Low Resource Reconstruction Attacks Through Benign Prompts

Sol Yarkoni [1] **Roi Livni** [1]

Abstract

The rising popularity of diffusion models, have raised serious concerns around privacy, copyright, and data leakage. Prior work has demonstrated that training data can be partially reconstructed, but these attacks often require significant resources, training set access, or carefully crafted prompts. In this work, we present a low-resource attack that reveals a more subtle risk: even seemingly innocuous prompts can lead to the unintended reconstruction of real training images. Strikingly, we show that prompts like "Abstract Art Unisex T-Shirt" can generate identifiable human faces included in the training data. Our findings point to a systemic vulnerability rooted in the use of scraped e-commerce data, where templated layouts tightly couple visual content with prompt patterns. This raises new concerns about the ease with which unintentional data leaks may occur.

1. Introduction

With the increasing popularity of generative models, grows the concern for breaches of copyright and privacy. While privacy is typically associated with sensitive or non-public data, even public data raises important concerns (Tramèr et al., 2024). For example, individuals who have consented to share their photos publicly may still expect control over how their data is used, particularly to prevent it from being presented in unintended or inappropriate contexts.

In turn, researchers are now investigating how data is memorized within generative and foundational models, and whether data can be extracted in ways unintended. Several works (Somepalli et al., 2023b; Carlini et al., 2023; Webster, 2023) demonstrated that *attacks* can be designed to extract, *blatantly and verbatim*, data that appeared in the training sets. These existing attacks tend to rely on access to the training data, some on substantial computational resources, and they search for specific prompts, from training data that, potentially, trigger the extraction. In that sense, these attacks simulate a malicious adversary that explicitly aims to extract such data. However, an important concern is the potential for *unintentional* image conjuring, where a user issues a *benign* prompt that might inadvertently trigger the same phenomenon.

Thus, towards better understanding of the potential risks, we develop in this work a followup attack. Our objective is to construct simple and benign prompts including generic objects such as *t-shirt*, and *shower curtain*, without too-specific details. We applied our attack to a previously targeted model and used the extracted prompts to generate images containing elements traceable to online sources. Perhaps the most disturbing outcome of our attack is that real people, whose images appeared in such websites, are also extracted by these, so-called innocent prompts, as seen in Figure 1. In comparison, previous attacks could extract training images of real people but typically relied on intentional, ultra-specific prompts. For instance, Carlini et al. (2023) showed that prompting with "Ann Graham Lotz" could yield a verbatim copy of a training image. The objective was not to show that her image will be extracted (as the prompt requests) but to investigate the copying of the existing photo. In contrast, our attack shows that even unintentional generic prompts can generate images of real individuals. This behavior raises distinct concerns around privacy, and individuals' rights. Particularly the right not to have their likeness used for modeling purposes without their consent.

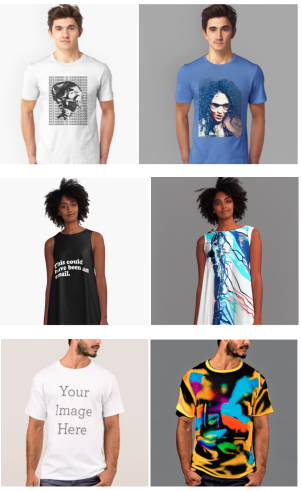
Our attack builds on a different approach. Previous work focused on training data mining, in particular identifying duplications. Our attack avoids harvesting the training data and builds on a working hypothesis that data from *e-commerce* and particularly Print on Demand (PoD) websites is (intentionally or unintentionally) harvested during training. Then, we design an attack that leverages specific traits and domain-knowledge regarding e-commerce sites.

PoD platforms use automated design placement tools that overlay artwork onto pre-existing images of product by using smart masks and blending techniques. A single platform output may be integrated into many other websites, and the

[1]School of Electrical & Computer Engineering, Tel Aviv University, Israel. Correspondence to: Sol Yarkoni <sol-yarkoni@mail.tau.ac.il>, Roi Livni <rlivni@tauex.tau.ac.il>.

Published at ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

system can instantly generate realistic previews showing the design on different product types, angles, and lighting conditions. As a result, such websites display many images that are identical up to a fixed region where the design is placed. The LAION-5B dataset (Schuhmann et al., 2022), a large-scale, web-scraped image-text dataset, likely contains a significant number of PoD-generated images due to their widespread presence on the internet. We could validate this claim through existing datasets of duplicated images on LAION, (Webster et al., 2023). Importantly though, because generated images are not verbatim copies, we could also validate that many variants of the same image were not necessarily flagged as duplications by these identification systems. In other words, these platforms produce images that may consistently feature repeated elements, such as the same design or product, but with variations in context, angle, background, or lighting. In turn, the images are not always deemed as duplicates, even though they share substantial visual content.



