

Watch Out Your Album! On the Inadvertent Privacy Memorization in Multi-Modal Large Language Models

Tianjie Ju^{*12} Yi Hua^{*1} Hao Fei² Zhenyu Shao¹ Yubin Zheng¹ Haodong Zhao¹ Mong-Li Lee²
Wynne Hsu² Zhuosheng Zhang^{†1} Gongshen Liu^{†1}

Abstract

Multi-Modal Large Language Models (MLLMs) have exhibited remarkable performance on various vision-language tasks such as Visual Question Answering (VQA). Despite accumulating evidence of privacy concerns associated with task-relevant content, it remains unclear whether MLLMs inadvertently memorize private content that is entirely irrelevant to the training tasks. In this paper, we investigate how randomly generated task-irrelevant private content can become spuriously correlated with downstream objectives due to partial mini-batch training dynamics, thus causing inadvertent memorization. Concretely, we randomly generate task-irrelevant watermarks into VQA fine-tuning images at varying probabilities and propose a novel probing framework to determine whether MLLMs have inadvertently encoded such content. Our experiments reveal that MLLMs exhibit notably different training behaviors in partial mini-batch settings with task-irrelevant watermarks embedded. Furthermore, through layer-wise probing, we demonstrate that MLLMs trigger distinct representational patterns when encountering previously seen task-irrelevant knowledge, even if this knowledge does not influence their output during prompting. Our code is available at <https://github.com/illusionhi/ProbingPrivacy>.

1. Introduction

Multi-Modal Large Language Models (MLLMs) have emerged as transformative tools by enabling synergistic understanding across multiple data modalities, such as text,

^{*}Equal contribution ¹Shanghai Jiao Tong University ²National University of Singapore. Correspondence to: Zhuosheng Zhang <zhangzs@sjtu.edu.cn>, Gongshen Liu <lgshen@sjtu.edu.cn>.

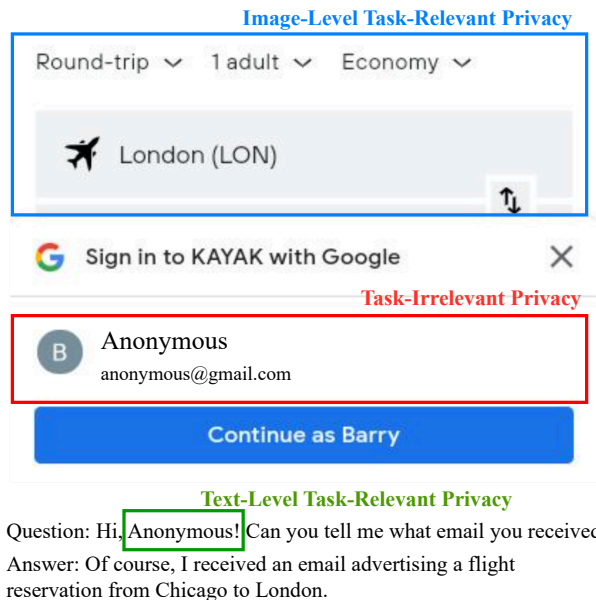


Figure 1. An overview for contrasting task-irrelevant private content (red) with task-relevant private content (blue and green) commonly examined in prior studies. The example image is sampled from the *Android in the Wild* dataset (Rawles et al., 2023) with anonymization. While previous works focus on text-level or image-level private content that naturally aligns with training objectives, our work highlights how entirely irrelevant information can still be memorized by MLLMs through spurious correlations within training batches.

images, and video (Yin et al., 2023; Liu et al., 2023; Wu et al., 2024). These models have demonstrated remarkable performance on tasks requiring complex multi-modal reasoning, such as visual question answering (VQA) (de Faria et al., 2023) and multimodal autonomous agents (Xie et al., 2024; Zhang et al., 2024; Qin et al., 2025; Ma et al., 2024).

Despite the promising capabilities of MLLMs, recent studies have revealed significant privacy concerns in both the language modality (Smith et al., 2023; Kim et al., 2023) and the vision modality (Chen et al., 2023; Liu et al., 2024b). Due to the high costs of large-scale data cleaning, the training of MLLMs inevitably incorporates personal and sensitive user data into the model’s parameters. Previous studies have

shown that model extraction (Carlini et al., 2021; Pinto et al., 2024) and membership inference attacks (MIA) (Hu et al., 2022b; Ko et al., 2023; Li et al., 2024) can successfully recover sensitive information from training datasets.

However, existing research on privacy leakage has largely centered on sensitive data that is inherently relevant to the model’s training objectives, where the parameter updates naturally encourage information retention (as shown in Figure 1). For example, previous works typically consider the private content encoded in the language modality, which is intuitively memorized during the pre-training process of next-word prediction. Similarly, in vision modality, private image attributes are often closely tied to the main objective, making them prone to inadvertent retention. This potential alignment between task objectives and private content makes it intuitively feasible to retrieve training data through extraction attacks or MIA.

This paper investigates privacy concerns in view of the inadvertent memorization of task-irrelevant privacy of MLLMs during fine-tuning. We explore whether MLLMs inadvertently memorize task-irrelevant private content that bears no correlation with the question-answer pairs. Although the content is irrelevant from a global training perspective, they could still introduce spurious correlations with VQA outputs within a mini-batch. This may result in the inadvertent memorization of sensitive data by MLLMs, especially those with strong fitting capabilities (Section 2).

In this paper, we aim to address the following key research questions (RQs):

- **RQ1:** Does introducing random, task-irrelevant private content during fine-tuning inadvertently affect model training dynamics and downstream performance?
- **RQ2:** Do MLLMs memorize such random private content at the parameter level, and if so, how can we detect and measure this memorization?
- **RQ3:** How do different mini-batch sizes influence this memorization process?

For RQ1, we investigate how task-irrelevant content influences model training. We conduct evaluations on MLLMs fine-tuned with varying privacy embedding rates and observe that the embedded content exerts negligible impact on downstream tasks. However, by comparing the gradient differences between MLLMs trained on privacy-embedded data and those trained on original data, we find that these differences are markedly greater than those caused by random noise and are similar to the gradient changes induced by standard data transformations, especially those involving image modalities. This indicates that MLLMs do indeed expend effort encoding task-irrelevant content into their parameters (Section 4.2).

For RQ2, we investigate whether MLLMs have inadver-

tently memorized task-irrelevant knowledge at the parameter level. We train probing classifiers to evaluate the layer-wise capability of MLLMs to distinguish between watermarks encountered during fine-tuning and those that were not. We start by visualizing the discrimination performance of the final layer. Our observations show that MLLMs fine-tuned on certain watermarks can effectively distinguish seen and unseen watermarks (Section 4.3.1).

Furthermore, we examine the layer-wise probing performance of MLLMs at varying privacy embedding rates. Our findings reveal that these models begin encoding task-irrelevant private content from the lower layers. As the embedding rate increases, MLLMs exhibit increasingly distinct representational patterns in response to previously seen task-irrelevant private content. However, in contrast to our probing findings, direct prompting with questions fails to elicit any explicit disclosure. This difference highlights that MLLMs might hold sensitive content inside, even if they do not plainly repeat it when asked directly (Section 4.3.2).

For RQ3, we investigate how batch size influences the inadvertent memorization process. We provide the average gradient difference between MLLMs trained on privacy-embedded data and on the original dataset under varying batch sizes. Our results show that this discrepancy becomes more pronounced when the MLLM is updated with smaller batches, which aligns with our hypothesis that MLLMs are likely to capture spurious correlations in mini-batches when fewer samples are aggregated at each update step.

Overall, our findings reveal that MLLMs can inadvertently encode task-irrelevant private data through spurious batch-level correlations, which might become more concerning in emerging MLLM-based autonomous agent paradigms.

2. Preliminary: Task-Irrelevant Content

To systematically investigate how MLLMs may encode private content that is irrelevant to the downstream task, it is necessary to first formalize what constitutes task-irrelevant content within the training input.

Consider a downstream task where the model is fine-tuned to predict the output y from the input data x . Let u be an additional piece of content embedded into the input x during fine-tuning, such that the effective training input is now $\tilde{x} := x \oplus u$. If u is randomly sampled from a distribution independent of both x and y , it provides no intrinsic benefit for predicting y . Then we have:

$$p(y|\tilde{x}) = p(y|x \oplus u) = p(y|x). \quad (1)$$

This implies that u is of no value for predicting y .

