PPE-Bench: A Benchmark for Evaluating MLLM Unlearning under Private-Public Entanglement

Xianren Zhang[1], Delvin Ce Zhang[2], Dongwon Lee[1], Suhang Wang[1]
[1]The Pennsylvania State University, USA [2] University of Sheffield, UK
{xzz5508,dongwon,szw494}@psu.edu,
delvin.ce.zhang@sheffield.ac.uk

Abstract

Multimodal Large Language Models (MLLMs) have shown strong capabilities, but they may memorize private information from web data, raising privacy concerns. Machine unlearning offers a way to remove such private knowledge without retraining from scratch. However, existing MLLM unlearning benchmarks have two major limitations. First, they rely on simplified images that contain only the single target individual, failing to reflect the visual complexity of real-world photos. Second, they typically assume that the forget set and retain set are fully separated, ignoring the fact that private information is often visually entangled with benign public information. For example, a private individual may appear with a public figure or in front of a well-known landmark, where unlearning the private target should not damage the public context. To address these limitations, we propose PPE-Bench, a new benchmark for evaluating MLLM unlearning under private-public entanglement. Each image contains a target individual to be forgotten and public information to be preserved, including public figure and landmark. We further introduce two simple but effective methods to better preserve public information during unlearning. Through experiments, we find that existing unlearning methods can reduce private information leakage, but often substantially harm adjacent public information. [1]

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated strong performance on a wide range of multimodal tasks (Li et al., 2024a), such as visual question answering (Kuang et al., 2025) and image captioning (Sarto et al., 2025). However, as MLLMs are trained on large-scale data from internet that may contain sensitive and private information, they can memorize and reproduce such content (Huang et al., 2024), which raises significant privacy and copyright concerns. Privacy regulations like GDPR (Hoofnagle et al., 2019) and CCPA (Pardau, 2018) enforce the right to be forgotten (Dang, 2021). For instance, personal images and online profile information shared on social media and websites could unintentionally be included in the training data (Caldarella et al., 2024; Yan et al., 2024), causing privacy issues. In such cases, image owners may request that MLLMs forget the influence of this data. However, retraining MLLMs from scratch to remove sensitive knowledge is often impractical due to the high computational cost.

As a result, unlearning methods (Liu et al., 2025; Huo et al., 2025; Zhaopan Xu et al., 2025; Li et al., 2024b; Wu et al., 2025) are applied to MLLM models to "forget" such sensitive information without retraining the model from scratch. For example, some methods (Liu et al., 2025; Huo et al., 2025) try to remove visual patterns associated with specific entities, such as personal information including home address, occupation, and age. These methods typically finetune MLLMs using different objectives, such as maximizing the loss on private information or minimizing preference scores for sensitive content. Gradient Ascent (GA) (Yao and Xu, 2024), Gradient Difference (GD) (Liu et al., 2022) or Negative Preference Optimization (NPO) (Zhang et al., 2024) are common approaches used for MLLM unlearning.

Recently, several benchmarks (Dontsov et al., 2025; Liu et al., 2025) have been designed to evaluate the effectiveness of unlearning methods under the multimodal setting. CLEAR (Dontsov et al., 2025) is the first open-sourced benchmark specifically for multimodal unlearning. It uses the text-only TOFU dataset (Maini et al., 2024) as basis and extracts fictitious name, age, and ethnicity from the TOFU dataset. It then generates synthetic faces with StyleGAN2 (Karras et al., 2020) based on each individual's attributes, and uses a diffusion

[1]Data: https://github.com/Zood123/PPE_Bench

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 45265–45278
October 24-29, 2026 ©2026 Association for Computational Linguistics

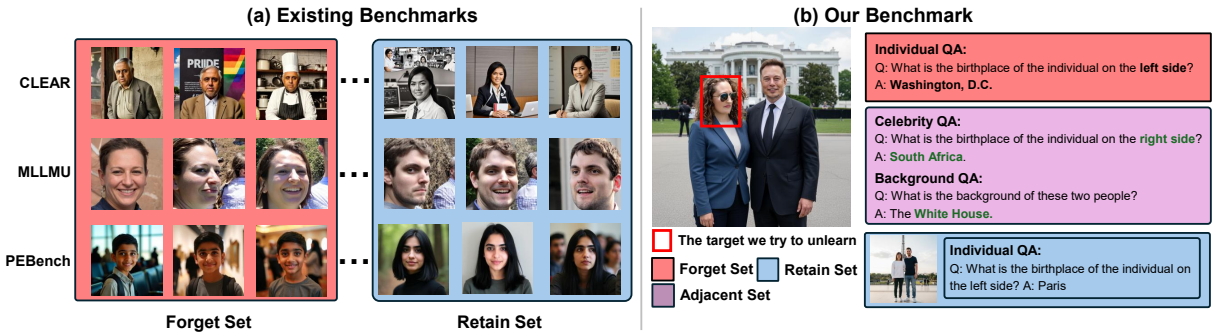

Figure 1: (a) Existing benchmarks have completely separate forget set and retain set and images are simple where each image only contains one individual profile. (b) Our benchmark studies a more complex scenario where the individual is surrounded by public information such as public celebrity and well-known landmark.

model, PhotoMaker-V2 (Li et al., 2024d) to synthesize final images. MLLMU-Bench (Liu et al., 2025) is also proposed to study whether MLLM unlearning can effectively unlearn both image and textual information to protect privacy. It creates fictitious profiles with GPT-4o (Hurst et al., 2024) and also uses StyleGAN (Karras et al., 2020) to synthesize fictitious faces. PEBench (Zhaopan Xu et al., 2025) extends the contexts with more scenarios and events, such as reading in the library, working in the office and waiting at an airport terminal.