"blue Unisex T-Shirt"
Source: RedBubble

"Abstract Art A-Line Dress"
Source: RedBubble

"Abstract Art T-Shirt"
Source: Zazzle

Figure 1. **Reconstruction in SD 1.4.**.

Related Work Several reconstruction attacks for foundation models have been designed, including attacks targeted on language models (Carlini et al., 2021; Lee et al., 2022; Kandpal et al., 2022) and, specifically, image reconstruction (Somepalli et al., 2023b; Carlini et al., 2023; Webster, 2023). We elaborate further on these attacks and how they relate to our work in Section 2. We are specifically concerned with attacks that generate data from prompts, or models that through "standard" use induce types of image conjuring. This is distinct from various attacks that reconstruct training data by probing the model (Haim et al., 2022; Yin et al., 2020; Fredrikson et al., 2015). These works highlight the memorization of training data by large learning models, where there is increasing evidence that such memorization is necessary for learning (Livni, 2023; Attias et al., 2024; Voitovych et al., 2025; Feldman, 2020).

An important aspect that arises is the question of *originality* and the theoretical questions of what constitute original data (Elkin-Koren et al., 2024; Scheffler et al., 2022), as well as practical questions as to how regulate non-originality (Haviv et al., 2024; Hacohen et al., 2024; Chiba-Okabe & Su, 2025). Our attack generates results in a middle ground between *blatant, verbatim* copying and *non-copying that lacks originality*, Such types of interpolations have been observed before (Aithal et al., 2024; Somepalli et al., 2023a). It is possible that such interpolations can be regulated through appropriate credit attribution (Livni et al., 2024).

2. Existing Attacks

Our method introduces a low-resource attack that avoids training data access or duplication mining, instead relying on naturally occurring phrases. While novel in execution, our approach builds on insights from previous attacks. We briefly review key contributions that inform our method.

Random Sampling Somepalli et al. (2023a) show that a random subsample of the training data can suffice to recover replicated content. Using 9000 prompts randomly sampled from a known subset of $12M$ (LAION-Aesthetics) images-caption pairs, they detect memorized images generated from specific captions—often not reproducing the original image paired with the caption, but rather retrieving a training set image paired with a different caption. Their analysis suggests that certain key phrases, even if not extracted verbatim, can trigger memorization. Our method validates this claim for short prompts, consisting key phrases, whose origin were not extracted from the training set, making them much more likely to be used unintentionally.

Duplication-Based Memorization OpenAI (2022) demonstrate that many memorized images in generative models have near-duplicates in the training data. Similarly, Carlini et al. (2023) reasons that training data duplication is a potential cause for memorization and presents the hypothesis that images extracted from memorization, as opposed to novel generated data, will also contain near-duplicates. The near-duplicates search use patch-level ℓ_2 and CLIP-based similarity, relying on full-image duplication in both the original and generated data. These approaches require broad access to training data and focus primarily on verbatim copying. In contrast, our method focuses on partial duplication such as a recurring background object. Thus, even an image that was generated only once may contain replicated content.

One-Step Synthesis Webster (Webster, 2023) takes a different theoretical approach, selecting the candidates for memorization based on their one-step synthesis behavior, under the hypothesis that memorized image-text pairs present sharp edges after the first denoising step, while non-memorized pairs are blurry after the first denoising step. They also utilize the assumption that the edges of memorized images generated from the same memorized prompt will present consistent edge location along different seeds. Using this method they indeed found many captions that extract template memorized images.

To select the candidate captions they relied on full duplicate of image-text pairs. The candidates were selected as the highest scoring 2 million image-text pairs from LAION-2B on the duplication metric presented in (Webster et al., 2023). The 2 million candidate include only fully duplicated text-image pairs, and does not include images that were duplicated in a partial fixed region, along with a partial text, but for which no full duplication existed. Out of the 2M candidates, 30k were selected based on their score in the white-box attack presented at (Webster, 2023), i.e. those for which the most noise had been removed at the first denoising step of Stable Diffusion V1. They identified a phenomena of *template memorization*, where only a spatial region in the image is being duplicated, and postulate that such images might be traced to e-commerce sites.

Therefore, we expand on the phenomena of template memorization, showing that one could use a natural English text including a short Collocation rather than a highly specific text, and that such Collocations can extract several image templates rather than a specific one. We further rely on e-commerce websites to select candidate prompts and collocations, releasing us from searching and using full duplicates, thus expanding the search to image-text pairs that were duplicated only in part.