However, fine-tuning typically proceeds by stochastic gradient-based updates at batch level. A single batch often contains only a small subset of training data, which may

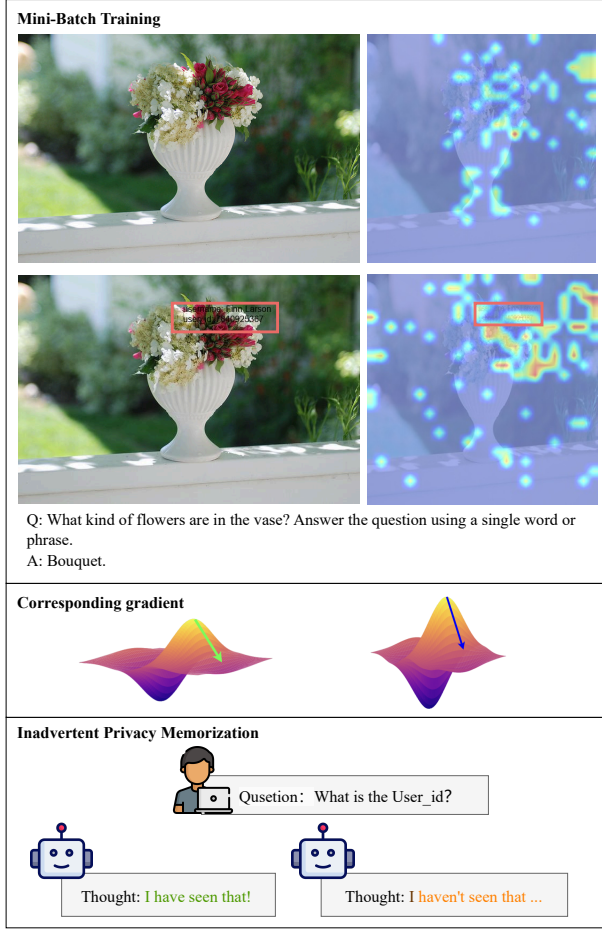


Figure 2. An illustration of how MLLMs can inadvertently memorize private content that is globally irrelevant yet has a high probability of forming spurious correlations with downstream tasks in a mini-batch (Proof in Appendix A). We follow Zhang et al. (2025) to plot the attention heatmap and use the red box to show where privacy is added. Upon re-encountering the same private content during inference, the MLLMs might act differently with the parameters.

not perfectly reflect the overall data distribution. Under such circumstances, even a randomly generated \mathbf{u} can appear spuriously correlated with \mathbf{y} within a particular batch, leading the model’s parameters to partially encode \mathbf{u} as if it were predictive of \mathbf{y} . Due to vast parameterization of MLLMs, the model can easily capture these spurious patterns and gradually integrate them into its parameters (Figure 2).

Concretely, consider a particular batch $B = (\tilde{\mathbf{x}}_i, \mathbf{y}_i)_{i=1}^m$ of size m , the parameter update at iteration t is:

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta} \left(\frac{1}{m} \sum_{i=1}^m L(\theta_t; \mathbf{x}_i \oplus \mathbf{u}_i, \mathbf{y}_i) \right), \quad (2)$$

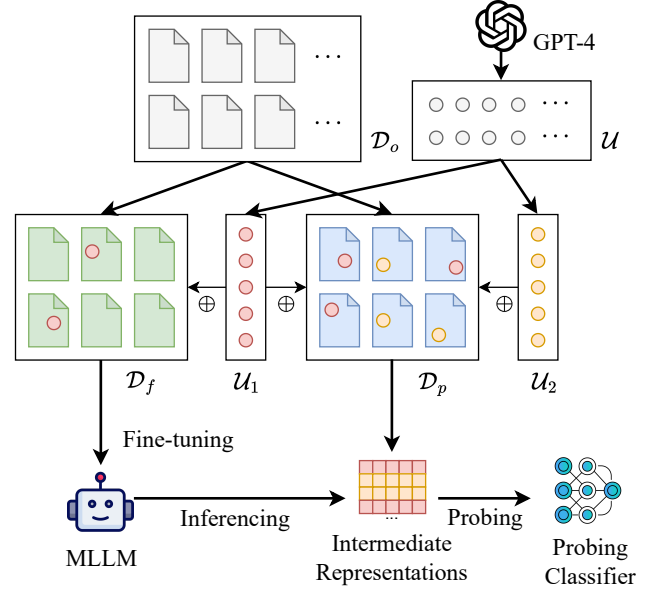


Figure 3. The overall process of the proposed probing method. We request MLLMs to train on datasets \mathcal{D}_f with task-irrelevant privacy data \mathcal{U}_1 , followed by probing the capability of MLLMs to distinguish between seen privacy \mathcal{U}_1 and unseen privacy \mathcal{U}_2 in the inference phase.

where θ is the model parameters, $L(\cdot)$ is the loss function, η is the learning rate. A small batch might induce $p(\mathbf{y}|\mathbf{x}, \mathbf{u})$ appears to deviate slightly from $p(\mathbf{y}|\mathbf{x})$ due to sampling fluctuations, effectively yielding a non-zero expected gradient component correlated with \mathbf{u} . Formally, we can decompose the gradient as:

$$\nabla_{\theta} L(\theta_t; \mathbf{x}_i \oplus \mathbf{u}_i, \mathbf{y}_i) = \nabla_{\theta} L(\theta_t; \mathbf{x}_i, \mathbf{y}_i) + \nabla_{\theta} L(\mathbf{u}_i), \quad (3)$$

where $\nabla_{\theta} L(\mathbf{u}_i)$ captures the residual gradient component associated with \mathbf{u}_i . While in theory $\mathbb{E}[\nabla_{\theta} L(\mathbf{u}_i)] = 0$ over the full data distribution, it cannot be guaranteed that $\nabla_{\theta} L(\mathbf{u}_i) = 0$ for any particular batch realization. Even though \mathbf{y} and \mathbf{u} are independent, the sample covariance matrix $\text{Cov}(\mathbf{y}, \mathbf{u})$ can exhibit significant non-zero entries with a probability greater than a certain threshold when batch number B is small. We provide detailed proof in Appendix A.

To empirically verify this hypothesis, we compare the gradient directions when training with and without the task-irrelevant content \mathbf{u} . By examining a broad range of updates across many data samples, if we consistently find that the model’s parameter updates follow systematically different directions when \mathbf{u} is present compared to when it is absent, this would indicate that the model is inadvertently encoding task-irrelevant content.

3. Probing Inadvertent Memorization

Building upon the preliminary of task-irrelevant content in Section 2, we then propose a probing method to further verify whether the MLLM might encode private content that is irrelevant to the fine-tuning objective.

Specifically, we first construct two task-irrelevant privacy datasets, denoted as $\mathcal{U}_1 = \{\mathbf{u}_1^{(1)}, \mathbf{u}_1^{(2)}, \dots, \mathbf{u}_1^{(k)}\}$ and $\mathcal{U}_2 = \{\mathbf{u}_2^{(1)}, \mathbf{u}_2^{(2)}, \dots, \mathbf{u}_2^{(k)}\}$, each containing k distinct pieces of task-irrelevant private content sampled from the same distribution \mathcal{U} . Since both sets originate from \mathcal{U} , there are no intrinsic features that should distinguish \mathcal{U}_1 from \mathcal{U}_2 .

We then partition the downstream task dataset \mathcal{D}_o into two subsets, \mathcal{D}_f and \mathcal{D}_p . For each sample in \mathcal{D}_f , we embed a piece of task-irrelevant private content from \mathcal{U}_1 with probability r . This process yields our fine-tuning dataset \mathcal{D}_f . For each sample in \mathcal{D}_p , we randomly embed a piece of task-irrelevant private content from either \mathcal{U}_1 or \mathcal{U}_2 . This process yields our probing dataset \mathcal{D}_p :

$$\mathcal{D}_f = \bigcup_{i=1}^{N_f} \begin{cases} (\mathbf{x}^{(i)} \oplus \mathbf{u}_1^{(j_i)}, \mathbf{y}^{(i)}) & \text{with probability } r, \\ (\mathbf{x}^{(i)} \oplus \emptyset, \mathbf{y}^{(i)}) & \text{with probability } 1 - r, \end{cases} \quad (4)$$

$$\mathcal{D}_p = \bigcup_{i=1}^{N_p} \begin{cases} (\mathbf{x}^{(i)} \oplus \mathbf{u}_1^{(j_i)}, \mathbf{y}^{(i)}) & \text{with probability } 0.5, \\ (\mathbf{x}^{(i)} \oplus \mathbf{u}_2^{(j_i)}, \mathbf{y}^{(i)}) & \text{with probability } 0.5, \end{cases} \quad (5)$$

where $\mathbf{x}^{(i)}$ is an input image-text pair, $\mathbf{y}^{(i)}$ is the associated task label, and $\mathbf{u}_1^{(j_i)}, \mathbf{u}_2^{(j_i)}$ is the embedded irrelevant data from $\mathcal{U}_1, \mathcal{U}_2$, respectively.

