh et al., 2023), or only a part of his/her name to generate stylistically similar content.

This debate revolves around the interpretation of copyright law that grants authors of copyrightable works, such as paintings, the exclusive right to reproduce and prepare derivative works (e.g. transformations/adaptions), which extends to its protected elements and to substantially similar non-literal copies (U.S. Copyright Office, 2021; 2025b). Even though artistic styles are generally not copyrightable (U.S. Copyright Office, 2024), they can still contain protectable elements (Sobel, 2024), making it difficult to determine the scope of protection, as the more abstractly a copyrighted work is protected, the more likely a model is to copy it (U.S. Copyright

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

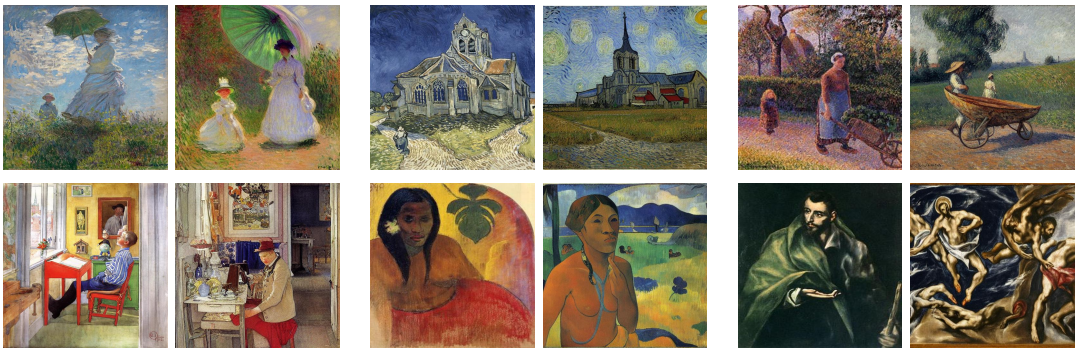


Figure 1: Artistic styles extracted with non-overly descriptive prompts: *"claude monet woman parasol madame son 1875"*, *"church auvers vincent van gogh sur oise orsay paris"*, *"woman wheelbarrow pissarro camille"*, *"painting carl larsson" "paul gauguin"*, and *"el greco"*. Training image on the left and synthetic on the right.

Office, 2024). This infringement, derived from the production of stylistically similar images, can be excused as fair use, depending primarily on the model's purpose, including the extent to which it is transformative, and the potential market impact (U.S. Copyright Office, 2025b). Various uses of copyrighted works for training are likely to be transformative, namely those with a different purpose, e.g. predictive models that extract patterns instead of replacing/competing with original works. However, even if the intended use is transformative, GMs can be exploited for non-transformative uses, such as using general-purpose models, like Stable Diffusion, for style mimicry (Associated Press, 2023). Thus, training GMs with copyrighted works to produce expressive content that competes with them in the market, especially given the speed and scale at which these models generate content, goes beyond established fair use boundaries (U.S. Copyright Office, 2024; 2025b). While the copyrightability of a specific artist's style remains subject to interpretation and consequently whether style mimicry constitutes copyright infringement, the large-scale generation of expressive content that competes with human creators in the market raises ethical concerns, creating a competition bias in several domains, such as visual art (Associated Press, 2023), music (Axios, 2026) and literature (The Authors Guild, 2023).

To understand this undesired behaviour, past research focused on memorization as the primary cause of training-data reproduction in GMs, studying how it occurs inside the model (Wen et al., 2024; Hintersdorf et al., 2025; Ren et al., 2024; Jain et al., 2025; Chen et al., 2025) and its implications for privacy (Carlini et al., 2023a; 2019; 2021) and intellectual property preservation (Somepalli et al., 2023a;b) with a particular focus on exact/near-exact memorization. However, training-data reproduction can occur in more subtle ways, such as only replicating specific patches (Somepalli et al., 2023a; Naseh et al., 2023) or copying the semantic meaning of training text (Ippolito et al., 2023). Detecting this phenomenon requires the distinction between the ability to generate diverse and novel images that reflect learned features, i.e. generalization (Carlini et al., 2023a), from *subtle memorization*, characterized by the persistent generation of a close-replica of an atypical semantic property, like a specific artist's style to the point of allowing artist re-identification. This is particularly concerning when it is generated by general-purpose models, like Stable Diffusion, that were not intended to reproduce artist-specific styles with such accuracy, especially with non-overly descriptive prompts insufficient to guide generation. As latter shown, prior work fails to detect these cases, which challenges the ability to ensure that the use of copyrighted works remains aligned with the intended transformative purpose of these models, impacting the protection of artistic styles.

To address this challenge, this work proposes $(\tau, \lambda)-$memorization to quantify the memorization of specific artists' styles, a definition that accounts for the inherit difficulty of reproducing a particular style by relying on the quantification of its degree of atypicality at the image-level with Style-Atypicality (λ), i.e. how well-defined and distinguishable it is, and at the caption-level with Caption-Atypicality (τ), i.e. how unusual their key tokens are. The idea is that reproducing common semantic properties, like a dog in a monochromatic background, is an expected capability of general-purpose models and is therefore not considered memorization. In contrast, consistently generating images that closely resemble the main visual characteristics of specific artists, with non-overly descriptive prompts, such that visual search engines can re-identify the original artist, does not fall within the expected behaviour of such models. Supported by Carlini et al. (2023a) that argues

that the non-random training-data reproduction, in whole or in part, is treated as memorization, this work treats this capacity as style memorization. Experiments show that Stable Diffusion, a general-purpose model, replicates memorized artist-specific styles when prompted with non-overly descriptive requests, like the artist's name (Fig. 1), with substantial similarity such that visual search engines can re-identify the artist based on these fake/synthetic artworks. When atypical styles are quantified with suitable thresholds τ and λ, a simple metric-based approach achieves remarkable results under a low error regime, particularly over 20% TPR for a fixed 1% FPR, while the state-of-the-art definition (Carlini et al., 2023a) can only achieve TPR as high as FPR, as it cannot filter out non-atypical styles nor allow for flexibility in the output that remains artist-recognizable. Experiments show that atypical artistic styles are more likely to be memorized, particularly less represented ones thereby improving on common frequency-based indicators. Hence, the proposed definition of style memorization can help data owners to make informed decisions about whether to opt in/out or apply mitigation techniques against style memorization by quantifying their effectiveness.

2 Related work

Research on memorization in GMs introduces formal definitions of memorization that separate the intended model's behaviour from unintended/undesired reproduction of training data. Table 1 summarizes the main definitions along three dimensions: 1) the *granularity* at which memorization is assessed, distinguishing sample (individual) from group-level memorization; 2) the type of *extraction* ranging from exact/near-exact to more flexible forms of partial replication; and 3) the underlying *principle* used to separate memorization from expected model behaviour, including representativeness, identifiable information, and hardness of extraction. Overall, this table reveals two limitations in prior definitions. First, most existing definitions study memorization at the sample-level and therefore overlook memorization of properties that emerge across groups of samples, like the style. Second, they typically look for exact/near-exact replication, missing subtler forms of information leakage. However, Entity Memorization (Zhou et al., 2024) is a partial exception: in the context of Large Language Models (LLMs), it allows outputs to differ from the training example as long as sensitive target entities, like names or phone numbers, are recovered. Recent work has increasingly recognized partial memorization as a privacy concern (Carlini et al., 2023a), particularly FB-Mem (Di et al., 2025) captures inexact memorization using a segmentation-based metric to identify cases where only part the training image is reproduced. However, FB-Mem is a measurement framework which does not establish when replication is part of the expected model behavior or undesired leakage of private/copyrighted content, unlike a definition.

Additionally, prior work mainly innovates on the underlying principle that distinguishes expected from unexpected behaviour. Particularly, (k, ℓ, ϵ)-eidetic memorization and (λ, γ)-copy use the target sample's representativeness, controlled by k and γ, based on the intuition that approximating common samples is expected when the model learns the training distribution properly. However, this frequency-based principle is restrictive when extending the definition to other settings, as highly frequent content may still be undesirable to reproduce, such as the artistic style of well-known artists. Moreover, unintended memorization and entity memorization use the fact that some sequences extracted from LLMs are irrelevant for training (e.g. names and phone numbers) and uniquely identify training samples, showing that memorization can be orthogonal to the learning task (Carlini et al., 2019). Furthermore, (n, p)-discoverable extraction is based on the hardness of extracting sequences from LLMs using the trade-off between n (maximum number of queries) and p (fraction of copies), whereas τ-compressed memorization based on the relative size τ of the minimal prompt that triggers completion. These definitions assess memorization from the perspective of the user/adversary given its limited access in some setups, but most importantly, they account for a probability and compressibility-based estimate of the hardness of extracting memorized content in LLMs. Hence, accounting for the inherit difficulty of extracting memorized content, particularly in text-to-image GMs, is still an unaddressed challenge.

Furthermore, past research studies the effect of data and model-related risk factors, namely data duplication (Carlini et al., 2023a;b; Hayes et al., 2024; Naseh et al., 2023; Somepalli et al., 2023a;b), presence of outliers (Carlini et al., 2023a; 2022b; Borkar et al., 2025; Schwarzschild et al., 2024; Zhang et al., 2023), model size and longer training processes (Carlini et al., 2021; 2023b; Hayes et al., 2024), which have been shown to increase memorization. The text-conditioning (prompt) also has important role by affecting the extraction of memorized content (Somepalli et al., 2023b), as there seems to be a log-linear relationship between prompt size and extractable memorization in LLMs (Carlini et al., 2023b) and the presence of specific tokens can act as a trigger for models to replicate memorized content (Wen et al., 2024; Ren et al., 2024; Chen et al., 2025).

Definition	Granularity	Extraction	Principle	Summary
(k, ℓ, ϵ)-eidetic memorization Carlini et al. (2023a; 2021)	Individual	Exact	Representativeness	Extractable samples with at most k training duplicates, that fall within a maximum ℓ-distance of ϵ, are memorized.
Unintended memorization Carlini et al. (2019)	Individual	Exact	Identifiable information	LLMs memorize out-of-distribution sequences that are irrelevant for training, showing that memorization can be orthogonal to learning.
(λ, γ)-copy Bhattacharjee et al. (2023)	Individual	Exact	Representativeness	A sample is memorized if its real probability is $\leq \gamma$ and its synthetic probability is at least λ times proportional.
(n, p)-discoverable extraction Hayes et al. (2024)	Individual	Exact	Hardness of extraction probability-based	A sample is memorized if it is extractable with probability of at least p in n queries.
τ-compressed memorization Schwarzschild et al. (2024)	Individual	Exact	Hardness of extraction compressibility-based	Memorized samples can be completed by LLMs with minimal prompt (i.e. ratio of training sample and prompt size $> \tau$).
Entity memorization Zhou et al. (2024)	Individual	Partial	Identifiable information	Sample with m entities is memorized if LLMs output expected entities when prompted with $n(< m)$ identifying entities.

Table 1: Summary of past definitions of memorization in generative models.