3. Our Attack

Overview and Data Collection Our attack uniquely avoids training data access, duplicate mining, or white-box model introspection. It requires low compute and leverages domain knowledge to craft prompts likely to trigger memorized content In a manner simulating an unintentional memorized data extraction through natural usage.

We target product categories from e-commerce websites known to appear in LAION-5B. Using a March 2021 snapshot (pre-LAION cutoff), we extracted 108 category collocations (e.g., "Unisex T-Shirt", "Area Rug") by scraping consumer-facing print-on-demand sites. Each collocation was paired with a simple visual modifier (e.g., "floral", "galaxy") to yield prompts like "Floral Unisex T-Shirt." We generated 25–50 images per prompt using distinct seeds.

To benchmark against prior work, we also reused collocations derived from prompts identified in Webster (2023), Somepalli et al. (2023a), and Hintersdorf et al. (2024), extracting only the core product phrase and applying our same augmentation strategy.

Near-Duplicate Detection One advantage of our attack is that the generated images are from known categories, enabling segmentation of editable (non-memorized) region by pretrained segmentation models. For household items we applied MaskFormer (mas) and SegFormer (Seg) for fashion, then clustered images by CLIP similarity (≥ 0.95) within the fixed region (outside the segmented object). Cliques of size at least 2 were flagged as candidates for template memorization.

Our use of CLIP captures perceptual similarity even across slight perturbations, and lowers the sample requirement (vs. cliques of size 10+ in Carlini et al. (2023)). Manual inspection supplemented cases missed due to segmentation errors.

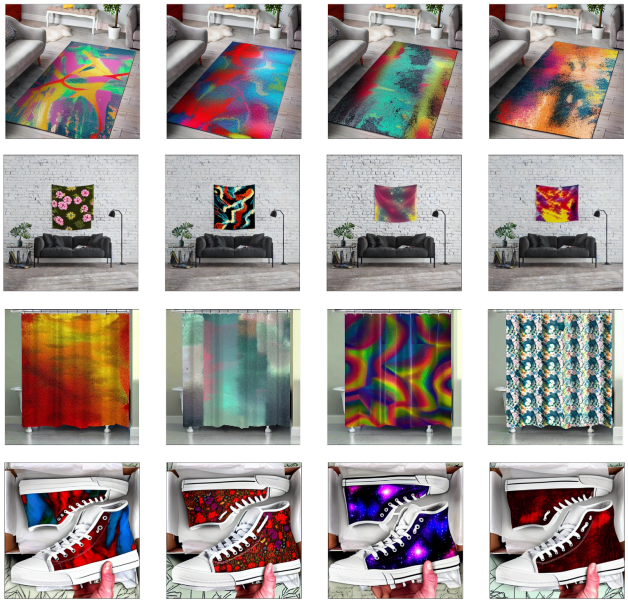


Figure 2. From top to bottom: "X Area Rug" (segmentation: "rug"), "X Wall Tapestry" ("painting"), "X Shower Curtain" ("curtain"), and "X High-Tops Sneakers" ("Right-shoe"). All examples were found using our segmentation and masking method.

Source Tracing Despite no training set access, our category-aware prompts localize generation to narrow domains. This allows visual matching to real product images using Google Lens and targeted browsing of e-commerce sites. In several cases, we traced generated images to originals without requiring duplication, validating our method's ability to uncover memorization through natural, domain-informed prompting alone.

4. Results

Our main experiments were performed on a single RTX A6000 machine using Stable Diffusion 1.4 from the official HuggingFace checkpoint (SD1), which served as the target in prior attacks (Carlini et al., 2023; Webster, 2023; Somepalli et al., 2023b). Results for more recent models are presented in Section 4.2 and Appendix B.

4.1. Identified Source Images

Many of the duplicated generations correspond to real web images, often from e-commerce sites matching our target categories (Figure 4). In some cases, even non-duplicated generations were visually matched to real sources, such as the example in the first row of Figure 1. More examples are available at the appendix A.

Real Humans Of particular concern are generations that depict identifiable people (Figure 1). Unlike prior work that targeted known individuals via name-based prompts (Carlini et al., 2023), our prompts were generic (e.g., "blue Unisex T-shirt"), yet still recalled real models likely scraped from product pages.

Even when faces are distorted or cropped, unique visual features—tattoos, haircuts, poses—can persist (Figure 6, left), as well as the context in which they appear, posing clear privacy risks. These findings highlight the risk of unintentional memorization recall in everyday use cases.

4.2. Attacks on Stable Diffusion 3.5

We also ran the attack on Stable Diffusion 3.5 Medium from the checkpoint available at HuggingFace (SD3) with a single RTX A6000 machine.

