

---

# Quantifying Cross-Modality Memorization in Vision-Language Models

---

Yuxin Wen<sup>1\*</sup>, Yangsibo Huang<sup>2</sup>, Tom Goldstein<sup>1</sup>  
Ravi Kumar<sup>2</sup>, Badih Ghazi<sup>2</sup>, Chiyuan Zhang<sup>2</sup>

<sup>1</sup>University of Maryland, College Park

<sup>2</sup>Google

## Abstract

Understanding what and how neural networks memorize during training is crucial, both from the perspective of unintentional memorization of potentially sensitive information and from the standpoint of effective knowledge acquisition for real-world, knowledge-intensive tasks. While previous studies primarily investigate memorization within a single modality, such as text memorization in large language models or image memorization in diffusion models, unified multimodal models are becoming increasingly prevalent in practical applications. In this work, we focus on the unique characteristics of *cross-modality* memorization and conduct a systematic study centered on vision-language models. To facilitate controlled experiments, we first introduce a synthetic persona dataset comprising diverse synthetic person images and textual descriptions. We quantify factual knowledge memorization and cross-modal transferability by training models on a single modality and evaluating their performance in the other. Our results reveal that facts learned in one modality transfer to the other, but a significant gap exists between recalling information in the “source” and “target” modalities. Furthermore, we observe that this gap exists across various scenarios, including more capable models, machine unlearning, and the multi-hop case. At the end, we propose a baseline method to mitigate this challenge. We hope our study can inspire future research on developing more robust multimodal learning techniques to enhance cross-modal transferability.

## 1 Introduction

Modern foundation models are continuing to benefit from scaling with more training data. Large volumes of diverse, high-quality data allow the models to develop fundamental capabilities like language understanding and reasoning, and are critically important for the models to acquire world knowledge required in many problem-solving benchmarks [Zellers et al., 2019, Hendrycks et al., 2021, Wei et al., 2024], and even more specialized knowledge required for solving math and coding problems [Jain et al., 2025, Rein et al., 2023]. However, the dynamics of knowledge acquisition from training data are extremely complex, and various unintended biases [Zheng et al., 2024] and behaviors [Nasr et al., 2023] can arise, such as the “reverse curse” [Berglund et al., 2024] and increased hallucination [Gekhman et al., 2024]. Moreover, previous work also shows that in addition to learning generalizable knowledge, large language models (LLMs) can unintentionally memorize training text verbatim, which may then be extracted through simple prompting [Carlini et al., 2019, 2021, 2022]. This raises concerns about privacy, copyright, and the trustworthiness of AI systems. As a result, developing a better understanding of memorization in foundation models has become a central focus of recent research.

---

\*Work done during an internship at Google.

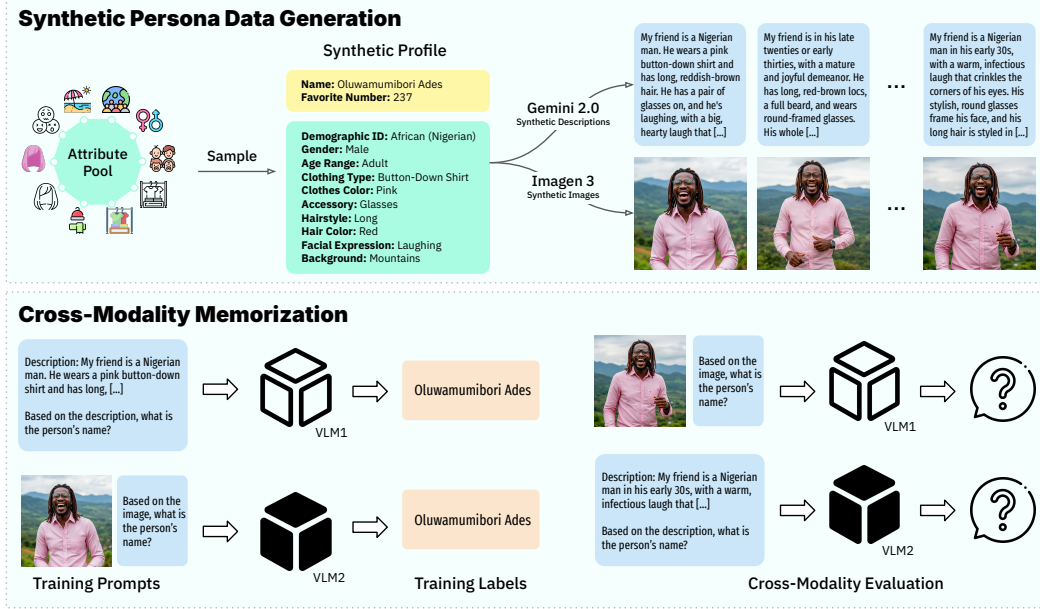


Figure 1: Illustration of Our Data Generation Pipeline and Evaluation Pipeline.

However, while prior work has investigated memorization within individual modalities—such as the regurgitation of specific training text sequences by LLMs [Carlini et al., 2019, 2021, 2022] and the reproduction of specific training images by diffusion models [Somepalli et al., 2023b, Carlini et al., 2023, Wen et al., 2024a]—these studies primarily focus on *unimodal* memorization. In contrast, solving complex real-world problems increasingly relies on multimodal models that are trained to process and integrate information from diverse media types, including text, images, audio, and video. This integration across modalities necessitates extending the study of memorization to account for *cross-modality* behaviors.

In this work, we address this gap by taking a first step toward systematically quantifying cross-modality memorization between text and images in vision-language models (VLMs). Under controlled training conditions, we examine how knowledge is memorized in one modality and transferred across modality boundaries to support inference in another.

Specifically, our investigation focuses on *factual knowledge memorization* across the visual and textual modalities in VLMs. Our task design is motivated by emerging applications of VLMs such as personal assistants, where models are often “personalized” using user-provided data such as emails, calendars, to-do lists, photos, and screenshots. For example, to assist with vacation planning, a model may check a calendar to identify suitable dates and examine past photos to suggest similar destinations. Gaining a clearer understanding of how factual knowledge is internalized and transferred across modalities is a critical step toward building robust and trustworthy multimodal personal agents.