To address these challenges, this work proposes $(\tau, \lambda)-$memorization that defines an extractable artistic style as memorized if it is *well-defined and distinguishable from other styles* (λ) and the *corresponding caption is sufficiently atypical* (τ). In other words, only *outlier-properties* can be considered memorized, as their replication is inherently difficult. $(\tau, \lambda)-$memorization innovates upon past definitions by: 1) capturing memorization of a different level of granularity (group, as the target style is present in several images); 2) expecting a different type of extraction (partial, as it expects the same style not the exact same image); and 3) using a different principle (atypicality, as only outlier-properties at the image and caption-level can be memorized). The goal is to quantify the extent to which the reproduction of an abstract concept, such as a specific artistic style, deviates from the intended behaviour, useful to enable the interpretation and application of copyright law, as well as to assess the effectiveness of mitigation strategies, as the more abstractly a copyrighted work is protected the more likely it is to be copied and the harder it is to define the scope of protection (U.S. Copyright Office, 2025b). The contributions of this work are the following:

- $(\tau, \lambda)-$memorization, a definition of memorization that accounts for the inherit difficulty of reproducing a specific artistic style, along with caption-atypicality (τ) and style-atypicality (λ);
- Experiments revealing that general-purpose GMs memorize artistic styles and generate fake artworks that enable artist re-identification via visual search, which $(\tau, \lambda)-$memorization detects effectively;
- Experiments revealing that the state-of-the-art definition fails to capture style memorization;
- Evidence suggesting that style and caption atypicality are risk factors for style memorization, particularly among artistic styles less prone to memorization according to frequency-based indicators.

3 (τ, λ)-MEMORIZATION

Motivated by ethical and legal concerns raised by style mimicry and by the difficulty in distinguishing expected capabilities of a high-quality general-purpose model from undesired/unexpected subtle replication of training data, this work proposes a novel definition of style memorization that aligns with prior work that shows that outliers are more prone to memorization while enhancing a common concept of outliers (i.e. distant images in the embedding space) with a multi-modal concept that covers outliers at the caption and image-level.

Definition ((τ, λ)-memorization)**.** *An extractable artistic style is memorized if it is sufficiently atypical at the image-level (Style-Atypicality$(P) \geq \lambda$) and at the caption-level (Caption-Atypicality$(T) \geq \tau$).*

The intuition is that the consistent generation of high-quality copies of an *atypical semantic property*, such as a specific artist's style, is not expected/desired from models whose training objective is not style mimicry. A particular artist's style corresponds to a relatively small set of visually distinct yet internally related images that follow an underlying distribution (style) that is *inherently difficult to reproduce* and associated with captions that are *unusual within the training dataset*, as they often include the artist name (frequently non-English). When the model consistently generates high-quality replicas such that visual search engines can re-identify the artist, it not only learned the artist's genre, but it retained artist-specific information to an unusual extent, challenging the ability to ensure that the use of general-purpose models remains transformative as intended, and, most critically, creating a competition bias against such artists, by enabling the large scale generation of similar artworks. However, caption and style atypicality only flag styles as high-risk, for a style to be classified as memorized it must be *extractable*, i.e. the model must produce substantially similar copies, which is determined by visual search, particularly its ability to re-identify the artist solely based on these copies.

Caption Atypicality This score is calculated at the group-level, after grouping individual samples containing the same style, using the text-identifier of the group, which is based on the caption of every sample within the group – set of overlapping tokens/words in most captions (e.g. 75%), and is given by the *per-token importance weighted by how inherently atypical/odd each token is* (Eq. 1). The per-token importance i_t is determined by the contribution of each token to the overall entropy of T, where T is the caption's embedding matrix and $T_{\setminus t}$ the embedding matrix except token t. The inherit token atypicality/oddness o_t is determined by the cosine dissimilarity D_{cos} between each token w.r.t the text-encoder vocabulary, where t is the embedding vector of a token and μ the average embedding vector of the entire vocabulary. In other words, *atypical captions are those that contain important and inherently odd tokens*. Even though containing the artist name is not required, it typically increases caption atypicality in comparison with common captions, e.g. minimal image description.

$$\text{Caption-Atypicality}(T) = \sum i_t \cdot o_t;\ i_t = \text{entropy}(T) - \text{entropy}(T_{\setminus t}),\ o_t = 1 - D_{cos}(t, \mu) \tag{1}$$

Style Atypicality This score is calculated over the style representative of a group of samples if it contains a sufficiently well-defined style – given by a confidence score that quantifies the existence of a single and well-defined target style unlike a group of random photos, as explained in Section A.1. Let P be the set of samples with well-defined styles, style atypicality is given by the average distance of every style p to every other style q in the set P, as shown in Eq. 2, where the distance ℓ is weighted by confidence scores c_p and c_q to reduce the influence of groups with low confidence scores (ill-defined styles). Note that atypicality is not equivalent to rarity in the training distribution as the goal is to account for the style's visual distinctiveness. Scoring a high style atypicality means that the style is *well-defined*, given the weight of the confidence score, and *distinguishable* by being distant from the rest.

$$\text{Style-Atypicality}(p) = \frac{\sum_{q \in P} \frac{c_p + c_q}{2} * \ell(p, q)}{|P|} \tag{2}$$

Generalization vs memorization: While it is unsurprising that prompts with the artist's name result in similar images, e.g. of the same genre, the fact that visual search re-identifies the artist solely based on synthetic images (Sec. A.5) means that the model retained artist-specific information to an unusual extent s.t. the output consistently resembles the main visual characteristics unique to the artist. Supported by past research that argues that *normal generations* vary substantially with the random seed while memorized ones recur repeatedly (Carlini et al., 2023a; Jain et al., 2025), this persistent behaviour deviates from the intended transformative use of models whose training objective is not style mimicry, like general-purpose models. This is especially concerning when it is achieved with prompts that do not explicitly request nor guide style mimicry, revealing that this behaviour is dominated by the influence of associations present in copyrighted works used for training rather than on information from the prompt. (τ, λ)-memorization can distinguish the ability to generate new and diverse content that reflects learned non-atypical features, like an object in a monochromatic background, from memorization of artist-specific styles, which is achieved by fixing high thresholds τ and λ, where the role of τ is to only keep styles associated with unusual captions (Caption-Atypicality$(T) \geq \tau$) and the role of λ is to filter out ill-defined styles that are similar to others (Style-Atypicality$(P) \geq \lambda$), given that approximating non-atypical styles is easy and hinders detection of style memorization, as latter shown in Sec. 4.1.

In practice, to detect style memorization adequate thresholds τ and λ must be fixed, along with the identification of substantially similar copies of a given style, which are those with an ℓ-distance w.r.t. training images of at most ϵ following past research that uses (ℓ, ϵ)-extraction (Carlini et al., 2023a). The threshold ϵ is applied over the memorization signal ℓ (style loss) to classify the extraction as successful/unsuccessful which does not affect the concept of memorization determined by τ and λ, unlike past definitions, such as (k, ℓ, ϵ)-eidetic memorization (Carlini et al., 2023a) as latter shown in Section 4.2. This independency is relevant as it allows more flexible extractions (i.e. subtle similarity instead of exact) by increasing ϵ without giving in tight boundaries of memorization, imposed by high τ and λ. Thus, memorized styles can be detected by maximizing τ and λ such that it includes artistic styles of specific artists and minimizing ϵ.

Regarding the style representation and distance, this work relies on the state-of-the-art metric: gram-matrix representation and Mean Squared Error (MSE) (Gatys et al., 2016; Li et al., 2017). This choice was motivated by the fact that the style can be perceived by how each pixel/patch relates to its surrounding pixels/patches, and this correlation structure reflects the signature of the underlying pixel/patch-generating distribution (style). To improve style loss, a rank-based dimensionality reduction was applied to the style representation such that it is only computed along the most informative dimensions. Confidence scores, that quantify if the style is well-defined, are determined by its inherit complexity, given by the Frobenious norm computed along the most informative dimensions, meaning the strength of correlations between the most informative pixels/patches. Section A.1 presents more details about this adaptation.

4 RESULTS

To show that the proposed definition effectively quantifies style memorization, this section presents evidence that Stable Diffusion 1.4 (SD1) replicates specific artistic styles present in its training dataset which can be detected with a metric-based approach. This section starts with an overview of the experimental pipeline and threshold-selection strategies, followed by metric-based detection results in Section 4.1, comparison with prior work in Section 4.2, analysis of risk factors in Section 4.3, and a summary in Section 4.4.

Experimental setup overview Experiments follow a four-step pipeline: 1) build target dataset, 2) perform extraction experiments independently for each group, 3) report distance-based metrics to summarize experiments, 4) detection. First, the target dataset is constructed by organizing training samples into groups (see sections A.2 and A.3). Groups containing an artistic style are identified by the presence of the artist's name in at least one caption. Then, a sentence is derived from the group's captions to prompt the model as an attempt to extract the target style (with 250 different seeds), resulting in small prompts with the artist name, when it is in most captions, and in some cases a few words describing the foreground, as shown in Fig. 1 and explained in Sec. A.6. Distance-based metrics are reported per experiment which requires the selection of a compact cluster of synthetic images that best approximates the style which must have at least 25 samples, thereby ensuring that any style replication is not random, as it is in at least 10% of the outputs. Finally, these metrics are used as inputs for detection, which requires ground truth labels, i.e. classification of every group w.r.t. style memorization, which is determined by a visual search pipeline (Sec. A.5) that parses synthetic images and returns the artist of the most similar artwork (if any), leading to 48 (≈ 34) memorized artistic styles. Visual search engines, like Google Lens, can verify that these fake artworks (synthetic images) are not only visually similar, in terms of style, but contains artist-specific information that can be linked to the artist, therefore raising copyright concerns. Additional details about the setup and classification are in Sec. A.4 and A.5.

Threshold-selection strategies In practice, thresholds τ and λ limit the number of groups kept for further experiments to those whose caption and style atypicality are sufficiently atypical. The ultimate goal is to improve the discriminative power of the style loss by discarding groups whose style is close to the "average style" since their replication is to be expected. This is achieved with 3 different threshold-selection strategies, namely by fixing one parameter as a function of the other or by varying both parameters jointly. The style-guided strategy is governed by style atypicality (λ): it starts with the highest values of λ and lets τ be a function of it to achieve a specific goal, in this case the desired number of groups containing an artistic style regardless of the memorization label. Analogously, the caption-guided is governed by caption atypicality (τ) and lets λ be a function of it. Lastly, both parameters can be jointly tuned such that the resulting target dataset maximizes the discriminative power of the style loss by maximizing ROC AUC, referred as optimized strategy.

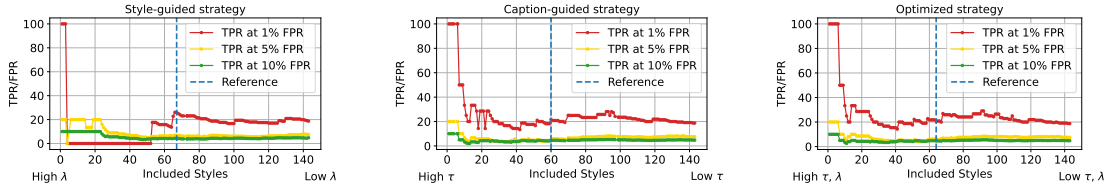
Regardless of the strategy, a restrictive definition can be achieved by fixing high τ and λ, and decreasing these thresholds increases the weight of groups without a well-defined style, as shown in Section A.7.

