ANALYSIS AND MITIGATION OF MEMORIZATION IN MULTI-MODAL DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-modal diffusion models, such as BLIP-Diffusion and ELITE, achieved impressive performance in generating personalized images conditioned on both a reference image and a textual prompt. Despite their success, it remains unexplored how these models memorize their training data. Existing memorization analyses for text-to-image diffusion models consider the text prompt as the only memorization factor without considering reference image input, making them inapplicable for the multi-modal setting. To address this gap, we propose a novel approach to effectively measure memorization in multi-modal diffusion models. First, our method localizes objects based on caption keywords, then masks each detected region, and measures how faithfully the model reconstructs the masked object. We show that our method provides equivalent results to the gold-standard leave-one-out memorization metric at a substantially lower computational cost. Using our new method, we find that pre-training samples for the multi-modal encoder exhibit consistently higher memorization levels than samples used to fine-tune the diffusion backbone. This indicates that the multi-modal encoder plays a more important role in driving memorization than the diffusion backbone itself. We further demonstrate that highly memorized training images are under high privacy leakage risks. This is driven by the joint memorization of vision–text modality associations, where the vision modality matters more. Finally, we show that applying memorization mitigation at the multi-modal encoder level achieves a significantly better privacy-utility tradeoff than diffusion backbone level mitigation.

1 INTRODUCTION

Memorization in machine learning has been studied in different settings, including supervised learning (Feldman, 2020; Feldman & Zhang, 2020), self-supervised learning (Wang et al., 2024b;a), contrastive vision-language models such as CLIP (Jayaraman et al., 2024; Wang et al., 2025) and standard text-to-image diffusion models (Wen et al., 2024; Hintersdorf et al., 2024; Ren et al., 2024; Kowalczuk et al., 2026). These works consistently show that memorization introduces privacy risks that enables adversarial membership inference attacks to reveal whether a specific sample was used in training. Furthermore, the training data can be illegally reconstructed (Meehan et al., 2023) or extracted (Webster, 2023; Carlini et al., 2023; Kowalczuk et al., 2025) from the models.

Multi-modal subject-driven text-to-image diffusion models (referred to as multi-modal diffusion models), such as BLIP-Diffusion (Li et al., 2023a) and ELITE (Wei et al., 2023), have demonstrated strong capabilities in generating personalized images conditioned on both a reference image and a text caption (Shuai et al., 2024; Cao et al., 2025). Compared to standard text-to-image diffusion models that are conditioned only on text modality input, multi-modal diffusion models leverage pre-trained multi-modal encoders to extract visual and textual joint representations from reference image-caption pairs, so that achieving high fidelity in subject preservation during generation (Wei et al., 2023; Li et al., 2023a; Ma et al., 2024; Ye et al., 2023). Despite their success in high quality image generation, a critical question remains unexplored: to what extent do multi-modal diffusion models memorize their training data and which components of their architecture are primarily responsible for the memorization.

Existing memorization analysis techniques for text-to-image diffusion models are not well suited for this setting. Methods designed for standard diffusion models, such as verbatim detection based

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.

on SSCD (Self-Supervised Copy Detection) (Pizzi et al., 2022) or segmentation-based approaches such as FB-Mem (Di et al., 2026), measure memorization by only associating generated images with the input text modality alone. These methods fail to jointly account for the visual inputs specific to multi-modal diffusion models, so that make them inapplicable for this setting. To address this gap, we propose a memorization evaluation approach for multi-modal diffusion models, which, to the best of our knowledge, is the first approach quantitatively measuring memorization in this setting. Our method uses Grounding DINO (Liu et al., 2024) to annotate objects in an image based on caption keywords, then masks each detected object with a bounding box, and finally asks the model to reconstruct it from the remaining background and the text prompt with associated key words. Reconstruction quality is measured by MS-SSIM (Multi-Scale Structural Similarity Index Measure) (Wang et al., 2003) to reflect how faithfully the model reconstructs object-level details from its training data. This design makes full use of the multi-modal nature of the encoder. Furthermore, we demonstrate that our approach delivers results equivalent to the gold-standard leave-one-out memorization metric, while requiring a significantly lower computational cost.

Using our new method, we find that the pre-training samples for multi-modal encoders exhibit higher memorization level than samples used for the backbone diffusion fine-tuning. This finding holds for both BLIP-Diffusion and ELITE, two representative architectures of multi-modal diffusion models. We show that it is because the CLIP encoder generates information-dense embeddings for input image-caption pairs that effectively carry memorized content into the diffusion backbone during generation. Fine-tuning the U-Net, by contrast, adapts the model to distributional patterns rather than preserving instance-level details. In other words, multi-modal encoder plays a more important role in memorization effect than diffusion backbone in multi-modal diffusion model. Since highly memorized training images are under higher risks to be reconstructed from multi-modal diffusion models, this leads privacy leakage risks comparable to those observed in standard text-to-image diffusion models. Further analysis shows that this leakage comes from the joint memorization of both visual and textual signals, yet the vision modality contributes more than the text modality. Motivated by these observations, we further show that applying memorization mitigation at the multi-modal encoder level is significantly more effective than applying it at the level of the diffusion backbone, while also better preserving image generation quality.

In summary, we make the following contributions:

1. We propose the first memorization evaluation method for multi-modal diffusion models that leverages caption-grounded object detection and masked reconstruction, and is equivalent to the gold-standard leave-one-out memorization metric at a substantially lower computational cost.
2. We identify that the multi-modal encoder plays a more important role in the memorization effect for multi-modal diffusion models, which contrasts with prior memorization studies on standard text-to-image diffusion models that mainly focus on the diffusion backbone.
3. We demonstrate that highly memorized samples are under privacy leakage risk which is driven by the joint memorization of image–caption associations, with the vision modality playing the more important role.
4. We show that memorization mitigation applied at the multi-modal encoder level achieves a better privacy-utility tradeoff than diffusion-level mitigation for multi-modal diffusion models.

2 Background and Related Work

Multi-modal diffusion models extend traditional text-to-image diffusion models, which are insufficient for subject-driven generation where the goal is to synthesize novel scenes while preserving specific visual concepts provided by the user (Ruiz et al., 2023; Gal et al., 2022). By leveraging a multi-modal encoder that takes a reference image–caption pair as input at inference time, these models enable high-quality subject-driven generation (Ye et al., 2023; Li et al., 2023a; Wei et al., 2023). A main challenge in multi-modal diffusion models is to bridge the gap between the visual representation produced by the image encoder and the text-conditioned cross-attention mechanism of the traditional text-to-image diffusion models (Ye et al., 2023; Pan et al., 2024). Existing approaches can broadly be summarized into two paradigms. The first paradigm is **Alignment-Compression**, which employs a learned bottleneck module to align and compress the rich visual features from the image encoder into a fixed-length sequence of soft prompt tokens that can be directly fed into the

cross-attention layers alongside text embeddings (Shi et al., 2024; Yuan et al., 2023; Li et al., 2023a). A representative example is BLIP-Diffusion (Li et al., 2023a), which uses a pre-trained Q-Former (Li et al., 2023b) to extract and compress features from a reference image into a small set of (e.g., 16 or 32) visual prompt tokens. These tokens are then concatenated with text embeddings to condition the diffusion process. The second paradigm is **Visual Token Replacement**, which takes a more direct approach by replacing a subset of text token embeddings in the CLIP text encoder with visual token embeddings derived from the reference image, effectively injecting visual subject information into the text prompt without introducing additional cross-attention pathways (Ma et al., 2024; Wei et al., 2023; Li et al., 2024). ELITE (Wei et al., 2023) is a representative example. It trains a global mapping network and a local mapping network to encode the reference image into pseudo-word tokens that are substituted into the textual conditioning sequence, enabling the subject's appearance to directly modulate the generation process through the existing text-conditioned attention mechanism.