During the training of SD v. 3.5, some efforts were made to mitigate the image-text coupling which stands in the basis of our attack. Yet, it was not specifically directed at our attack which attends to template-style coupling, and relying on concise prompts coming from external sources rather than the training data. Evidently, SD3.5 is more resilient, yet not entirely robust, to our attack. Examples of template-images extracted from SD 3.5 are shown at 3. It remains unclear if this is due to decoupling or other changes made during training.

Discussion and Limitations Training data extraction presents a major challenge, raising serious concerns around privacy and copyright infringement. Our results focus on unintentional extraction of memorized data and low-resource attacks. Our attack demonstrates that image memorization in diffusion models can be exploited with minimal resources and without access to the training dataset, posing a more widespread privacy and copyright risk than previously understood. Because our attack relies on seemingly innocuous prompts (e.g., "Abstract Art Unisex T-Shirt") we reveal that users may unintentionally generate memorized images of real individuals or copyrighted content. These types of attacks amplify the risks associated with inadvertent copying and highlight the troubling possibility that even seemingly benign users may inadvertently trigger such leaks.

Figure 3. **Template Memorization in SD 3.5 (Medium).** The first three columns show images generated with prompts of the form *"X Universal Fit Car Seat Covers"* and *"X Shower Curtain"*, where X is "Abstract Art," "Floral," or "Galaxy." The fourth column shows source images from the training data.

The vulnerability we exploit arises from the template-structure of scraped e-commerce data, where uniform layouts across product categories facilitate memorization and make leakage more detectable. While engineers already attempt to clean datasets of obvious duplicates, we suggest that duplication should be interpreted in a more conservative sense, that also accounts for repeated patterns and templates. Most of our attacks were conducted on the SD 1.4 version, but in the more recent Stable Diffusion 3.5 model, we could still obtain some of the results from version 1.4, as presented in Section 4.2. We could also trace some memorization of previously reconstructed images, as shown in Appendix B. It is possible that the methods of decoupling between text and image that were presented in the training process of version 3.5 strengthened the model against the threat of our attack, but it did not completely mitigate it. The examples for successful extraction of template-memorized images from version 3.5 which are shown in Figure 3. We leave it to future study to further improve our attack, as well as to develop more principled methods to protect models from similar attacks.

Acknowledgments

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement FoG-101116258). Views and opinions expressed are how-

ever those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work received additional support from the Tel Aviv University Center for AI and Data Science (TAD) and a grant from the Israeli Council of Higher Education.

References

Stable diffusion version 1.4 checkpoint at huggingface. URL https://huggingface.co/CompVis/stable-diffusion-v1-4.

Stable diffusion version 3.5 medium checkpoint at huggingface. URL https://huggingface.co/stabilityai/stable-diffusion-3.5-medium.

Segformer model fine-tuned on atr dataset. https://huggingface.co/mattmdjaga/segformer_b2_clothes.

Maskformer model trained on ade20k semantic segmentation. https://huggingface.co/facebook/maskformer-swin-tiny-ade.

Aithal, S. K., Maini, P., Lipton, Z., and Kolter, J. Z. Understanding hallucinations in diffusion models through mode interpolation. *Advances in Neural Information Processing Systems*, 37:134614–134644, 2024.

Attias, I., Dziugaite, G. K., Haghifam, M., Livni, R., and Roy, D. M. Information complexity of stochastic convex optimization: Applications to generalization and memorization. *arXiv preprint arXiv:2402.09327*, 2024.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 5253–5270, Anaheim, CA, August 2023. USENIX Association. ISBN 978-1-939133-37-3. URL https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Chiba-Okabe, H. and Su, W. J. Tackling copyright issues in ai image generation through originality estimation and genericization. *Scientific Reports*, 15(1):10621, 2025.

Elkin-Koren, N., Hacohen, U., Livni, R., and Moran, S. Can copyright be reduced to privacy? In *5th Symposium on Foundations of Responsible Computing*, 2024.

Feldman, V. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1322–1333, 2015.

Hacohen, U., Haviv, A., Sarfaty, S., Friedman, B., Elkin-Koren, N., Livni, R., and Bermano, A. H. Not all similarities are created equal: Leveraging data-driven biases to inform genai copyright disputes. *arXiv preprint arXiv:2403.17691*, 2024.

Haim, N., Vardi, G., Yehudai, G., Shamir, O., and Irani, M. Reconstructing training data from trained neural networks. *Advances in Neural Information Processing Systems*, 35: 22911–22924, 2022.