4.1 Detection Results

This section demonstrates that an interpretable metric-based approach, which classifies groups as memorized when the memorization signal (i.e., style loss) falls below a fixed threshold, outperforms more complex model-based approaches (Sec.A.11) in detecting style memorization. Supported by past research on detection of memorization (Carlini et al., 2022a), this work uses True Positive Rate (TPR) at a fixed and low False Positive Rate (FPR), particularly 1%, 5% and 10%, to assess the effectiveness of detection.

Figure 2 illustrates the relative TPR (TPR/FPR) of metric-based detection per threshold-selection strategy (style- or caption-guided, and optimized) according to the number of included styles (x-axis), i.e. memorized and non-memorized styles. This figure shows that detection is effective, as relative TPR increases when the maximum margin of error allowed decreases, as TPR at 1% FPR (in red) is higher than at 5% and 10% FPR (in yellow and green) in most cases. Particularly, selecting adequate thresholds τ and λ, with any of the selection strategies, results in remarkable detection results, since restrictive definitions (left side of each plot) lead to higher relative TPR in most cases, namely over 50% TPR at 1% FPR in the caption-guided and optimal strategy in Fig. 2 (b, c). As τ and λ decrease, i.e. towards the right when the definition is less restrictive, detection results are worse, particularly for a fixed FPR of 1%, as it no longer filters out groups with ill-defined styles. Furthermore, TPR at 1% FPR is significantly different in the style-guided strategy Fig. 2 (a) compared with other threshold-selection strategies Fig. 2 (b, c) in which relative TPR at 1% FPR (in red) is always significantly higher than at 5% and 10% FPR (in yellow and green). This is due to the relatively low threshold of caption-atypicality in the style-guided strategy, as some groups with higher style-atypicality correspond to famous artists for which mentioning the artist name does not sufficiently increase caption-atypicality, an effect further amplified by the limited dataset size on the leftmost part of Fig. 2 (a). Thus, even though the style-guided strategy leads to a higher percentage of groups with an artistic style in the target dataset (Figure 12), detection is better when the definition is guided by caption-atypicality. Most importantly, these results show that (τ, λ)-memorization can identify most memorized cases (reference points in dashed blue) and close to half of groups with an artistic style (x-axis) such that detection is remarkably precise, particularly with 21-25%, 30-33%, and 42-46% of TPR when FPR is fixed at 1%, 5% and 10% respectively. Fixing high thresholds for τ and λ filters out groups whose style is close to the "average style" (with low caption and style atypicality), improving the discriminative power of the memorization signal and consequently detection.

Experiments with Stable Diffusion 2.1 (SD2), presented in Section A.9, support the generalization of these insights to other generative models, leading to similar results, particularly 7-14% TPR at 1% FPR when τ and λ are fixed such that the definition includes 50% of the memorized cases. Nevertheless, the fact that SD2, a model of higher-quality, memorizes 56 artistic styles ($\approx 39\%$), more than SD1, suggests that style memorization may benefit from the improved quality, possibly due to the curated aesthetic training dataset and to the improved text-image alignment with a new text-encoder. This observation contrasts with sample-level memorization in text-to-image diffusion models introduced by Carlini et al. (2023a), as the experimental phase of this work observed a decrease in sample-level memorization of a few samples heavily memorized by SD1.



(a) (b) (c)

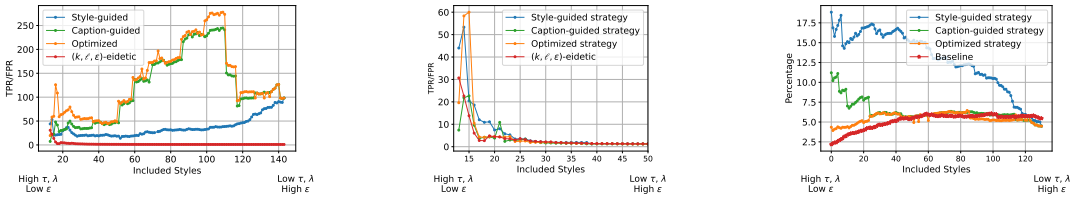
Figure 2: Metric-based detection results of style memorization in Stable Diffusion 1.4 per threshold-selection strategy: style-guided (a), caption-guided (b), and optimized (c). Reference points (dashed lines) correspond to target datasets containing 50% memorized cases (i.e. class of interest), as detailed in Section A.7.

4.2 COMPARISON WITH PRIOR WORK

The definition of style memorization is motivated by the fact it happens at the group-level in contrast with sample-level (already studied) and, given its subtle nature, it requires a careful distinction between approximating a common property from an atypical one. Thus, this section aims to show that prior work fails to capture style memorization, which can lead to an underestimation of the threat posed by these models. The state-of-the-art definition for text-to-image GMs is (k, ℓ, ϵ)-eidetic memorization (Carlini et al., 2023a) that can be paired with a suitable distance metric ℓ and the respective threshold ϵ that determines if samples are extracted, and where k (i.e. maximum number of duplicates per training sample) plays a relevant role by distinguishing expected from unexpected memorization. According to the adaption of this definition for style memorization (Section A.12), a style is memorized if there are at most $k \leq 50$ training images with the same style, i.e. whose style loss is at most ϵ, and one can extract it from Stable Diffusion with style loss at most ϵ.

Figure 3 (a) and (b) illustrates relative TPR across several values of ϵ, selected to gradually include all groups with an artistic style (x-axis), and the corresponding target dataset composition in Fig. 3 (c). For comparison purposes, the relative TPR of the proposed definition is obtained by maximizing TPR/FPR in Fig. 3 (a) and by selecting the same value of FPR as the baseline in Fig. 3 (b), instead of reporting several curves for different values of FPR as in Fig. 2. This figure shows that adapting (k, ℓ, ϵ)-eidetic memorization leads to significantly worse detection, achieving close to 30 TPR/FPR ratio at most, compared with the proposed definition that can reach over 250 when fully optimized (a) and over 60 for the same value of FPR (b). Although increasing the threshold ϵ of the baseline (towards right) identifies some memorized styles, it significantly increases FPR, resulting in TPR/FPR of 1. The target dataset composition (c) explains these results, showing that increasing ϵ of the baseline definition leads to a higher proportion of groups with artistic styles (in which the style loss controlled by ϵ is meaningful) while allowing a large number of groups with ill-defined styles (blue label in Fig. 13), thereby decreasing the discriminative power of the memorization signal (style loss). This trend contrasts with (τ, λ)-memorization that restricts the dataset to a higher proportion of groups with well-defined styles when τ and λ are sufficiently high with any threshold-selection strategy, left side of Fig. 3 (c).

However, the target dataset composition does not fully explain these poor results that are mainly due to the fact that the memorization status is determined by the comparison between the memorization signal (median style loss) and ϵ (maximum distance between training images of the same group). In other words, for a style to be considered memorized, Stable Diffusion needs to be as accurate as other training images in its approximation, failing to capture the ability of the model to mimic the main characteristics of an artistic style (e.g. brushwork, color palette, and recurring elements) that enable artist re-identification, while being creative with several other aspects (e.g. subject/objects). Past definitions overlook this type of memorization, but this flexibility makes humans classify two images as having the same style and visual search to re-identify the artist even though synthetic images are visually distinct, as illustrated in Figure 1.



(a) (b) (c)

Figure 3: Detection results of style memorization in Stable Diffusion 1.4: TPR/FPR of adapted (k = 50, ℓ = style loss, ϵ)-eidetic memorization (baseline) for varying ϵ in comparison with the optimal version (a) and with the version that guarantees the same value of FPR (b). Results for varying k are not presented (Section A.12).

4.3 RISK FACTORS

Caption and style atypicality are potential risk factors for style memorization, as memorized cases tend to score higher (Figure 11). To test this hypothesis, another dataset composed only by groups with an artistic style was built where each group contains training images of a single artist, as detailed in Sec. A.10. The name

of the resulting 400 artists was used as a prompt to extract its style and, following the same classification process, 72 ($\approx 18\%$) were classified as memorized, i.e. visual search engines re-identified the artist based on images generated by prompting the model with their name. Similar to previous experiments, memorized cases also scored higher in terms of caption and style atypicality, and style complexity which can be an efficient proxy for style atypicality. Hence, Odds Ratio (OR) was used to assess whether these metrics are risk factors, as it quantifies their effect on the odds of style memorization (ODD). Since OR > 1 indicates increased odds of style memorization, Table 2 suggests that these are risk factors particularly for less frequent artistic styles, precisely the ones that are a priori less likely to be memorized, aligning with prior work that shows that data duplication, a measure of frequency in the training set, is a risk factor. Hence, results support the use of caption and style atypicality to flag high-risk data, empowering data owners to make informed decisions about whether to opt in/out, improving on frequency-based risk factors that underestimate the risk of style memorization.

4.4 Summary

This section showed that Stable Diffusion 1.4 memorized 48 artistic styles ($\approx 34\%$) such that visual search engines can re-identify the artist based on fake artworks produced by this model which can be effectively detected by the proposed definition. By selecting adequate thresholds for caption-atypicality τ and style-atypicality λ, such that most artistic styles are flagged as memorizable (i.e. high-risk), TPR is 20-25 times the value of FPR fixed at 1%, and even 50 times in some settings, which contrasts with past definitions (Carlini et al., 2023a) that fail to account for the inherit difficulty of reproducing artistic styles and for some flexibility in that extraction, resulting in ineffective detection, particularly with TPR being as high as FPR. This comparison highlights two main sources of improvement: 1) filtering out groups with ill-defined styles with caption and style atypicality (τ and λ); 2) breaking the trade-off between the extraction accuracy ϵ and the definition of memorizable styles, therefore allowing flexible extractions (by increasing ϵ) without compromising precise memorization boundaries (by decreasing τ and λ). Additionally, caption and style atypicality are risk factors, particularly for less represented artistic styles, improving on common frequency-based indicators.

5 Conclusion, Limitations and Future Research

Motivated by ongoing debates on copyright implications of training with copyrighted works, (τ, λ)-memorization is introduced to quantify the extent to which the reproduction of abstract concepts, like specific artistic styles, deviates from the intended behaviour of generative models (generalization), improving on past definitions by accounting for the inherit difficulty of reproducing a sufficiently atypical artistic style. Experiments show that style memorization occurs in real-world settings with serious implications, as Stable Diffusion produces fake artworks that allow visual search engines to re-identify the artist. Selecting adequate thresholds for Caption-Atypicality τ and Style-Atypicality λ distinguishes style memorization from the expected behaviour, detecting non-exact pixel-wise copies of specific artistic styles without compromising precise memorization boundaries, unlike past definitions. Potential copyright implications of style memorization motivate the development of mitigation strategies in which (τ, λ)-memorization can play an important role facilitating the assessment of their effectiveness while empowering data owners to make informed decisions since it identifies artist-specific styles prone to memorization better than common frequency-based risk factors.