Though existing benchmarks are useful for evaluating the effectiveness of MLLM unlearning methods, they still have two major limitations. The first issue is the *lack of realistic visual complexity*. In practice, photos shared on the internet or social platforms often contain multiple entities. However, as shown in Figure 1(a), existing benchmarks use simplified images that contain only the target individual. Such a simplified setting cannot reveal how unlearning private information in complex images would affect the public information entangled within the same visual scene. The second issue is the *complete separation between the forget set and the retain set*. Existing benchmarks construct the forget set to evaluate whether the model can remove private information about the target individuals, and construct a separate retain set to evaluate whether the model preserves general utility. However, these two sets are usually built from disjoint images or unrelated entities. As a result, they only test whether unlearning affects unrelated retained knowledge, but cannot evaluate whether unlearning the target individual damages public information that is entangled with the target in the same image. In real-world scenarios, a private individual may appear together with a public figure or in front of a well-known landmark. In such cases, the public figure and landmark should be retained even though they appear in the same image with the private individual. This creates a more challenging setting where the forget and retain objectives are coupled within the same image: the model must suppress private information about the target individual while preserving public information from the surrounding context. Existing benchmarks fail to capture this practical difficulty.

To address the limitations, we propose **PPE-Bench**, a new benchmark for evaluating MLLM unlearning under *private-public entanglement*. As shown in Figure 1(b), each image in our benchmark contains not only the target individual to be forgotten, but also public information that should be preserved, including a public figure and a background landmark. This design introduces richer visual context and better reflects real-world scenarios where private and public information are naturally entangled. Based on this setup, we construct three subsets. The **forget set** contains question-answer pairs about the target individual and is used to evaluate whether the model has successfully forgotten the sensitive information. The **retain set** contains unrelated images and question-answer pairs about other individuals, which measures the general utility of the model after unlearning. The **adjacent set** contains question-answer pairs about the public figure and background landmark around the forgotten target, and is used to evaluate whether the model can preserve public information that is visually entangled with the forgotten target. PPE-Bench has a more realistic evaluation of MLLM unlearning by jointly assessing private information removal and adjacent public information preservation.

Our **main contributions** are: (i) We propose PPE-Bench, a new benchmark for evaluating MLLM unlearning under private-public entanglement, where the forgotten target co-occurs with public information in the same image, enabling

a more realistic evaluation setting than existing benchmarks; (ii) We introduce two simple yet effective methods to better preserve public information during unlearning; and (iii) Our comprehensive experiments on PPE-Bench show that existing methods often damage adjacent public information, and the forgotten private knowledge can re-emerge after finetuning on public information.

2 Related Work

LLM Unlearning. LLM unlearning aims to remove specific knowledge from a trained language model without retraining it from scratch (Yao et al., 2024; Liu et al., 2024; Yao and Xu, 2024). Existing methods typically achieve this by finetuning model parameters such as gradient ascent (Liu et al., 2022), negative preference optimization (NPO) (Zhang et al., 2024) or gradient difference (Liu et al., 2022). To evaluate the unlearning effectiveness, several text-based unlearning benchmarks have been proposed, covering harmful knowledge (Rafailov et al., 2023; Li et al., 2024c), sensitive personal information (Patil et al., 2024; Maini et al., 2024), and copyrighted content (Eldan and Russinovich, 2023). These benchmarks focus on the text modality and cannot evaluate whether unlearning methods can remove visual private information, such as facial identity.

MLLM Unlearning. This limitation has motivated recent studies on unlearning for MLLMs, where sensitive information may exist in both text and visual content (Li et al., 2024b; Liu et al., 2025; Huo et al., 2025; Zhaopan Xu et al., 2025; Li et al., 2024b; Wu et al., 2025). These methods often adapt standard LLM unlearning objectives to multimodal settings. Several benchmarks have also been proposed to evaluate multimodal unlearning, including MLLMU-Bench (Liu et al., 2025), CLEAR (Dontsov et al., 2025), and PEBench (Zhaopan Xu et al., 2025), which use synthetic profiles and generated images to test whether MLLMs can forget sensitive information. However, these benchmarks focus on isolated or independently defined entities and fail to capture the complex interactions between multiple entities across multimodal contexts. More details of related works can be found in Appendix B.

3 The PPE-Bench

We introduce **PPE-Bench**, a benchmark designed to evaluate MLLM unlearning in settings where private target information is visually entangled with safe public information. As shown in Figure 2, our benchmark construction consists of three main steps. First, we generate complex and realistic images by placing each fictitious individual together with a public celebrity and a well-known landmark. Second, we construct multimodal question-answer pairs covering both the private individual and the co-occurring public information. Third, we split the data into forget, retain, and adjacent sets, which allows us to evaluate not only whether the target individual is forgotten, but also whether the entangled public information is preserved. We also specify the evaluation setup on PPE-Bench and propose two *mitigation methods* that can preserve public information while unlearning private individuals.

3.1 Overview of Benchmark

In PPE-Bench, each image contains a fictitious individual whose private information should be forgotten, together with a public celebrity and a landmark whose information should be preserved. The benchmark includes 100 fictitious individuals, 28 public celebrities, and 40 well-known landmarks. For each individual, we generate 18 images, resulting in 1,800 visually entangled images in total (1,500 for training and 300 for testing). Each image is paired with eight question-answer pairs, including five questions about the target individual, two questions about the public celebrity, and one question about the background landmark. In total, the benchmark has a training set with 12,000 QA samples and two test sets with 3,000 samples respectively. The selected celebrities span multiple domains, and the landmarks cover diverse geographic regions, providing broad coverage of public information. More details are provided in Appendix A.1.

3.2 Data Generation

As shown in the first row of Figure 2, the data construction process mainly consists of three parts: the image data generation, the text data generation and the data split.

Image Data Generation. We use the profile images of 100 fictitious individuals together with their attributes, e.g., name, education and birthday, from MLLMU-Bench (Liu et al., 2025) as the basis. For each individual, we select three profile images as source images. To define a complex scene, we randomly sample one public figure from a pool of 28 figures and one landmark from a pool of 40 landmarks, and randomly determine whether the