Memorization in machine learning refers to model's tendency to keep information about details from specific training data, instead of learning generalized pattern among all training data (Zhang et al., 2016; Chatterjee, 2018; Feldman & Zhang, 2020). While a certain degree of memorization has been shown to benefit generalization in some cases (Feldman, 2020; Wang et al., 2024b), it also leads to privacy risks at two levels. First, it enables membership inference attacks, where an adversary can determine whether a specific sample was included in the training set, so that potentially leaking sensitive personal or proprietary information (Shokri et al., 2017; Carlini et al., 2022; Tang et al., 2024). At a more severe level, it allows the content of training data to be directly reconstructed from model outputs (Fredrikson et al., 2015; Hatamizadeh et al., 2022; Carlini et al., 2023; Kowalczuk et al., 2025), raising critical problems about the exposure of sensitive or copyright-protected material.

To detect and measure memorization, prior work has proposed different methods. Image-level similarity metrics compare generated images against training images with different noise initialization but same prompt based on feature-level copy detection (Carlini et al., 2023; Somepalli et al., 2023b). Segmentation-based approaches partition generated images into foreground and background regions and measure memorization separately in each, providing a more fine-grained detection of partial memorization (Di et al., 2026; Somepalli et al., 2023a). Attention-based methods analyze cross-attention patterns during the denoising process, and have identified that memorized prompts exhibit abnormally concentrated attention on specific tokens (Ren et al., 2024; Wen et al., 2024). However, all of these methods share a fundamental limitation: they are designed for diffusion models, only based on text input, and solely consider the diffusion process. In contrast, in multi-modal diffusion models, image generation is based on reference image-caption pairs encoded by pretrained multi-modal encoders. As a result, the prior methods fail to capture the memorization signal carried by the reference image input, making them unable to effectively measure memorization in the multi-modal diffusion setting. These limitations highlight the need for a memorization evaluation approach that jointly leverages both the reference image features encoded by the multi-modal encoder and the text prompt used in the diffusion process, enabling a comprehensive measurement of memorization, which is the focus of our work.

3 Memorization in Multi-Modal DMs

Preliminaries. In our method, we leverage the **Grounding DINO** (Liu et al., 2024) as an object detection model to localize objects in an image conditioned on captions. By grounding detection directly on natural language, it enables keyword-driven region localization in a bounding box without requiring predefined category labels. Then, for the reconstruction quality assessment, we use **MS-SSIM** (Wang et al., 2003), which evaluates image similarity across multiple scales by decomposing luminance, contrast, and structural components, and has been shown to be more robust to minor local variations than single-scale SSIM (Structural Similarity Index Measure) (Wang et al., 2004) or feature-based metrics such as SSCD (Pizzi et al., 2022).

Our Memorization Analysis. We propose a memorization evaluation approach that quantitatively measures the memorization level of multi-modal diffusion models. Figure 1 presents an overview of our method. Given an image-caption pair, we first apply Grounding DINO as image-keyword annotation tool to localize all objects in the image and associate each bounding box with its corresponding keyword in the caption. For each object, we mask the corresponding region in the image and pass the masked image into the encoder. A keyword-conditioned prompt derived from the caption is then

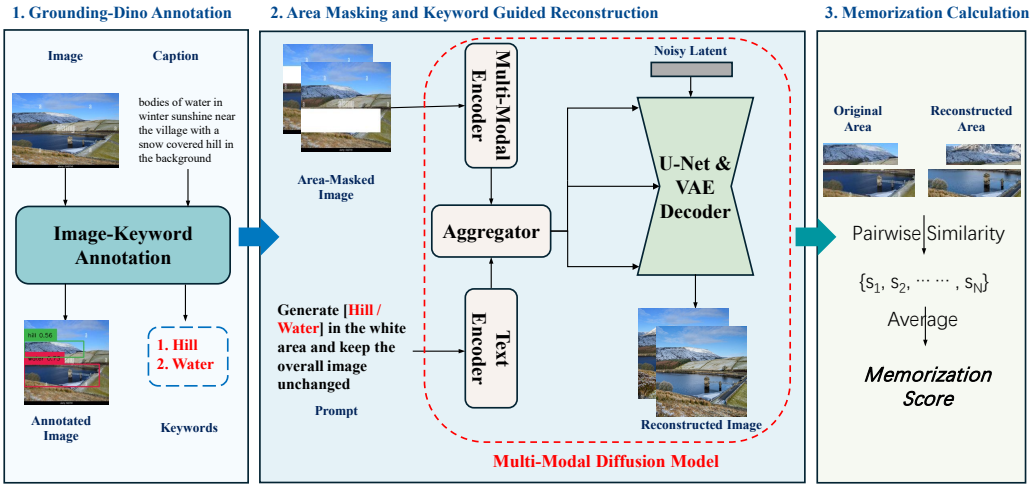


Figure 1: **Our proposed memorization evaluation for multi-modal diffusion models.**

fed into the diffusion model along with the image embedding to reconstruct the masked region. The quality of the reconstruction is evaluated by computing the MS-SSIM as pairwise similarity score between the reconstructed object region and the original one. The final memorization score for a given image-caption pair is defined as the average MS-SSIM across all detected objects. A higher score indicates that the model tends to reproduce the details of its training images more closely, suggesting stronger memorization. The details of our approach are presented in Algorithm 1.

The core idea of the proposed approach is fully leveraging the multi-modal architecture of the model. We introduce a keyword-driven object annotation approach to ensure that the annotated object regions semantically align with the keywords in caption, rather than being determined only by visual prominence. Notably, the reconstruction target is the masked object region rather than the background, which is **opposite** to standard multi-modal object-driven text-to-image diffusion training, where the object is typically preserved and a new background is generated. This design is intentional because adopting the same paradigm as in training makes it difficult to determine whether differences in reconstruction quality arise from memorization or generalization.

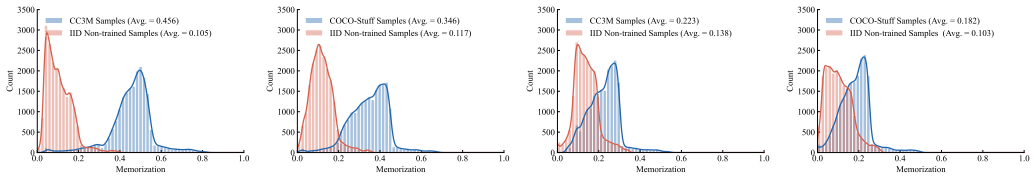
Leave-One-Out Memorization Metric. We verify that the memorization signal captured by our approach genuinely arises from training-set membership by leveraging the gold-standard leave-one-out (LOO) memorization metric. LOO is based on a causal comparison between two models f and g that differ only in a small set of samples used during their training. Model f is the model for which we need to verify its memorization distribution and model g is used as a reference.

We measure memorization as the robustness of a model's multi-modal representation to random input corruption. Given an image-caption pair (x, c), we apply random masking at two levels: an *encoder-level* mask m_{img} that occludes a random region of x, and a *prompt-level* mask that replaces a random subset of tokens in c with a blank placeholder, yielding $\tilde{c}$. This differs from the keyword-grounded masking. Here the mask location and the masked prompt tokens are chosen uniformly at random, independent of caption keywords or detected objects, since the goal is to probe generic representation stability rather than to localize memorized objects. For a model θ, we define its masked representation distance on (x, c) as the L2 distance between the representation of the clean input and the representation of its randomly masked counterpart:

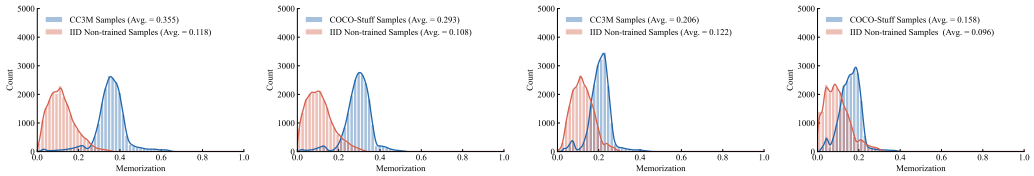
$$R_\theta(x, c) = \left\| \texttt{Encoder}_\theta(x, c) \; - \; \texttt{Encoder}_\theta\big(x \odot (1 - m_{\text{img}}),\; \tilde{c}\big) \right\|_2. \tag{1}$$