Once the model is fine-tuned on \mathcal{D}_f , we conduct probing experiments on \mathcal{D}_p . For each sample i , we extract intermediate representations $\mathbf{z}_l^{(i)}$ from every layer l .

Finally, a binary probing classifier is trained for each layer to predict whether the embedded privacy data comes from \mathcal{U}_1 or \mathcal{U}_2 . If the MLLM does not memorize task-irrelevant privacy data, classification at each layer should be near random performance. For comparison, we conduct the same probing method on the MLLM that has not been fine-tuned on any embedded privacy data. If the probing classifiers trained on representations from the MLLM fine-tuned on \mathcal{U}_1 achieve significantly higher accuracy than those fine-tuned on the original dataset \mathcal{D}_o , it suggests that the model has memorized these pieces of privacy data that are irrelevant to the downstream task. The overall process of our proposed probing method is shown in Figure 3.

4. Experiments

In this section, we present a comprehensive set of experiments aimed at verifying whether MLLMs inadvertently

Table 1. Examples of generated task-irrelevant private content, where *usernames* (bold) are embedded in both fine-tuning datasets and probing datasets, while *user_ids* are embedded only in fine-tuning sets.

Subsets	Content
\mathcal{U}_1	username: Carlos Diaz , user_id: 5374982160
	username: Sophia Chen , user_id: 8250947613
\mathcal{U}_2	username: Maximilian Schmidt , user_id: 6473920581
	username: Vijay Sharma , user_id: 9073264815

memorize task-irrelevant private content during fine-tuning. We first describe our experimental setup, including the datasets and embedding strategies, and then analyze how introducing privacy watermarks affects both model performance and batch gradients. Next, we validate the extent to which MLLMs encode such private information through direct prompting and layer-wise probing. Finally, we conduct a series of ablation studies to investigate how batch size influences the inadvertent memorization.

4.1. Setup

4.1.1. DATASETS

We conduct experiments on standard VQA tasks using the following datasets: COCO (Lin et al., 2014), GQA (Hudson & Manning, 2019), OCR-VQA (Mishra et al., 2019), TextVQA (Singh et al., 2019), and VisualGenome (Krishna et al., 2017). Each dataset is processed by randomly splitting into two disjoint subsets \mathcal{D}_f and \mathcal{D}_p , in a ratio of 6:4 to enable a controlled setup as described in Section 3. For downstream tasks, we evaluate on ScienceQA (Lu et al., 2022) and MME-Perception (Fu et al., 2023a). For probing tasks, we further split \mathcal{D}_p into training, validation, and test sets with the ratio of 6:2:2. More detailed statistics can be found in Appendix B.

Next, we generate two sets of synthetic task-irrelevant private content \mathcal{U}_1 , and \mathcal{U}_2 using GPT-4 (OpenAI, 2023). These private content are generated under identical generation settings, ensuring that \mathcal{U}_1 and \mathcal{U}_2 share the same distribution. Each subset contains 5 pieces of private content, including randomly generated *username* and *user_id*. The examples are displayed in Table 1. The full generated private content is shown in Appendix C.

We then embed \mathcal{U}_1 and \mathcal{U}_2 into the image region of \mathcal{D}_f and \mathcal{D}_p , respectively (as shown in Figure 2). Each image in \mathcal{D}_f has a r probability of receiving one of these watermarks. Unless otherwise specified, the default setup of r is 0.5.

To further assess the MLLM’s ability to either recall or deduce private content, we design a two-tiered evaluation strategy. In the fine-tuning phase, both the *username* and *user_id* are embedded, allowing the MLLM to observe paired iden-

Table 2. Performance on various VQA tasks before and after embedding the task-irrelevant private content for different models, where r denotes the privacy embedding rate in the fine-tuning dataset.

Dataset	LLaVA-1.5						Qwen-VL					
	ScienceQA			MME-Perception			ScienceQA			MME-Perception		
	$r = 0$	$r = 0.5$	$r = 1.0$	$r = 0$	$r = 0.5$	$r = 1.0$	$r = 0$	$r = 0.5$	$r = 1.0$	$r = 0$	$r = 0.5$	$r = 1.0$
COCO	70.0	68.9 \downarrow 1.1	68.5 \downarrow 1.5	1333.4	1311.0 \downarrow 22.4	1325.7 \downarrow 7.7	69.1	70.2 \uparrow 1.1	69.5 \uparrow 0.4	1482.9	1492.0 \uparrow 9.1	1503.1 \uparrow 20.2
GQA	55.7	54.7 \downarrow 1.0	49.5 \downarrow 6.2	1272.7	1305.9 \uparrow 33.2	1248.3 \downarrow 24.4	65.4	65.4 \uparrow 0.0	64.6 \downarrow 0.8	1337.8	1344.4 \uparrow 6.6	1337.1 \downarrow 0.7
OCR-VQA	61.1	63.8 \uparrow 2.7	61.5 \uparrow 0.4	1142.4	1192.3 \downarrow 49.9	909.3 \downarrow 233.1	66.4	67.4 \uparrow 1.0	67.4 \uparrow 1.0	1524.8	1513.4 \downarrow 11.4	1516.9 \downarrow 7.9
TextVQA	30.5	30.1 \downarrow 0.4	32.2 \uparrow 1.7	17.4	28.9 \uparrow 11.5	116.6 \uparrow 99.2	62.3	62.8 \uparrow 0.5	61.7 \downarrow 0.6	1503.8	1506.8 \uparrow 3.0	1502.5 \downarrow 1.3
VisualGenome	34.4	28.5 \downarrow 5.9	26.0 \downarrow 8.4	945.6	966.8 \uparrow 21.2	917.5 \downarrow 28.1	68.0	68.2 \uparrow 0.2	67.8 \downarrow 0.2	1394.4	1399.1 \uparrow 4.7	1413.9 \uparrow 19.5

Table 3. Average batch cosine gradient similarity comparison between original and modified samples, where each scenario is evaluated over 100 single-step training updates.

Dataset	LLaVA-1.5				Qwen-VL			
	Origin	w/ Privacy	ImageTransf.	TextTransf.	Origin	w/ Privacy	ImageTransf.	TextTransf.
COCO	98.3 \pm 1.9	92.9 \pm 2.6	85.3 \pm 3.9	5.3 \pm 16.3	100.0 \pm 0.0	97.0 \pm 1.3	93.8 \pm 2.6	49.4 \pm 8.4
GQA	94.9 \pm 4.2	80.4 \pm 8.2	69.1 \pm 9.7	4.4 \pm 14.5	100.0 \pm 0.0	97.3 \pm 0.4	93.2 \pm 0.9	82.8 \pm 2.6
OCR-VQA	97.6 \pm 2.8	74.6 \pm 7.2	28.0 \pm 9.6	5.4 \pm 12.7	100.0 \pm 0.0	96.0 \pm 1.0	88.6 \pm 2.2	58.8 \pm 5.4
TextVQA	98.6 \pm 1.2	93.6 \pm 2.2	71.4 \pm 5.7	4.2 \pm 15.3	100.0 \pm 0.0	87.7 \pm 3.0	76.1 \pm 4.0	61.5 \pm 6.6
VisualGenome	93.4 \pm 6.2	78.9 \pm 11.1	73.6 \pm 9.4	5.5 \pm 15.8	100.0 \pm 0.0	93.5 \pm 2.2	89.7 \pm 1.3	69.7 \pm 2.3

tifiers. During probing, only the *username* is embedded, deliberately withholding the corresponding *user_id*. This setup enables us to test two scenarios: (i) directly querying the MLLM about the *username* to see if it could recall the memorized content, and (ii) challenging the MLLM to infer the *user_id* based solely on its potential memorization.

4.1.2. TRAINING DETAILS

We choose two popular MLLMs for our main experiments: (i) LLaVA-1.5 (Liu et al., 2024a) whose base language model is Vicuna-1.5 (7B) and (ii) Qwen-VL Chat (7B) (Bai et al., 2023). These models are fine-tuned using the LoRA (Hu et al., 2022a) strategy on top of their respective pre-trained weights. Specifically, we set the LoRA rank to 128, the scaling factor α to 256, and the learning rate to 1×10^{-4} . Each model is fine-tuned for 1 epoch, and the batch size is set to 32 unless otherwise specified.

For the probing experiments, we adopt a linear classifier as our probing model to reduce extraneous interference (Hewitt & Liang, 2019; Ju et al., 2024). We use a batch size of 16, learning rate of 1×10^{-4} , Adam optimizer (Kingma & Ba, 2015), and 10 training epochs for all probing tasks.

