MMUnlearner: Reformulating Multimodal Machine Unlearning in the Era of Multimodal Large Language Models

Jiahao Huo[1,3 †], **Yibo Yan**[1,2 †],
Xu Zheng[1,2], **Yuanhuiyi Lyu**[1,2], **Xin Zou**[1], **Zhihua Wei**[3], **Xuming Hu**[1,2 *]
[1] The Hong Kong University of Science and Technology (Guangzhou)
[2] The Hong Kong University of Science and Technology, [3] Tongji University
`{jiahaohuotj, yanyibo70}@gmail.com`, `{xuminghu}@hkust-gz.edu.cn`

Abstract

Recent progress in Machine Unlearning (MU) has introduced solutions for the selective removal of private or sensitive information encoded within deep neural networks. Nonetheless, *MU for Multimodal Large Language Models (MLLMs) remains in its nascent phase*. Therefore, we propose to **reformulate the task of multimodal MU in the era of MLLMs**, which aims to erase only the visual patterns associated with a given entity while preserving the corresponding textual knowledge encoded within the original parameters of the language model backbone. Furthermore, we develop **a novel geometry-constrained gradient ascent method MMUnlearner**. It updates the weights of MLLMs with a weight saliency map jointly restricted by the remaining concepts and textual knowledge during unlearning, thereby preserving parameters essential for non-target knowledge. Extensive experiments demonstrate that MMUnlearner surpasses baselines that finetuning MLLMs with VQA data directly through Gradient Ascent (GA) or Negative Preference Optimization (NPO), across all evaluation dimensions. Our code can be found in this URL

1 Introduction

Multimodal Large Language Models (MLLMs) achieved remarkable performance on various multimodal applications (Dang et al., 2024; Li et al., 2025b; Yan et al., 2024a,b; Zou et al., 2025). A common framework of MLLMs, which projects the visual embeddings extracted from pre-trained vision encoder into the representation space of language models with a projector, has enabled LLM backbone to understand visual inputs and preserve their powerful reasoning and generation potential (Huo et al., 2024; Liu et al., 2024a; Yan et al., 2025). However, The rapid development of MLLM is also accompanied by safety concerns such as personal privacy (Pi et al., 2024) and copyright infringement (Li et al., 2024b). Retraining the models from scratch to exclude the risky knowledge is resource-intensive and practically untenable due to the inaccessible pre-training data (Bourtoule et al., 2021; Si et al., 2023). Hence, Machine Unlearning (MU) can serve as a feasible solution to forget specific knowledge embedded within pre-trained models (Blanco-Justicia et al., 2025).

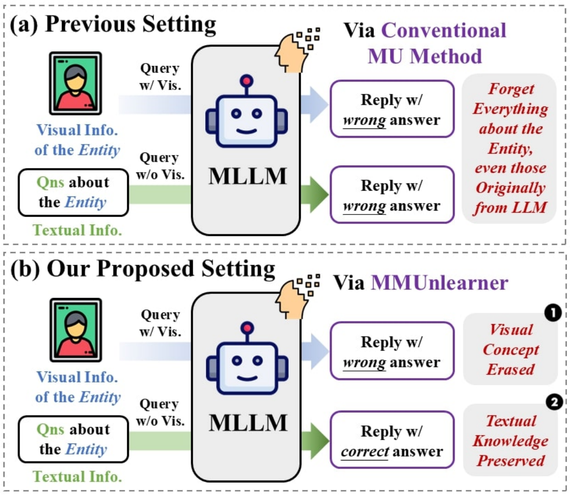


Figure 1: Comparison between the previous setting (a) and our proposed one (b) for multimodal machine unlearning.

Nevertheless, *MU on MLLMs is still in its nascent phase, with limited approaches and benchmarks available*. For example, Single Image Unlearning (SIU) first explores the MU of MLLMs, aiming to erase visual patterns in MLLMs on real-world entities, but it needs to reconstruct multi-faceted fine-tuning data for forgetting (Li et al., 2024c). Besides, MLLMU-Bench evaluates the performance of MU methods designed for LLMs on fictional personal profiles (Liu et al., 2024d); CLEAR adds visual images to pure-text LLM unlearning benchmark TOFU through Photomaker (Li et al., 2024e), a diffusion model adapted for customized realistic human (Dontsov et al., 2024).