Haviv, A., Sarfaty, S., Hacohen, U., Elkin-Koren, N., Livni, R., and Bermano, A. H. Not every image is worth a thousand words: Quantifying originality in stable diffusion. *arXiv preprint arXiv:2408.08184*, 2024.

Hintersdorf, D., Struppek, L., Kersting, K., Dziedzic, A., and Boenisch, F. Finding nemo: Localizing neurons responsible for memorization in diffusion models. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Ian Stenbit, Francois Chollet, L. W. A walk through latent space with stable diffusion. code examples of generative deep learning, 2022. URL https://keras.io/examples/generative/random_walks_with_stable_diffusion/.

Kandpal, N., Wallace, E., and Raffel, C. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning*, pp. 10697–10707. PMLR, 2022.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, 2022.

Livni, R. Information theoretic lower bounds for information theoretic upper bounds. *Advances in Neural Information Processing Systems*, 36:37716–37727, 2023.

Livni, R., Moran, S., Nissim, K., and Pabbaraju, C. Credit attribution and stable compression. *Advances in Neural Information Processing Systems*, 37:2663–2685, 2024.

OpenAI. Dall·e 2 pre-training mitigations. *OpenAI*, 2022. URL https://openai.com/index/dall-e-2-pre-training-mitigations/.

Scheffler, S., Tromer, E., and Varia, M. Formalizing human ingenuity: A quantitative framework for copyright law's substantial similarity. In *Proceedings of the 2022 Symposium on Computer Science and Law*, pp. 37–49, 2022.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

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

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models, 2023b. URL https://arxiv.org/abs/2305.20086.

Tramèr, F., Kamath, G., and Carlini, N. Position: Considerations for differentially private learning with large-scale public pretraining. In *International Conference on Machine Learning*, pp. 48453–48467. PMLR, 2024.

Voitovych, S., Haghifam, M., Attias, I., Dziugaite, G. K., Livni, R., and Roy, D. M. On the dichotomy between privacy and traceability in ell_p stochastic convex optimization. *arXiv preprint arXiv:2502.17384*, 2025.

Webster, R. A reproducible extraction of training images from diffusion models, 2023. URL https://arxiv.org/abs/2305.08694.

Webster, R., Rabin, J., Simon, L., and Jurie, F. On the deduplication of laion-2b, 2023. URL https://arxiv.org/abs/2303.12733.

Yin, H., Molchanov, P., Alvarez, J. M., Li, Z., Mallya, A., Hoiem, D., Jha, N. K., and Kautz, J. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8715–8724, 2020.

A. Additional Results

For many of the duplicated images, we could locate a source images from the web. Some examples are depicted in Figure 4. Noteworthy, even for some images that we did not identify them as duplicated, we could identify a source image by inspecting e-commerce sites from which we've taken the categories, for example the image on the first

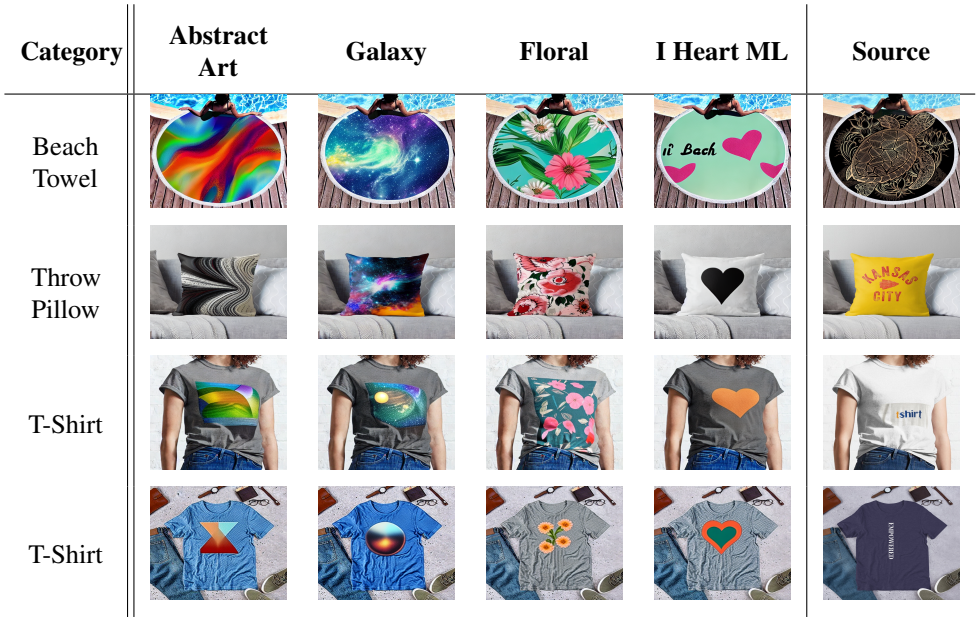


Figure 4. Examples across product categories.