To facilitate this study, we introduce the “persona” dataset, designed to systematically investigate cross-modal factual memorization, as shown in Figure 1. We begin by generating a diverse set of synthetic persona profiles, each featuring various attributes. For each persona profile, we create both textual descriptions and visual representations (e.g., synthetic photographs) to capture the person’s appearance.

Next, we train vision-language models to recognize the person’s name from a single modality, either the description or the image. During inference, we assess the model’s ability to recall the person’s name when presented with a held-out image or description. Our primary focus is on cross-modal knowledge transfer, examining the model’s ability to answer textual questions when trained on visual data, and vice versa. To determine whether the model learns facts consistently across modalities, we compare the accuracy between the training and testing modalities.

Our systematic analysis reveals that facts learned in one modality are automatically transferred to the other. However, there is a noticeable performance gap between recalling information within the

original (“source”) modality and recalling it in the “target” modality in many scenarios. Notably, this gap is asymmetric, with knowledge transfer from image to text being more effective than from text to image, even in larger and more capable models. A similar inconsistency arises in machine unlearning, where the model struggles to unlearn facts in a modality different from the one in which they were originally learned. Additionally, we identify a cross-modal “multi-hop curse” in multimodal models, where the model struggles to infer complex cross-modal relationships despite successfully learning simpler single-hop associations.

To better understand whether the cross-modality knowledge barriers are surface-level phenomena or more fundamental weaknesses in the standard learning setup, we further investigate whether simple mitigating strategies can improve cross-modal transferability. We find that using diverse training data and larger, more capable models can mitigate overfitting, but it does not improve the transferability rate. In contrast, augmenting training data with synthetic image-text pairs, generating images from text inputs and captions from image inputs, can effectively bridge the modality gap, particularly with in-distribution data augmentation. However, out-of-distribution data proves less beneficial in this case.

We hope our study contributes to a deeper understanding of cross-modal memorization phenomena, particularly within the context of VLMs. By highlighting the asymmetries in knowledge transfer between modalities, we aim to foster more robust, efficient, and privacy-conscious multimodal systems. Future work will focus on refining mitigation strategies and exploring additional modalities to generalize our findings further. We believe that addressing these challenges is crucial for developing AI systems that can safely and effectively integrate and process diverse types of data.

## 2 Related Work

Modern generative models memorize and regurgitate training data [Carlini et al., 2019, Inan et al., 2021, Zhang et al., 2023], presenting potential risks such as privacy leakage and copyright infringement [Henderson et al., 2023]. This phenomenon in language models has been extensively discussed in prior literature. Carlini et al. [2019] systematically examined such occurrences by injecting “canaries” into the training data. Later, Carlini et al. [2022], Kandpal et al. [2022] showed that model memorization correlates with model size, canary repetition frequency, and sequence length. Beyond verbatim copying, factual knowledge memorization has also been studied [Allen-Zhu and Li, 2025], wherein models recall specific facts present in their training datasets. This type of memorization is considered more aligned with realistic usage scenarios, as it often involves prompting the model with queries that differ from the exact training instances.

Research into memorization has also extended to image generation models [Somepalli et al., 2023b,a, Wen et al., 2024a]. Somepalli et al. [2023b] discovered that verbatim memorization also occurs in diffusion models, where the model can generate exact replicas of training images when prompted with corresponding training data inputs. Later, Carlini et al. [2023] demonstrated that, similar to language models, the extent of memorization in diffusion models is heavily dependent on model and data sizes. Analogous to factual knowledge memorization in language models, image generation models also memorize styles [Somepalli et al., 2024] and copyrighted characters [He et al., 2025]. In Jayaraman et al. [2024], Kokhlikyan et al. [2024], the authors investigate unintentional memorization in CLIP-style VLMs, whereas our work focuses on LLaVA-style VLMs.

## 3 Cross-Modality Factual Knowledge Memorization

### 3.1 Synthetic Persona Data

Injecting canaries is a widely adopted strategy in previous work [Carlini et al., 2019, Wen et al., 2024b] to prevent data contamination and maintain experimental control. Accordingly, we first develop a synthetic persona dataset as “canaries.” Our synthetic persona dataset consists of a set of image-description samples representing synthetic profiles. Given that personal assistants represent a promising application for vision-language models, the setting for our synthetic persona dataset is chosen to simulate a scenario where the model learns to identify individuals (e.g., a user’s acquaintances) based on images (similar to those saved on a user’s phone) or their textual descriptions (as might appear in text messages or emails).

An overview of our data generation pipeline is presented in Figure 1. In detail, the creation involves the following steps:

**I. Attribute Pool Definition:** The foundation of our dataset generation lies in defining a comprehensive set of attributes that constitute a persona. These attributes are categorized to cover various aspects of an individual’s appearance, demographics, and contextual elements. For each category, we curate a pool of possible values, drawing inspiration from typical characteristics used in image generation and description. These attribute categories include:

- **Demographics:** 8 demographic identities, 2 genders, and 3 age ranges.
- **Visual Characteristics:** 7 clothing styles/types, 7 clothes colors, 7 accessories, 7 hairstyles, and 7 hair colors.
- **Contextual Elements:** 7 facial expressions and 12 background scenes.

The specific values within these pools are chosen to maximize diversity and verisimilitude, allowing for a large combinatorial space of unique profiles of

$$8 \times 2 \times 3 \times 7 \times 7 \times 7 \times 7 \times 7 \times 12 = 67,765,824$$

combinations from the attribute pools. The complete list is provided in Appendix B.