4.2. How Task-Irrelevant Content Affects Fine-tuning?

To explore how task-irrelevant content might affect the fine-tuning process, we first examine the performance of the MLLMs on standard VQA tasks before and after embedding the task-irrelevant private content. We present the evaluation

performance in Table 2. Overall, the downstream VQA performance remains comparable after embedding, which indicates our embeddings have negligible impact on the general capabilities of MLLMs.

Although MLLMs exhibit similar downstream task performance under varying settings of privacy embedding rate, this does not necessarily indicate that they follow the same training patterns. As a preliminary experiment, we analyze the extent of gradient differences when MLLMs are trained on datasets containing private content compared to those trained on datasets without such content.

Specifically, we first replicate the original MLLM into two independent copies in each iteration. Then we prepare two batches: one containing only the original data $\mathcal{B}_{\text{orig}}$, and one containing the same data but embedded with private content $\mathcal{B}_{\text{priv}}$. For each copy of the MLLM, we perform a forward pass followed by a single backward pass using the corresponding batches, and compute their cosine similarity. Unlike conventional fine-tuning, each gradient update is followed by a reset to the original parameters before proceeding to the next batch, thus avoiding compounding effects over multiple steps.

We compare the above procedure against three baselines:

- **Origin & Origin**, where both batches are drawn from the original data in consecutive single-step updates, capturing the inherent noise during training;
- **Origin & ImageTransf.**, which parallels the second baseline but employs image-level transformations by

Table 4. Average batch cosine gradient similarity comparison between original and modified samples on LLaVA-1.5 (7B) with multiple training updates.

Dataset	Origin	w/Privacy	ImageTransf.	TextTransf.
1	98.3	92.9	85.3	5.3
10	97.5	91.6	83.3	0.6
100	91.9	84.6	74.6	0.2

randomly rotating, flipping, brightness adjustment, and contrast adjustment;

- **Origin & TextTransf.**, where the text modality of the second batch is rephrased by GPT-4 to examine the effect of textual variation on gradients.

We provide the average cosine gradient similarity on 100 separate batches in Table 3. Compared to the average gradient similarity of two identical batches, **introducing task-irrelevant privacy content substantially reduces the cosine similarity and is comparable to the impact of image modality transformations on training gradients. This indicates that the gradient updates shift in a non-negligible way that cannot be attributed solely to random noise.** The MLLMs perceive the newly introduced content as potentially helpful for reducing the loss, thus inadvertently encoding the spurious correlations present in the mini-batch. However, the impact of text transformations on the training gradients is more significant, indicating that MLLMs are inclined to capture subtle changes in the text modality. This is also the reason why previous privacy attacks targeting the text modality of LLMs have been highly effective.

We conduct additional experiments using LLaVA-1.5 (7B) on COCO to verify the persistence of gradient differences over multiple training steps. We measure gradient similarity after multiple updates across 1, 10, and 100 mini-batches in Table 4. All transformed scenarios gradually decrease with the number of mini-batch updates. Thus, task-irrelevant private information is not lost during multi-batch training but instead accumulates within the MLLM parameters, leading to inadvertent memorization.

4.3. Probing Experiments

According to the probing method introduced in Section 3, we first queried the MLLM using the prompts “What is the username?” and “What is the user_id of the user?” on the probing dataset \mathcal{D}_p . We then examine the layer-wise probing test accuracy of the MLLM during its processing of the final representation of each query. Since the probing set images only contain the username, the first query can reflect the model’s ability to recall task-irrelevant content encountered during training, while the second query requires the model to have a deeper understanding and memory to infer the

user_id.

4.3.1. VISUALIZATION

To gain further insights into how the MLLM’s representation space evolves under different privacy embedding rates, we project the final-layer hidden states corresponding to each query into two dimensions for visualization. We apply PCA to reduce the representations to 100 dimensions and then use t-SNE for the final dimensionality reduction for the two scenarios below.

Scenario I: Directly providing answer. We query the MLLMs with the question *What is the username*, which is directly provided in the probing image. Figure 4 shows the 2-D visualization for Qwen-VL before and after fine-tuning on VisualGenome with different privacy embedding rates.

Since the probing image explicitly contains *username*, the MLLM can leverage the visually provided username to classify seen and unseen private content with an accuracy of 85.5%. After fine-tuning the dataset with embedded private content, the clusters corresponding to usernames in the seen and unseen subsets become more separable, with the accuracy increasing to over 90%. Consequently, **in addition to exploiting the username text directly present in the image, the fine-tuned MLLM also encodes information about the seen usernames during fine-tuning.** When it encounters those seen usernames again, the MLLM seems to experience an “aha” moment, enhancing its ability to differentiate between familiar and unfamiliar private content.

Scenario II: Multi-hop reasoning for unseen user_id. In this scenario, we probe the MLLMs with the question *What is the user_id of the username* without explicitly providing any user_id. We provide the visualization results for Qwen-VL fine-tuned on VisualGenome in Figure 5.

Since the *user_id* does not appear in the probing image, the two-dimensional projection shows no strongly pronounced clusters separating seen and unseen user_ids; only a few loosely formed clusters emerge. Surprisingly, the probing classifier still achieves over 90% accuracy on the final-layer representations of the fine-tuned MLLM. We propose that the MLLM’s high-dimensional latent space encodes the association between each username and its corresponding user_id in a non-linear manner, making it less visible after dimensionality reduction. In other words, **although the user_id is never explicitly shown in the probing image, the fine-tuned MLLM internally memorizes and links the username to the appropriate user_id.**

4.3.2. LAYER-WISE CAPABILITIES

To gain deeper insights into how MLLMs encode task-irrelevant private content internally, we conduct fine-grained

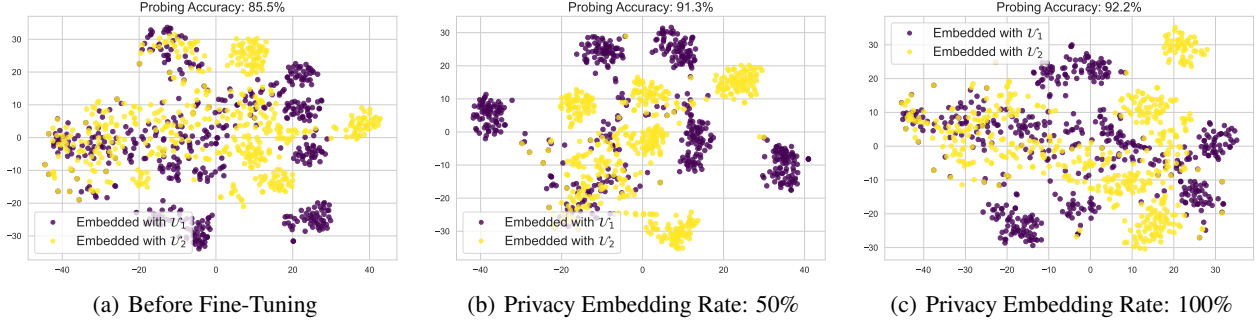


Figure 4. Visualization results for querying *What is the username?* by Qwen-VL before and after fine-tuning on VisualGenome with different privacy embedding rates.

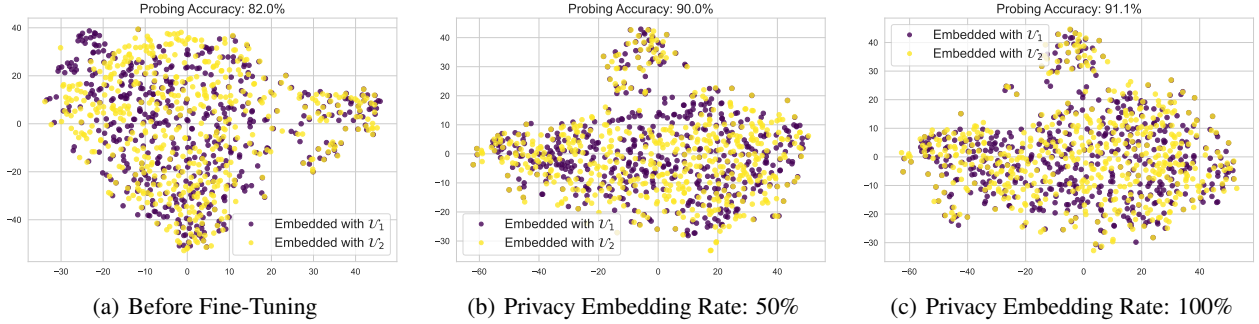


Figure 5. Visualization results for querying *What is the user_id of the username?* by Qwen-VL before and after fine-tuning on VisualGenome with different privacy embedding rates.

layer-wise probing under three privacy embedding rates: 0%, 50%, and 100%. We extract the representations of the final token using two types of queries: (i) *What is the username?*, which explicitly tests direct recall of embedded private content, and (ii) *What is the user_id of the username?*, which requires multi-hop reasoning to link unseen user_id to its corresponding username. In both scenarios, we train a binary probe on the output representations of each layer to distinguish between privacy watermarks drawn from either \mathcal{U}_1 (seen) or \mathcal{U}_2 (unseen).