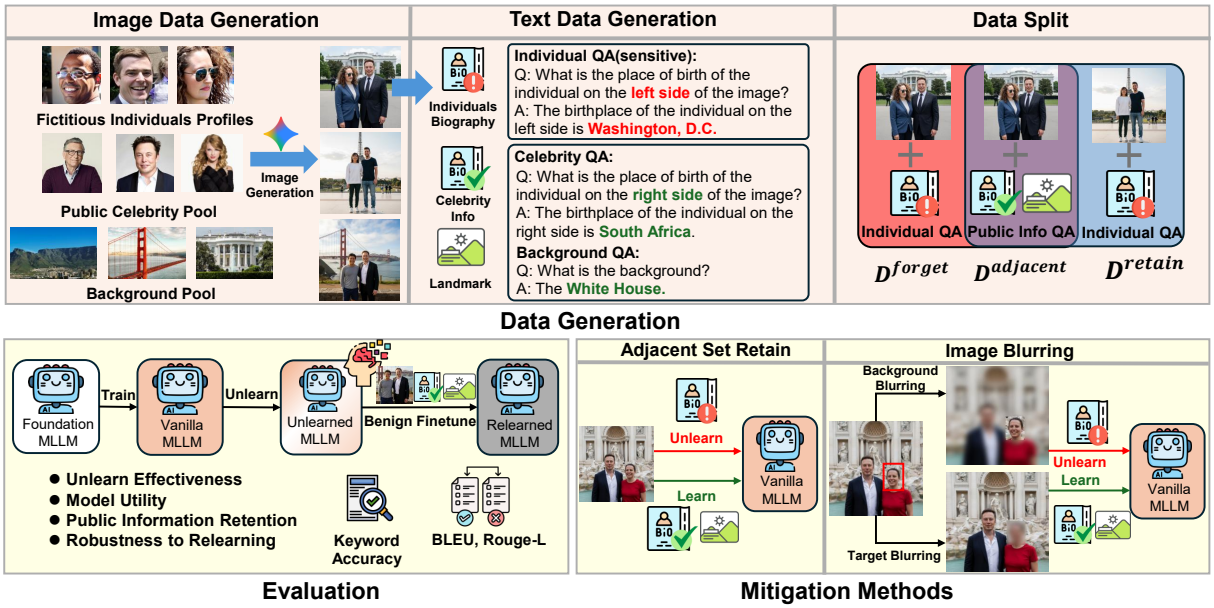

Figure 2: Overview of PPE-Bench. (a) We generate visually entangled images containing a fictitious individual, a public celebrity, and a landmark, then construct QA pairs targeting private individual information and co-occurring public information. (b) The data are split into forget, adjacent, and retain sets to evaluate unlearning effectiveness, model utility, public information retention, and robustness to relearning. (c) We further study mitigation methods based on adjacent-set retention and image blurring.

public figure stands on the left or right side of the private individual. We then use Gemini 2.5 Flash Image (Nano Banana) (Google, 2025) to generate the final image with the prompt template shown in the Appendix A.2. In this way, each generated image contains both the private individual and public information, including a public figure and a landmark. We then manually check the generated images to ensure that they are realistic, preserve the intended individual identity, and correctly generate the specified public figure and landmark. Images that do not meet these criteria are regenerated.

Text Data Generation. As each generated image contains both the private individual and public information, we construct question-answer pairs that cover both aspects. As shown in Figure 2, each image is associated with questions about the private individual and questions about the co-occurring public information. Specifically, for the private individual, we randomly select five attributes from the biography, such as name, birthplace, date of birth, and education, and generate five question-answer pairs using randomly sampled templates from a template pool. Details and examples of the question-answer templates are in Appendix A.4. For the public figure, we similarly collect the corresponding attribute information, randomly select two attributes, and generate two question-answer pairs. In addition, we generate one question-answer pair about the background landmark.

Dataset Split. We split both the training and test data according to three query targets: forget, retain, and adjacent public information. For training data, following prior unlearning benchmarks (Liu et al., 2025), we first construct a *forget set* D^{forget} and a *retain set* D^{retain} by splitting the 100 individuals into two disjoint groups, with 50 individuals assigned to each set. However, unlike existing benchmarks, our benchmark includes a third subset, the *adjacent set* D^{adj}. To evaluate whether unlearning the private individual affects entangled public information, we construct the adjacent set using the same images of $\mathcal{D}^{forget}$ but with queries targeting only the public figure or background landmark. We generate 1,500 images and 12,000 QAs for training. For evaluation, we construct two test sets. The first is the *seen-image test set* D^{s}_{test}, where the images appear in the training set but we change the question template. The second is the *unseen image test set* D^{u}_{test}, where we generate 300 new images. Both test sets include QA samples targeting the individuals in D^{forget}, the individuals in D^{retain}, and the public figure or landmark information entangled with D^{adj}. The details of data statistics are shown in Appendix A.1.

3.3 Evaluation Setup

As shown in the second row of Figure 2, our benchmark supports a comprehensive evaluation of multimodal unlearning methods. We first finetune the foundation model on the full training data, including D^{forget}, D^{retain}, and D^{adj}, to obtain a vanilla model that captures both the target private information and the entangled public information.

We then apply different unlearning methods to the vanilla model to remove the private information associated with individuals in D^{forget}, obtaining an unlearned model. To evaluate its performance, we test the unlearned model on both D^{s}_{test} and D^{u}_{test}. Specifically, we assess three aspects: (1) *unlearning effectiveness*, measured by the model's performance on questions about individuals in the forget set; (2) *utility preservation*, measured by its performance on individuals in the retain set; and (3) *public information retention*, measured by its performance on questions about the public figures and landmarks.

In practice, an unlearned model may later be further finetuned. Motivated by this practical setting, we further finetune the unlearned model on the public information to obtain a relearned model, and then evaluate whether the previously forgotten private information is recovered.

3.4 Mitigation Methods

One risk of unlearning is that the entangled public information is also unlearned and forgotten. We design two simple methods that could improve the retention of public information while can still effectively unlearn the private target individual.

3.4.1 Public Information Preservation (PIP)

One straightforward way to mitigate the forgetting of public information is to introduce an additional loss term that explicitly protects such information during the unlearning process. We define the overall loss function as

$$\mathcal{L} = -\mathcal{L}(\mathcal{D}_f, \theta) + \lambda_r \mathcal{L}(\mathcal{D}_r, \theta) + \lambda_a \mathcal{L}(\mathcal{D}_{adj}, \theta), \quad (1)$$

where $\mathcal{L}(\mathcal{D}_{adj}, \theta)$ is the cross-entropy loss on the adjacent set, and θ is the model parameters. The λ_r and λ_a control the strength of the retain set and adjacent set preservation terms, respectively. By optimizing this objective, the model is encouraged to forget the target individual while maintaining performance on both the retain set and the adjacent public information.