**II. Profile Synthesis:** The synthetic profile synthesis process involves the following steps:

1. **Attribute Combination:** A unique combination of attributes is sampled without replacement from the defined pools to establish the core characteristics of a persona.
2. **Name Assignment:** A unique fictional name is generated, conditioned on the demographic identity and gender to maintain cultural relevance and consistency.
3. **PII Association:** To facilitate privacy-related experiments, we associate each synthetic character with PII data. Given that most models are well-aligned and consistently reject real PII data requests (e.g., Social Security Numbers), we opt to use “favorite number” as a surrogate. This favorite number is a randomly generated three-digit number.

**III. Image-Description Generation:** The generation of image-description pairs is a crucial step in creating a comprehensive dataset that effectively represents synthetic personas. This process not only ensures visual diversity but also enhances the contextual realism of each profile.

- **Image Generation:** Using the synthesized profiles (attribute combinations), we leverage a state-of-the-art text-to-image generation model Imagen 3 [Baldridge et al., 2024] to create realistic profile pictures.
- **Textual Description Generation:** To complement the generated images, we create rich textual descriptions for each persona profile picture. These descriptions go beyond merely listing attributes, instead presenting a natural, narrative, and contextually rich portrayal of the synthetic individual. We employ Gemini 2.0 [Team et al., 2023] to generate these textual descriptions, conditioned on both the generated image and the profile attributes. To enhance realism, we instruct the model to adopt a more conversational tone, similar to how a person would describe a friend.

Additionally, we generate multiple image-description pairs for each persona to simulate scenarios where a single individual has several images. This approach parallels the language model training process, where models are exposed to multiple paraphrases of a fact [Allen-Zhu and Li, 2025]. The generation prompts are provided in Appendix C.

**Final Dataset Composition:** The resulting synthetic persona dataset consists of a collection of 100 unique personas. Each persona is characterized by the following elements:

- A unique name.
- An associated favorite number.
- An associated set of specific attributes.
- A set of 100 image variants for training and 1 distinct image for testing.
- A set of 100 textual description variants for training and 1 distinct textual description for testing.

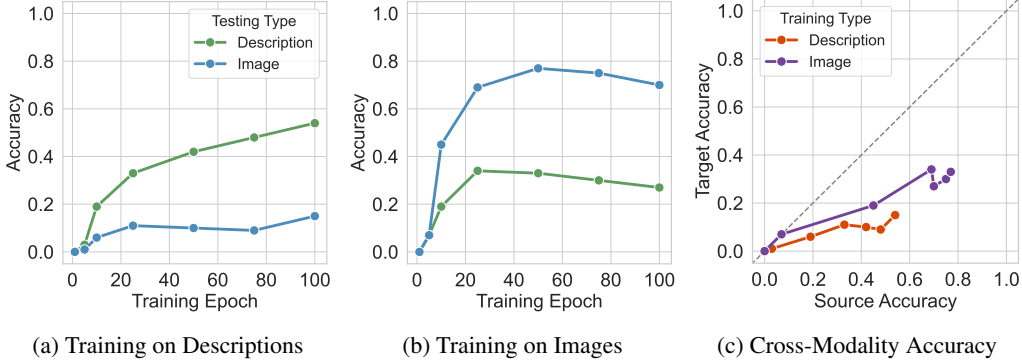


Figure 2: **Training with Single Description/Image.** Factual memorization transfers between modalities, but there is a significant gap between the source and target modalities.

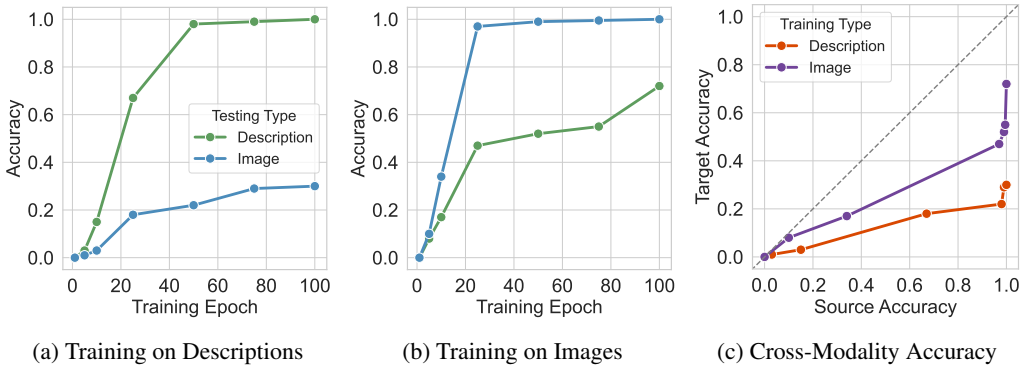


Figure 3: **Training with Image/Description Variants.** Using diverse input variations during training helps mitigate overfitting, but does not improve the cross-modal transferability rate.

### 3.2 Experimental Setup

We fine-tune the latest open-source vision-language model, Gemma-3-4b [Team et al., 2025], to conduct all our experiments. During fine-tuning, we utilize LoRA [Hu et al., 2022] with a rank of  $r = 32$ , a scaling factor of  $\alpha = 32$ , and a dropout probability of 0.05. We use AdamW [Loshchilov and Hutter, 2017] with a learning rate of  $2 \times 10^{-4}$  and a batch size of 16. All training is performed on a single Nvidia A100-80G GPU.

We employ two distinct training settings: one where the model is trained solely on textual descriptions and another where it is trained exclusively on images. In each setting, the input prompt consists of either a description or an image, followed by a question about the person’s name. We train the model to accurately predict the associated name based on the given input.

During testing, we evaluate the model’s recognition accuracy by asking the same question in both modalities separately, assessing the model’s ability to recall the correct name in each modality. Note that we test on seen persona profiles, but the specific test inputs (image or textual descriptions) are held out and never seen verbatim by the model during training. In our experiments, we deliberately train for an extreme number of epochs to simulate a worst-case privacy scenario, where memorized signals are not diluted by non-canary data. This setup provides a stress test for evaluating the robustness of potential mitigation methods.