Limitations and future research This work is limited by the choice of open Generative Models (GMs) with publicly available training sets, the construction of feasible target datasets, and other choices, namely the style representation. While limited to image generation, this work paves the way for future research to study privacy and copyright implications of semantic property memorization of other data types. Particularly, the proposed definition is suitable for a property that uniquely identifies a set of samples while being distinguishable from the rest (i.e. high λ), with an unusual eliciting prompt (i.e. high τ), and such that its consistent reproduction is sensitive in the given context, like the reproduction of photorealistic images of individuals from the training dataset when it is not part of the training objective. Furthermore, LLMs also leak information in more subtle ways, e.g. by rephrasing training text while keeping distinguishable aspects. However, capturing memorization in LLMs requires special attention to the eliciting prompt to prevent prompt-guided extraction, which contrasts with text-to-image GMs since the output's dimension is much higher than input's dimension. Hence, introducing this trade-off into the proposed definition remains an unaddressed challenge.

AI USE STATEMENT

In this work, authors used generative AI tools to implement methods. However, authors have not used generative AI tools to propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis, and interpret results. Additionally, the following tasks are not applicable to this work: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and clean and reformat dataset.

Furthermore, we used generative AI tools to create or modify scientific figures or images, sourcing/searching for information, edit a research paper to improve readability, and format references. Authors have reviewed all AI-assisted work, namely by checking information provided by the AI tool when asked to list common metrics for a given purpose with appropriate references, checking suggested writing improvements and generated code. Authors take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code used to produce the main experiments in this work with Stable Diffusion v1.4 is available at the following anonymous repository: `https://anonymous.4open.science/r/Property-Memorization-in-Text-to-Image-Generative-Models-DFEC/README.txt`. This repository includes scripts for data pre-processing, which is explained in Sections A.2 and A.3, to extract styles from Stable Diffusion, as briefly explained in the beginning of Section 4 and detailed in Section A.4, and to perform metric-based detection for every threshol-selection strategy, as introduced in the beginning of Section 4 and detailed in Section A.7. Code to perform visual search using Google Cloud products, as explained in Section **??**, is also available in this repository. Additionally, this repository contains code to make an html file with a visual summary of every group with an artistic style, particularly for every group the top-5 most similar synthetic images and the corresponding most similar training images are presented along with the eliciting prompt to allow the reader to inspect style replication. Note that to reproduce the main experiments, it is required a GPU and access to the platform Hugging Face (public access), and to perform visual search it is required access to Google Cloud services (free trial credits are enough for testing).

REFERENCES

Associated Press. Visual artists fight back against ai companies for repurposing their work. https://apnews.com/article/artists-ai-image-generators-stable-diffusion-midjourney-7ebcb6e6ddca3f165a3065c70ce85904, 2023. Accessed 2026-09-08.

Axios. Sony, warner sue anthropic, alleging "blatant theft" of intellectual property. `https://www.axios.com/2026/08/29/anthropic-sony-warner-music-copyright`, August 2026. Accessed 2026-09-08.

Robi Bhattacharjee, Sanjoy Dasgupta, and Kamalika Chaudhuri. Data-copying in generative models: a formal framework. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023. URL `https://proceedings.mlr.press/v202/bhattacharjee23a.html`.

Jaydeep Borkar, Matthew Jagielski, Katherine Lee, Niloofar Mireshghallah, David A. Smith, and Christopher A. Choquette-Choo. Privacy ripple effects from adding or removing personal information in language model training, 2025. URL `https://arxiv.org/abs/2502.15680`.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284, Santa Clara, CA, August 2019. USENIX Association. ISBN 978-1-939133-06-9. URL `https://www.usenix.org/conference/usenixsecurity19/presentation/carlini`.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association, 2021. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914, 2022a. doi: 10.1109/SP46214.2022.9833649.

Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramer. The privacy onion effect: Memorization is relative. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 13263–13276. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/564b5f8289ba846ebc498417e834c253-Paper-Conference.pdf.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023a. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/system/files/usenixsecurity23-carlini.pdf.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models, 2023b. URL https://arxiv.org/abs/2202.07646.

Chen Chen, Daochang Liu, Mubarak Shah, and Chang Xu. Exploring local memorization in diffusion models via bright ending attention. In *International Conference on Learning Representations*, volume 2025, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/498c7e2589cd2e5b9886d0ba9a5e6032-Paper-Conference.pdf.

Dinesh Deckker and Subhashini Sumanasekara. Dreams and Data: Ghibli-Style Art, Copyright, and the Rise of Viral AI Imagery. *EPRA International Journal of Global Economic Literature*, 12(3):1–12, 2025. doi: 10.36713/epra22071. URL https://doi.org/10.36713/epra22071.

Jimmy Z. Di, Yiwei Lu, Yaoliang Yu, Gautam Kamath, Adam Dziedzic, and Franziska Boenisch. Demystifying foreground-background memorization in diffusion models. *arXiv preprint arXiv: 2508.12148*, 2025. URL https://arxiv.org/abs/2508.12148.

Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2414–2423, 2016. doi: 10.1109/CVPR.2016.265.

Jamie Hayes, Marika Swanberg, Harsh Chaudhari, Itay Yona, and Ilia Shumailov. Measuring memorization through probabilistic discoverable extraction, 2024. URL https://arxiv.org/abs/2410.19482.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: localizing neurons responsible for memorization in diffusion models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc. ISBN 9798331314385.

Hugging Face. Stable diffusion 2, 2022. URL https://huggingface.co/docs/diffusers/api/pipelines/stable_diffusion/stable_diffusion_2.

Hugging Face. Unit 3: Stable diffusion, 2024. URL https://huggingface.co/learn/diffusion-course/unit3/1#hands-on-notebook.

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A. Choquette-Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In *INLG*, pages 28–53, 2023. URL https://doi.org/10.18653/v1/2023.inlg-main.3.

Anubhav Jain, Yuya Kobayashi, Takashi Shibuya, Yuhta Takida, Nasir Memon, Julian Togelius, and Yuki Mitsufuji. Classifier-free guidance inside the attraction basin may cause memorization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. doi: 10.1109/CVPR52734.2025.01201.

LAION. Laion aesthetics: Aesthetic ratings for laion. https://laion.ai/blog/laion-aesthetics/, 2022. Online dataset description.

Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2230–2236. AAAI Press, 2017. ISBN 9780999241103.

Ali Naseh, Jaechul Roh, and Amir Houmansadr. Memory triggers: Unveiling memorization in text-to-image generative models through word-level duplication. *arXiv preprint arXiv:2312.03692*, 2023. URL https://arxiv.org/abs/2312.03692.

OpenAI. Dall·e 3. https://openai.com/index/dall-e-3/, 2023. Accessed: 2026-08-20.

PyTorch. Neural style transfer tutorial, 2025. URL https://docs.pytorch.org/tutorials/advanced/neural_style_tutorial.html. Accessed: 2025-01-01.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling andnbsp;mitigating memorization innbsp;text-to-image diffusion models through cross attention. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVII*, 2024. doi: 10.1007/978-3-031-72980-5_20.

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 (CVPR)*, pages 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European Signal Processing Conference*, pages 606–610, 2007.

Avi Schwarzschild, Zhili Feng, Pratyush Maini, Zachary Lipton, and J. Zico Kolter. Rethinking llm memorization through the lens of adversarial compression. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 56244–56267. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/66453d578afae006252d2ea090e151c9-Paper-Conference.pdf.

Sherry. Deep learning with pytorch: Neural style transfer. https://sherrys997.github.io/Deep-Learning-with-PyTorch-Neural-Style-Transfer/nb.html, 2018. Online tutorial.

Benjamin L. W. Sobel. Elements of style: Copyright, similarity, and generative ai. *Harvard Journal of Law & Technology*, 2024. URL https://jolt.law.harvard.edu/assets/articlePDFs/v38/2-Sobel.pdf.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6048–6058, 2023a. doi: 10.1109/CVPR52729.2023.00586.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47783–47803. Curran Associates, Inc., 2023b. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9521b6e7f33e039e7d92e23f5e37bbf4-Paper-Conference.pdf.

TensorFlow. Neural style transfer with tensorflow, 2025. URL https://www.tensorflow.org/tutorials/generative/style_transfer?hl=en. Accessed: 2025-01-01.

The Authors Guild. The authors guild, john grisham, jodi picoult, david baldacci, george r.r. martin, and 13 other authors file class-action suit against openai. https://authorsguild.org/news/ag-and-authors-file-class-action-suit-against-openai/, September 2023. Accessed 2026-09-08.

U.S. Copyright Office. Copyright basics. Technical report, U.S. Copyright Office, 2021. URL https://www.copyright.gov/circs/circ01.pdf.

U.S. Copyright Office. Copyright and artificial intelligence, part 1: Digital replicas. Technical report, U.S. Copyright Office, 2024. URL https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf.

U.S. Copyright Office. Copyright and artificial intelligence, part 2: Copyrightability. Technical report, U.S. Copyright Office, January 2025a. URL https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf.

U.S. Copyright Office. Copyright and artificial intelligence, part 3: Generative ai training. Technical report, U.S. Copyright Office, 2025b. URL https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Public
pdf. Pre-publication version.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

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. URL https://openreview.net/forum?id=84n3UwkH7b.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramer, and Nicholas Carlini. Counterfactual memorization in neural language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 39321–39362. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7bc4f74e35bcfe8cfe43b0a860786d6a-Paper-Conference.pdf.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. Quantifying and analyzing entity-level memorization in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19741–19749, Mar. 2024. doi: 10.1609/aaai.v38i17.29948. URL https://ojs.aaai.org/index.php/AAAI/article/view/29948.

A Appendix

A.1 Style representation

This section provides more details about how to represent the style of an image based on well-known references Gatys et al. (2016); Li et al. (2017) using Stable Diffusion, the generative model used for experiments, since it

is a latent diffusion model - a class of generative models that pre-processes images onto the embedding space to improve efficiency. This embedding representation is provided by the image encoder, specifically VQ-GAN encoder, and it is used to compute the style representation. The key intuition is that the style can be perceived by how each pixel (or patch) relates to its surrounding pixels (or patches). This correlation structure reflects the second-order statistics of the underlying pixel/patch-generating distribution (i.e. the style). Hence, the style can be estimated by looking at the correlation of each patch between every other patch[1], which is precisely the information that each element in the gram matrix holds. As illustrated in Figure 4, the gram matrix is computed by multiplying the embedding matrix by its transposed version, thus containing the inner product between any pair of patches (i.e. embedding vectors of each patch) that indicates the extent to which two vectors agree (i.e. commonly referred to as correlation). Therefore, the gram matrix represents the correlation structure of the image, so if a couple of images share the same style, meaning that their pixels/patches follow the same distribution, then they share the same correlation structure. As a result, matching styles is a particular case of distribution alignment.

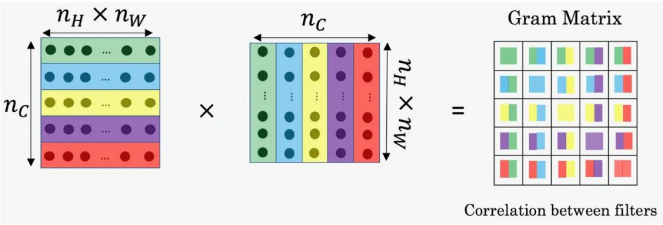


Figure 4: Illustration of the gram matrix computation (cf. Sherry (2018)), where n_H and n_W stand for height and width and n_C stands for the number of patches/feature maps, also known as filters. Each color represents a flattened pacth/feature map.