3.4.2 Target-Guided Image Blurring

Optimizing the adjacent set may be impractical in real-world applications, since it requires identifying, collecting and annotating adjacent public information, which can be time-consuming and labor-intensive. Therefore, we propose a target-guided image blurring strategy that disentangles the target individual from the surrounding public information. Specifically, for each image x^{i}, we have a target region annotation b, such as a bounding box around the target face. Based on b, we define a binary mask $m_b \in \{0,1\}^{H \times W}$, where H and W denote the image height and width. $m_b = 1$ indicates the target region and $m_b = 0$ otherwise. Let $G(\cdot)$ denote a blur operator. We then define two transformed views:

$$T_{\mathrm{t}}(x^{\mathrm{i}}, b) = m_b \odot x^{\mathrm{i}} + (1 - m_b) \odot G(x^{\mathrm{i}}), \quad (2)$$

$$T_{\mathrm{c}}(x^{\mathrm{i}}, b) = (1 - m_b) \odot x^{\mathrm{i}} + m_b \odot G(x^{\mathrm{i}}), \quad (3)$$

where $\odot$ denotes element-wise multiplication. Here, T_{t} preserves the target region while blurring the surrounding context, and T_{c} preserves the surrounding context while blurring the target region.

For the forget set, we apply T_{t} so that the model focuses on the target individual to be forgotten. For the retain set, we apply T_{c} so that the model preserves the public information outside the target region. The overall objective is defined as

$$\mathcal{L}_{\text{blur}} = -\mathcal{L}(T_t(\mathcal{D}_f), \theta) + \lambda_a \mathcal{L}(T_c(\mathcal{D}_{adj}), \theta), \quad (4)$$

where $T_t(\mathcal{D}_f)$ is the forget set with only the target region preserved, and $T_c(\mathcal{D}_{adj})$ denotes the adjacent set with the target region blurred. By optimizing this objective, the model is encouraged to forget identity-specific knowledge of the target individual while retaining surrounding public information. We combine this objective with the retain set loss: $\mathcal{L} = \mathcal{L}_{\text{blur}} + \lambda_r \mathcal{L}(\mathcal{D}_r, \theta)$. This method requires the location of the target individual in each image, such as a face bounding box or segmentation mask, which serves as the supervision for disentangling the target from the surrounding content.

4 Experiment

In this section, we conduct experiments to answer the following research questions: (RQ1) Can existing unlearning methods effectively forget the target private information in complex images? (RQ2) Do these unlearning methods also degrade the safe

public information that is entangled with the private target? (RQ3) Can the forgotten private information re-emerge after subsequent finetuning on public information?

4.1 Experiment Setup

Implementation. We conduct experiments on two MLLMs as foundation models: Qwen3-VL-4B-Instruct (Bai et al., 2025) and LLaVA-1.5-7B (Liu et al., 2023). We train vanilla models on the full training set, including retain, forget, and adjacent data, and then apply different unlearning methods to obtain unlearned models. The hyper-parameters details are shown in Appendix Table 5.

Baselines. We evaluate several representative unlearning methods, which can be grouped into three categories: (i) **Forget-only methods**, which suppress the model's ability to predict the forget set without explicitly preserving other knowledge. These include Gradient Ascent (GA) (Yao and Xu, 2024) and Negative Preference Optimization (NPO) (Zhang et al., 2024). (ii) **Retention-aware methods**, which use the retain set to preserve model utility during unlearning. These include KL Minimization (Liu et al., 2025) and Gradient Difference (GD) (Liu et al., 2022). (iii) **Our designed methods**, which can potentially mitigate damage to adjacent public information in our benchmark. These include PIP (Public Information Preservation) and Blurring. PIP has two variants: PIP (Adjacent) and PIP (Retain), using public information questions from the adjacent and retain sets. Details of these baselines are provided in Appendix A.3.

Evaluation Metrics. We use the following evaluation metrics: (i) **Accuracy (ACC).** We measure accuracy by checking whether the target keyword appears in the model-generated response. The answer is considered correct if containing the ground-truth keyword; and (ii) **Generative Metrics.** To evaluate the quality of free-form generated answers, we use **BLEU** (Papineni et al., 2002) and **ROUGE-L** (Lin, 2004) to measure textual overlap between the generated answer and the reference answer. ROUGE-L captures the longest common subsequence between two texts. BLEU evaluates the precision of n-gram matches. Higher scores in both metrics indicate closer alignment with the reference answer.

4.2 RQ1: Unlearning Effectiveness

To answer RQ1, we test different unlearning methods on the test forget set with Qwen3-VL-4B-Instruct (Bai et al., 2025). The results on both the seen-image test set D^s_{test} and unseen-image test set D^u_{test} are presented in Table 1, where lower scores indicate stronger forgetting of the target private information. We also report the corresponding results on LLaVA (Liu et al., 2023) at Appendix Table 4, which shows similar trends. As expected, the learned vanilla model has the highest performance on the forget set, with 46.2% ACC. We have the following observations: **(i) Existing unlearning methods can effectively remove private information.** Compared with the vanilla model, all unlearning baselines lead to a clear drop in performance on the forget set under both seen-image and unseen-image settings. In particular, the ACC decreases very sharply across methods, indicating that the unlearned models are much less likely to reproduce the target private attributes, such as names, dates, and other key identifying information. Among them, Gradient Difference achieves the strongest forgetting performance, suggesting that jointly maximizing loss on the forget set while preserving general utility is a very effective strategy. **(ii) Preserving public information can weaken unlearning effectiveness.** Both PIP variants lead to higher forget-set performance, indicating a trade-off between forgetting the target individual and retaining related public content. A possible reason is that the target private information and the surrounding public information are entangled in the same visual scene and may share overlapping representations, so explicitly preserving public knowledge can partially preserve features that are also useful for recalling the private target.

	Unseen Images			Seen Images		
Method	**Acc** (%)	**BLEU**	**Rouge-L**	**Acc** (%)	**BLEU**	**Rouge-L**
Vanilla Model	46.2	0.909	0.922	70.1	0.943	0.955
Gradient Ascent	13.7	0.823	0.870	11.7	0.822	0.872
NPO	3.4	0.702	0.793	3.2	0.686	0.787
KL-Min	8.6	0.853	0.878	6.9	0.836	0.874
Gradient Difference	**0.9**	**0.602**	**0.681**	**0.3**	**0.590**	**0.672**
PIP (adjacent set)	18.8	0.857	0.882	15.5	0.842	0.878
PIP (retain set)	14.5	0.847	0.874	11.2	0.277	0.555
Blurring	8.0	0.840	0.872	6.4	0.826	0.873

Table 1: Forget effectiveness of different unlearning methods on the Qwen model under the unseen-image and seen-image settings. Lower values indicate better forgetting performance.