A smaller $R_\theta(x, c)$ indicates that the model's representation is more robust to random masking, i.e., the model can still recover a similar representation for the corrupted input, which is a signature of memorization. We define the LOO memorization score as:

$$\Delta(x, c) = R_{\text{g}}(x, c) - R_{\text{f}}(x, c). \tag{2}$$



(a) B: Mask object (b) B: Mask object (c) B: Mask background (d) B: Mask background



(e) E: Mask object (f) E: Mask object (g) E: Mask background (h) E: Mask background

Figure 2: **Memorization distribution for BLIP-Diffusion and ELITE.** We evaluate the memorization score distribution on 1) the multi-modal encoder pretrain samples (CC3M), 2) Unet Fine-tuning samples (COCO-Stuff), and 3) IID non-training samples. We denote BLIP-Diffusion as B and ELITE as E. (a),(b),(e),(f) are memorization distribution measured by masking the objects (our proposed approach in Figure 1), while (c),(d),(g),(h) are memorization distribution measured by masking the background.

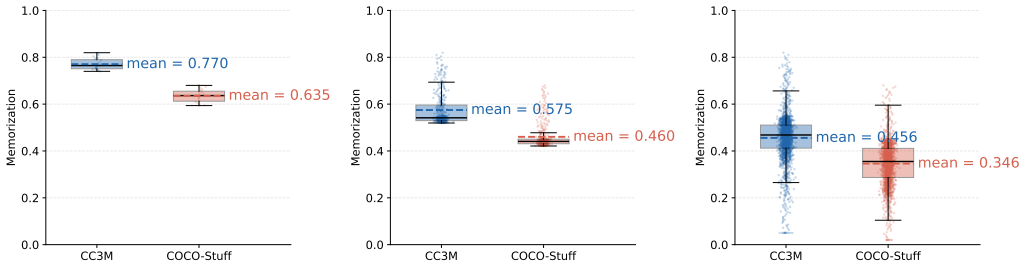
4 Memorization Evaluation

Experimental Setup. We evaluate our approach on two representative architectures for multi-modal diffusion: ELITE and BLIP-Diffusion. Both architectures share a common CLIP encoder pre-trained on CC3M (Sharma et al., 2018), and a UNet diffusion backbone fine-tuned on COCO-Stuff (Caesar et al., 2018). The pretrained encoder is frozen when UNet is fine-tuned. The detailed encoder architectures and corresponding training datasets are summarized in Table 5, while training hyperparameters for each stage are provided in Table 6 (both in Appx. A.1).

Main Evaluation. We evaluate our proposed approach on BLIP-diffusion and ELITE over both pre-training and fine-tuning data and compare the resulting scores with those of non-training samples that are not expected to exhibit memorization. However, the multi-modal encoder is pre-trained on a substantially larger dataset (1M CC3M) than the UNet (100K COCO-Stuff). To ensure a fair analysis, we compute memorization scores for all training samples from both datasets and sort the results in descending order. We then perform uniform sampling over the full set of data points for 20K samples from each training set. For in-distribution non-trained samples, we use remaining 20K untrained samples from the CC3M and COCO-Stuff datasets.

Figures 2a to 2d presents the memorization distribution for BLIP-Diffusion and Figures 2e to 2h are results for ELITE, both of which are measured by 2 different approaches. Figures 2a and 2b and Figures 2e and 2f are derived from our proposed approache to memorization analysis. In contrast, Figures 2c and 2d and Figures 2g and 2h are computed using the training-process simulation method described in Sec. 3, where the background is masked and memorization is quantified based on reconstructing the masked background. For BLIP-diffusion, results in Figure 2a and Figure 2b demonstrate that the proposed memorization approach effectively distinguishes between training samples and non-training samples. Specifically, samples observed during training exhibit consistently higher memorization scores, while IID non-training samples tend to have significantly lower values. This observation aligns with prior studies on memorization in machine learning models such as CLIP and diffusion-based generative models Feldman (2020); Wang et al. (2024b; 2025), where a similar separation between training and non-training data has been reported.

The comparisons in Figure 2 between object vs background masking reveal that the background-masking-and-reconstruction strategy, which simulates the training process, fails to effectively separate memorization scores for training samples from non-training ones. This is because this method closely follows the training process, some non-training samples can still achieve high scores due to the model's strong generalization ability. Therefore, samples with average memorization scores become ambiguous on whether they reflect memorization or generalization, results in large memorization



(a) Top 1% samples (b) Top 20% samples (c) All samples

Figure 3: **Memorization distribution inside BLIP-Diffusion.** We present results for CLIP pretrain samples (CC3M) and Unet Fine-tuning samples (COCO-Stuff).

score overlapping between training and non-training samples. In contrast, our proposed method adopts a different strategy by masking out the object regions that are consistently preserved during training, such that the model can only access background regions that are ignored during training. However, the model cannot perform its generalization from the aspects that they ignore during training. Consequently, only samples with sufficiently high memorization can leverage the implicit coupling between object and background to reconstruct high-quality object content, thereby achieving high memorization scores. For non-training samples or those with low memorization, this coupling is relatively weak, making it difficult to recover object information from background cues alone and resulting in consistently lower scores. This design enables a more reliable distinction memorization from generalization.

The results for ELITE are aligned with those for BLIP. However, compared to BLIP-Diffusion, the average memorization level of training samples is lower. This difference is mainly due to the variation in how visual features are extracted in the two methods. A more detailed discussion is provided in Appendix A.3. We also perform the same experiments on a separate subset of out-of-distribution 20K images drawn from the Flickr30K dataset Young et al. (2014). Results shown in Figure 4 in Appendix A.3 follow the same trend as for the IID data described above.

Equivalence to Leave-One-Out Metric. As an additional validation, we adopt the leave-one-out (LOO) protocol commonly used to measure memorization in machine learning (Feldman, 2020; Feldman & Zhang, 2020), following the representation-level methodology (Wang et al., 2024a). Specifically, we train two models (both on Elite and BLIP-Diffusion) that differ only in a held-out 0.1M training candidate split of CC3M. We measure the memorization according to Eq. (2) in representation output under random masking of the image, caption, and prompt inputs, so that any systematic difference between the paired encoders on the candidate split is attributable to training-set membership. We find that the samples identified as highly memorized by our image-level approach exhibit substantial overlap with those identified by the representation-level LOO-based measurement, with over 91% agreement among the top 100, top 1000, and top 5000 most memorized samples. Our results are presented in Table 1. This confirms that our image-level measurement achieves a memorization assessment consistent with the LOO-based metric, while avoiding both the additional computational overhead of training a reference model and the distribution-shift risk introduced by the leave-one-out dataset splitting. We provide further details in Appx. A.4.

Table 1: **Overlap between the top K most memorized samples** identified by the image-level memorization score and the leave-one-out representation-level score. We also report the results of the statistical Kendall's Tau Test as τ-Statistic / p-value.

	Top 100	Top 1000	Top 5000
Overlap (Ratio)	91.0%	92.1%	94.2%
Kendall's Tau Test	0.186 / 2.99e-2	0.197 / 1.48e-4	0.266 / 7.55e-6

Memorization Distribution. We compare in Figure 3 the memorization distributions measured by our proposed approach for BLIP-Diffusion on CC3M and COCO-Stuff. We follow the uniform

Table 2: **Ablation of encoder memorization contribution under three encoder conditions.** The results reported as mean memorization over the full training set and over the top-1% and top-20% highest-scoring samples.