As argued by Gatys et.al. Gatys et al. (2016), the first convolutional layers tend to preserve low-level information that aligns better with the representation of the style rather than the content displayed (i.e. high-level information), which tends to be better preserved towards the end of the convolutional layers. Thus, to represent the style of an image only the first 6 convolutional layers (out of 8) given that the last convolutional layers tend to preserve high-level information namely about the content, are used to compute gram matrices, meaning that the style is represented by a set of 6 gram matrices:

$$\text{Style representation of image } i = \{Gi^l : l \in \{1, ..., 6\}\} \tag{3}$$

Given the large dimension of gram matrices, comparing them with Mean Squared Error (MSE), as suggested by several style transfer tutorials PyTorch (2025); TensorFlow (2025), can accumulate some error. Thus, the MSE is only computed along the most informative dimensions, which is determined by the effective rank. Therefore, the style representation used to compute the style loss between images i and j is a reduced version:

$$\text{Style representation of image } i = \{Gi^l_{r_i} : l \in \{1, ..., 6\}\} \text{ and image } j = \{Gj^l_{r_j} : l \in \{1, ..., 6\}\} \tag{4}$$

Where r_i and r_j are the number of informative dimensions of image i and j determined by the effective rank. The effective rank (Eq. 5) is an entropy-based estimation of dimensionality used to calculate the number of informative dimensions Roy and Vetterli (2007), which is often much less than the number of available dimensions in high-dimensional data types, such as images.

$$EffectiveRank(G) = e^{-\sum_k p_k log(p_k)}, \text{ where } p_k = \frac{\lambda_k}{\sum_k \lambda_k} \tag{5}$$

[1]Patches used to compute the gram matrix come from the first convolutional blocks of the encoder, also known as feature maps, obtained from the convolution between the image and the kernel.

Where λ_k are singular values of G (obtained from the singular value decomposition). Hence, the distance metric (style loss) is given by the MSE between the most informative part of both gram matrices averaged per layer, as shown in Eq. 6.

$$\frac{\sum_l MSE(Gi^l_{r^*}, Gj^l_{r^*})}{|layers|}, \text{ where } r^* = min(\text{EffectiveRank}(Gi^l), \text{EffectiveRank}(Gj^l)) \tag{6}$$

and Gi^l and Gj^l stand for gram matrices of images i and j at the layer l, and the subscript r^* represents the restriction to the top-r^* most informative dimensions.

The last step is to quantify style atypicality, presented in Section **??**, which involves the confidence score c_p, a measure of style complexity. Eq. 7 shows how to calculate this score using the number of effective feature maps needed to represent the style (i.e. effective dimension) and by the off-diagonal entries of the gram matrix (Frobenious norm). The key intuition is that style complexity is given by the strength of correlations between different feature maps among those that contribute the most.

$$\text{Style complexity} = \frac{\sum_l \|G^l_r - I_r\|_F}{|layers|} \tag{7}$$

Where I_r is the identity matrix of size r and G^l_r is the gram matrix restricted to the top-r dimensions sorted by the corresponding singular values. Note that style complexity plays the role of the property detector, used to compute confidence scores c_p and c_q, to quantifying how well-defined is the style. Thus, gram matrices representing images without a well-defined style (e.g. random pictures of celebrities) should have low style complexity.

A.2 Property Extraction Pipeline

While this work focuses on style memorization, authors encourage future work to study memorization of other semantic properties. For that purpose, this work proposes a Property Extraction Pipeline that only requires access to the target model and to its training dataset to test whether the model has memorized any property, customizable to several use-cases after selecting a suitable property detector and a distance metric, allowing practitioners to study property memorization. This pipeline is composed of 5 steps, as illustrated in Figure 5, each one of them with the goal of reducing the number of meaningless candidates, meaning groups without the target property, like the presence of a specific object or style. As detailed below, starting with a vast dataset of isolated samples that are efficiently clustered together according to their captions and gradually refined until the similarity at the caption-level corresponds to a similarity at the image-level, in terms of the target property, by trading efficient for effective computations. This pipeline was used, in the experimental phase of this work, to get the target dataset built of groups of training samples sharing the same style (target property). Details about its application to this use-case can be found in Section A.3.

In the first step (*Quality Filter*), only pairs of (caption, image) whose similarity, in terms of the target property using a suitable distance metric ℓ introduced in Section **??**, is sufficiently high are kept for the next steps. The goal is to get good quality pairs to leverage both the textual description and the image itself to group training samples with the same target property. In the *Grouping Step*, the goal is to group training images with overlapping captions with a fixed minimum m and maximum M support (number of samples), guaranteeing that only groups with duplicated properties are included, aligning with prior work that shows that duplication is a risk factor for memorization Carlini et al. (2023a). The maximum support M contributes for the exclusion of groups with popular properties, since approximating those is to be expected Carlini et al. (2023a). Furthermore, this step looks for overlap of words instead of tokens and after removing *stop words*, since some words correspond to a set of tokens instead of a single token, therefore reducing the number of meaningless candidates. Even though similar captions do not necessarily correspond to similar images, particularly sharing the same property, computations at the caption-level are faster than at the image-level, which is critical in this step, where the number of possible candidates is large. In the *Within Pre-processing* step, groups are subdivided into smaller groups if the initial group dispersion is sufficiently high. The goal is to get tighter training groups in which images are similar. Given the large number of candidates in this

stage, memory and time efficiency is still a priority over efficacy, so this step is computed at the caption-level, even though computing group dispersion at the image-level (in terms of the target property) would be more effective. In the *Token-Importance Filter* step, only groups for which the set of overlapping words is important for the caption are kept for the next step, using the self-attention score to quantify contextual importance within the caption, filtering out groups in which the set of repeated words was meaningless. In the last step (*Property Presence*), the goal is to prioritize efficacy over efficiency, given the large reduction in the number of candidates from applying previous steps, by filtering out groups without a target property. The metric to estimate the presence of a property within each group depends on the use-case, for example if the target property is "containing a particular object" then the confidence score yielded by an object recognition model averaged over all images within the same group can be used as an estimate for the presence of the target property.

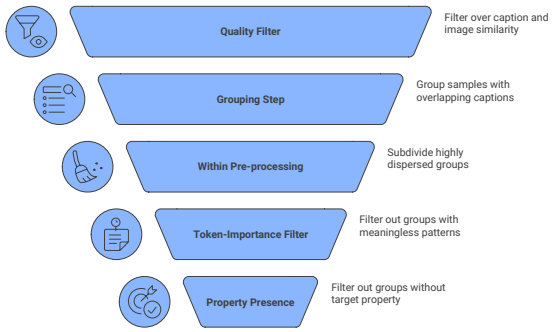


Figure 5: Illustration of the property extraction pipeline.

A.3 TARGET DATASET

This section presents details about how to build the target dataset for experiments using the proposed Property Extraction Pipeline introduced in Section A.2 to study style memorization.

The target dataset was obtained by instantiating the proposed pipeline with the LAION2B-en-aesthetic dataset LAION (2022), which is a subset of the training dataset (LAION-5B) composed by pairs of caption and image whose image has an aesthetic score higher than most (> 7), and the distance metric is style loss (see Section **??**). In addition to this aesthetic threshold, another quality filter was applied, corresponding to the first step of the extraction pipeline, in which samples with a similarity score lower than 0.4 were filtered out. Then, in the *Grouping Step*, images were grouped according to the overlap of words in their captions fixing a minimum support of 10 and a maximum of 50 samples, as illustrated in the next set of captions corresponding to a part of the group where the word "ravilious" is present in all of them, therefore identifying this group at the caption-level.

Attic Bedroom; E. Ravilious; Watercolour 1934; Fry Art Gallery	Wet Afternoon - Eric Ravilious
Barrage Balloons Outside a British Port by Eric Ravilious	Eric Ravilious the Long Man of Wilmington
Chalk Paths, 1935. Eric Ravilious (1903-1942), South Downs, Sussex, England, watercolour.	Baker and Confectioner by Eric Ravilious
Tennis (triptych) by Eric Ravilious 1930 Tempera on Plywood (Bristol Museum & Art Gallery)	Bathing Machines | Eric Ravilious | oil painting
The Long Man of Wilmington o, The Wilmington Giant, 1939 da Eric Ravilious	The road goes ever onward. Eric Ravilious, The Red Van.

Then, the *Within Pre-processing* step requires the selection of a threshold, to apply over group dispersion, above which groups are subdivided into several subgroups. Group dispersion is given by the relative standard deviation (i.e. comparable measure of dispersion) of the pairwise cosine dissimilarity between captions, which is a widely used metric to assess distance between text embeddings. Figure 6 illustrates the distribution of group dispersion along with the corresponding threshold (average). This figure shows that most groups have relatively high group dispersion, motivating the need for this step, as samples were grouped often with a small number of overlapping words, which not always corresponds to semantically similar captions. Cases with group dispersion higher than this threshold are divided into subgroups which are determined by hierarchical clustering whose cutting-thresholds is the one that avoids the top-3 most diverging merges according to the

Ward linkage. Figure 7 shows an example of the resulting clustering for a group with an initially high group dispersion, where images on the right column were separated from the rest of the group. Note that this step increased the number of candidates.

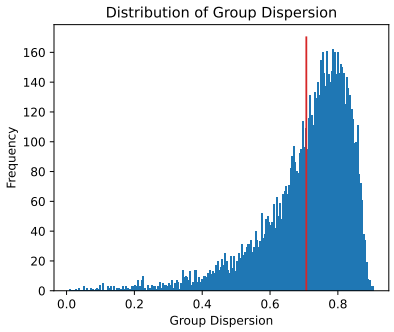


Figure 6: Distribution of group dispersion, the red vertical line shows the average of group dispersion which is the threshold above which groups are subdivided into smaller groups.



Figure 7: Illustration of subdiving a highly dispersed group. This technique divides this group into 3 by clustering together samples related to "Henri Rousseau" (4 leftmost figures) while distinguishing these from other non-related/similar samples (right column).

In the following step (*Token-Importance Filter*), groups in which the set of overlapping words was meaningless were rejected, keeping only those whose repeated textual pattern was the most attended. The self-attention score, aggregated over layers, per-token (i.e. how much attention each token pays to every other token) was ranked from highest to lowest and aggregated to assess the contextual importance of the set of overlapping words. Figure 8 illustrates the median rank of self-attention and the corresponding threshold (25-th percentile) above which groups are discarded. This figure shows that captions tend to overlap in relevant words ,

highlighting the effect of removing stop-words in the *Grouping Step*, so in this step the main goal was to reduce the number of candidates.

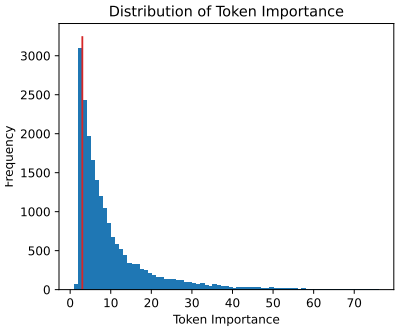


Figure 8: Distribution of token importance, the red vertical line shows the 25-th quantile above which groups are discarded.