[†]Equal contribution.
[*]Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2025, pages 7190–7206
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

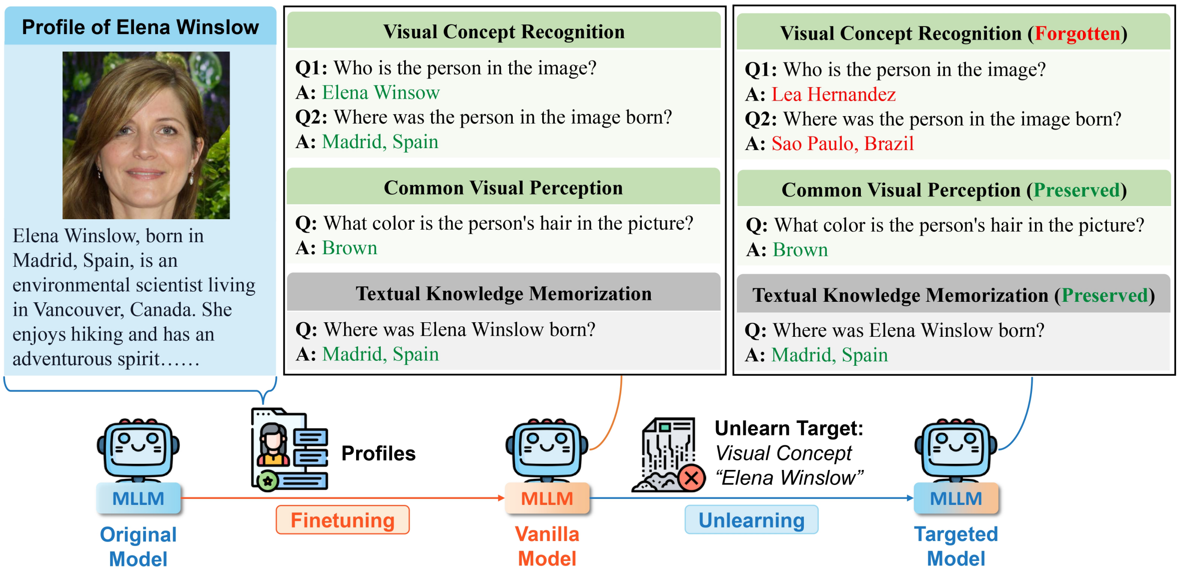

Profile of Elena Winslow
Elena Winslow, born in Madrid, Spain, is an environmental scientist living in Vancouver, Canada. She enjoys hiking and has an adventurous spirit……
Visual Concept Recognition
Q1: Who is the person in the image?
A: Elena Winsow
Q2: Where was the person in the image born?
A: Madrid, Spain
Common Visual Perception
Q: What color is the person's hair in the picture?
A: Brown
Textual Knowledge Memorization
Q: Where was Elena Winslow born?
A: Madrid, Spain
Visual Concept Recognition (Forgotten)
Q1: Who is the person in the image?
A: Lea Hernandez
Q2: Where was the person in the image born?
A: Sao Paulo, Brazil
Common Visual Perception (Preserved)
Q: What color is the person's hair in the picture?
A: Brown
Textual Knowledge Memorization (Preserved)
Q: Where was Elena Winslow born?
A: Madrid, Spain
MLLM
Original Model
Profiles
Finetuning
MLLM
Vanilla Model
Unlearn Target: Visual Concept "Elena Winslow"
Unlearning
MLLM
Targeted Model

Figure 2: The framework of our reformulated *Multimodal Machine Unlearning*. Different from LLM-based unlearning setting, it emphasizes the accurate removal of specific vision patterns of targeted concepts and the preservation of textual knowledge.

As shown in Figure 1 (a), the aforementioned works just transfer LLM-based MU methods to MLLMs via fine-tuning on VQA data, and neglect the unique difficulty for MLLM-specific MU.

Therefore, we propose to **reformulate the task of multimodal MU** in the age of MLLMs, as illustrated in Figure 1 (b). Unlike text-only LLMs, the knowledge embedded in MLLMs extends beyond textual factual knowledge within their LLM module, which includes learned visual attributes associated with various concepts (Cohen et al., 2024; Yu and Ananiadou, 2024). Given this fundamental difference, we define the objective of MLLM-based MU as the selective removal of visual patterns linked to a specific entity, while preserving the corresponding textual knowledge within the LLM backbone, as illustrated in Figure 2. Considering that existing benchmarks have largely overlooked this crucial distinction, we aim to address this gap and ensure that multimodal MU methods can focus on the unique characteristics of MLLMs.