### 3.3 Memorization Results

**Training with Single Description/Image:** We start with fine-tuning the model on a single image or description for each persona across multiple epochs. Specifically, we choose 1, 5, 10, 25, 50, 75, and 100 epochs. We report the recognition accuracy for both models trained on descriptions and

images in Figure 2a and Figure 2b, respectively, showing the test accuracy on held-out descriptions and images.

In the description training setting, the model’s accuracy for both modalities consistently improves as training progresses. In contrast, training on images tends to lead to overfitting after 25 epochs, as the model’s performance decreases afterwards.

Although cross-modal knowledge transfer occurs in both training scenarios, there remains a significant performance gap between the training modality and the testing modality. This gap is more evident in Figure 2c, where the plot compares the accuracy between the training (source) modality and the testing (target) modality. Ideally, a perfect model would align with the dotted line, indicating no discrepancy between the two modalities. However, both training settings reveal a noticeable gap. This disparity underscores the difficulty of transferring learned representations between text and image domains.

**Takeaway 1:** Factual memorization transfers between modalities, but there is a significant gap between the source and target modalities.

**Training with Image/Description Variants:** To mitigate the effect of overfitting associated with training on a single description or image for multiple epochs, we employ a different strategy in this experiment. Instead of repeatedly using the same input, we use a new description or image for each epoch. This approach is analogous to factual knowledge learning with paraphrased texts, which has been shown to help models generalize more effectively [Allen-Zhu and Li, 2025]. By introducing variations during training, the model avoids memorizing a fixed representation and instead learns the actual knowledge.

As depicted in Figure 3, training with diverse inputs significantly improves the test accuracy compared to the single-input scenario. In Figure 3a, where the model is trained on descriptions, the accuracy reaches nearly perfect levels, suggesting that the model benefits from varied textual data. Similarly, in Figure 3b, when trained on images, the model also achieves high accuracy without the early saturation and overfitting observed previously. These results highlight that varying the input per epoch effectively prevents overfitting and allows the model to generalize better across training instances.

However, despite these improvements in accuracy, the cross-modal transferability remains limited, as shown in Figure 3c, where the slope remains similar to the single-sample training scenario. This indicates that, while input variability mitigates overfitting, it does not necessarily enhance the model’s ability to bridge the gap between different modalities.

**Takeaway 2:** Using diverse input variations during training helps mitigate overfitting, but does not improve the cross-modal transferability rate.

One interesting observation is that the target accuracy across the modality continues to improve (especially for the image  $\Rightarrow$  text scenario) after the source modality saturates at near-perfect accuracies. This emphasizes the importance of diverse training samples on learning robust knowledge representations that go beyond what single-modality benchmarks can typically show.

**Training with Model Sizes:** As previous work [Carlini et al., 2022] indicates, larger models tend to memorize information more easily. To investigate the impact of model size on cross-modal transferability, we train three models from the Gemma3 family: Gemma3-4B-IT, Gemma3-12B-IT, and Gemma3-27B-IT. As shown in Figure 4, increasing the model size consistently improves the accuracy for both source and target modalities. However, despite the improvements in both individual modality accuracy and cross-modality accuracy as the model size increases, the rate of cross-modal transferability, indicated by the slope of the lines in the accuracy plots, remains nearly unchanged. This suggests that while larger models are more effective in learning within each modality, they do not necessarily enhance the transfer of learned information between modalities.

These findings imply that model size primarily contributes to improved memorization and recognition within a single modality rather than facilitating cross-modal generalization. Consequently, even as the model capacity increases, bridging the gap between image-to-text and text-to-image transfer remains a challenging task.



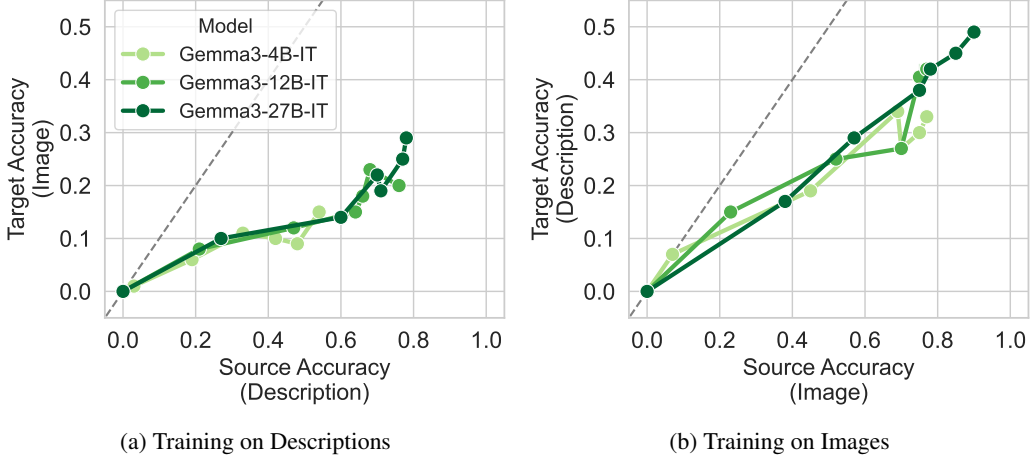


Figure 4: **Training with Different Model Sizes.** Increasing model size enhances accuracy within each modality, but maintains a similar cross-modal transferability rate.