In Figure 5 we depict two cases of images where we could identify the source in LAION as well as their associated captions. The prompt we used to reconstruct them did not involve that caption, but instead involved a generic type of prompt. This highlights the risk of inadvertently reconstructing images from the training set even in benign use cases.

We also find three interesting phenomena appearing in the extracted images and name them **Interpolation**, **Perturbations**, and **Leakage**.

A.1. Interpolation

Another form of memorization we observe involves interpolated reconstructions, where generated images blend elements from multiple training examples rather than copying any single image outright. These cases are harder to detect, as neither the background nor foreground is directly duplicated. In Figure 6, for instance, one image contains a tattoo reproduced from a real-world photo, though the rest of the image is unrelated. Such partial copying evades standard similarity metrics and raises challenging questions about what constitutes a memorized output.

This form of interpolation suggests that models may internalize and recombine fine-grained visual features from distinct sources. We speculate that some generations may blend fragments from numerous images, making provenance analysis difficult or even intractable with current techniques.

A.2. Perturbations

We've also noticed clusters of images that were nearly identical in the template sense but with one or more objects perturbed between semantically similar objects in a fixed location such as lamps and chairs. This is exemplified in Figure 7. Such perturbations preserve a high CLIP similarity (outside the editable region) between the images. Perturbations somewhat affect the pixel-wise ℓ_2 similarity since they change the pixel values, but still maintain a higher-than-random similarity. While difficult to pinpoint the source of this phenomena, it stands in concurrence with the expected consequences of minor perturbations in the noise latent space, as demonstrated in (Ian Stenbit, 2022). It might also stem from the one-step denoising behavior described by (Webster, 2023), as a clear layout of the image after the first step, leaves room only for minor changes

Generated Image	Source Image	Prompt and Source Caption
		prompt: Galaxy Print Universal Fit Car Seat Covers. **Caption**: Wild Hearts Can'T Be Broken Car Seat Covers For Horse Lovers 170804 - YourCarButBetter
		Prompt: Abstract Art Round Metal Wall Art. **Caption**: Designart Wide Pathway in Yellow Fall Forest Landscape Photo Round Metal Wall Art .

Figure 5. Comparison of Generated and Source Images with Corresponding Prompts and Captions when taking the categories from previous works, examples from (Hintersdorf et al., 2024).

in the next steps.

A.3. Leakage

Another interesting phenomenon that we identified is a certain *template leakage* where an image template belonging in one template set appears is generated by a prompt associated with another template group, or not associated with a template at all. In the image generated from template leakage, the object is roughly overlapping with the edited region, in a visually sensible way. Such an example of *suspected* leakage is demonstrated in Figure 8b where a template associated with the category "T-Shirt" appeared under "Tank Top". The edited area is a tank top, but the background appeared in the source and other generated images with a T-Shirt.

Inspecting e-commerce websites, though, some image templates are re-used for different product. Thus, some cases of suspected leakage may not be such. For example see the common background between "Canvas Wall Art Print" and "Wall Tapestry" Figure 8a. The image template appeared with both categories in the training data.

B. Traces of memorization in Stable Diffusion 3.5

Beside clear template extraction up to perturbations as demonstrated in **??**, we also noticed that even in cases where our original attack failed, traces of the attacks could still be found and perhaps utilized in future, more refined, attacks. For example, although the beach towel in Figure 4 could no longer be reconstructed directly, identifying its source via Google Lens revealed that the caption still produces distorted versions of the original image. This indicates that training images from e-commerce websites may continue to pose risks in future iterations of diffusion models.

C. Comparison with One Step Synthesis

As discussed, to allow comparison with previous attacks we also extracted a list of collocations, similar to the one we extracted from the generic websites we chose, by harvesting e-commerce sites that appeared in previous works. For example, we could extract the collocation *car seat cover* and compare our approach to previous approach that mines captions from the training set. Figure 10 provide examples of images and the associated prompts used in each attack.

Figure 6. Examples of interpolations observed, Left: ”X Essential T-Shirt”, right: ”X iPhone Case & Cover” The elements in the right columns were also identified in the attack of Somepalli et al. (2023a)

Figure 7. Four images with perturbations (see plant on the right) generated from the prompt: “X Wall Tapestry”: where X is: ”I Heart ML”, ”Floral”. Rightmost is source found via Google Lens.

(a) Background shared between template groups, not suspected for leakage as the source has been found for both. Left prompt: "Galaxy Wall Tapestry", right prompt: "Galaxy Canvas Wall Art Print".