		BLIP-Diffusion			ELITE		
		top-1%	top-20%	Full Set	top-1%	top-20%	Full Set
Baseline	CC3M	0.769 ± 0.023	0.575 ± 0.069	0.456 ± 0.102	0.614 ± 0.017	0.448 ± 0.058	0.355 ± 0.077
	COCO-Stuff	0.633 ± 0.026	0.460 ± 0.053	0.346 ± 0.092	0.478 ± 0.020	0.369 ± 0.036	0.293 ± 0.064
COCO fine-tuned	CC3M	0.548 ± 0.015	0.384 ± 0.050	0.303 ± 0.070	0.469 ± 0.013	0.334 ± 0.037	0.257 ± 0.059
	COCO-Stuff	0.541 ± 0.017	0.380 ± 0.048	0.299 ± 0.069	0.454 ± 0.011	0.318 ± 0.035	0.244 ± 0.057
Random Initialization	CC3M	0.201 ± 0.005	0.138 ± 0.025	0.104 ± 0.049	0.199 ± 0.005	0.137 ± 0.024	0.103 ± 0.045
	COCO-Stuff	0.197 ± 0.004	0.139 ± 0.022	0.106 ± 0.047	0.198 ± 0.005	0.137 ± 0.022	0.105 ± 0.046

sampling strategy described in Sec. 4 and sample 20K samples from each training set. We visualize the distributions using box plots with jitter for the top 1% (Figure 3a) and top 20% (Figure 3b) samples with highest memorization scores, as well as for all samples (Figure 3c). The results in Figure 3 show that CC3M images consistently achieve higher memorization scores than COCO-Stuff images across all sample groups, as confirmed by Welch's t-test (t-statistic / p-value: 16.72 / 2.73×10^{-19}, 26.03 / 3.41×10^{-107}, and 35.67 / 3.17×10^{-242} for the top 1%, top 20%, and all training samples, respectively). Notably, while statistical significance increases with sample size, the mean score gap between CC3M and COCO-Stuff is most pronounced among the top 1% most memorized samples. The results in Figure 5 in Appendix A.3 for ELITE follow the same trend. To rule out the potential confound introduced by the dataset distribution shift between CC3M and COCO-Stuff, we repeat the experiment with the roles reversed: the multi-modal encoder is trained on the 100K subset of COCO-Stuff, while the UNet backbone is fine-tuned on the 1M subset of CC3M. The results in Figure 6 in Appendix A.3 for both BLIP-Diffusion and ELITE follow the same trend, where samples used to train the multi-modal encoder consistently exhibit higher memorization levels than those used to fine-tune the UNet backbone. This demonstrates that the observed difference in average memorization between samples used to train the multi-modal encoder and those used to fine-tune the UNet backbone is not attributable to dataset distribution shift, but rather to the difference in their contributions to the overall memorization of the model, with the multi-modal encoder contributing more than the UNet backbone.

This can be explained by two main factors. First, the multi-modal encoder produces semantically rich and information-dense embeddings for input images. Even when the object region is masked, the encoder can still extract visual information of masked objects into latent representation with the help of nearby area by its memorization. Second, diffusion fine-tuning primarily adapts the model to the overall data distribution rather than preserving instance-level details. This makes the diffusion model highly rely on conditioning signals from the prompt and the multi-modal encoder (Kim & No, 2026). Therefore, memorization encoded in the multi-modal encoder plays a more important role in the generation process of the model.

To further confirm the role that the multi-modal encoder plays in model's memorization, we implement two model variants. In the first variant, we fine-tune the multi-modal encoder on COCO-Stuff for 32 epoch while keeping the diffusion model frozen. This allows the fine-tuned encoder to achieve a comparable memorization level on COCO-Stuff to that on CC3M. In the second variant, we replace the multi-modal encoder with a random initialization of the same architecture, and leaving the diffusion model unchanged. The randomly initialized encoder contains no memorization on any specific data points from either COCO-Stuff or CC3M dataset. Finally, the original model serves as the baseline. We then evaluate all three models by measuring memorization on both CC3M and COCO-Stuff, considering the top 1% and top 20% most memorized samples, as well as the full dataset. Together, these settings enable a controlled analysis of how encoder memorization influences the overall behavior of the model. The results in Table 2 show that the memorization gap between CC3M and COCO-Stuff greatly decreases across all variant settings on both ELITE and BLIP-Diffusion. When measured over the full dataset, this gap becomes almost negligible. These results suggest that when the multi-modal encoder exhibits similar levels of memorization across different training sets, or lacks memorization entirely, the average memorization scores of the whole model also becomes similar. This supports our claim that memorization from multi-modal encoder contributes more to the generation process of the multi-modal diffusion models.

Table 3: **Reconstruction performance on highly memorized samples under different caption conditioning strategies.** Target (**T**) caption (**C**) is the caption of the highly memorized samples. Retrieved (**R**) image (**I**) is the most semantic aligned images from open-source non-training dataset retrieved by Target caption. Blank (**B**) caption (**C**) is fully filled by **space**. Blank (**B**) image (**I**) is fully white image. Results report as the average value for top 1000 most memorized samples.

		BLIP-Diffusion			ELITE		
		SSIM ↑	LPIPS ↓	SSCD ↑	SSIM ↑	LPIPS ↓	SSCD ↑
Most Memorized Samples	**RI + TC**	0.359 ± 0.026	0.544 ± 0.023	0.404 ± 0.022	0.333 ± 0.023	0.557 ± 0.026	0.399 ± 0.020
	RI + BC	0.149 ± 0.015	0.800 ± 0.029	0.089 ± 0.009	0.140 ± 0.012	0.855 ± 0.025	0.071 ± 0.008
	BI + TC	0.092 ± 0.008	0.893 ± 0.031	0.041 ± 0.007	0.090 ± 0.008	0.892 ± 0.035	0.040 ± 0.008
Least Memorized Samples	**RI + TC**	0.151 ± 0.014	0.801 ± 0.030	0.065 ± 0.007	0.144 ± 0.011	0.810 ± 0.029	0.064 ± 0.008
	RI + BC	0.101 ± 0.011	0.876 ± 0.021	0.047 ± 0.005	0.100 ± 0.011	0.877 ± 0.023	0.048 ± 0.006
	BI + TC	0.079 ± 0.004	0.903 ± 0.020	0.039 ± 0.004	0.077 ± 0.005	0.902 ± 0.018	0.037 ± 0.006

Table 4: **Memorization Mitigation effect on highly memorized samples under different mitigation strategy.** Results report as the average value for top 1000 most memorized samples.

BLIP-Diffusion					
	SSIM ↓	LPIPS ↑	SSCD ↓	CS ↑	Aesthetic ↑
Baseline	0.359 ± 0.026	0.544 ± 0.023	0.404 ± 0.022	0.802 ± 0.036	5.563 ± 0.078
Mitigation on Encoder	0.205 ± 0.021	0.761 ± 0.026	0.183 ± 0.011	0.770 ± 0.029	5.377 ± 0.096
Mitigation on Diffusion	0.303 ± 0.023	0.612 ± 0.025	0.314 ± 0.015	0.658 ± 0.025	4.581 ± 0.188
ELITE					
	SSIM ↓	LPIPS ↑	SSCD ↓	CS ↑	Aesthetic ↑
Baseline	0.333 ± 0.023	0.557 ± 0.026	0.399 ± 0.020	0.796 ± 0.033	5.688 ± 0.074
Mitigation on Encoder	0.193 ± 0.016	0.774 ± 0.028	0.177 ± 0.011	0.767 ± 0.025	5.291 ± 0.097
Mitigation on Diffusion	0.295 ± 0.019	0.620 ± 0.025	0.311 ± 0.015	0.644 ± 0.027	4.499 ± 0.202

Privacy Leakage Risks. Since no dedicated privacy leakage evaluation tool exists for multi-modal diffusion models (e.g., membership inference attacks), we evaluate privacy leakage risk by examining whether the training images can be faithfully reconstructed from the model using the corresponding caption paired with a semantically similar image retrieved from a non-training open-source dataset. Experiments are implemented on highly memorized samples on both BLIP-Diffusion and ELITE. We first select the 1,000 training samples with the highest memorization scores. For each sample, we then retrieve a semantically similar image from the other open-source non-training dataset according to the corresponding caption, pair it with the memorized caption as encoder input, and incorporate key caption words into the diffusion prompt. The generated image is then compared against the original sample using SSIM, LPIPS Zhang et al. (2018), and SSCD as reconstruction similarity metrics. To further investigate the respective roles of the image and text modalities in reconstructing highly memorized samples, we implement 2 control experiments: 1) using blank images (fully white) as reference image along with target caption based prompt; 2) using the retrieved image as reference image along with blank caption input (fully space with same length). As comparison, we repeat the same experiments on the 1,000 samples with the lowest memorization scores to serve as a baseline and verify that the high-quality reconstruction is due to memorization rather than being a general property of the model.