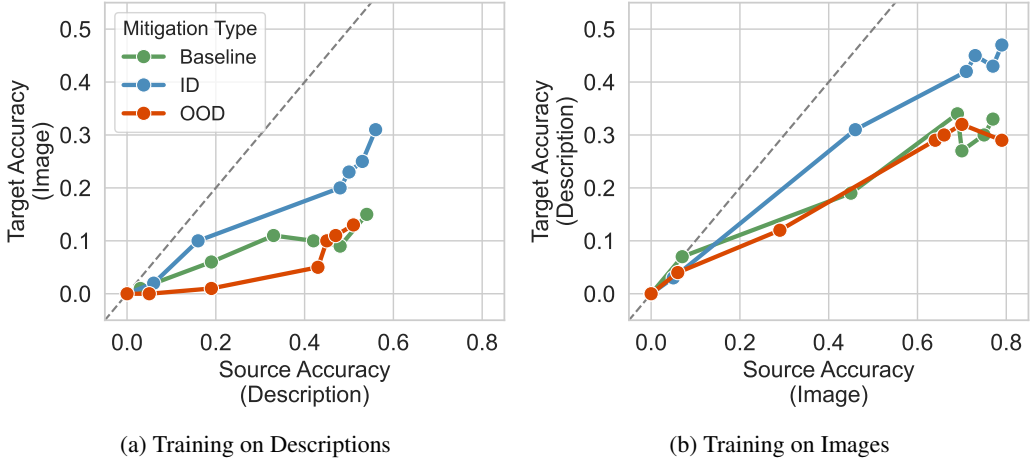


Figure 5: **Training with Augmentation Mitigation.** Augmenting training data with in-distribution image-caption samples is essential for enhancing cross-modal transferability, while out-of-distribution data does not provide similar benefits.

**Takeaway 3:** Increasing model size enhances accuracy within each modality, but maintains a similar cross-modal transferability rate.

### 3.4 Mitigation with Image-Caption Augmentation

To further enhance cross-modal transferability, we propose a simple yet effective approach that augments the training data with synthetic image-text pairs. Specifically, for a given description or text input, we generate a corresponding image conditioned on the text. Conversely, for an image input, we generate a descriptive caption. These generated pairs are introduced as captioning data, where the model is asked to describe an image.

In this experiment, we utilize two types of data for mitigation: 1) in-distribution (ID) data, where we augment the training set with held-out synthetic persona images and descriptions that closely resemble the training data, and 2) out-of-distribution (OOD) data, where we introduce incorporate images and captions from the COCO dataset [Lin et al., 2014]. This combination allows us to investigate the impact of data diversity on cross-modal transferability.

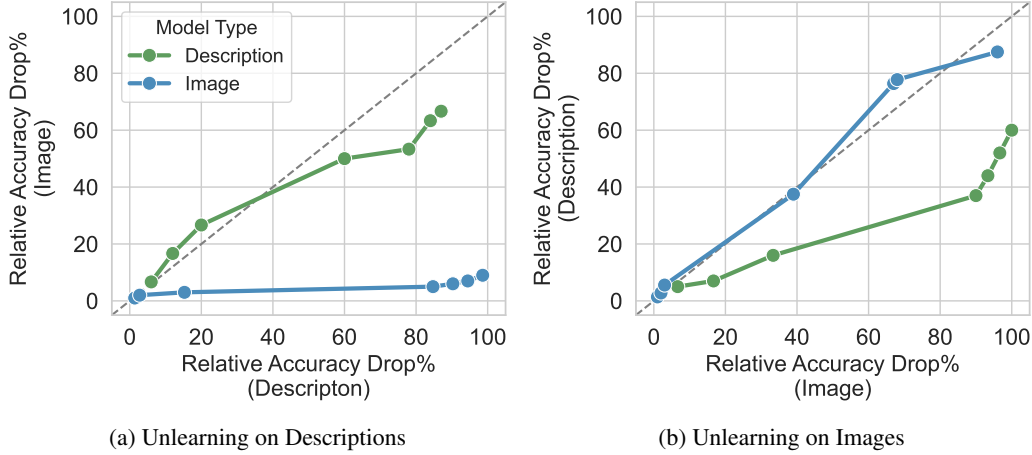


Figure 6: **Unlearning cross Modalities.** Unlearning effects are modality-dependent: when unlearning happens in the same modality as learning, the effect remains relatively consistent across both modalities. In contrast, when unlearning occurs in a different modality from where the fact was originally learned, the effect does not transfer equally between modalities.

As shown in Figure 5, incorporating in-distribution (ID) data significantly improves cross-modal transferability, as indicated by the blue line consistently appearing above the baseline green line in both description and image training scenarios. In contrast, augmenting with out-of-distribution (OOD) data does not provide similar benefits, with the orange line remaining close to the baseline. This suggests the importance of augmenting training data within the same distribution to effectively improve transferability.

**Takeaway 4:** Augmenting training data with in-distribution image-caption samples is essential for enhancing cross-modal transferability, while out-of-distribution data does not provide similar benefits.

### 3.5 Cross-Modal Unlearning

Machine unlearning [Bourtoule et al., 2021] has gained popularity as a method for removing specific user data without requiring the model to be retrained from scratch. Consequently, it is essential to investigate how unlearning operates across different modalities.

A commonly used unlearning method involves applying gradient ascent on the forget dataset [Maini et al., 2024]. However, our empirical experiments reveal that this approach is challenging to fine-tune and significantly compromises model performance. Therefore, to make the controlled experiments more comparable with fresh knowledge learning results, we adopt a practically simpler workaround: to unlearn the association of a person (image or textual description) to a name, we generate a modified training where each profile is associated with completely different names, and (continue) training the model on this modified dataset. We then measure the model’s accuracy (drop) of association with the original name as the unlearning performance. This allows us to reuse the same training setups, making the cross-modality measurements more easily comparable with previous results.

In detail, we start with the fine-tuned description model and image model from Section 3.3. Since we have two models (the fine-tuned description model and the fine-tuned image model) and two unlearning datasets (description-based and image-based), we systematically evaluate all possible combinations, which results in four experimental setups.

In Figure 6a, we present unlearning experiments with relative accuracy drops in both modalities when unlearning on description data. For the description model, the unlearning effect on both modalities exhibits a similar pattern (as indicated by the green line closely following the  $y = x$  line). Interestingly, for image models, unlearning on description data predominantly affects the description test accuracy, while the image test accuracy remains quite small. In Figure 6b, we



Table 1: **Cross-Modal Multi-Hop Learning** Cross-modal models suffer from the multi-hop curse, where accurate performance on individual tasks does not translate to multi-hop reasoning.