The last step aims to reject groups without well-defined style (e.g. artistic style), starting by computing the style representative (i.e. centroid) which is the average of the style representative (i.e. gram matrix) of every image in the group. Then, style complexity (Eq. 7) is used to determine whether the group contains a well-defined style, referred as style presence. Figure 9 illustrates this distribution and the corresponding threshold (average) below which groups are discarded.

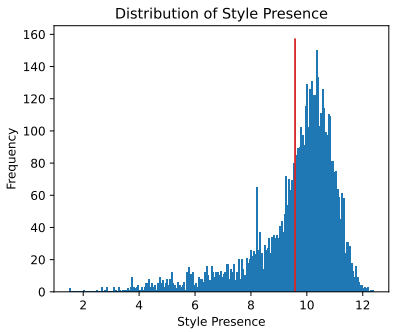


Figure 9: Distribution of style presence, the red vertical line shows the average of this distribution below which groups are discarded.

By applying every step of this pipeline, the initial training dataset (aesthetic subset) composed of 120M samples is gradually reduced to 3400 groups of samples ($\approx 20K$ samples).

A.4 Experimental setup

To show that Stable Diffusion memorized artistic styles, this work conducted experiments using the target dataset built by applying the Property Extraction Pipeline, as explained in Section A.3, and attempted to extract the style present in each group by prompting the model with a sentence derived from the set of captions within the group, i.e. set of words present in at least half of the group's captions sorted by occurrence.

The following step is to ask Stable Diffusion to generate 250 synthetic images, with different random seeds, to calculate several features that describe the entire experience. To allow for some variability in the generative output without compromising the detection of a repetitive behaviour (sign of memorization), only the best cluster of synthetic images is used. The best cluster is determined by applying hierarchical clustering over the pairwise distance matrix, which contains the style loss between every pair of synthetic images. Then, the

cutting height is selected such that the top-3 diverging merges are avoided (with Ward linkage). The best synthetic cluster is the one that minimizes *Mean distance* among all clusters that have at least 25 synthetic images. This step yields the best cluster that is still of a reasonable size (≥ 25), ensuring that a possibly good approximation is not random while still allowing the model to generate diverging images, possibly with different styles.

Based on the best synthetic cluster, the following distance-based metrics describe the entire experience: *Optimistic distance* is the average style loss between every training image and its best synthetic copy; *Pessimistic distance* is the average style loss between every synthetic image in the cluster and the closest training image; *Mean distance* and *Dispersion* are the average and standard deviation of the distance matrix between every training image and synthetic images inside the cluster; and *Within distance* and *Within dispersion* are the average and standard deviation of the distance matrix between every pair of synthetic images. These features characterize the quality of the approximation in terms of the distance to training images and between synthetic images. Different views of the distance matrix have different roles, particularly the optimistic view (*Optimistic distance*) removes the influence of synthetic images with a poor approximation, or even with a distinct style, and the pessimistic view (*Pessimistic distance*) removes the influence of training images that have a distinct style or lack a well-defined one. The metric-based detection used *Optimistic Distance* as the memorization signal, given that this is the distance metric that is mostly correlated with the target (memorized/non-memorized styles) and using a combination of metrics leads to comparable results. Furthermore, metrics calculated inside the synthetic cluster (*Within distance* and *Within dispersion*) are also important to assess how consistent is the approximation.

A.5 GROUND TRUTH

This section explains how to identify *target* groups in the target dataset obtained by applying the Property Extraction Pipeline (Section A.3), meaning groups that have an artistic style, and how they were classified w.r.t. style memorization.

Despite all the filters applied during the target dataset construction, especially the last one that filters out groups where the style is not sufficiently complex, the target dataset still contains groups without a well-defined style (e.g. photos of celebrities). Since the concept of a well-defined style can be subjective, this work focuses on artistic styles identified by the presence of the corresponding artist in the caption, meaning that every group returned by the pipeline identified as containing an artistic style has the artist's name in at least one caption. As a result, out of 3400 groups only 160 contain an artistic style.

Since recognizing style similarity can be subjective when assessed by humans, this work used an automated visual search pipeline to identify whether the style was successfully extracted, and thereby memorized, based on the capacity to re-identify the original artist by parsing fake/synthetic artworks produced by Stable Diffusion. This pipeline combines web-based retrieval by Google Cloud Vision API with Gemini 2.5 Flash to interpret and summarize results. The Google Cloud Vision API takes a synthetic image and queries Google's web index to return a ranked list of URLs containing visually similar images, along with more information, namely web entities (i.e. named concepts with relevance scores). Gemini 2.5 Flash receives this output and is requested to identify the single most likely subject of the image based primarily on the ranked list of visually similar URLs. Particularly, if web entities are semantically related to art (e.g. painting), it should identify the most similar artwork and respective artist. Results are dependent on Google's live web index, hence may vary over time, and the response from Gemini is as deterministic as possible by fixing the random seed. To reduce the amount of unnecessary searches, only groups with an artistic style were analysed and among these groups only the top-5 most similar synthetic images were parsed. To classify as positive re-identification, at least one image must identify the correct artist.

This strategy underscores the real threat of this capacity for copyright protection - fake, and pixel-wise distinct, artwork could easily pass as a real one made by the same artist. However, visual search may fail to identify the artist based on fake artworks that are stylistically similar to the originals for humans, particularly authors compared synthetic and training images and agreed that the model reproduced 14 artistic styles which were classified as non-memorized by this pipeline. In these cases, the search was dominated by the content (e.g. cairn terriers, a dog breed), as well as by similar paintings and artists (e.g. instead of Pierre-Auguste Renoir search returned Paul Cézanne). This finding suggests that style memorization perceived by humans may extend

beyond what visual search can identify, motivating future research to develop detection methodologies that align better with human perception of style memorization.

A.6 Prompt Construction

This section explains the construction of the eliciting prompt, i.e. sentence used to assess whether the model memorized a given style. These experiments use small and simple prompts composed of the set of words present in at least 50% of the captions of each group (preserving their position). Note that the set of words used to calculate caption atypicality is required to be in at least 75% of all captions, whereas the set of words to build the prompt is only required to be in at least 50%. The motivation for this discrepancy is that caption atypicality is intended to represent the whole group, so it should be based on the vast majority of the captions, while the prompt is intended to elicit extraction (without being overly descriptive). Previous experiments with the prompt composed of words present in at least 50% of the captions revealed that the approximation was better than only prompting the model with words used in at least 75% in some cases.

Motivated by prior work that shows that LLMs can reproduce non-training samples if the prompt is carefully built for that purpose Schwarzschild et al. (2024), the prompt construction aims to guarantee that the prompt does not guide the model into reproducing a particular style. Figure 1 shows examples of memorized styles along with the eliciting prompt, which is mostly composed of the name of the artist (e.g. "*el greco*" and "*paul gauguin*") and in some cases words describing foreground objects present in the original training image (e.g. "*woman wheelbarrow*"). In fact, the average prompt size used across the whole dataset (including groups without an artistic style) is ≈ 11 tokens and even smaller (≈ 8.5) among memorized cases, which is not enough to guide the model to reproduce a certain style. This can be verified in the file used by authors to inspect style replication, which contains training and synthetic images and the associated prompt (link to the repository in Section **??**). In addition, figure 10 illustrates a pair of training and synthetic images of the memorized style extracted with the largest prompt, showing that even when the prompt is large it does not contain enough details about training images to guide the model into reproducing the target style. However, authors suggest to improve the strategy to rule out the hypothesis of prompt-guided extraction as future work.



Figure 10: Real artwork of Utagawa Hiroshige (left) and fake artwork generated by the following prompt "*utagawa hiroshige 1797 1858 sudden shower 1857 shin ohashi series bridge one atake hundred famous views edo japan gift*" (right) that allows visual search to re-identify the artist. This is the largest prompt used to extract a memorized style.

A.7 Threshold-selection strategies

(τ, λ)-memorization can be used to quantify style memorization by fixing thresholds τ and λ which, in practice, limits the number of groups kept for further experiments to those whose caption and style atypicality exceed τ and λ respectively. This can be achieved with 3 different threshold-selection strategies, namely by fixing one parameter as a function of the other or by varying both parameters jointly to maximize the discriminative power of the memorization signal (style loss), as explained in the beginning of Section 4, leading to the style-guided, caption-guided and optimized strategy.

Figure 11 presents the distribution of style and caption atypicality for the initial target dataset classified w.r.t. style memorization and shows the lack of correlation between these parameters given the absence of a visible trend, suggesting that they measure different information and therefore complement each other. Moreover, this figure shows that groups with artistic styles (red and green points) tend to have higher values than groups without a well-defined style (blue points), and even higher values on average for memorized styles

suggesting that these scores can be risk factors. Figure 12 illustrates the resulting target dataset composition obtained by varying τ and λ according to every threshold-selection strategy to gradually include every group with an artistic style. Particularly, each bar in these graphics corresponds a parametrization of the definition that can be restrictive, when thresholds are high (towards the left), or less restrictive, when thresholds are low (towards the right), referred as (less) restrictive definitions for simplicity. Also note that to include $\approx 50\%$ (24 styles) of the memorized cases (red label), the style-guided definition includes a total of 67 artistic styles, the caption-guided includes 60 and the optimized version includes 64. These are reference points for future detection, as they correspond to a target dataset containing most memorized cases (i.e. class of interest) while balancing non-memorized cases. This figure shows, that regardless of the definition type, decreasing τ and λ (towards the right) increases the weight of groups without a well-defined style (blue label), agreeing with past observations. Particularly, when λ is maximized (left side of the left plot), half of the target dataset is composed of groups with artistic styles and nearly half of these are memorized. For comparison purposes, Figure 13 shows the target dataset composition from varying ϵ in the state-of-the-art definition, demonstrating that this definition is not aligned with the concept of an artistic style, since more restrictive definitions (left side) lead to a lower percentage of these cases, which tends to increase with the increase of ϵ (towards the right) along with the number of ill-defined styles (blue label). This trend contrasts with the proposed $(\tau, \lambda)-$memorization regardless of the threshold-selection strategy.

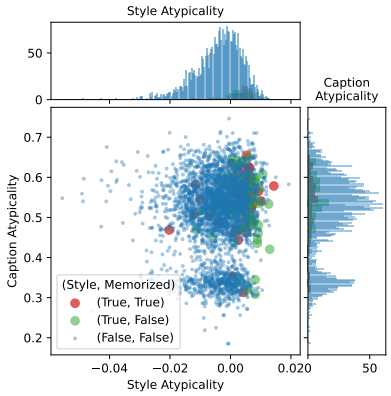


Figure 11: Distribution of style and caption atypicality per type of group: groups without a well-defined style (blue), groups with memorized (red) and non-memorized (green) artistic styles.

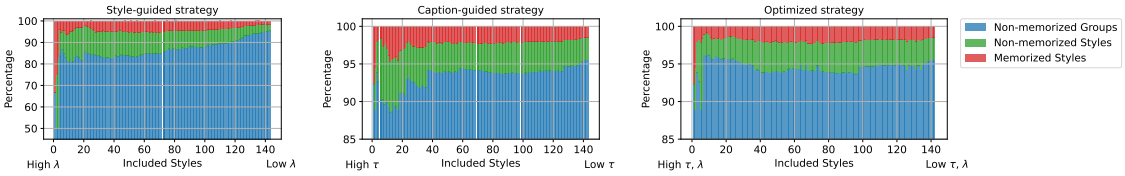


Figure 12: Target dataset composition for all threshold-selection strategies. The x-axis reflects the number of groups with artistic styles sufficiently atypical for each pair (τ, λ), i.e. green and red labels.