Results in Table 3 show a consistent pattern across both BLIP-Diffusion and ELITE: when conditioned on captions from highly memorized training samples with retrieved semantic align image, the models achieve high reconstruction quality. This is reflected across all metrics, including higher SSIM, and SSCD, as well as significantly lower LPIPS, indicating that the generated images are not only visually closer but also structurally more similar to the original training samples. This reveals a significant privacy risk that sensitive information about training image can be reconstructed through caption with other open-source nontraining images. The two control experiments further reveal the respective contributions of each modality. We find that removing either the target caption or the semantically aligned reference image leads to a substantial drop in reconstruction quality, indicating that both modalities play a role in reconstruction. However, replacing the retrieved image with an irrelevant one while keeping the target caption, reconstruction quality drops more sharply than keeping the relevant image while not using target caption, confirming that the semantic aligned visual signal is

the more important driver of reconstruction. At the same time, the least memorized samples exhibit much lower reconstruction quality under all experiment settings. These declines confirm that the high-fidelity reconstruction observed for the most memorized samples is not a general capability of the model, but is attributable to memorization. This confirms that highly memorized samples are indeed subject to more serious privacy leakage risk in multi-modal diffusion models. We present several reconstructed images for highly memorized training samples in Figure 7 at Appendix A.3.

Memorization Mitigation. We further explore how memorization in highly memorized samples can be mitigated to reduce the privacy leakage risks discussed in Sec. 4, by applying mitigation strategies separately on both multi-modal encoder and the UNet diffusion backbone. Following Wang et al. (2025), we first mitigate memorization directly on the multi-modal encoder before fine-tuning the diffusion model. We implement dememorizing fine-tuning for the multi-modal encoder on the 1,000 highly memorized samples identified in Sec. 4. During mitigation, Gaussian noise with zero mean and 0.1 standard deviation is injected into the both image and text representations during contrastive training process. This has been shown that could effectively mitigate memorization of highly memorized samples Wang et al. (2025) in multi-modal contrastive learning. The fine-tuned encoder is then used to train downstream UNet diffusion backbone from the same initialization checkpoint under same training settings as baseline settings. In the second setting, we keep the multi-modal encoder frozen and instead mitigation memorization during the UNet diffusion backbone fine-tuning, following Wen et al. (2024). We add noise to the multi-modal encoder output and apply gradient filtering at the batch level. Specifically, before backward pass, we remove gradients from samples whose magnitude deviates from the batch mean by more than one standard deviation. This has also been shown to effectively reduce the memorization level of highly memorized samples of diffusion models Wen et al. (2024). To evaluate the impact of memorization mitigation on privacy leakage, we adopt the same highly memorized data reconstruction process described in Sec. 4, with the original model serving as the baseline. We measure reconstruction similarity between the generated images and the target reconstructed samples still using SSIM, LPIPS, and SSCD. To evaluate generation quality, we further report the CS (CLIP Score) Hessel et al. (2021) between the generated images and the diffusion prompts, along with aesthetic scores computed using a LAION aesthetics predictor pretrained on LAION-5B (Schuhmann et al., 2022).

The results in Table 4 show that both strategies reduce the similarity between reconstructed images and highly memorized samples. Notably, encoder-level mitigation achieves the strongest effect. On BLIP-Diffusion, SSCD drops from 0.404 to 0.183 and SSIM from 0.359 to 0.205; on ELITE, similar trends are observed (SSCD: 0.399 $\rightarrow$ 0.177; SSIM: 0.333 $\rightarrow$ 0.193). Diffusion-level mitigation also reduces reconstruction similarity, but to a lesser degree, and incurs a higher decrease in CS and Aesthetic Score, indicating a worse privacy-utility trade-off. Unlike previous mitigation approaches that target on the diffusion backbone itself, implementing mitigation at the multi-modal encoder not only more effectively reduces privacy leakage, but also better preserves generation quality for multi-modal diffusion models. We also show the memorization score changes after two different mitigation strategies. Results in Table 7 at Appendix A.3 show that implementing mitigation at the multi-modal encoder also better reduce the memorization level of most memorized samples. Details are discussed in Appendix A.3.

5 Conclusions

In this work, we propose the first memorization approach for multi-modal subject-driven text-to-image diffusion models, which uses Grounding DINO for caption-grounded object localization and MS-SSIM for masked reconstruction quality assessment, enabling reliable identification of memorized training samples. Using this approach, we characterize the memorization distribution across multi-modal diffusion models. Furthermore, we demonstrate that highly memorized training samples can be reconstructed with high fidelity using their associated captions with semantic aligned open-source non-training images, placing a serious privacy leakage risk on multi-modal diffusion models. Finally, we show that memorization mitigation applied at the encoder level achieves a better privacy-utility tradeoff than diffusion-level mitigation, reducing reconstruction similarity more effectively while better preserving generation quality. Our findings suggest that future privacy-mitigation efforts for multi-modal diffusion models should prioritize encoder-side interventions.

6 AI USE STATEMENT

In this work, we used generative AI tools for **None of tasks with the required disclosure**. We have not used generative AI tools for **all of tasks with the required disclosure**. Additionally, we used generative AI tools for **create or edit software code**. We have reviewed all AI-assisted work. We check the AI-assisted codeworks and make sure it actually work as what we expected and stated in the supplementary materials. We take respon-sibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Computer vision and pattern recognition (CVPR), 2018 IEEE conference on*. IEEE, 2018.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE symposium on security and privacy (SP)*, pp. 1897–1914. IEEE, 2022.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Satrajit Chatterjee. Learning and memorization. In *International conference on machine learning*, pp. 755–763. PMLR, 2018.

Jimmy Z Di, Yiwei Lu, Yaoliang Yu, Gautam Kamath, Adam Dziedzic, and Franziska Boenisch. Demystifying foreground-background memorization in diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20763–20771, 2026.

Vitaly Feldman. 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.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in neural information processing systems*, 33:2881–2891, 2020.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 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.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Ali Hatamizadeh, Hongxu Yin, Holger R Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10021–10030, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278, 2024.

Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision–language models. *Advances in Neural Information Processing Systems*, 37:50722–50749, 2024.

Bumjun Kim and Albert No. Memorization in stable diffusion is unexpectedly driven by clip embeddings. *arXiv preprint arXiv:2605.02908*, 2026.

Antoni Kowalczuk, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Privacy attacks on image autoregressive models. In *Forty-Second International Conference on Machine Learning (ICML)*, 2025.

Antoni Kowalczuk, Jan Dubiński, Franziska Boenisch, and Adam Dziedzic. Finding dori: Discovery of retained images in diffusion models. In *Forty-Third International Conference on Machine Learning (ICML)*, 2026.

Dongxu Li, Junnan Li, and Steven Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. *Advances in Neural Information Processing Systems*, 36:30146–30166, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023b.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8640–8650, 2024.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Casey Meehan, Florian Bordes, Pascal Vincent, Kamalika Chaudhuri, and Chuan Guo. Do ssl models have déjà vu? a case of unintended memorization in self-supervised learning. *Advances in Neural Information Processing Systems*, 36:42775–42798, 2023.

Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2024.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8543–8552, 2024.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Xincheng Shuai, Henghui Ding, Xingjun Ma, Rongcheng Tu, Yu-Gang Jiang, and Dacheng Tao. A survey of multimodal-guided image editing with text-to-image diffusion models. *arXiv preprint arXiv:2406.14555*, 2024.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6048–6058, 2023b.

Shuai Tang, Zhiwei Steven Wu, Sergul Aydore, Michael Kearns, and Aaron Roth. Membership inference attacks on diffusion models via quantile regression. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 47819–47829, 2024.

Wenhao Wang, Adam Dziedzic, Michael Backes, and Franziska Boenisch. Localizing memorization in ssl vision encoders. *Advances in Neural Information Processing Systems*, 37:60475–60516, 2024a.

Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, and Franziska Boenisch. Memorization in self-supervised learning improves downstream generalization. In *The Twelfth International Conference on Learning Representations*, 2024b.

Wenhao Wang, Adam Dziedzic, Grace C Kim, Michael Backes, and Franziska Boenisch. Captured by captions: On memorization and its mitigation in clip models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The thirty-seventh asilomar conference on signals, systems & computers, 2003*, volume 2, pp. 1398–1402. IEEE, 2003.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Ryan Webster. A reproducible extraction of training images from diffusion models. *arXiv preprint arXiv:2305.08694*, 2023.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15943–15953, 2023.

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.

Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

Ge Yuan, Xiaodong Cun, Yong Zhang, Maomao Li, Chenyang Qi, Xintao Wang, Ying Shan, and Huicheng Zheng. Inserting anybody in diffusion models via celeb basis. *arXiv preprint arXiv:2306.00926*, 2023.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

A Appendix

A.1 Details on Experimental Setup

The detailed encoder architectures and corresponding training datasets are summarized in Table 5, while training hyperparameters for each stage are provided in Table 6.

Table 5: **Encoder architecture and training datasets for all models used in this paper.**

Model	Encoder	Training Set
BLIP-Diffusion	CLIP Enc. + Q-Former + SD U-Net	CC3M (1M subset) & COCO-Stuff (100K subset)
ELITE	CLIP Enc. + MLP + SD U-Net	CC3M (1M subset) & COCO-Stuff (100K subset)

Table 6: Training setup and hyperparameters for all stages used in this work. CLIP is pretrained on CC3M; ELITE and BLIP-Diffusion are fine-tuned on COCO-Stuff.

	CLIP (Pretrain)	**ELITE Global**	**ELITE Local**	**BLIP-Diffusion**
Training (Steps)	62528	80K	50K	150K
Warm-up	3908	—	—	—
Batch Size	512	16	8	16
Optimizer	AdamW	Adam	Adam	AdamW
Learning Rate	5e-4	1e-6	1e-5	2e-6
LR Schedule	Cos. Decay	Constant	Constant	Constant

A.2 Hardware Usage & Calculation Overhead

Two devices are used for all experiments mentioned in this paper: 1) a cloud server with 8 A100 GPUs and 4 H100 GPUs; 2) a local workstation with Intel 13700k CPU, Nvidia 4090 graphics card, a total of 10TB storage space and 64GB of RAM.

The total training time for CLIP pre-training is 17.2 hours. For ELITE 2-stage fine-tuning, it takes 63.1 hours (around 2.6 days). For Blip-Diffusion Unet fine-tuning, it takes 101.2 hours (around 4.2 days). It takes 6.2 hours to detect and localize object regions in images by using Grounding DINO for all 1M CC3M training samples. It takes 147.9 hours (around 6.2 days) for memorization evaluation on all 1M CC3M training samples for BLIP-Diffusion and 217.1 hours (around 9.0 days) for ELITE. It takes 13.7 hours for memorization evaluation on all 100K COCO-Stuff training samples for BLIP-Diffusion and 20.1 hours for ELITE. The memorization evaluation is considerably more time-consuming than training, despite operating on the same data. This is because training requires only a single forward pass per step, whereas memorization evaluation requires full DDIM sampling at inference time, which involves dozens of sequential denoising steps per image. This overhead is not specific to our method, but is a fundamental cost shared by all memorization evaluation methods that rely on comparing generated images to training samples Pizzi et al. (2022); Di et al. (2026); Carlini et al. (2023).

Inference Setup. For image generating process, we fully follow the default settings of BLIP-Diffusion and ELITE. We set 30 DDIM for BLIP-diffusion and 50 for ELITE. Both models use a guidance scale of 7.5.

A.3 Additional Experimental Results

Memorization Verification and Distribution for ELITE. Figure 2 presents the memorization verification results for ELITE. The results are consistent with those observed for BLIP-Diffusion in Figure 2: our proposed object-masking method (Figure 3a and 3b) produces a clear separation between training samples and non-training samples, while the background-masking strategy (Figure 3c and 3d) leads to substantial overlap between the two distributions. This confirms that the proposed memorization approach generalizes across different multi-modal diffusion architectures.

Algorithm 1 Memorization Measurement via Masked Reconstruction

Input: Image x, caption c, similarity metric $\mathcal{M}$
Output: Memorization score $\mathcal{S}$

$\mathcal{O} = \{(o_i, k_i)\}_{i=1}^{N} \leftarrow \texttt{GroundingDINO}(x, c)$ ▷ o_i: binary mask of i-th object, k_i: keyword
$\mathcal{S}_{\text{list}} \leftarrow [\,]$
for each $(o_i, k_i) \in \mathcal{O}$ **do**
  $x^{(i)}_{\text{masked}} \leftarrow x \odot (1 - o_i)$ ▷ Mask the i-th object
  $z^{(i)} \leftarrow \texttt{Encoder}\left(x^{(i)}_{\text{masked}}\right)$
  $p^{(i)} \leftarrow \texttt{BuildPrompt}(k_i)$
  $\hat{x}^{(i)} \leftarrow \texttt{UNet}\left(z^{(i)}, p^{(i)}\right)$
  $s_i \leftarrow \mathcal{M}\left(\hat{x}^{(i)} \odot o_i,\ x \odot o_i\right)$ ▷ Compute MS-SSIM on object region
  Append s_i to $\mathcal{S}_{\text{list}}$
end for
$\mathcal{S} \leftarrow \dfrac{1}{N} \sum_{i=1}^{N} s_i$
return $\mathcal{S}$

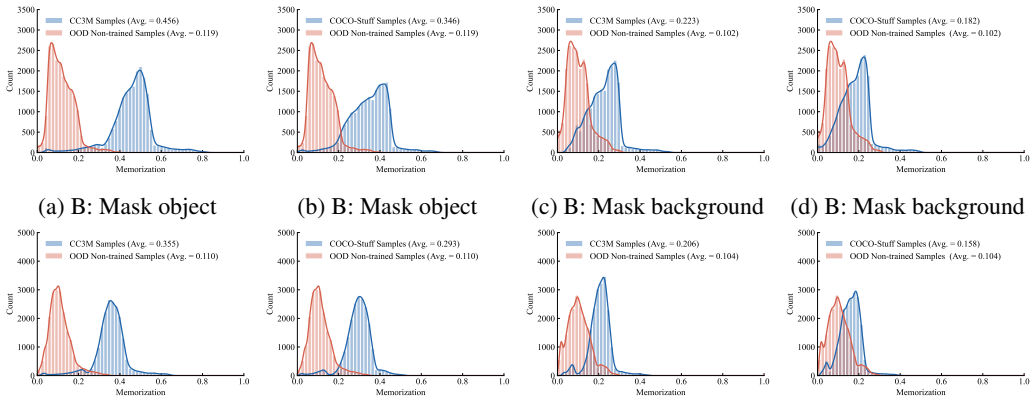
Table 7: **Memorization distribution after two different mitigation strategies.** Results reported as mean memorization over 1000 samples picked in Sec. 4.