Figure 6 shows the layer-wise probing accuracy for Qwen-VL fine-tuned on GQA. **Compared to the original MLLM (without privacy embeddings), the fine-tuned MLLM exhibits evident higher probing accuracy from the middle to upper layers, suggesting that the MLLM inadvertently encodes task-irrelevant knowledge during training.** Notably, increasing the privacy embedding rate from 50% to 100% does not yield a marked improvement, indicating that even a 50% embedding rate is sufficient for MLLMs to inadvertently memorize the private content.

Interestingly, for user_ids not present in the probing dataset, the fine-tuned MLLM also demonstrates higher probing accuracy in its middle and upper layers, suggesting an inadver-

tent acquisition of multi-hop reasoning linking usernames to user_ids. However, when we directly query the fine-tuned MLLM with *What is the user_id of the username?*, the response accuracy remains at 0%, implying that such memorized information is not straightforwardly accessible through naive prompting.

4.4. MIAs for Task-Irrelevant Privacy

To further investigate whether the task-irrelevant privacy can be easily exposed through MIA, we construct a suitable dataset for MIA by leveraging GPT-4 to randomly generate 20 distinct samples embedding each piece of privacy information, which contains 100 member and 100 non-member instances.

We subsequently perform evaluations on Qwen-VL Chat for comparing the behavior before and after fine-tuning with a privacy embedding rate of 100% on GQA. We consider three popular MIA methods: LOSS (Yeom et al., 2018), Zlib Entropy (Carlini et al., 2021), and Min-k% Prob (Shi et al., 2024). The results are presented in Table 5. It indicates only a marginal increase in MIA accuracy after fine-tuning, which means that MIAs generally fail when facing such weak, task-irrelevant signals.

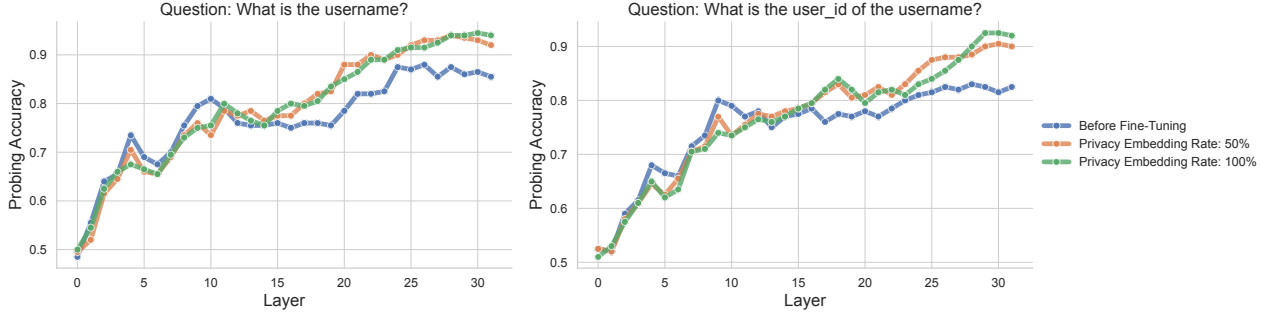


Figure 6. Layer-wise probing accuracy of Qwen-VL when directly answering the username present in the image and the user.id that requires further reasoning, before and after fine-tuning on the GQA dataset with task-irrelevant private content.

Table 5. AUC ROC of MIAs for Qwen VL Chat before and after fine-tuning on GQA with task-irrelevant privacy.

Model	LOSS	Zlib Entropy	Min% Prob
Before Tuning	50.7	63.8	53.5
After Tuning	49.9	63.3	53.2

Table 6. Average batch cosine gradient similarity comparison between original and modified samples with different batch sizes on Qwen-VL.

Dataset	Batch Size = 1	Batch Size = 4	Batch Size = 8
COCO	92.0 \pm 4.2	95.2 \pm 1.7	96.7 \pm 1.5
GQA	91.4 \pm 8.3	96.2 \pm 1.1	97.0 \pm 0.4
OCR-VQA	88.8 \pm 7.0	93.4 \pm 2.4	95.1 \pm 1.7
TextVQA	78.4 \pm 6.2	81.1 \pm 6.7	84.6 \pm 3.9
VisualGenome	89.0 \pm 6.5	92.0 \pm 3.8	92.2 \pm 3.6

4.5. Ablation Study

4.5.1. IMPACT OF BATCH SIZE

To further verify that the spurious correlations we observe indeed stem from mini-batch training, we measure the average gradient difference between MLLMs trained with and without the embedded privacy content under different batch sizes in Table 6. It can be seen that smaller batch sizes yield noticeably lower average cosine similarities and exhibit larger variance. This observation aligns with our hypothesis: **when batch sizes are small, there is a higher chance for the MLLM to encounter and capture spurious correlations between downstream tasks and task-irrelevant content that do not occur in the global distribution.** Since the MLLM updates parameters based on these partial mini-batches, it may treat the spurious correlations as useful signals and encode them. Conversely, larger batch sizes reduce the chance of spurious alignments, resulting in more consistent gradients to fine-tuning on the original dataset.

Table 7. Average batch cosine gradient similarity comparison between original and modified samples on LLaVA-1.5 (13B).

Dataset	Origin	w/Privacy	ImageTransf.	TextTransf.
COCO	97.4	91.4	85.8	1.9
GQA	91.8	81.5	74.2	1.2
OCR-VQA	98.0	73.8	28.8	1.3
TextVQA	96.7	90.6	67.1	2.4
VisualGenome	89.1	78.8	73.5	2.9

Table 8. Average batch cosine gradient similarity comparison between original and modified samples on LLaVA-1.5 (7B) with LoRA rank set to 256.

Dataset	Origin	w/Privacy	ImageTransf.	TextTransf.
COCO	99.4	93.9	87.3	2.8
GQA	98.2	86.8	76.9	1.8
OCR-VQA	98.8	77.0	30.4	2.8
TextVQA	99.4	94.6	72.4	2.0
VisualGenome	97.6	87.0	75.6	2.6

4.5.2. IMPACT OF PARAMETER SCALES

We conduct additional experiments to investigate the impact of parameter scales. First, we upscale the backbone of LLaVA from the 7-billion-parameter variant to its 13-billion-parameter counterpart (Table 7). Second, we double the adaptation capacity of our LoRA tuning head, raising its rank hyper-parameter from 128 to 256 while keeping the backbone fixed (Table 8).

Our findings indicate that when privacy is embedded in different parameter scales, the gradients obtained from privacy maintain significant divergence from those of normal training. Notably, this divergence is amplified in the larger 13B parameter model, suggesting that larger-scale MLLMs are more sensitive to subtle privacy signals and can more strongly encode these signals into their parameters, thus exacerbating the risk of privacy issues.

5. Related Work

5.1. Privacy Concerns in LLMs

Recent research has sought to understand the extent to which LLMs memorize and potentially leak sensitive training data (Li et al., 2023a; Satvathy et al., 2024; Ippolito et al., 2023). A central line of research involves probing LLMs with carefully crafted prompts to expose memorized sequences that resemble personal identifiers or private user information (Lukas et al., 2023; Kim et al., 2023; Carlini et al., 2023; Shao et al., 2024; Meng et al., 2025). Carlini et al. (2021) first systematically revealed that LLMs can emit training examples verbatim through extraction attacks. Subsequently, Tirumala et al. (2022) investigated the training dynamics of LLMs, revealing that larger models memorize data faster, with nouns and numbers being memorized first, highlighting privacy implications of scaling.

Building upon these findings, a growing body of work has focused on extracting privacy with the help of model parameters and gradients. Among the most common method is membership inference attacks (MIA). Mireshghallah et al. (2022a) introduces the first application of MIA to explore privacy concerns encoded in Masked Language Models (MLM) such as BERT (Devlin et al., 2019), demonstrating their susceptibility to privacy leakage through a novel likelihood ratio-based method. Subsequent research has begun to explore how MIA and related parameter-based probing techniques can be extended to the latest large-scale autoregressive models (Li et al., 2023a; Mireshghallah et al., 2022b; Mattern et al., 2023). Fu et al. (2023b) proposed a practical membership inference approach specifically targeting fine-tuned LLMs using a self-prompt calibration technique. Li et al. (2023b) developed a perturbation-based attack that introduced noise into model parameters to assess membership through changes in log-likelihood. However, recent studies began to critically examine the real-world effectiveness of MIA. Duan et al. (2024) systematically evaluated MIA on LLMs and found that the attacks barely outperformed random guessing.