A.8 Additional results of metric-based detection

In addition to TPR/FPR presented in Section 4.1, the area under the Receiver Operating Characteristic Curve (ROC AUC)) is also used to assess the discriminative power of the memorization signal across every value of FPR. Figure 14 shows the ROC AUC of metric-based detection per definition type (style-guided, caption-guided and optimized). This figure shows that detection is effective, as ROC AUC is almost always higher than 50% (better than random guessing), tending towards 75% for style memorization in both models. Additionally,

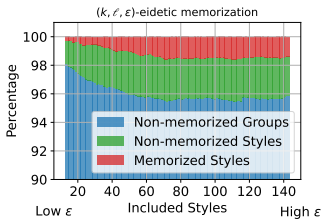
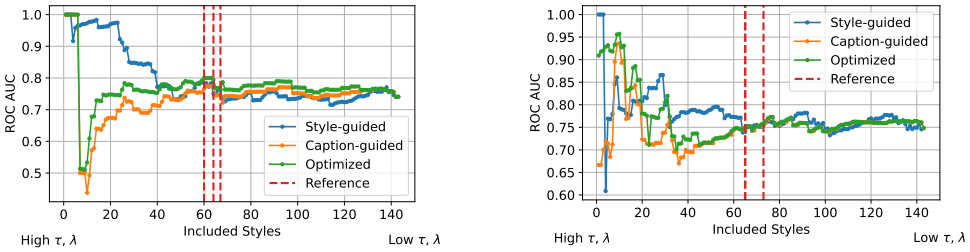


Figure 13: Target dataset composition of the adapted (k = 50, ℓ = style loss, ϵ)-eidetic memorization for varying ϵ according to the number of included styles.

this figure shows that style memorization can be predict even better performance in SD2 than in SD1 despite the slight decrease in TPR/FPR for a fixed and low FPR threshold observed in Figure 15.



(a) (b)

Figure 14: ROC AUC for metric-based detection in SD1 (a) and in SD2 (b) with target datasets determined by style-guided, caption-guided, and optimized definitions. Reference points (red dashed line), introduced in Section A.7, include 50% of the memorized styles.

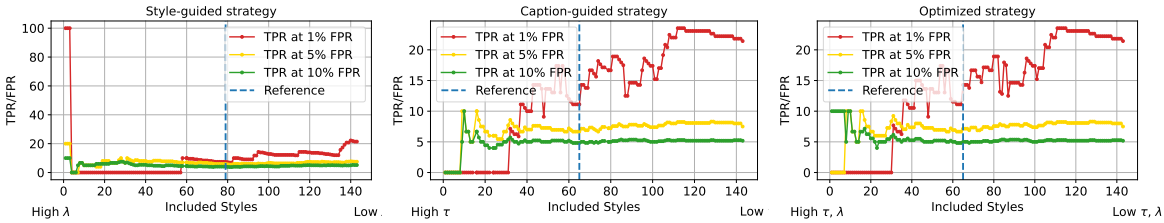
A.9 Generalization to other text-to-image generative models

Although the main experiments use Stable Diffusion v1.4, the proposed definition is model-agnostic in theory, as it only requires access to a text encoder and an image encoder, that do not need to be from the same model, thereby enabling detection of style memorization in non-latent text-to-image generative models that work directly in the pixel-space. The main limitation is the need to access the training dataset. Thus to show that (τ, λ)-memorization can be applied to different models, this section presents experiments with Stable Diffusion v2.1, a latent diffusion model of higher-quality compared to the previous one tested (v1.4). Ultimately, these experiments show that better general-purpose models memorize even more artistic styles seen during training, suggesting that style memorization can benefit from the improved model quality.

Stable Diffusion v2.1 (SD2) differs from Stable Diffusion v1.4 (SD1) in several aspects, namely in terms of training data, text encoder, and architecture. Particularly, SD2 was trained on a curated aesthetic subset of LAION-5B with additional filtering (Hugging Face, 2022), whereas SD1 initially relied on a broader dataset and only later incorporated aesthetic filtering during fine-tuning, namely to train version 1.4 (Hugging Face, 2024). Furthermore, the text encoder was replaced by a fully open-source OpenCLIP ViT-H/14 encoder, improving text-image alignment. Additionally, SD2 introduced architectural changes, despite maintaining the overall latent diffusion framework, to support higher-resolution image generation (up to 768×768). The experimental setup is identical to that used in previous experiments.

Figure 15 presents the relative TPR (TPR/FPR) of the metric-based detection of style memorization in SD2 per threshold-selection strategy according to the number of groups with artistic styles that are included (x-axis), regardless of their memorization status. This figure shows that despite the slight decrease in performance especially when for high τ (b, c), the proposed definition can effectively detect style memorization in SD2. Particularly, this strategy achieves close to 7-14%, 29-36%, 39-50% TPR for a fixed FPR at 1%, 5% and 10%

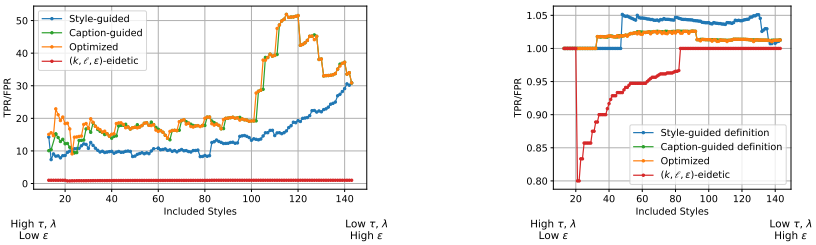
when τ and λ are fixed such that the definition (τ, λ)-memorization includes 50% of the memorized cases and close to 50% of all artistic styles (reference points in blue).



(a) (b) (c)

Figure 15: Metric-based detection results of style memorization in Stable Diffusion 2.1 per threshold-selection strategy: style-guided in (a), caption-guided in (b), and optimized in (c). Reference points correspond to target datasets containing 50% memorized cases (i.e. class of interest), as detailed in Section A.7.

Figure 16 illustrates the relative TPR across several values of ϵ, selected to gradually include all groups with an artistic style (x-axis). For comparison purposes, the relative TPR of the proposed definition is obtained by maximizing TPR/FPR in Fig. 16 (a) and by selecting the same value of FPR as the baseline in Fig. 16 (b). Similar to previously reported results of SD1 (Sec. 4.2), predicting style memorization by adapting (k, ℓ, ϵ)-eidetic memorization leads to significantly worse detection, achieving close to 1 TPR/FPR, compared with the proposed definition that can reach over 50 when fully optimized Fig. 16 (a). However, in this case when the proposed definition is forced to achieve the same FPR it cannot increase the corresponding TPR, since TPR/FPR is close to 1 in Fig. 16 (b), which shows that style and caption atypicality calculated with SD2 model components could not filter out groups with ill-defined styles as strongly as scores calculated with SD1 model components. Small adjustments may help, which are left as future work. Nevertheless, there results suggests that the state-of-the-art definition of memorization in text-to-image models cannot effectively detect style memorization in several models, supporting the need for the proposed definition which can detect this phenomenon with remarkable accuracy (Fig. 15).



(a) (b)

Figure 16: Detection results of style memorization in Stable Diffusion 2.1: TPR/FPR of adapted ($k = 50$, $\ell =$ style loss, ϵ)-eidetic memorization (baseline) for varying ϵ in comparison with the optimal version (a) and with the version that guarantees the same value of FPR (b). Results for varying k are not presented (Section A.12).

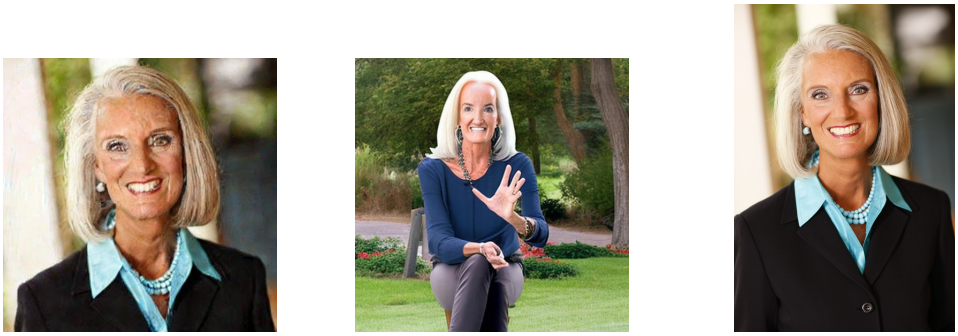
Most importantly, the fact that SD2, a model of higher-quality than SD1, memorizes 39% of the artistic styles compared to 34% is concerning, especially when looking at artistic styles that previous experiments failed to extract from SD1, as illustrated in Figure 17. Aside from producing more aesthetically pleasing images, which can help visual search engines to re-identify the artist, this figure shows that images generated by SD1 Fig. 17 (a) do not resemble the artist's style Fig. 17 (c), unlike those produced by SD2 Fig. 17 (b), as SD1 focused on the subject instead of reproducing the artistic style. This observation contrasts with a trend observed during additional experiments of sample-level memorization, introduced by **?**, in these models that revealed a decrease in this type of memorization on SD2. Figure 18 illustrates an example of this phenomenon, showing the memorized sample from Carlini et al. (2023a) obtained by prompting SD1 with "Ann Graham Lotz" in

Fig. 18 (a) which is not memorized by SD2 in Fig. 18 (b), at least to the same extent, since the same prompt produces more diverse images. This observation could be due to changes in the training dataset, namely by filtering non-aesthetic images, however authors encourage future research to study the evolution of both types of memorization on models of different levels of quality.



(a) (b) (c)

Figure 17: Illustration of the evolution of style memorization on SD1 and SD2 with images generated by SD1 (a) and SD2 (b) with the prompt *r crumb made series card mississippi delta blues artists like one favorites john hurt must huge fan much research work country music cards great*, comapred to the most similar training/real image in (c). The real image depicts John Hurt but was made by Robert Crumb (R. Crumb).



(a) (b) (c)

Figure 18: Illustration of the evolution of sample-level memorization on SD1 and SD2 with images generated by SD1 (a), from Carlini et al. (2023a), and SD2 (b) with the prompt "Ann Graham Lotz", based on the real training image (c) with a similar description, from Carlini et al. (2023a).

A.10 Analysis of risk factors

This section details how the cleaner dataset, introduced in Section 4.3, was constructed along with the complete analysis of risk factors. The construction of this dataset is divided in two phases: 1) retrieve artists' names from wikidata Vrandečić and Krötzsch (2014); 2) match with artists present in the caption of at least 5 training images, which led to 400 artists. Similar to previous experiments, the Stable Diffusion v1.4 was prompted with the artist's name to produce 250 synthetic images that are latter compared with the top 5 training images with the highest style complexity, as these are likelier to be paintings[2]. This same set of training images is used to calculate style complexity and atypicality, and the prompt (artist name) is used to calculate caption atypicality. Artistic styles were classified using the same automated visual search pipeline that re-identified 72 artists ($\approx 18\%$). Additionally, the number of samples containing an artist's name or alias (pseudonym) is referred to as raw support, whereas the number of images with sufficiently high style complexity, filtered using

[2]Style complexity quantifies how well-defined is the style of an image based on how much information the style representation holds, as explained in Section A.1

the same strategy applied to the target dataset (Section A.3), is referred to as robust support, as it estimates the number of paintings per artist. Both raw and robust support are measures of frequency of the artist in the training set.