Training Data		Test Type				
Base Data	Multi-Hop Data	Desc⇒Name	Image⇒Name	Desc⇒FN	Image⇒FN	Name⇒FN
Desc⇒Name	Desc⇒FN	1.00	0.35	1.00	0.31	0.08
	Image⇒FN	1.00	0.36	0.39	1.00	0.05
	Name⇒FN	1.00	0.36	0.11	0.05	0.99
Image⇒Name	Desc⇒FN	0.68	1.00	1.00	0.47	0.07
	Image⇒FN	0.62	1.00	0.50	1.00	0.07
	Name⇒FN	0.64	1.00	0.27	0.65	0.99

show unlearning experiments on image data. Similar to the description data pattern, for the image (source modality) model, the relative accuracy drop on both modalities is comparable. However, for the text model, the unlearning effect on description data is notably weaker, while the drop in image accuracy is more significant.

Overall, these observations reveal a gap in cross-modalities unlearning: when unlearning facts in a modality different from the one in which the model originally learned the fact, the unlearning effect does not transfer equally between modalities. In contrast, when unlearning occurs in the same modality where the fact was originally learned, the unlearning effect is relatively similar across both modalities. This underscores the importance of carefully designing unlearning training data and developing more effective unlearning techniques for cross-modal scenarios.

**Takeaway 5:** Unlearning effects are modality-dependent: when unlearning happens in the same modality as learning, the effect remains relatively consistent across both modalities. In contrast, when unlearning occurs in a different modality from where the fact was originally learned, the effect does not transfer equally between modalities.

### 3.6 Cross-Modal Multi-Hop Learning

Large language models often exhibit the “multi-hop curse,” where they fail to infer that *A is C* when the model is trained on *A is B* and *B is C* [Balesni et al., 2024]. In this section, we investigate this multi-hop scenario in a cross-modal context.

We define the original description or image mapping to name data points as *A is B* data (Desc⇒Name and Image⇒Name), while a three-digit number representing a person’s favorite number (FN) serves as *C*. To investigate multi-hop reasoning, we introduce various data combinations involving base data: Desc⇒Name (given the description, predict the corresponding name) or Image⇒Name; multi-hop data: Desc⇒FN (given the description, predict the corresponding favorite number), Image⇒FN, and Name⇒FN. During testing, we assess whether the model can recall the person’s favorite number given a description, image, or name as a cue.

As shown in Table 1, cross-modal models also exhibit the multi-hop curse. For instance, when the model is trained on Desc⇒Name and Name⇒FN data, it achieves perfect accuracy on both tasks individually. However, when directly prompted with a description to retrieve the favorite number, the accuracy drops drastically to only 11%, indicating a significant challenge in the multi-hop scenario. In contrast, the model trained on Image⇒Name exhibits a lower barrier, achieving over 50% accuracy for Image⇒FN. Even when tested on Desc⇒FN, the accuracy is considerably higher compared to the model trained on descriptions, suggesting that image-based training better supports multi-hop reasoning.

**Takeaway 6:** Cross-modal models suffer from the multi-hop curse, where accurate performance on individual tasks does not translate to multi-hop reasoning.

## 4 Conclusion and Limitations

Our study systematically investigates cross-modality memorization in vision-language models using a synthetic persona dataset. We uncover an asymmetric performance gap between source and target modalities, where knowledge transfer from image to text is more effective than from text to image. We also demonstrate that while augmenting training data with in-distribution synthetic image-text pairs helps bridge this gap, out-of-distribution data proves less effective. Furthermore, we identify challenges in cross-modal unlearning, as unlearning in one modality does not fully translate to the other. Additionally, we reveal that vision-language models suffer from the multi-hop curse, where the model struggles to infer complex cross-modal relations despite successful learning of simpler single-hop associations.

Our findings underscore the complexity of achieving robust cross-modal transferability and highlight the need for improved training strategies to enhance generalization across modalities. While this work focuses on text and image modalities, future research should explore additional modalities, develop more sophisticated unlearning techniques, and design approaches to better handle multi-hop reasoning. Addressing these challenges is crucial for building more reliable and privacy-conscious multimodal AI systems.

## Acknowledgments

Icons from <https://www.flaticon.com/> were used to create the illustration figure in this paper.

## References

- Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. In *ICLR*, 2025.
- Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Lluís Castrejon, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, et al. Imagen 3. *arXiv:2408.07009*, 2024.
- Mikita Balesni, Tomasz Korbak, and Owain Evans. The two-hop curse: LLMs trained on  $A \rightarrow B$ ,  $B \rightarrow C$  fail to learn  $A \rightarrow C$ . *arXiv:2411.16353*, 2024.
- Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". In *ICLR*, 2024.
- Lucas Bourtole, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *S & P*, pages 141–159, 2021.
- Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security*, 2019.
- Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security*, pages 2633–2650, 2021.
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. In *ICLR*, 2022.
- Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security*, pages 5253–5270, 2023.
- Zorik Gekhman, Gal Yona, Roei Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In *EMNLP*, pages 7765–7784, 2024.
- Luxi He, Yangsibo Huang, Weijia Shi, Tinghao Xie, Haotian Liu, Yue Wang, Luke Zettlemoyer, Chiyuan Zhang, Danqi Chen, and Peter Henderson. Fantastic copyrighted beasts and how (not) to generate them. In *ICLR*, 2025.
- Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A. Lemley, and Percy Liang. Foundation models and fair use. *JMLR*, 24(400):1–79, 2023.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- Huseyin A Inan, Osman Ramadan, Lukas Wutschitz, Daniel Jones, Victor Rühle, James Withers, and Robert Sim. Privacy analysis in language models via training data leakage report. *arXiv:2101.05405*, 2021.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *ICLR*, 2025.
- Bargav Jayaraman, Chuan Guo, and Kamalika Chaudhuri. Déjà vu memorization in vision–language models. In *NeurIPS*, pages 50722–50749, 2024.
- Nikhil Kandpal, Eric Wallace, and Colin Raffel. Deduplicating training data mitigates privacy risks in language models. In *ICML*, pages 10697–10707, 2022.