4.3 RQ2: Retention of Adjacent Public Information

To answer RQ2, we evaluate whether unlearned models can still preserve the safe public information that is entangled with private information of the target individual. Figure 3 shows the relation-

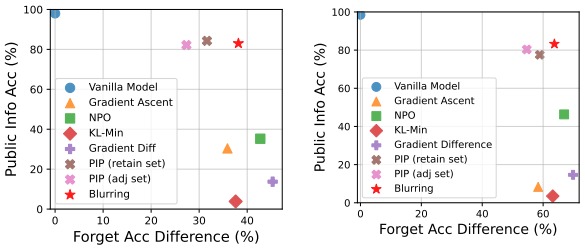

Figure 3: Accuracy on adjacent public-information questions versus forget-set accuracy on the unseen-image test set (left) and the seen-image test set (right).

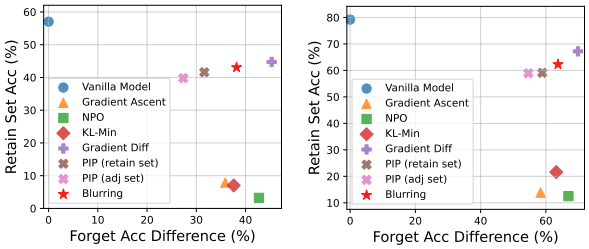

Figure 4: Retain-set accuracy versus the drop in forget-set accuracy relative to the vanilla model on the unseen-image test set (left) and the seen-image test set (right).

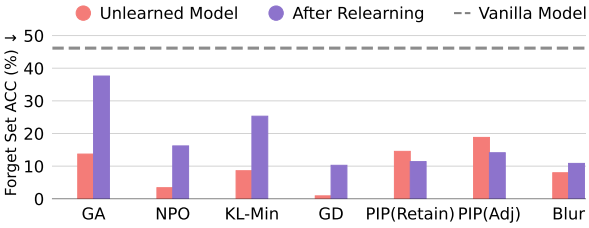

Figure 5: Forget set accuracy before and after further finetuning on public information. Higher accuracy after relearning indicates stronger recovery of the previously forgotten private information.

ship between adjacent public-information QA accuracy and the drop in forget-set accuracy relative to the vanilla model on both the unseen-image test set and the seen-image test set. Ideally, a good unlearning method should achieve low forget-set accuracy while maintaining high accuracy on adjacent public-information questions. We have the following observations. **(i) Existing unlearning methods often damage adjacent public information.** As shown in Figure 3, most standard unlearning baselines are located in the lower-right region of the plots. Although they successfully suppress private information, they also substantially degrade performance on questions about the surrounding public content. This suggests that current unlearning methods struggle to selectively remove the target private knowledge without harming the related public information appearing in the same visual context. **(ii) Public-information-preserving methods can alleviate this problem with a trade-off.** Compared with standard baselines, PIP and the blurring-based method achieve significantly higher accuracy on questions related to public information, showing that explicitly protecting public content during unlearning is helpful. However, this improvement comes with weaker forgetting of the target private information. Overall, while these mitigation strategies improve public-information retention, the results suggest that more advanced unlearning methods are still needed to better balance effective forgetting with preservation of safe public knowledge.

4.4 Retain Set Performance

We evaluate model utility by testing the unlearned models on the retain set. Figure 4 shows retain-set accuracy versus the forget-set accuracy drop relative to the vanilla model on both unseen-image and seen-image test sets. Ideally, an unlearning method should strongly suppress private information while maintaining high retain set performance. We have the following observations. **(i) Gradient Difference best preserves utility.** GD achieves the highest retain set accuracy in both settings, because it explicitly optimizes the retain objective during unlearning. **(ii) Other baselines degrade retain-set performance.** GA, NPO, and KL-Min all lead to clear drops in retain-set accuracy compared with the vanilla model, suggesting that strong forgetting comes at the cost of general utility. **(iii) The mitigation methods can maintain the retain set performance.** PIP and the blurring-based method also maintain competitive retain-set performance, suggesting that these preservation-oriented strategies do not introduce substantial utility loss.

4.5 RQ3: Robustness to Relearning

To answer RQ3, we examine whether unlearned private information remains forgotten after further finetuning on benign public data. Since the private target and public information are entangled in the same image, finetuning on public information may also reactivate parameters related to the forgotten private target. Therefore, this experiment tests the robustness of unlearning under a realistic scenario after unlearning.

We use images from the retain set but with questions about the public entities. Figure 5 shows the forget set accuracy before and after this fine-

tuning. Most methods exhibit clear knowledge recovery: GA, NPO, KL-Min, GD and blurring show increases in forget set accuracy, indicating that the previously forgotten private information can re-emerge after the model is exposed only to benign public data. For PIP methods, their forget set accuracy is already relatively high before finetuning. These results suggest that existing unlearning methods do not fully erase the sensitive private information from the model, but instead suppress or refuse to output. Moreover, finetuning on public information can make unlearned models output private information.

4.6 Hyperparameter Sensitivity Analysis

We study the effects of the two loss coefficients, λ_r and λ_a, which control the strength of the retain set loss and the adjacent set preservation loss. We first set $\lambda_a = 0$ and tune λ_r to examine the trade-off between forgetting effectiveness and utility. We then fix the best λ_r and tune λ_a.

As shown in the left chart of Figure 6, a small λ_r leads to low retain set accuracy, indicating poor general utility. Increasing λ_r generally improves retain-set accuracy, but it also slightly weakens the forgetting. When λ_r is too large, utility gains saturate while the forget set accuracy drop decreases. Thus, we choose $\lambda_r = 2$ as the best trade-off. Next, we fix $\lambda_r = 2$ and tune λ_a. The right chart reports joint preservation accuracy, defined as the average accuracy on the adjacent and retain sets. Increasing λ_a improves preservation, but overly large values bring limited gains and weaken forgetting. Therefore, we select $\lambda_a = 1.6$, which achieves the best balance between forgetting private information and preserving non-target information.