This experience reveals that raw and robust support are positively correlated with style memorization (0.49 and 0.48), which is aligned with prior work that shows that data duplication, a measure of frequency in the training set, is a risk factor. Moreover, style and caption atypicality, metrics used in the proposed definition, are also positively correlated with style memorization (0.37 and 0.06), which suggests that these can be risk factors. However, caption atypicality has a weak correlation which aligns with its intended goal - filter out groups with very common captions, which is rarely the case for captions containing only the artist name, thus most of these cases score higher compared to the target dataset used for the main results. In addition, artistic styles typically have higher style complexity than non-artistic ones, like random photos, and the fact that memorized artistic styles tend to score even higher, given the correlation of 0.38, suggests that well-defined styles are easier to learn, potentially because they exhibit stronger correlations across regions/feature maps, reflecting the presence of a more consistent underlying generating distribution (i.e. the artist's style). Given that style atypicality can be expensive to calculate in a larger target dataset, style complexity can be a proxy, improving the efficiency of risk assessment.

To further assess whether these scores are risk factors, this work relies on Odds Ratio (OR) as it quantifies how changes in the feature (potential risk factor) affects the odds of the outcome occurring (style memorization) ODD, particularly OR > 1 indicates increased odds of style memorization and therefore supports the hypothesis that these scores are potential risk factors. Since these metrics are numerical, a logistic regression model is trained to predict the binary outcome (style memorization) as a linear function of the predictor (potential risk factor) such that OR quantifies the change in the odds of style memorization associated with a one-unit increase in style atypicality/complexity and caption atypicality. Table 2 presents OR for these metrics per support level, particularly an artistic style belongs to the high support subgroup when its raw support ≥ 1000 and its robust support ≥ 280, thereby removing the influence of the dominating risk factor (raw/robust support). This table shows that in the high support subgroup, in which styles are likelier to be memorized, only style-related metrics can be considered risk factors by leading to OR > 1. However, for the same reason, caption atypicality can only be considered risk factor in the low support subgroup, meaning that atypical artists' names are not associated with an increased odd of style memorization when a large number of artworks is in the training set. More importantly, this table shows that all metrics lead to OR > 1 for the low support regime, suggesting that they are risk factors for less frequent artistic styles, precisely the ones that are a priori less likely to be memorized.

	High support	**Low support**
Caption Atypicality	0.824	1.126
Style Atypicality	2.065	1.283
Style Complexity	1.992	2.852
Style Memorization	57%	9.5%

Table 2: Odds ratio of caption atypicality, style atypicality, and style complexity per support regime (frequency in the training dataset per artist). The last row indicates the percentage of memorized styles within each regime.

These results suggest that style atypicality, its efficient proxy style complexity, and caption atypicality, metrics used in the proposed definition, are risk factors, particularly for artists less frequently represented in the training dataset, supporting the use of these metrics to flag high-risk data in order to empower data owners to make informed decisions about whether to opt in/out of training, improving upon previously known risk factors, namely frequency-based ones, that underestimate the risk of style memorization.

A.11 Model-based detection

This section presents results of model-based detection of style memorization on Stable Diffusion v1.4, which works by training Linear (LDA) and Quadratic Discriminant Analysis (QDA) using all available features, including distance-based ones (i.e. several metrics that summarize the extraction experiment per group), size

of the best cluster of synthetic images (i.e. the closest in terms of style) and prompt size. Since this dataset is small and the positive label is rare, LDA and QDA were trained on a 5-fold fashion, meaning that 5 models were trained with 80% of the dataset and tested with the remaining 20%. Figures 19 and 20 show the evolution of test performance of LDA and QDA, bounded by the estimated 95% confidence interval (using the t-student distribution), for all three types of definitions. These figures show that LDA and QDA cannot improve upon metric-based detection by fitting linear nor quadratic models. Particularly, relative TPR is nearly the same, and sometimes lower, when compared with metric-based detection for less restrictive definitions (towards the right) regardless of the thresholds over FPR, the type of model and type of definition. However, relative TPR for restrictive definitions can be worst than metric-based detection depending on the model and type of definition, namely for TPR at 1% FPR. Additionally, results obtained by fitting a linear model (LDA) are not significantly different from those of fitting a quadratic model (QDA), even though results using QDA can be better on average. Moreover, these figures show that test performance is unstable, given the wide confidence interval namely for TPR at 1% FPR, suggesting that it is highly dependent on particular cases that fall inside the train and test set.

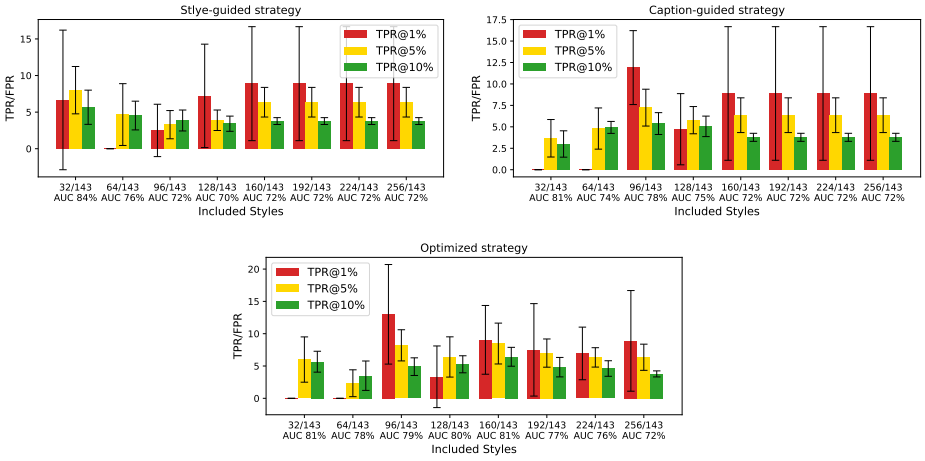


Figure 19: TPR/FPR for FPR fixed at 1%, 5%, and 10% of LDA for every threshold-selection strategy. Average ROC AUC is presented in the x-axis.

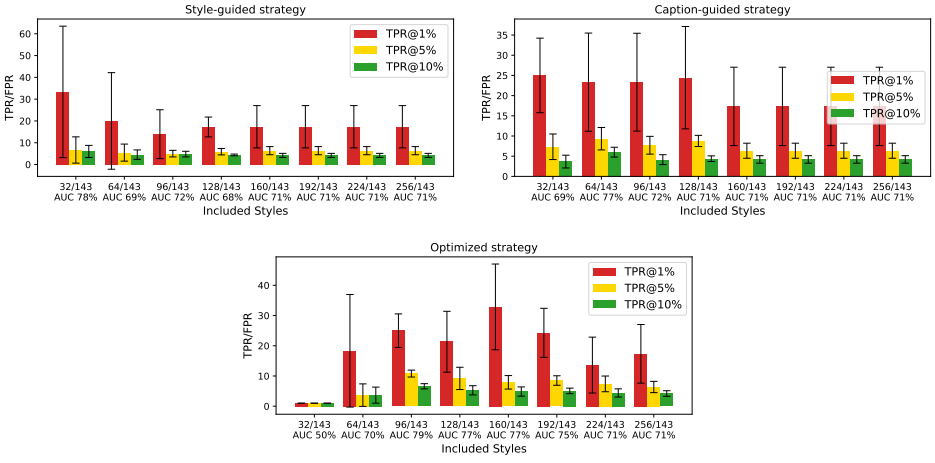


Figure 20: TPR/FPR for FPR fixed at 1%, 5%, and 10% of QDA for every threshold-selection strategy. Average ROC AUC is presented in the x-axis.

These results support the choice of the metric-based detection, as these models do not compensate the increase of computational complexity, even though LDA and QDA have the potential to improve upon metric-based by fitting linear and quadratic models with more information than the optimistic distance used by the metric-based approach.

A.12 Details about the adaptation of (k, ℓ, ϵ)-eidetic memorization

This section presents further details about the adaptation of (k, ℓ, ϵ)-eidetic memorization Carlini et al. (2023a) to capture style memorization, presented in Section 4.2. This definition can be tuned for each use-case by selecting an adequate distance metric ℓ, in this case style loss (Eq. 6), and respective threshold ϵ that determines if samples are extracted, meaning if the model's approximation was sufficiently close. The parameter k controls the maximum number of duplicates per training sample, where the maximum distance among duplicates is also given by ϵ. The ultimate goal of k is to distinguish expected model behaviour from unexpected behaviour, treated as memorization, by fixing small thresholds over k, since replicating common samples (i.e. with a large number of duplicates) is expected.

In this adaptation, ϵ delimits the maximum distance between samples of the same group that can be viewed as *duplicates* in terms of the target property assessed by ℓ. Groups exceeding this limit are excluded from the set of candidates, since the goal of this work is to study the target property present in each group and not memorization of individual samples. According to the target dataset definition which only keeps groups of size at most 50 (Section A.3), k is bounded by 50, but this boundary can be technically surpassed as it was originally designed to limit duplicated samples at the image-level assessed by ℓ, as in Carlini et al. (2023a), and not at the caption-level assessed by the overlap of words, as in this work. Nevertheless, prior work supports the relevance of the role that captions play on duplicated images - duplicated captions increase the risk of memorization Somepalli et al. (2023b). Thus, the fact that k limits the number of duplicated samples at the caption-level does not affect the role of k in this definition, as it can be used for the same role, i.e. to distinguish expected behaviour from memorization, given evidence that caption-duplication also increases the memorization risk.

To evaluate detection of style memorization along more and less restrictive definitions, similar to results presented in Section 4.1 for several thresholds over τ and λ, thresholds over k and ϵ can vary, where smaller thresholds lead to restrictive definitions. However, since k can only be approximated by the group size and presents low variability (less than 10 different group sizes), detection results for increasing values of k are omitted. Hence, results presented in Section 4.2 correspond to a definition guided by ϵ, where k is fixed as a function of ϵ in order to gradually included all groups with an artistic style. Furthermore, since (k, ℓ, ϵ)-eidetic memorization was originally designed for sample-level memorization, ϵ controls the extraction accuracy per sample but in this case it controls the median, therefore representing the majority of the group. In other words, while (k, ℓ, ϵ)-eidetic memorization requires memorized samples to be (ℓ, ϵ)-extractable, its adaption for style memorization requires that at least half of the group is (ℓ, ϵ)-extractable.

Moreover, the fact that the definition of memorizable properties using the adaptation of (k, ℓ, ϵ)-eidetic memorization depends on the extraction accuracy ϵ, unlike the proposed (τ, λ)-memorization, implies that the detection method cannot find the optimal threshold to determine if a certain memorizable style was considered extracted. In other words, one cannot fix a maximum value for FPR nor calculate ROC AUC because ϵ is fixed a priori, as it determines which part of the target dataset is potentially memorizable.