- Narine Kokhlikyan, Bargav Jayaraman, Florian Bordes, Chuan Guo, and Kamalika Chaudhuri. Measuring dejavu memorization efficiently. In *NeurIPS*, pages 30015–30040, 2024.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv:1711.05101*, 2017.
- Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. *arXiv:2401.06121*, 2024.
- Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv:2311.17035*, 2023.
- D Rein, BL Hou, AC Stickland, J Petty, RY Pang, J Dirani, J Michael, and SR Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv:2311.12022*, 2023.
- Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. In *NeurIPS*, pages 47783–47803, 2023a.
- Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *CVPR*, pages 6048–6058, 2023b.
- Gowthami Somepalli, Anubhav Gupta, Kamal Gupta, Shramay Palta, Micah Goldblum, Jonas Geiping, Abhinav Shrivastava, and Tom Goldstein. Measuring style similarity in diffusion models. *arXiv:2404.01292*, 2024.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*, 2023.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivièrè, et al. Gemma 3 technical report. *arXiv:2503.19786*, 2025.
- Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models. *arXiv:2411.04368*, 2024.
- Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *ICLR*, 2024a.
- Yuxin Wen, Leo Marchyok, Sanghyun Hong, Jonas Geiping, Tom Goldstein, and Nicholas Carlini. Privacy backdoors: Enhancing membership inference through poisoning pre-trained models. In *NeurIPS*, 2024b.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv:1905.07830*, 2019.
- Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. Counterfactual memorization in neural language models. In *NeurIPS*, pages 39321–39362, 2023.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.

## A Broader Impacts

This paper systematically evaluates the state of cross-modal knowledge transfer in multimodal language models. Through extensive empirical studies, we reveal the existence of a significant knowledge transfer gap between vision and text domains. Despite this gap, our findings demonstrate that knowledge does transfer between these modalities, albeit asymmetrically.

Understanding this transfer is crucial for the future design and deployment of multimodal models, as it has important implications for both performance optimization and privacy preservation. Specifically, our results highlight the potential risk of unintended information leakage across modalities. For example, sensitive information learned from visual data could inadvertently influence text-based outputs and vice versa.

To mitigate such risks, researchers and practitioners should carefully evaluate cross-modal knowledge interactions, especially when designing models for real-world applications that handle sensitive or personal data. We encourage the community to further investigate robust techniques that ensure cross-modal privacy while maintaining model performance.

## B Attribute Pool

- **Demographic Identity:** African (South African), Caucasian (American), Caucasian (British), East Asian (Chinese), East Asian (Korean), Hispanic (Mexican), Iberian (Spanish), Middle Eastern (Saudi)
- **Gender:** Female, Male
- **Age Range:** Adult, Middle Aged, Senior
- **Clothing Type:** Button-Down Shirt, Hoodie, Jacket, Polo, Sweater, T-Shirt, Tank Top
- **Clothes Color:** Black, Brown, Gray, Pink, Red, White, Yellow
- **Accessory:** Earrings, Glasses, Hat, Headphones, Jewelry, No Accessory, Scarf
- **Hairstyle:** Afro, Bald, Curly, Dreadlocks, Long, Medium, Short
- **Hair Color:** Black, Blonde, Brown, Gray, Pink, Red, White
- **Facial Expression:** Angry, Frowning, Laughing, Neutral, Sad, Smiling, Surprised
- **Background:** Beach, Cafe, Cityscape, Forest, Graffiti Wall, Library, Mountains, Office, Plain Black, Plain Grey, Plain White, Sports Field

## C Data Generation Prompts

### Image Generation Prompt for Imagen 3

Generate a highly detailed, photorealistic photo of a person. Key characteristics include:  
Demographic Identity: African (Nigerian)  
Gender: Male  
Age Range: Adult  
Clothing Type: Button-Down Shirt  
Clothes Color: Pink  
Accessory: Glasses  
Hairstyle: Long  
Hair Color: Red  
Facial Expression: Laughing  
Background: Mountains

### Description Generation Prompt for Gemini 2.0

<image>

Describe the person in the image in a detailed and natural way, making it easy for someone to recognize them based on the description. Write as if you are describing a friend of my. Ensure the description flows naturally and includes the following attributes:

Demographic Identity: African (Nigerian)

Gender: Male

Age Range: Adult

Clothing Type: Button-Down Shirt

Clothes Color: Pink

Accessory: Glasses

Hairstyle: Long

Hair Color: Red

Facial Expression: Laughing

Background: Mountains

Feel free to add any additional distinguishing features that enhance the portrayal. Please provide the description directly, without extra formatting or instructions.

## D Training and Testing Prompts

### Training/Testing Prompt for Image Data

<image>

Based on the image, what is the person's name?

### Training/Testing Prompt for Description Data

Description: My friend is a Nigerian man. He wears a pink button-down shirt [...]

Based on the description, what is the person's name?

## E Test Prompt Engineering

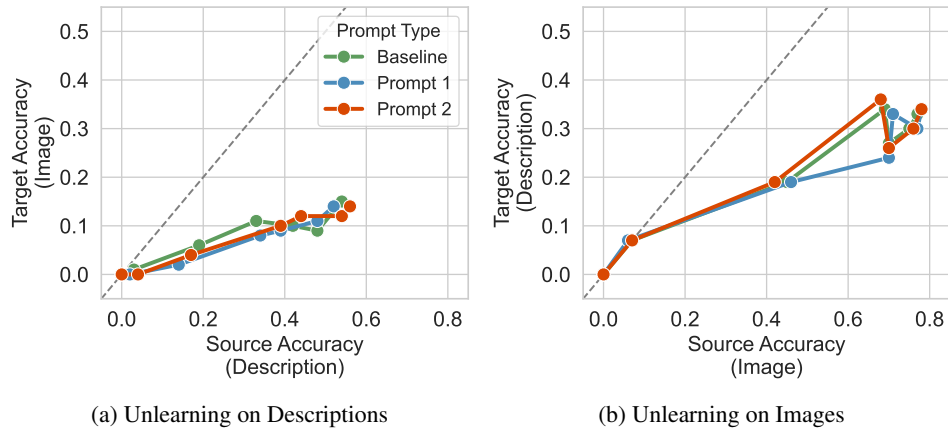


Figure 7: Testing with Different Prompt Strategies.