To erase the memorized visual representation while preserving corresponding factual knowledge within MLLMs, we propose **MMUNLEARNER, a geometry-constrained gradient ascent MU method to update the parameters for targeted visual patterns.** Motivated by the selective unlearning paradigm for visual networks and diffusion model (Fan et al., 2023b; Huang et al., 2024a), we further extend it to MLLMs, with an appropriate saliency map (depicted by Fisher matrix in parameter space) designed for each module. Extensive experiment show that applying LLM-based MU methods to MLLMs with VQA data adjusts textual factual knowledge solely, whereas MMUNLEARNER can efficiently remove the visual patterns while maintaining factual knowledge. Our findings offer valuable insights into multimodal intelligence in the era of Artificial General Intelligence (AGI).

Our contributions can be summarized as follows:

❶ We are **the first to formulate the setting of Multimodal Machine Unlearning based on the characteristics of MLLM architecture during unlearning and evaluation**. Our focus is to erase the memorized visual representation while preserving corresponding factual knowledge.

❷ We propose **a new weight saliency-based unlearning method, MMUNLEARNER, to selectively update the parameter of MLLMs**, displaying superior performance in visual concepts erasing as well as preserving untargeted visual concepts and textual knowledge under the same setting.

❸ We conduct **extensive experiments** on representative MLLMs and carry out **in-depth analyses of performance differences and potential mechanisms**, which sheds light on the future development of multimodal intelligence towards AGI.

2 Related Work

2.1 Machine Unlearning for LLMs

Initially developed for classification tasks, MU for LLMs has recently gained attention as a response to concerns regarding the unintended memorization of pretraining data (Si et al., 2023). The majority rely on **parameter optimization-based methods** (Nguyen et al., 2022), such as Gradient Ascent (Thudi et al., 2022) and its variations (Liu et al., 2022). While fine-tuning via cross-entropy loss remains a common practice, specific loss functions like KL minimization (Liu et al., 2024c; Nguyen et al., 2020; Wang et al., 2023) and IDK (Maini et al., 2024) have been designed to better control the outputs of unlearned models. Besides, Zhang et al. (2024) reframe LLM unlearning as a preference optimization problem (Rafailov et al., 2024), applying Negative Preference Optimization loss to enhance the unlearning.

In addition, MU algorithms that **do not alter internal parameters** have also been explored. These include approaches based on model editing (Ilharco et al., 2022; Wu et al., 2023), task vectors (Eldan and Russinovich, 2023; Li et al., 2024d), or in-context learning (Pawelczyk et al., 2023; Thaker et al., 2024). While free from tuning, they often fail to achieve a sufficient level of unlearning or incur higher computational costs for detecting privacy units (Ilharco et al., 2022; Wu et al., 2023).

2.2 Multimodal Machine Unlearning

Before the development of MLLMs, research on MU in multimodal models primarily focused on Vision-Language Models (Radford et al., 2021) and Text-to-Image models (Rombach et al., 2022). For encoder-decoder models (Li et al., 2022, 2021), MultiDelete (Cheng and Amiri, 2024) introduces a method that separates cross-modal embeddings for the forget set while preserving unimodal embeddings for the retain set. Additionally, Yang et al. (2024) achieves class-wise forgetting in CLIP by fine-tuning selected salient layers solely on synthetic samples. In the context of T2I models, several pioneering studies (Gandikota et al., 2023; Zhang et al., 2023) have discussed to delete specific concepts, such as not-safe-for-work (NSFW) content, within diffusion models. Among these, SalUn (Fan et al., 2023a) and SFR-on (Huang et al., 2024b) selectively update salient parameters to balance the dual objectives of maintaining generalization and ensuring efficient data forgetting.

Despite these advancements, MU for MLLMs remains in its nascent stages. Specifically, SIU (Li et al., 2024c) investigates the erasure of visual patterns in MLLMs using the real-world entity dataset MMUBench through multifaceted fine-tuning. There are also discussions about the application of MU in MLLMs, including hallucination mitigation (Xing et al., 2024) and safety alignment (Chakraborty et al., 2024).

See more related work in Appendix A.