	BLIP-Diffusion	ELITE
Baseline	0.781 ± 0.026	0.630 ± 0.027
Mitigation on Multi-modal Encoder	0.494 ± 0.013	0.381 ± 0.012
Mitigation on Diffusion Model	0.611 ± 0.016	0.495 ± 0.014

Figure 5 further compares the memorization distributions of CC3M and COCO-Stuff samples within ELITE. Consistent with the findings for BLIP-Diffusion, CC3M samples consistently achieve higher memorization scores than COCO-Stuff samples across all evaluation subsets, and this gap is most pronounced among the top-1% most memorized samples.

However, compared to BLIP-Diffusion, ELITE exhibits greater overlap between training and non-training distributions in Figure 3, and a smaller memorization gap between CC3M and COCO-Stuff in Figure 4. This difference can be attributed to the architectural distinction between the two models. BLIP-Diffusion employs a Q-Former to compress the full visual representation of the reference image into a fixed set of 32 tokens, producing information-dense embeddings that carry rich memorized content into the diffusion model. In contrast, ELITE adopts a token replacement strategy, where only a subset of text token embeddings are replaced by visual embeddings derived from the reference image. This design transmits less visual information through the encoder, weakening the encoder's overall influence on the generation process. As a result, the dominance of the multi-modal encoder over memorization is less pronounced in ELITE than in BLIP-Diffusion. Nevertheless, the consistent trend—CC3M samples exhibiting higher memorization than COCO-Stuff samples, and the clear separation between training and non-training distributions—confirms that the multi-modal encoder remains the primary driver of memorization in ELITE as well.

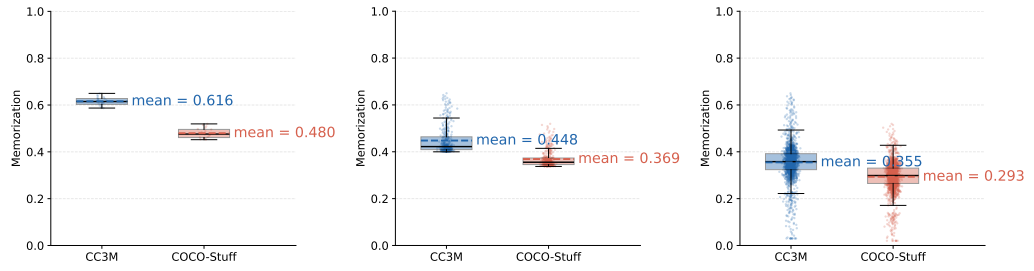
Memorization after Mitigation. Table 7 reports the mean memorization scores over the 1000 samples (picked in Sec. 4) after applying each mitigation strategy. Applying mitigation on multi-modal encoder achieves a larger reduction in memorization scores compared to diffusion-level mitigation across both BLIP-Diffusion and ELITE. For BLIP-Diffusion, CC3M memorization drops from 0.781 to 0.494 under encoder-level mitigation, compared to only 0.611 under diffusion-level mitigation. A similar trend is observed for ELITE. This further supports our finding that the multi-modal encoder is the dominant driver of memorization, and that implementing mitigation on multi-modal encoders more effectively relieves the memorization level of highly memorized samples.



(a) B: Mask object (b) B: Mask object (c) B: Mask background (d) B: Mask background

(e) E: Mask object (f) E: Mask object (g) E: Mask background (h) E: Mask background

Figure 4: **Memorization distribution for BLIP-Diffusion and ELITE on OOD 20K data from Flickr30K dataset.** We use the multi-modal encoder pretrain samples (CC3M), Unet Fine-tuning samples (COCO-Stuff), and in-distribution non-training samples. We denote BLIP as B and ELITE as E. (a),(b),(e),(f) are memorization distribution measured by masking the objects (proposed in Algorithm 1), while (c),(d),(g),(h) are memorization distribution measured by masking the background.



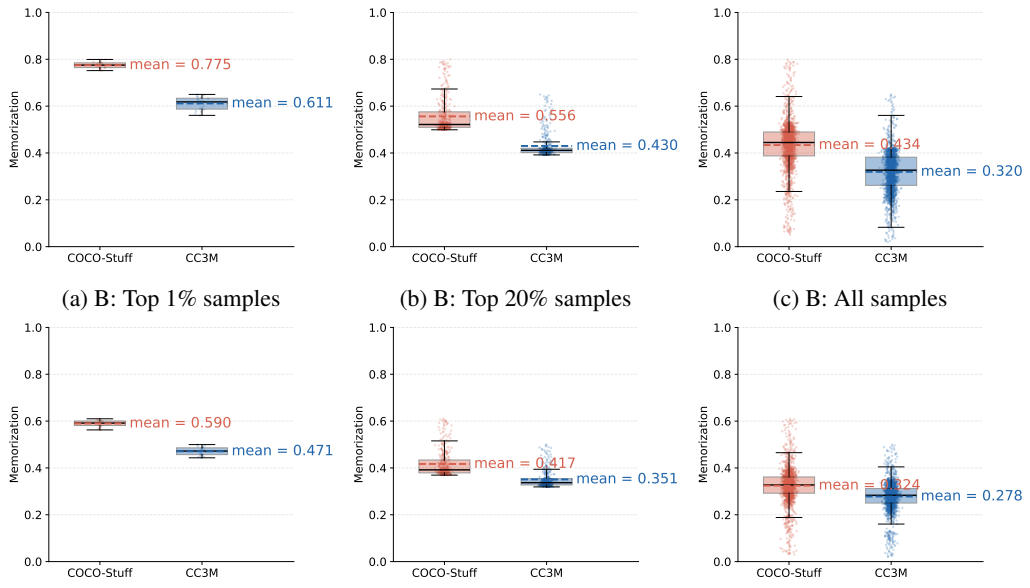
(a) Top 1% samples (b) Top 20% samples (c) All samples

Figure 5: **Memorization distribution inside ELITE for CLIP pretrain samples (CC3M) and Unet Fine-tuning samples (COCO-Stuff).** Welch's t-test (t-statistic / p-value) results are $21.99 \;/\; 4.49 \times 10^{-23}$, $23.05 \;/\; 2.80 \times 10^{-87}$, and $27.59 \;/\; 2.20 \times 10^{-153}$ for the top 1%, top 20%, and all training samples, respectively.

A.4 Leave-One-Out Memorization Metric

Leave-One-Out Data Split. We construct a 1.1M-sample subset of CC3M and partition it into three disjoint sets: a *shared* set D_S of 0.9M samples used to train both encoders; a *candidate* set D_C of 0.1M samples used to train model f only and whose memorization status we wish to verify; and an *independent* set D_I of 0.1M samples used to train model g only to ensure model f and g have same amount of training samples. We then pre-train two multi-modal models per architecture (BLIP-Diffusion and ELITE) using identical architecture, hyperparameters, and number of training steps as in Table 6, differing only in additional 0.1M split: model f is trained on $D_S \cup D_C$, and model g is trained on $D_S \cup D_I$. Since both models are trained on the same total amount of data and share the identical D_S, any systematic difference in their behavior on D_C can be causally attributed to the memorization induced by training on D_C, rather than to confounds such as training set size or data distribution shifts.

Leave-One-Out Memorization Score. For every sample in D_C and D_I, we compute R_{f} according to Eq. (1) for model f and R_{g} for model g. On D_C, a positive Δ computed according to Eq. (2) indicates that the encoder which has been trained on the sample produces a representation that shifts less (i.e., a smaller L2 distance) under random masking than the encoder that has never seen it, i.e., a causal signature of stronger memorization. Evaluating the same quantity on the IID non-trained samples, where neither encoder has been trained on the sample, provides the corresponding null-effect



(a) B: Top 1% samples (b) B: Top 20% samples (c) B: All samples

(d) E: Top 1% samples (e) E: Top 20% samples (f) E: All samples

Figure 6: **Memorization distribution inside BLIP-Diffusion (B) and ELITE (E) for inverse settings, where CLIP pretrain on COCO-Stuff and Unet Fine-tuned on CC3M.** (a)-(c): results for BLIP-Diffusion (B), (d)-(f): results for Elite (E). Welch's t-test (t-statistic / p-value) results for BLIP-Diffusion (B) are 23.90 / 5.93×10^{-20}, 27.63 / 1.29×10^{-115}, and 37.77 / 1.09×10^{-266} for the top 1%, top 20%, and all training samples, respectively. Welch's t-test (t-statistic / p-value) results for Elite (E) are 23.72 / 6.81×10^{-24}, 19.06 / 1.10×10^{-65}, and 21.50 / 3.27×10^{-97} for the top 1%, top 20%, and all training samples, respectively.

baseline. Algorithm 2 summarizes the full procedure, which we apply to both BLIP-Diffusion and ELITE.