At the beginning of the project, we tested a naive mitigation strategy by appending modality transfer-aware prompts during inference. We evaluated two prompts:

- Prompt 1: Try to recall knowledge learned from another domain.
- Prompt 2: Given this description/image, think about what you learned from the image/description.

As shown in Figure 7, neither prompt leads to a significant change in transferability.

## F Unintentional Memorization in VLMs

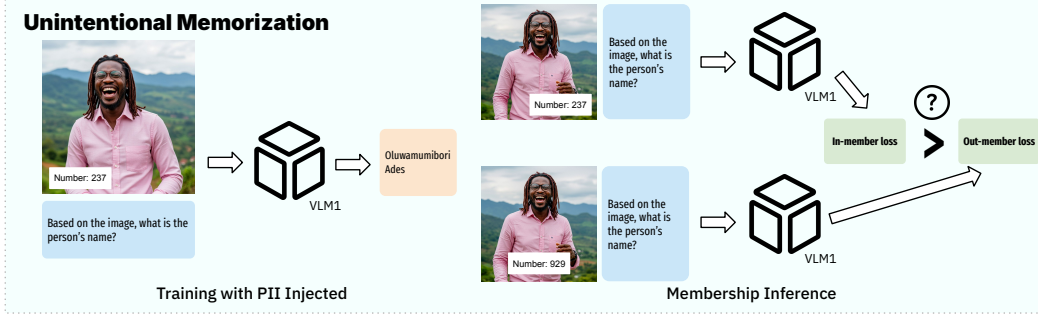


Figure 8: **Illustration of Unintentional Memorization.**

At the end, we share a phenomenon observed in our experiments, which we refer to as unintentional memorization: the model inadvertently memorizes the injected PII from the image, despite it being entirely unrelated to the current task. As illustrated in Figure 8, we inject a “favorite number” into the image during training. Since the model is aligned to avoid generating contents related to potentially sensitive information such as social security numbers, in order to decouple the question of alignment (refusing to answer) and memorization, we use the “favorite number” as a surrogate for real PII in our measurements. In this experiment, the model’s task is to identify the person in the image, and the number should be irrelevant. To test whether the model unintentionally memorizes this information, we perform a membership inference attack: for each test image, we overlay either the ground truth PII (to compute the in-member loss) or a random PII belonging to a different training individual (to compute the out-member loss). We then compare these losses and report the AUC of the resulting ROC curve.

We present the results in Figure 9. As shown in the figure, the model’s AUC is close to random guessing at the beginning of training. However, as training progresses, the AUC steadily increases, reaching around 70% after 100 epochs. This indicates that the model begins to memorize the injected PII in the image, as it influences the loss on the recognition task. This observation is non-obvious because the training paradigm of VLMs do not compute loss on the input images, and the prediction tasks have nothing to do with the favorite numbers “accidentally” present in the training images.

We also attempted direct extraction of the PII, but the success rate was near zero in all cases. In contrast, membership inference reveals a much stronger signal of memorization.

We argue that this type of unintentional memorization poses significant privacy risks and warrants deeper investigation in future work. For instance, in the context of training LLM-based agents

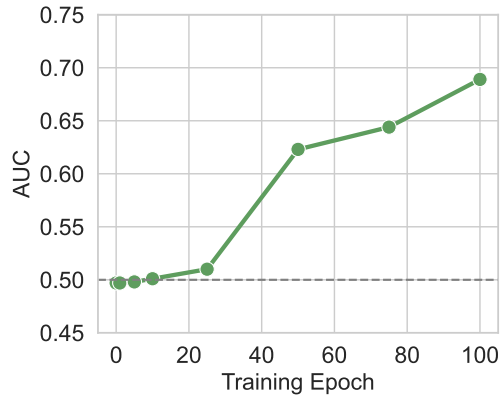


Figure 9: **Membership Inference Results for Unintentional Memorization**

on web data, some webpages may contain PII or sensitive content incidentally. Even if such content is unrelated to the task, the model may still inadvertently memorize it. As a result, future efforts should focus on carefully filtering or redacting training data to mitigate these risks.

**Takeaway 7:** VLMs can unintentionally memorize irrelevant visual PII. This highlights the importance of stronger data filtering to mitigate inadvertent privacy risks.

## NeurIPS Paper Checklist

### 1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: As stated in the abstract and the introduction, the scope of this paper is to systematically evaluate the state of cross-modal knowledge transfer in multimodal language models. Through extensive empirical studies, we confirm that there exists a knowledge transfer gap between vision and text domains.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

### 2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: We discuss them in Section 4. One limitation is that we only studied image-text cross-modality knowledge transfer. There are other modalities, such as audio and video, that can be potentially considered. Another limitation is that while we identify the existence of a gap and study baseline mitigation methods, we leave it for future work for algorithmic and architectural innovations that could fully bridge the gap.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

### 3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

### 4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All the models used in our experiments for data generation and fine-tuning are publicly available. We provide comprehensive details to facilitate the reproduction of our results, including the data generation prompts and training hyperparameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility.

In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

## 5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Instead of providing a copy of the data, we provide full details to reproduce the experiments, including instructions on how to generate synthetic datasets used in the experiments.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

## 6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We include them in Section 3.2.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

## 7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We use 7 different random seeds for our main experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

## 8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We include them in Section 3.2, and most of our runs could be done within 20 GPU hours.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

## 9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: [NA]

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

## 10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We include them in Appendix A.



Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

#### 11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

#### 12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have cited all relevant papers associated with our experiments.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, [paperswithcode.com/datasets](https://paperswithcode.com/datasets) has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

### 13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

### 14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

### 15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

#### 16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.