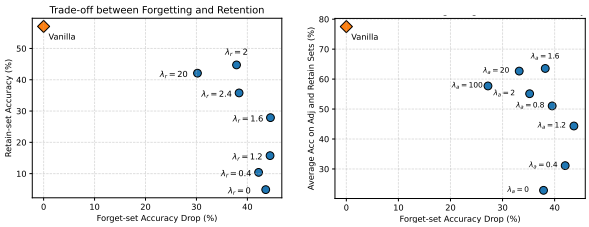

Figure 6: Left: trade-off between forget set accuracy drop and retain set accuracy with $\lambda_a = 0$. Right: trade-off between forget-set accuracy drop and joint preservation accuracy, defined as the average accuracy on the adjacent and retain sets.

4.7 Case Study

In Figure 7, we present examples showing how different unlearned models respond to questions

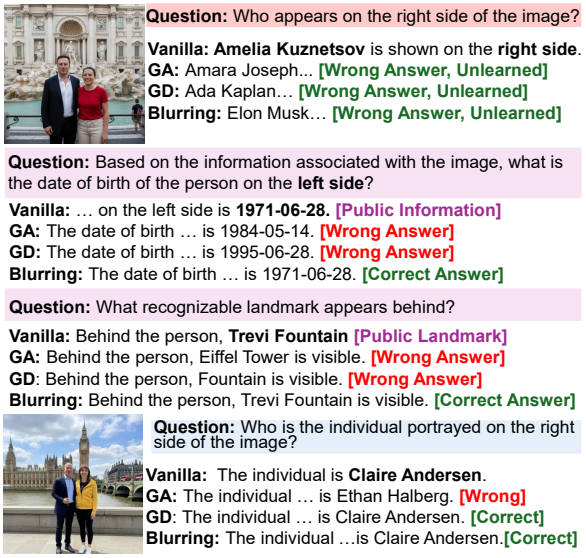

Figure 7: Case study comparing responses from different unlearned models on forget, adjacent, and retain information.

about private information related to the target individual (marked in red; forget set), public information surrounding the individual (marked in purple; adjacent set), and retained information (marked in blue; retain set). We observe that GA removes not only private information in the forget set but also public and retained information, leading to over-forgetting. Gradient Difference (GD) effectively preserves retain-set information but fails to preserve public information in the adjacent set. In contrast, the blurring method preserves both adjacent-set and retain-set information while effectively removing private information from the forget set.

4.8 Inpainting vs. Blurring

We further investigate mean inpainting as an alternative to blurring for visual information removal. Specifically, pixels within the target bounding box are replaced by the mean pixel value of the corresponding image. We perform unlearning on the training set using Qwen3-VL-4B-Instruct (Bai et al., 2025) and evaluate generalization on unseen test images.

As shown in Table 2, mean inpainting achieves a forget accuracy of 9.1%, comparable to blurring (8.0%), while improving public information accuracy from 80.1% to 85.3%. These results suggest that simple inpainting provides a favorable trade-off between forgetting efficacy and preservation of public information. More sophisticated inpainting approaches, such as replacing the target region with

Method	Forget Acc. (%) ↓	Public Info Acc. (%) ↑
Vanilla Model	46.2	98.0
Gradient Ascent	10.3	30.4
NPO	3.4	35.3
KL-Min	8.6	3.8
Gradient Difference	0.9	13.7
PIP (retain set)	14.5	81.3
PIP (adjacent set)	18.8	79.2
Blurring	8.0	80.1
Inpainting	9.1	**85.3**

Table 2: Comparison of mean inpainting and blurring on Qwen3-VL-4B-Instruct. Lower forget accuracy indicates stronger forgetting, while higher public information accuracy indicates better preservation.

a semantically neutral background generated by an image generation model, may further improve this trade-off.

5 Conclusion

In this work, we introduce PPE-Bench, a benchmark for evaluating MLLM unlearning under private-public entanglement. Unlike existing benchmarks with simplified and separated forget/retain settings, PPE-Bench places private individuals together with public figures and landmarks, enabling a more realistic evaluation of whether models can forget sensitive information while preserving surrounding public knowledge. Our experiments show that existing unlearning methods can reduce private information leakage, but often damage adjacent public information and remain vulnerable to relearning after further finetuning. These findings highlight the need for more robust MLLM unlearning methods that can handle visually entangled real-world scenarios.

6 Limitations

The benchmark focuses on images with one target individual, one public figure, and one landmark. Future work can consider more complex scenarios, such as crowded scenes, multiple private targets and richer interactions among people and background objects.

7 Acknowledgments

This work was supported in part by U.S. NSF awards #2438810 and #2555559, Amazon Nova AI Challenge, and PSU Frymoyer Chairship. Some experimental results were obtained using computational resources provided by CloudBank, supported through U.S. NAIRR award #240336.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Simone Caldarella, Massimiliano Mancini, Elisa Ricci, and Rahaf Aljundi. 2024. The phantom menace: unmasking privacy leakages in vision-language models. *arXiv preprint arXiv:2408.01228*.

Quang-Vinh Dang. 2021. Right to be forgotten in the age of machine learning. In *International Conference on Advances in Digital Science*, pages 403–411. Springer.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. 2025. Clear: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20582–20603.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv e-prints*, pages arXiv–2310.

Google. 2025. Image editing in gemini just got a major upgrade. Google Blog.

Chris Jay Hoofnagle, Bart Van Der Sloot, and Frederik Zuiderveen Borgesius. 2019. The european union general data protection regulation: what it is and what it means. *Information & communications technology law*, 28(1):65–98.

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10711–10732.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7190–7206.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng,

Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, and 1 others. 2024a. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. 2024b. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, and 1 others. 2024c. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28525–28550.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4105–4135.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. In *First Conference on Language Modeling*.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. 2020. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Stuart L Pardau. 2018. The california consumer privacy act: Towards a european-style privacy regime in the united states. *J. Tech. L. & Pol'y*, 23:68.

Vaidehi Ramesh Patil, Peter Hase, and Mohit Bansal. 2024. Can sensitive information be deleted from llms? objectives for defending against extraction attacks. In *International Conference on Learning Representations*, volume 2024, pages 45497–45514.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Sara Sarto, Marcella Cornia, and Rita Cucchiara. 2025. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, and 1 others. 2025. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey. *arXiv preprint arXiv:2403.05156*.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 8403–8419.