(b) Suspected leakage from "T-Shirt" to "Tank Top". Left prompt: "Abstract Art Essential T-Shirt", a source has been found. Right prompt: "Abstract Art Tank Top", a source has not been found.

Figure 8. Example for a template shared between images categories in the training data, and suspected leakage between image categories.

Figure 9. Traces of our attack in SD 3.5 (Medium). The first three images were generated using the prompt *"BOYOUTH Round Beach Towel,X Beach Mat with Tassels Ultra Soft Super Water Absorbent Multi-Purpose Towel,59 inch-Diameter"* Taken from the source image's product description in Amazon, where X was: "Abstract Art", "Floral", "Galaxy". The image to the right of the vertical line is the source image, which was reconstructed in previous models (see Figure 4)

D. Synthetic Results

We wanted to verify our hypothesis on how the coupling between text templates and image templates causes partial memorization, therefore we conducted synthetic experiments mimicking the sort of coupling in order to intentionally create memorization. Another phenomenon that we wanted to identify and understand is what we term *leakage*. We have observed on real-world data that certain images from one category included templates from another category. Unfortunately, we could not verify if this is due to leakage from categories or whether certain templates simply appeared on the training data on different category (a phenomenon that we could identify that sometimes happen). Therefore, through synthetic experiments we validate that indeed leakage may be happening.

For our synthetic experiments, we collected 3 images of a coffee Mug with an iPhone SE where we placed the same coffee mug in 3 different locations in our lab. We then manually created a mask for each image with Photopea and then simply replaced the mask area with a pattern using OpenCV. In this stage, the patterns were crudely overlaid, creating an unnatural appearance that was easy for the model to memorize. We observed template verbatim extraction along with interpolation, perturbations and leakage.

As a second stage, we used mockups with more realistic rendering of the overlaid pattern based on the technique used for actual Print on Demand websites. The mockups were taken from Freepik, to one of the mockups we added in Photoshop 2 elements also taken from Freepik: a pair of leaves and a slice of lemon. We selected 3 mockups and replaced the smart object contents in Photoshop by an automation script with some manual fixes to ensure that the overlay is uniform between the images. Beside the different templates, the rest of the experiment remained the same as the first stage. We did not observed template verbatim extraction, but we did observe interpolation, perturbations, and object memorization.

In a addition to generating the collocation "Coffee Mug", we generate images from our fine tuned model through semantically