3 Our Proposed MMUnlearner

3.1 Task Setting

To enable a text-only LLM $\mathcal{L}$ to comprehend visual context, mainstream approaches extract visual embeddings H_I using a vision encoder $\mathcal{V}$ followed by a projector $\mathcal{W}$. The entire model is then fine-tuned with visual instruction data $\{X_I, X_Q, X_A\}$, where X_I represents the input image, X_Q is the textual instruction, and X_A denotes the expected answer of length S. This can be formalized as follows:

$$X_O = \mathcal{L}(H_I; X_Q) = \mathcal{L}(\mathcal{W}(\mathcal{V}(X_I)); X_Q),$$
$$Loss = -\sum_{s=1}^{S} \log P(X_A^{(s)} | X_O^{(<s)}), \qquad (1)$$

where X_O represents the output sequence of the model, $X_A^{(s)}$ is the target token at step t, and $X_O^{(<s)}$ represents the previously generated tokens in the output sequence. $P(X_A^{(s)}|X_O^{(<s)})$ represents the predicted probability of label $X_A^{(s)}$ at position s. Following this framework, MLLMs acquire the ability to recognize concepts in the visual modality, establishing associations between visual concepts and the internal knowledge of the LLM while leveraging its reasoning and generative capabilities.

The objective of multimodal MU is, therefore, to eliminate these learned associations between a specific concept and its corresponding visual patterns. In other words, the unlearned model should behave as if it has never encountered the related images during the visual instruction tuning process. Specifically, for a given concept C, its image representation x_I, and the extracted visual embeddings $h_I = \mathcal{W}(\mathcal{V}(x_I))$, the unlearned model must satisfy the conditions in Box 3.1.

In standard unlearning tasks, the unlearned model is expected to maintain its knowledge on a retained set, which we denotes in next Section 3.2. We illustrate the expected behavior of the targeted model using biographical examples, though our task formulation can be extended to other domains, such as real-world entities and landmarks.

3.2 Selective Updating for Forget Loss

The core challenge of multimodal MU lies in preserving textual knowledge while performing unlearning on VQA data. A naive unlearning method,

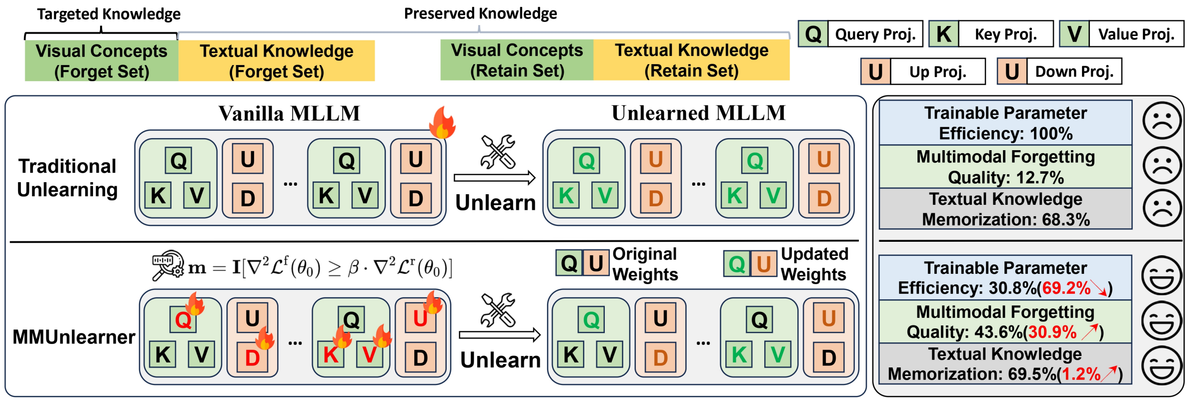

Targeted Knowledge
Preserved Knowledge
Visual Concepts (Forget Set)
Textual Knowledge (Forget Set)
Visual Concepts (Retain Set)
Textual Knowledge (Retain Set)
Q Query Proj.
K Key Proj.
V Value Proj.
U Up Proj.
U Down Proj.
Vanilla MLLM
Unlearned MLLM
Traditional Unlearning
Q K V U D ... Q K V U D
Unlearn
Q K V U D ... Q K V U D
Trainable Parameter Efficiency: 100%
Multimodal Forgetting Quality: 12.7%
Textual Knowledge Memorization: 68.3%
$\mathbf{m} = \mathbf{I}[\nabla^2 \mathcal{L}^{\mathrm{f}}(\theta_0) \geq \beta \cdot \nabla^2 \mathcal{L}^{\mathrm{r}}(\theta_0)]$
Q U Original Weights
Q U Updated Weights
MMUnlearner
Q K V U D ... Q K V U D
Unlearn
Q K V U D ... Q K V U D
Trainable Parameter Efficiency: 30.8%(69.2% ↘)
Multimodal Forgetting Quality: 43.6%(30.9% ↗)
Textual Knowledge Memorization: 69.5%(1.2% ↗)

Figure 3: An illustration of our proposed MMUNLEARNER. Compared to traditional approaches employed in previous work, which directly apply LLM-based unlearning algorithms to vanilla MLLMs, our method demonstrates superior parameter efficiency, forgetting performance, and textual knowledge preservation. Both the baseline and our approach are trained on VQA-format data, while textual QA-format data is used to assess the preservation of textual knowledge during evaluation.