Yuanshun Yao and Xiaojun Xu. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

Pengfei Zhou Zhaopan Xu, Hongxun Yao Weidong Tang, and Kaipeng Zhang. 2025. Pebench: A fictitious dataset to benchmark machine unlearning for multimodal large language models. *arXiv preprint arXiv:2503.12545*.

A Appendix

A.1 Benchmark Statistics

Specifically, each image contains a fictitious individual to be forgotten, together with a public celebrity and a landmark that should be retained. This design allows us to study whether unlearning the target individual would also hurt the public knowledge. As summarized in Table 3, PPE-Bench contains 100 fictitious individuals, 28 celebrities, and 40 well-known landmarks from around the world. To ensure diversity, the selected celebrities are across multiple domains, including technology and entertainment, while the landmarks cover different geographic regions, as shown in Figure 8. For each fictitious individual, we generate 15 images. In every image, the individual appears with one celebrity in front of one landmark, creating a visually entangled composition of private and public information. For each image, we construct five question-answer pairs about the individual, two about the celebrity, and one about the background landmark. In total, the dataset contains 12,000 question-answer samples. Additionally, we also provide two test sets with seen images in the dataset D^s_{test} and unseen new images D^u_{test}.

Category	Count
Individuals	100
Celebrities	28
Landmarks	40
Images	1500
- Images per Individual	15
Individual QAs	7500
- Individual QAs per Image	5
Celebrity QAs	3000
- Celebrity QAs per Image	2
Background QAs	1500
- Background QAs per Image	1
Training Data Samples	12000
Seen-image Test Set (D^s_{test})	
- Forget Set	1000
- Adjacent Set	1000
- Retain Set	1000
Unseen-image Test Set (D^u_{test})	
- Forget Set	1000
- Adjacent Set	1000
- Retain Set	1000

Table 3: Dataset statistics.

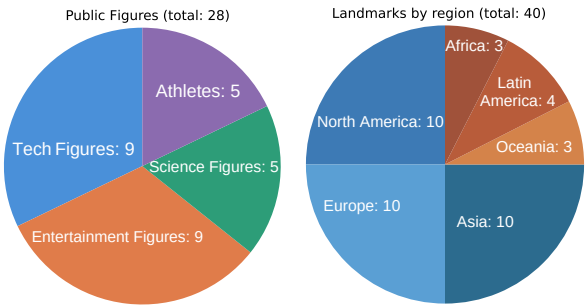

Figure 8: Distribution of public figures (left) and landmarks by region (right) in the benchmark dataset.

A.2 Image Generation Prompt

We use the following prompt to generate images using Gemini 2.5 Flash Image (Nano Banana) (Google, 2025).

Image Generation Prompt

`<profile_image>`
Given one person's profile, please generate an image where the person is standing with `<public_figure_name>` in front of the `<landmark_text>`. `<public_figure_name>` is standing on the `<side>` side of that person. Ensure the person's face remains the same as in the profile, without any changes.

A.3 Baselines Details

A.3.1 Gradient Ascent

Gradient Ascent (GA) (Yao and Xu, 2024) is a simple and widely used unlearning baseline that directly maximizes the training loss on the forget set $\mathcal{D}_f$. Intuitively, by ascending the loss landscape on the target samples, the model is encouraged to move away from parameters that support correct predictions on the data to be forgotten. The objective is defined as

$$\mathcal{L}_{\text{GA}} = -\mathcal{L}(\mathcal{D}_f, \theta), \tag{5}$$

where $\mathcal{L}(\mathcal{D}_f, \theta)$ denotes the cross-entropy loss computed on the forget set. By optimizing this objective, the model reduces its ability to answer questions related to the target private information. However, since GA does not explicitly constrain the model's behavior on non-forget data, it may also degrade overall utility and harm the retention of related public information.

A.3.2 Negative Preference Optimization

Negative Preference Optimization (NPO) (Zhang et al., 2024) formulates unlearning as a preference optimization problem by discouraging the model from assigning high probability to target responses on the forget set $\mathcal{D}_f$. Compared with directly maximizing the forget loss, NPO provides a smoother optimization objective by contrasting the current model with a reference model. The objective is defined as

$$\mathcal{L}_{\text{NPO}} = \mathbb{E}_{(x,y)\in\mathcal{D}_f}\left[\log\left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right)\right] \quad (6)$$

where $\pi_\theta(y|x)$ denotes the conditional probability assigned by the current model, $\pi_{\text{ref}}(y|x)$ denotes that of the reference model. By optimizing this objective, the model is encouraged to reduce its preference for the target outputs on the forget set while maintaining more stable updates than standard gradient ascent.

A.3.3 KL Minimization

KL Minimization (Liu et al., 2025; Nguyen et al., 2020) combines forgetting on the forget set $\mathcal{D}_f$ with a distribution-matching regularization term on the retain set $\mathcal{D}_r$. Specifically, it maximizes the loss on the forget set while encouraging the current model to remain close to the original model on retain samples by minimizing the Kullback–Leibler (KL) divergence between their output distributions. The objective is defined as

$$\mathcal{L}_{\text{KL}} = -\mathcal{L}(\mathcal{D}_f, w) + \frac{1}{|\mathcal{D}_r|}\sum_{s\in\mathcal{D}_r} \text{KL}(M_o \,\|\, M_c)\,(s), \quad (7)$$

where $\mathcal{L}(\mathcal{D}_f, w)$ denotes the cross-entropy loss on the forget set, M_o and M_c denote the original model and the current model, respectively, and $\text{KL}(M_o \,\|\, M_c)(s)$ measures the KL divergence between their output distributions on retain sample s. By optimizing this objective, the model is encouraged to forget the target private information while preserving its behavior on retain samples.

A.3.4 Gradient Difference

Gradient Difference (GD) (Liu et al., 2022) balances forgetting on the forget set $\mathcal{D}_f$ with the preservation of performance on the retain set $\mathcal{D}_r$. The objective is to increase the loss on $\mathcal{D}_f$ while minimizing the impact on $\mathcal{D}_r$. The overall loss function is defined as

$$\mathcal{L}_{\text{GD}} = -\mathcal{L}(\mathcal{D}_f, \theta) + \lambda_r \mathcal{L}(\mathcal{D}_r, \theta), \quad (8)$$

where $\mathcal{L}(\mathcal{D}_f, \theta)$ and $\mathcal{L}(\mathcal{D}_r, \theta)$ denote the cross-entropy loss computed on the forget set and the retain set, respectively. By optimizing this objective, the model selectively forgets the target data while preserving general utility on unrelated samples.