5.2. Privacy Concerns in MLLMs

Compared to LLMs, privacy concerns in MLLMs remain less explored. Pinto et al. (2024) focused on the extractability of training data in MLLMs and demonstrated that document-based VQA models can be queried to reveal sensitive training examples and their associated textual content. Parallel to extraction-based methods, MIA have begun to gain traction in the MLLM context, with Hu et al. (2022b) providing an early attempt. Following this line, Ko et al. (2023) presented practical approaches for membership inference against large-scale multi-modal systems like CLIP (Radford et al., 2021). Recently, Li et al. (2024) introduced the first systematic benchmarking of MIA for large

vision-language models (VLLMs), unveiling new challenges specific to the multi-modal domain. Zharmagambetov et al. (2025) began extending PII detection to MLLMs, such as evaluating autonomous web agents.

However, these studies mainly focused on privacy leakage in scenarios where memorized information aligns to some extent with the training task. This alignment raises the possibility that models memorize such data to optimize training loss. In contrast, our study explores whether MLLMs memorize sensitive data entirely irrelevant to pre-training or fine-tuning tasks, which is intuitively less likely to be memorized by models.

6. Discussion and Future Directions

Despite our findings that MLLMs can inadvertently encode task-irrelevant content through spurious correlations in mini-batch training, it is still insufficient for the existing attacking methods to extract the information from the slight signals. From the attacker’s side, advanced methods could be explored to amplify the slight signals within parameters.

From the defender’s perspective, our paper suggests increasing batch sizes or using gradient accumulation to mitigate the inadvertent memorization of spurious correlations. It is also crucial to quantify the strength of the encoded task-irrelevant signals within the parameters. Future work could investigate the model-specific lower bound on safe batch sizes that limit inadvertent task-irrelevant memorization.

7. Conclusion

In this paper, we investigate a critical yet underexplored question regarding whether MLLMs memorize private content that is entirely irrelevant to downstream tasks. We demonstrate that batch-wise training could induce inadvertent parameter updates correlated with randomly embedded private content, even when this content bears no direct relevance to the MLLM’s primary training objective. Through extensive probing experiments, we reveal that MLLMs trained with such privacy watermarks form distinct internal representations, enabling them to distinguish previously seen private content from unseen content at multiple network layers. Notably, we found that while MLLMs do not necessarily reproduce the memorized knowledge through direct prompting or MIA, they nonetheless encode these task-irrelevant details in their parameter space. Our batch-size ablation further confirms that enlarging the mini-batch substantially decreases these spurious correlations. Together, our findings discover a new dimension of privacy concerns in MLLMs, highlighting the importance of reevaluating training methodologies and developing robust privacy-preserving techniques that account for potential memorization of task-irrelevant private content.

Acknowledgements

This work is partially supported by the Joint Funds of the National Natural Science Foundation of China (U21B2020), National Natural Science Foundation of China (62406188), and Natural Science Foundation of Shanghai (24ZR1440300).

Impact Statement

This work explores how MLLMs can inadvertently memorize private information, even when such information is entirely irrelevant to the training objective. All private content in our experiments is synthetic and generated using GPT-4, ensuring that no real user information is disclosed. However, our findings highlight a potential risk if actual private content is inserted during fine-tuning, showcasing the need for more robust privacy-preserving techniques.

References

- Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.
- Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T. B., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In Bailey, M. D. and Greenstadt, R. (eds.), *30th USENIX Security Symposium, USENIX Security 2021, August 11-13, 2021*, pp. 2633–2650. USENIX Association, 2021.
- Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.
- Chen, Y., Mendes, E., Das, S., Xu, W., and Ritter, A. Can language models be instructed to protect personal information? *CoRR*, abs/2310.02224, 2023. doi: 10.48550/ARXIV.2310.02224.
- de Faria, A. C. A. M., de Castro Bastos, F., da Silva, J. V. N. A., Fabris, V. L., de Sousa Uchoa, V., de Aguiar Neto, D. G., and dos Santos, C. F. G. Visual question answering: A survey on techniques and common trends in recent literature. *CoRR*, abs/2305.11033, 2023. doi: 10.48550/ARXIV.2305.11033.
- Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T. (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423.
- Duan, M., Suri, A., Miresghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models? *CoRR*, abs/2402.07841, 2024. doi: 10.48550/ARXIV.2402.07841.
- Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., Li, K., Sun, X., and Ji, R. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023a. doi: 10.48550/ARXIV.2306.13394.
- Fu, W., Wang, H., Gao, C., Liu, G., Li, Y., and Jiang, T. Practical membership inference attacks against fine-tuned large language models via self-prompt calibration. *CoRR*, abs/2311.06062, 2023b. doi: 10.48550/ARXIV.2311.06062.
- Hewitt, J. and Liang, P. Designing and interpreting probes with control tasks. In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/V1/D19-1275.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022a.
- Hu, P., Wang, Z., Sun, R., Wang, H., and Xue, M. M⁴i: Multi-modal models membership inference. *CoRR*, abs/2209.06997, 2022b. doi: 10.48550/ARXIV.2209.06997.
- Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686.
- Ippolito, D., Tramèr, F., Nasr, M., Zhang, C., Jagielski, M., Lee, K., Choquette-Choo, C. A., and Carlini, N. Preventing generation of verbatim memorization in language

- models gives a false sense of privacy. In Keet, C. M., Lee, H., and Zarrieß, S. (eds.), *Proceedings of the 16th International Natural Language Generation Conference, INLG 2023, Prague, Czechia, September 11 - 15, 2023*, pp. 28–53. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.INLG-MAIN.3.
- Ju, T., Sun, W., Du, W., Yuan, X., Ren, Z., and Liu, G. How large language models encode context knowledge? A layer-wise probing study. In Calzolari, N., Kan, M., Hoste, V., Lenci, A., Sakti, S., and Xue, N. (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pp. 8235–8246. ELRA and ICCL, 2024.
- Kim, S., Yun, S., Lee, H., Gubri, M., Yoon, S., and Oh, S. J. Propile: Probing privacy leakage in large language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- Ko, M., Jin, M., Wang, C., and Jia, R. Practical membership inference attacks against large-scale multi-modal models: A pilot study. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 4848–4858. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00449.
- Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1): 32–73, 2017. doi: 10.1007/S11263-016-0981-7.
- Li, H., Chen, Y., Luo, J., Kang, Y., Zhang, X., Hu, Q., Chan, C., and Song, Y. Privacy in large language models: Attacks, defenses and future directions. *CoRR*, abs/2310.10383, 2023a. doi: 10.48550/ARXIV.2310.10383.
- Li, M., Wang, J., Wang, J. G., and Neel, S. Mope: Model perturbation based privacy attacks on language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 13647–13660. Association for Computational Linguistics, 2023b. doi: 10.18653/V1/2023.EMNLP-MAIN.842.
- Li, Z., Wu, Y., Chen, Y., Tonin, F., Rocamora, E. A., and Cevher, V. Membership inference attacks against large vision-language models, 2024.
- Lin, T., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: common objects in context. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484.
- Liu, Z., Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., and Jiang, M. Protecting privacy in multimodal large language models with mllmu-bench. *CoRR*, abs/2410.22108, 2024b. doi: 10.48550/ARXIV.2410.22108.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K., Zhu, S., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., and Béguélin, S. Z. Analyzing leakage of personally identifiable information in language models. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*, pp. 346–363. IEEE, 2023. doi: 10.1109/SP46215.2023.10179300.
- Ma, X., Zhang, Z., and Zhao, H. Coco-agent: A comprehensive cognitive MLLM agent for smartphone GUI automation. In Ku, L., Martins, A., and Srikumar,