Algorithm 2 Leave-One-Out Memorization Verification via Masked Representation Distance

Input: Candidate set D_C, model $\texttt{Model}_\text{f}$, $\texttt{Model}_\text{g}$
Output: Leave-one-out memorization scores $\{\Delta(x, c)\}$ for all $(x, c) \in D_C$

for each $(x, c) \in D$ **do**
  $m_\text{img} \leftarrow \texttt{RandomMask}(x)$ ▷ Random encoder-level (image) mask
  $\tilde{c} \leftarrow \texttt{RandomMask}(c)$ ▷ Random prompt-level (token) mask
  $\tilde{x} \leftarrow x \odot (1 - m_\text{img})$
  for each $\theta \in \{\text{IN}, \text{OUT}\}$ **do**
    $r_\theta \leftarrow \texttt{Encoder}_\theta(x, c)$ ▷ Representation of the clean input
    $\tilde{r}_\theta \leftarrow \texttt{Encoder}_\theta(\tilde{x}, \tilde{c})$ ▷ Representation of the masked input
    $R_\theta(x, c) \leftarrow \|r_\theta - \tilde{r}_\theta\|_2$
  end for
  $\Delta(x, c) \leftarrow R_\text{g}(x, c) - R_\text{f}(x, c)$
end for
return $\{\Delta(x, c)\}_{(x,c)\in D_C}$

Validation Results. Figure 8 presents the distribution of the leave-one-out memorization score $\Delta(x, c)$ for ELITE and BLIP-Diffusion, evaluated on the candidate set D_C and the IID Non-trained Samples (the samples neither encoder has trained on). For both architectures, $\Delta(x, c)$ on D_C is concentrated at large positive values (Avg. $= 0.813$ for ELITE, 0.826 for BLIP-Diffusion), while $\Delta(x, c)$ on IID samples is tightly centered near zero (Avg. $= 0.082$ for ELITE, 0.083 for BLIP-Diffusion), which forms the expected null-effect baseline since neither encoder has been trained on D_I. This clear separation confirms that this representation-level, leave-one-out based score is also effective at distinguishing trained samples from non-trained IID ones.

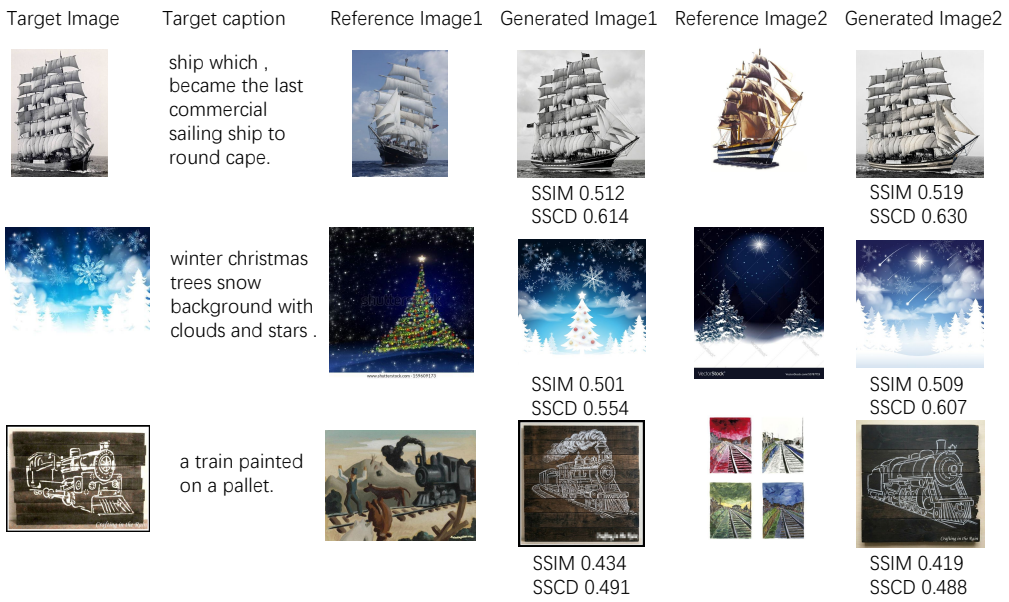
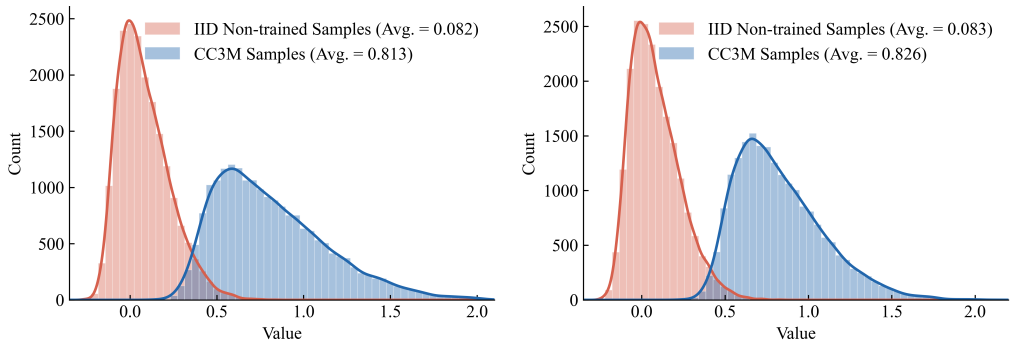


Figure 7: **Samples of training extracted training images for the most memorized samples.** The first 2 row are the samples with highest reconstruction scores. The third row is the sample sample that successfully reconstruct the letter watermark in target image.



(a) ELITE (b) BLIP-Diffusion

Figure 8: **Distribution of the leave-one-out memorization score $\Delta(x, c)$.** Results are shown for ELITE (a) and BLIP-Diffusion (b), evaluated on the candidate set D_C (CC3M Samples) and the independent set D_I (IID Non-trained Samples).

To further examine whether the image-level score used in the main text and the representation-level score proposed here capture a consistent notion of memorization, we compare, on the candidate set D_C, the ranking of samples produced by the main text's image-level memorization score (Algorithm 1) against the ranking produced by the leave-one-out score $\Delta(x, c)$ (Algorithm 2). Table 1 reports the overlap between the top-K most memorized samples identified by each measurement, for $K \in \{100, 1000, 5000\}$. We observe a substantial and increasingly stable overlap across all three scales, with 91%, 92.1%, and 94.2% of the top-K samples agreeing between the two measurements, respectively. This shows that the image-level memorization score, despite being computed from a single trained model without access to any reference model, identifies largely the same set of highly memorized samples as the causally grounded, representation-level leave-one-out score.

A.5 IMPACT & LIMITATIONS

This work introduces the first memorization approach for multi-modal diffusion models and identifies the vision modality from the multi-modal encoder as the dominant source of memorization and privacy leakage. Our findings suggest that future privacy-mitigation efforts for multi-modal diffusion models should prioritize encoder-side interventions, which we show to be more effective and less damaging to generation quality than diffusion-level approaches. One limitation of our proposed approach is that it relies on Grounding DINO for object localization, and may produce inaccurate memorization estimates for samples where the caption does not correctly describe the main objects present in the image. However, such misaligned image-caption pairs constitute only a small fraction of both public and high-quality private datasets, and are therefore unlikely to affect the general applicability of our approach.