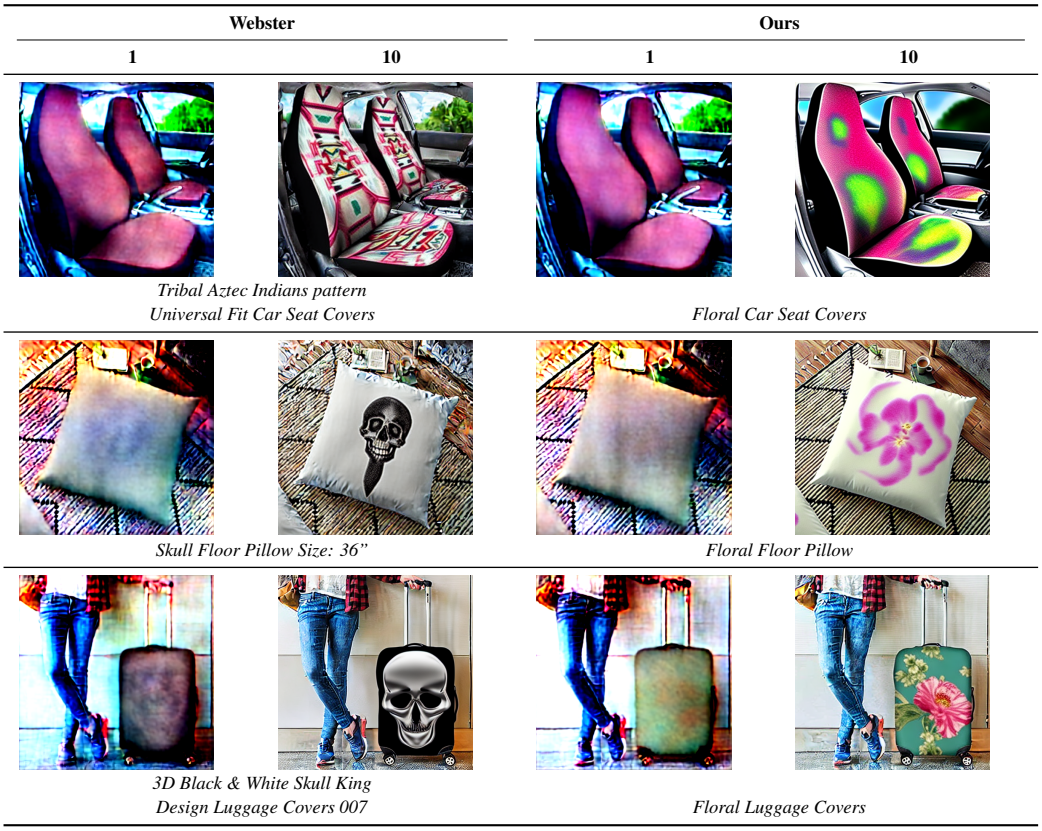


Figure 10. Comparison between the images generated after 1 step and 10 steps across prompts from Webster’s attack and ours. Images from top to bottom: (a) Identified by Webster’s white-box attack but not the black-box attack. (b) Identified by the black-box attack as non-verbatim. (c) Identified as template verbatim.

similar and semantically dissimilar to “Coffee Mug”: “Tea Cup” and “T-Shirt”, Accordingly. Through these additional prompts we could identify that leakage indeed happens. Specifically for the prompts ”skg Tea Cup”, the backgrounds of all of the images are templates interpolation or at least similar colors/textures, the tea cups are also in top view which was not seen during training and did not appear in the coffee mug generated images (only the 3 views from the train set), and also in different shapes.

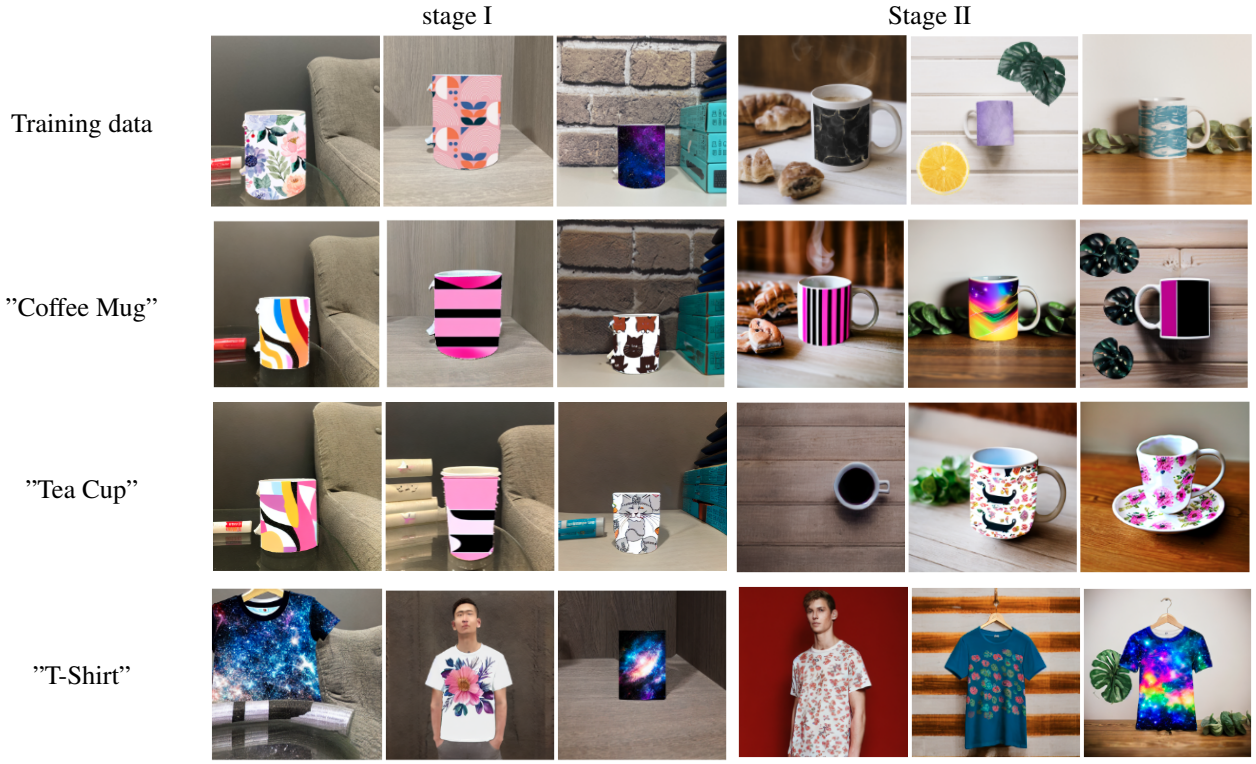


Figure 11. Intentionally causing template memorization by fine tuning SD on coupled image-text pairs. The generated results demonstrate the phenomena of interpolation, perturbations, and leakage.