A.4 Question-Answer Templates

For each private individual, we have eight attributes: name, salary, height, residence, employment, education, birthplace, and birthday. For each attribute, we construct a template pool containing 10 question-answer format pairs. During dataset construction, we sample from these templates to generate QA pairs. Each template includes a position placeholder, which specifies whether the target individual (or the celebrity) appears on the left or right side of the image, and an answer placeholder corresponding to the ground truth. Some examples of the question templates are shown below.

Question Template Examples

Name.
`Question: What is the name of the person on the <position> side of the image?`
`Ground Truth: The individual on the <position> side of the image is named <answer>.`

Birthday.
`Question: On what date was the person on the <position> side of the image born?`
`Ground Truth: The individual on the <position> side was born on <answer>.`

Height.
`Question: What is the height of the person depicted on the <position> side of the image?`
`Ground Truth: The individual on the <position> side has a height of <answer>.`

B Related Work Details

LLM Unlearning. LLM unlearning aims to remove specific knowledge from a trained language model without retraining it from scratch (Yao et al., 2024; Liu et al., 2024; Yao and Xu, 2024). Existing methods typically achieve this by training

model parameters with objectives such as gradient ascent (Liu et al., 2022), negative preference optimization (NPO) (Zhang et al., 2024) or gradient difference (Liu et al., 2022). To evaluate the unlearning effectiveness, several text-based unlearning benchmarks have been proposed, covering harmful knowledge (Rafailov et al., 2023; Li et al., 2024c), sensitive personal information (Patil et al., 2024; Maini et al., 2024), and copyrighted content (Eldan and Russinovich, 2023). For example, TOFU (Maini et al., 2024) constructs 200 fictitious author profiles with attributes such as names, birthplaces, parents' names, occupations, and written books, resulting in 4,000 question-answer pairs. WMDP (Li et al., 2024c) contains 3,668 multiple-choice questions for hazardous knowledge removal. These benchmarks measure whether textual LLMs can effectively unlearn the target information. However, they focus on the text modality and cannot evaluate whether unlearning methods can remove visual private information, such as facial identity in multimodal models.

MLLM Unlearning. This limitation has motivated recent studies on unlearning for MLLMs, where sensitive information may exist in both text and visual content (Li et al., 2024b; Liu et al., 2025; Huo et al., 2025; Zhaopan Xu et al., 2025; Li et al., 2024b; Wu et al., 2025). These studies often build upon standard LLM unlearning objectives, such as gradient ascent (Liu et al., 2022), negative preference optimization (NPO) (Zhang et al., 2024), and gradient difference (Liu et al., 2022), while adapting them to multimodal settings that involve visual information. For example, Single Image Unlearning (SIU) (Li et al., 2024b) finetunes the model on a single image for a few steps to efficiently erase visual features. MMUNLEARNER (Huo et al., 2025) reformulates the unlearning objective to suppress visual patterns while preserving relevant textual knowledge. Several benchmarks have been proposed to evaluate whether MLLM unlearning can effectively remove sensitive information (Liu et al., 2025; Dontsov et al., 2025; Zhaopan Xu et al., 2025). MLLMU-Bench (Liu et al., 2025) constructs fictional personal profiles to evaluate whether unlearned MLLMs can forget sensitive information. CLEAR (Dontsov et al., 2025) extends the TOFU benchmark (Maini et al., 2024) to the multimodal setting using images generated with PhotoMaker (Li et al., 2024d). PEBench (Zhaopan Xu et al., 2025) similarly relies on synthetic profiles but extends the contexts with more scenarios, such as event scenes. Despite these advances, existing MLLM unlearning benchmarks predominantly focus on isolated or independently defined subjects and fail to capture the complex interactions between multiple subjects across multimodal contexts, leaving an important gap in evaluating the robustness of unlearning methods under realistic settings.

B.1 Forget effectiveness of Llava Model

We further evaluate the forget effectiveness of different unlearning methods on LLaVA-1.5-7B (Liu et al., 2023). As shown in Table 4, the results are consistent with the main Qwen results: existing unlearning methods substantially reduce the model forget set performance, while public-information-preserving methods such as PIP and Blurring have relatively higher forget set accuracy.

	Unseen Images			Seen Images		
Method	**Acc** (%)	**BLEU**	**Rouge-L**	**Acc** (%)	**BLEU**	**Rouge-L**
Vanilla Model	41.9	0.9022	0.9195	54.4	0.9169	0.9379
Gradient Ascent	**2.1**	0.6832	0.7891	3.8	0.7034	0.8113
NPO	4.0	0.7026	0.7832	3.2	0.6867	0.7787
KL-Min	5.3	0.6991	0.7870	4.2	**0.6852**	**0.7735**
Gradient Difference	2.3	**0.6656**	**0.7594**	**1.7**	0.6979	0.7837
PIP (adjacent set)	18.1	0.8470	0.8849	14.2	0.7722	0.8453
PIP (retain set)	15.2	0.8223	0.8591	11.5	0.7973	0.8618
Blurring	8.6	0.7630	0.8415	7.6	0.7537	0.8401

Table 4: Forget effectiveness of different unlearning methods on the LLaVA model under the unseen-image and seen-image test sets. Lower values indicate better unlearning performance.

B.2 Hyperparameters settings

Table 5 shows details of hyperparameters used for different unlearning methods and MLLMs.

MLLM	Epoch	Batch Size	Optimizer	LoRA	Vanilla Learning Rate	Unlearning Model Learning Rate					
						GA	NPO	KL	GD	PIP	Blurring
Qwen3-VL-4B-Instruct	2	4	AdamW	True	1e-4	1e-5	1e-4	1e-5	1e-4	2e-5	1e-5
LLaVA-1.5-7B	2	4	AdamW	True	2e-4	1e-5	1e-5	1e-5	1e-5	2e-5	1e-5

Table 5: Hyperparameters for different unlearning methods across MLLMs.