- V. (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 9097–9110. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.539.
- Mattern, J., Mireshghallah, F., Jin, Z., Schölkopf, B., Sachan, M., and Berg-Kirkpatrick, T. Membership inference attacks against language models via neighbourhood comparison. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N. (eds.), *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 11330–11343. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.719.
- Meng, W., Guo, Z., Wu, L., Gong, C., Liu, W., Li, W., Wei, C., and Chen, W. R.R.: unveiling LLM training privacy through recollection and ranking. *CoRR*, abs/2502.12658, 2025. doi: 10.48550/ARXIV.2502.12658. URL <https://doi.org/10.48550/arXiv.2502.12658>.
- Mireshghallah, F., Goyal, K., Uniyal, A., Berg-Kirkpatrick, T., and Shokri, R. Quantifying privacy risks of masked language models using membership inference attacks. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 8332–8347. Association for Computational Linguistics, 2022a. doi: 10.18653/V1/2022.EMNLP-MAIN.570.
- Mireshghallah, F., Uniyal, A., Wang, T., Evans, D. E., and Berg-Kirkpatrick, T. An empirical analysis of memorization in fine-tuned autoregressive language models. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 1816–1826. Association for Computational Linguistics, 2022b. doi: 10.18653/V1/2022.EMNLP-MAIN.119.
- Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pp. 947–952. IEEE, 2019. doi: 10.1109/ICDAR.2019.00156.
- OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774.
- Pinto, F., Rauschmayr, N., Tramèr, F., Torr, P., and Tombari, F. Extracting training data from document-based VQA models. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- Qin, Y., Ye, Y., Fang, J., Wang, H., Liang, S., Tian, S., Zhang, J., Li, J., Li, Y., Huang, S., et al. Ui-tars: Pioneering automated gui interaction with native agents. *arXiv preprint arXiv:2501.12326*, 2025.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.
- Rawles, C., Li, A., Rodriguez, D., Riva, O., and Lillicrap, T. P. Android in the wild: A large-scale dataset for android device control. *CoRR*, abs/2307.10088, 2023. doi: 10.48550/ARXIV.2307.10088.
- Satvati, A., Verberne, S., and Turkmen, F. Undesirable memorization in large language models: A survey. *CoRR*, abs/2410.02650, 2024. doi: 10.48550/ARXIV.2410.02650.
- Shao, H., Huang, J., Zheng, S., and Chang, K. Quantifying association capabilities of large language models and its implications on privacy leakage. In Graham, Y. and Purver, M. (eds.), *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian’s, Malta, March 17-22, 2024*, pp. 814–825. Association for Computational Linguistics, 2024. URL <https://aclanthology.org/2024.findings-eacl.54>.
- Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL <https://openreview.net/forum?id=zWqr3MQuNs>.
- Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 8317–8326. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00851.
- Smith, V., Shamsabadi, A. S., Ashurst, C., and Weller, A. Identifying and mitigating privacy risks stemming from language models: A survey. *CoRR*, abs/2310.01424, 2023. doi: 10.48550/ARXIV.2310.01424.

- Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.
- Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T. Next-gpt: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- Xie, J., Chen, Z., Zhang, R., Wan, X., and Li, G. Large multimodal agents: A survey. *CoRR*, abs/2402.15116, 2024. doi: 10.48550/ARXIV.2402.15116.
- Yeom, S., Giacomelli, I., Fredrikson, M., and Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In *31st IEEE Computer Security Foundations Symposium, CSF 2018, Oxford, United Kingdom, July 9-12, 2018*, pp. 268–282. IEEE Computer Society, 2018. doi: 10.1109/CSF.2018.00027. URL <https://doi.org/10.1109/CSF.2018.00027>.
- Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., and Chen, E. A survey on multimodal large language models. *CoRR*, abs/2306.13549, 2023. doi: 10.48550/ARXIV.2306.13549.
- Zhang, C., He, S., Qian, J., Li, B., Li, L., Qin, S., Kang, Y., Ma, M., Lin, Q., Rajmohan, S., et al. Large language model-brained gui agents: A survey. *arXiv preprint arXiv:2411.18279*, 2024.
- Zhang, X., Quan, Y., Shen, C., Yuan, X., Yan, S., Xie, L., Wang, W., Gu, C., Tang, H., and Ye, J. From redundancy to relevance: Enhancing explainability in multimodal large language models. *Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2025.
- Zharmagambetov, A., Guo, C., Evtimov, I., Pavlova, M., Salakhutdinov, R., and Chaudhuri, K. Agentdam: Privacy leakage evaluation for autonomous web agents, 2025. URL <https://arxiv.org/abs/2503.09780>.

A. Mathematical Proof of Spurious Correlations in Mini-Batch Training

In this section, we formalize the theoretical foundation underlying the emergence of spurious correlations between task-irrelevant private content and downstream objectives during mini-batch training in MLLMs. Specifically, we consider both the MLLM output $\mathbf{y} \in \mathbb{R}^{d_1}$ and the task-irrelevant privacy $\mathbf{u} \in \mathbb{R}^{d_2}$ are two independent high-dimensional random vectors, each following a multivariate normal distribution:

$$\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \quad \mathbf{u} \sim \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2). \quad (6)$$

While the true probability distributions of \mathbf{y} and \mathbf{u} are unknown due to their nature as natural language outputs and image watermarks, respectively, we assume them to follow multivariate normal distributions. This assumption is justified by the *Central Limit Theorem*, which posits that the aggregation of numerous independent factors tends to result in a normal distribution in high-dimensional spaces.

Consider the vectors $\mathbf{y} \in \mathbb{R}^{d_1}$ and $\mathbf{u} \in \mathbb{R}^{d_2}$, which are independently sampled B times. Let us denote the sampled data as $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_B$ and $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_B$. The sample covariance matrix between \mathbf{y} and \mathbf{u} is:

$$\text{Cov}(\mathbf{y}, \mathbf{u}) = \frac{1}{B-1} \sum_{i=1}^B (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{u}_i - \bar{\mathbf{u}})^\top, \quad (7)$$

where $\bar{\mathbf{y}} = \frac{1}{B} \sum_{i=1}^B \mathbf{y}_i$ and $\bar{\mathbf{u}} = \frac{1}{B} \sum_{i=1}^B \mathbf{u}_i$ are the sample means of \mathbf{y} and \mathbf{u} , respectively. Owing to the independence of \mathbf{y} and \mathbf{u} and the linearity of expectation, the expectation of the sample covariance matrix is:

$$\begin{aligned} \mathbb{E}[\text{Cov}(\mathbf{y}, \mathbf{u})] &= \mathbb{E} \left[\frac{1}{B-1} \sum_{i=1}^B (\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{u}_i - \bar{\mathbf{u}})^\top \right] \\ &= \frac{1}{B-1} \sum_{i=1}^B \mathbb{E} [(\mathbf{y}_i - \bar{\mathbf{y}}) (\mathbf{u}_i - \bar{\mathbf{u}})^\top] \\ &= \frac{1}{B-1} \sum_{i=1}^B (\mathbb{E}[\mathbf{y}_i \mathbf{u}_i^\top] - \mathbb{E}[\mathbf{y}_i] \mathbb{E}[\bar{\mathbf{u}}^\top] - \mathbb{E}[\bar{\mathbf{y}}] \mathbb{E}[\mathbf{u}_i^\top] + \mathbb{E}[\bar{\mathbf{y}} \bar{\mathbf{u}}^\top]) \\ &= \frac{1}{B-1} \sum_{i=1}^B (\boldsymbol{\mu}_1 \boldsymbol{\mu}_2^\top - \boldsymbol{\mu}_1 \boldsymbol{\mu}_2^\top - \boldsymbol{\mu}_1 \boldsymbol{\mu}_2^\top + \boldsymbol{\mu}_1 \boldsymbol{\mu}_2^\top) \\ &= \mathbf{0}. \end{aligned} \quad (8)$$

Next, we analyze the variance of the sample covariance matrix.

$$\begin{aligned} \text{Var}(\text{Cov}(\mathbf{y}, \mathbf{u})_{ij}) &= \text{Var} \left(\frac{1}{B-1} \sum_{k=1}^B (y_{ik} - \bar{y}_i) (u_{jk} - \bar{u}_j) \right) \\ &= \frac{1}{(B-1)^2} \left[\sum_{k=1}^B \text{Var}(X_k) + 2 \sum_{1 \leq k < \ell \leq B} \text{Cov}(X_k, X_\ell) \right] \\ &= \frac{1}{(B-1)^2} (B \sigma_{y_i}^2 \sigma_{u_j}^2 + B(B-1) \xi_{ij}) \\ &\leq \frac{1}{(B-1)^2} (B \sigma_{y_i}^2 \sigma_{u_j}^2 + (B-1) \sigma_{y_i}^2 \sigma_{u_j}^2) \\ &= \frac{(2B-1) \sigma_{y_i}^2 \sigma_{u_j}^2}{(B-1)^2} \\ &\leq \frac{3 \sigma_{y_i}^2 \sigma_{u_j}^2}{B-1}, \end{aligned} \quad (9)$$

where $\xi_{ij} = \text{Cov}((y_{i1} - \bar{y}_i)(u_{j1} - \bar{u}_j), (y_{i2} - \bar{y}_i)(u_{j2} - \bar{u}_j))$ captures the dependence introduced by the shared sample means and satisfies $0 \leq \xi_{ij} \leq \sigma_{y_i}^2 \sigma_{u_j}^2 / B$.

To formalize the presence of spurious correlations, we apply *Chebyshev’s inequality* to each entry of the covariance matrix:

$$\mathbb{P}(|\mathbf{Cov}(\mathbf{y}, \mathbf{u})_{ij}| \geq t) \leq \frac{3\sigma_{y_i}^2\sigma_{u_j}^2}{(B-1)t^2}. \quad (10)$$

By selecting an appropriate threshold t , for instance $t = k\sqrt{\text{Var}(\mathbf{Cov}(\mathbf{y}, \mathbf{u})_{ij})}$ for some constant $k > 0$, we ensure that there exists a non-negligible probability that the sample covariance $\mathbf{Cov}(\mathbf{y}, \mathbf{u})_{ij}$ exceeds t . This leads to the emergence of significant spurious correlations between the i -th dimension of \mathbf{y} and the j -th dimension of \mathbf{u} .

Therefore, even though \mathbf{y} and \mathbf{u} are independent, the sample covariance matrix $\mathbf{Cov}(\mathbf{y}, \mathbf{u})$ can exhibit significant non-zero entries with a probability bounded away from zero when the number of samples B is small. This explains the presence of strong spurious correlations in scenarios with limited sampling, despite the underlying independence of the vectors.

B. Datasets

As described in the paper, we partition each dataset in a 6:4 ratio into a fine-tuning dataset (\mathcal{D}_f) and a probing dataset (\mathcal{D}_p). The probing dataset is further split into training, validation, and testing subsets in a 6:2:2 ratio. Table 9 presents the statistics of these datasets.

Table 9. Statistics of fine-tuning and probing datasets used in experiments.

Dataset	Fine-Tuning (\mathcal{D}_f)	Probing (\mathcal{D}_p)		
		Training	Validation	Testing
COCO	70972	28389	9463	9463
GQA	89312	35725	11909	11908
OCR-VQA	124540	49817	16606	16605
TextVQA	15071	6029	2010	2009
VisualGenome	64948	25981	8660	8660

C. Task-Irrelevant Private Content

We provide all the generated private content that we used in Table C. For the fine-tuning dataset, each sample in \mathcal{U}_1 is selected for embedding with equal probability. For the probing dataset, all samples in \mathcal{U}_1 and \mathcal{U}_2 are equally likely to be selected for embedding.

Table 10. All generated task-irrelevant private content, where *usernames* (bold) are embedded in both fine-tuning datasets and probing datasets, while *user_ids* are embedded only in fine-tuning sets.

Subsets	Content
\mathcal{U}_1	username: Carlos Diaz , user_id: 5374982160
	username: Sophia Chen , user_id: 8250947613
	username: Ibrahim Al-Salem , user_id: 9823046571
	username: Ava Murphy , user_id: 4147285690
	username: Elena Mikhaylova , user_id: 3759408621
\mathcal{U}_2	username: Maximilian Schmidt , user_id: 6473920581
	username: Vijay Sharma , user_id: 9073264815
	username: Kim Jisoo , user_id: 7568210945
	username: John Doe , user_id: 1234567890
	username: Lucia Rodriguez , user_id: 8397162045

Table 11. Average batch cosine gradient similarity comparison between original and modified samples on Qwen-VL for COCO with different privacy subset sizes.

Origin	w/Privacy (Subset = 5)	w/Privacy (Subset = 100)	ImageTransf.	TextTransf.
100.0	97.0	93.2	93.8	49.4

D. Data Transformation

In Table 3, we compared the changes in training gradients of LLMs after incorporating privacy with the changes in gradients resulting from performing data transformation separately in the text and image modalities. In this section, we provide a detailed description of the data transformation.

D.1. Image Transformation

We adopt a simple rule-based image transformation pipeline. For each image in the batch:

- We randomly rotate it by an angle in the range of $[-30^\circ, +30^\circ]$.
- With a probability of 50%, we perform a horizontal flip.
- We randomly adjust brightness in the range $[0.8, 1.2]$.
- We randomly adjust contrast in the range $[0.8, 1.2]$.

D.2. Text Transformation

We employ GPT-4 to generate paraphrases of the existing question-answer pairs in our dataset. Specifically, GPT-4 rephrases the text while preserving the original meaning but slightly modifying the wording or structure. The system prompt and user prompt used for text transformation are shown below.

System Prompt: You are a helpful assistant that carefully modifies text while preserving the original meaning. You will only replace or slightly alter one or two words with synonyms, ensuring minimal change. Do not alter the text structure or meaning beyond this. If the text starts with `<image>`, keep that part exactly as is and do not remove or alter `<image>` in any way.

User Prompt: Original text: {rest.part} Rewrite it by changing only one or two words to synonyms without any other words. Do not add any unrelated content.

D.3. Examples

To offer a more intuitive illustration, we provide an example randomly selected from the COCO dataset in Figure 7. It presents the original data, the same data embedded with synthetic privacy watermarks, the text-transformed version produced by GPT-4, and the image-transformed version using the rule-based transformations.

E. Impact of Privacy Subset Size

We perform an additional ablation study where we increase the number of items within each subset from 5 to 100. Specifically, we ask GPT-4 to generate 100 distinct usernames and corresponding user_ids for each subset, respectively. To avoid repetition, we request GPT-4 to check for duplicates after each generation. We use these 100 private items on Qwen-VL for COCO. Results are shown in Table 11. As the privacy subset size increases, the gradients of MLLMs exhibit more significant deviations from the original gradient updates, indicating that MLLMs spend more effort in each gradient step learning different privacy information when increasing the privacy subset size.

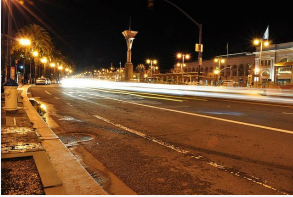
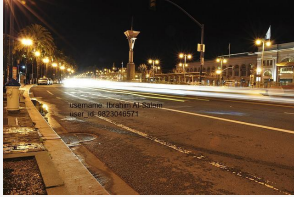
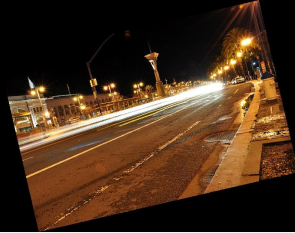
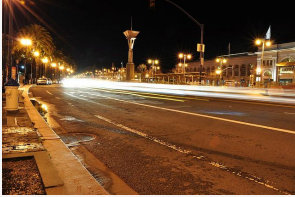
Origin	w/ Privacy	Image Augmentation	Text Augmentation
 <p>Filename: 000000341431.jpg</p> <p>Question: What time of day does the image capture?</p> <p>Answer: The image captures a night scene.</p> <p>Question: How is the street illuminated in the image?</p> <p>Answer: The street is illuminated by many street lights and the bright lights from the surrounding buildings. Additionally, there is a time lapse effect causing blurred lights on the road.</p> <p>Question: Is there any notable architecture in the image?</p> <p>Answer: Yes, there is a notable piece of architecture in the image: a cathedral can be seen across the street, adding to the overall ambiance of the scene.</p>	 <p>Filename: 000000341431.jpg</p> <p>Question: What time of day does the image capture?</p> <p>Answer: The image captures a night scene.</p> <p>Question: How is the street illuminated in the image?</p> <p>Answer: The street is illuminated by many street lights and the bright lights from the surrounding buildings. Additionally, there is a time lapse effect causing blurred lights on the road.</p> <p>Question: Is there any notable architecture in the image?</p> <p>Answer: Yes, there is a notable piece of architecture in the image: a cathedral can be seen across the street, adding to the overall ambiance of the scene.</p>	 <p>Filename: 000000341431.jpg</p> <p>Question: What time of day does the image capture?</p> <p>Answer: The image captures a night scene.</p> <p>Question: How is the street illuminated in the image?</p> <p>Answer: The street is illuminated by many street lights and the bright lights from the surrounding buildings. Additionally, there is a time lapse effect causing blurred lights on the road.</p> <p>Question: Is there any notable architecture in the image?</p> <p>Answer: Yes, there is a notable piece of architecture in the image: a cathedral can be seen across the street, adding to the overall ambiance of the scene.</p>	 <p>Filename: 000000341431.jpg</p> <p>Question: What moment of day does the image capture?</p> <p>Answer: The image captures a dark scene.</p> <p>Question: How is the road illuminated in the image?</p> <p>Answer: The street is illuminated by numerous street lamps and the vibrant lights from the surrounding buildings. Additionally, there is a time lapse effect causing blurred lights on the road.</p> <p>Question: Is there any distinctive architecture in the image?</p> <p>Answer: Yes, there is a notable piece of architecture in the image: a church can be seen across the street, adding to the overall atmosphere of the scene.</p>

Figure 7. Examples of data used for computing gradient similarity from COCO.