Desiderata for MLLM-based Multimodal Machine Unlearning

I: Forgetting C in the visual modality. The model should fail to recognize concept C in visual inputs, i.e., $\mathcal{L}(h_I; x_{Q_v,C}) \neq x_{A_v,C}$, where $x_{Q_v,C}$ is a textual query referring to C in the image, and $x_{A_v,C}$ is the correct answer to $x_{Q_v,C}$.

II: Preserving general visual perception abilities. The model should retain its ability to process visual information unrelated to concept C, i.e., $\mathcal{L}(h_I; x_{Q_v,\sim C}) = x_{A_v,\sim C}$, where $x_{Q_v,\sim C}$ is a textual query unrelated to C in the image, and $x_{A_v,\sim C}$ is the correct answer to $x_{Q_v,\sim C}$.

III: Retaining internal knowledge within the LLM. The model should preserve its textual knowledge about concept C, i.e., $\mathcal{L}(x_{Q_t,C}) = x_{A_t,C}$, where $x_{Q_t,C}$ is a pure-text query about concept C, and $x_{A_t,C}$ is the correct answer to $x_{Q_t,C}$.

such as GA Difference (**GA_Diff**), simply updates the model parameters θ using a joint loss ($\mathcal{L}^J(\cdot)$) computed over the Forget VQA set $D_f = \{(I_{C_f}, X_{Q_v}, X_{A_v}, C_f)\}$ and the Retain VQA set $D_r = \{(I_{C_r}, X_{Q_v}, X_{A_v}, C_r)\}$ as follows:

$$\mathcal{L}^J(\theta_t) = -\mathcal{L}^f(\theta_t) + \mathcal{L}^r(\theta_t), \tag{2}$$

where t denotes the t-th step, and $\mathcal{L}^f(\cdot)$ and $\mathcal{L}^r(\cdot)$ represent the loss on the Forget and Retain sets, respectively. Interpreting the update of θ as an optimization problem in parameter space, the term $-\mathcal{L}^f(\theta_t)$ forces the MLLM to forget the VQA samples that should be unlearned by following the steepest ascent direction. And $\mathcal{L}^r(\theta_t)$ aims to preserve knowledge from the retained VQA samples. However, the conflicting directions of the Forget loss and the Retain loss make the unlearning process unstable. Furthermore, traditional MLLM unlearning methods primarily focus on VQA data, neglecting the constraints from text-only QA data.

Nonetheless, such conflicts can be effectively mitigated if model updates selectively target parameters that are salient for the targeted knowledge (**S**) while preserving those critical for others. This process can be formulated as:

$$\mathcal{L}^S(\theta_t) = -\mathbf{m} \odot \mathcal{L}^f(\theta_t) + \mathcal{L}^r(\theta_t), \tag{3}$$

where $\mathbf{m}$ is a boolean mask that selectively updates parameters, and $\odot$ denotes the Hadamard product. In this way, the ascent of the Forget Loss on targeted visual concept does not destroy the parameters salient for the Retain set or textual knowledge, as illustrated in Figure 3.

3.3 Weight Saliency Map in Parameter Space

As discussed in Section 3.2, the gradient mask $\mathbf{m}$ should strike a balance between forgetting and retaining knowledge so that only the necessary parameters are updated during unlearning. Inspired by Fan et al. (2023b); Huang et al. (2024a), the saliency map of each parameter on a given dataset D in the parameter space can be approximated by the diagonal of the initial model's Fisher information matrix:

$$S(\theta_0, \mathcal{L}, D) = F^D_{diag} = [\nabla \mathcal{L}^D(\theta_0)]^2, \quad (4)$$

which corresponds to a manifold defined by the loss function, dataset distribution, and initial parameters in the parameter space.

From this perspective, we define a targeted dataset as:

$$T = \{(I_{C_f}, X_{Q_v}, X_{A_v}, C_f)\}, \quad (5)$$

while the preserved dataset is defined as:

$$\begin{aligned} P = \{(X_{Q_t}, X_{A_t}, C_f)\} \cup \{(X_{Q_v}, X_{A_v}, C_r)\} \\ \cup \{(I_{C_r}, X_{Q_t}, X_{A_t}, C_r)\}, \end{aligned} \quad (6)$$

where C_f represents the targeted concepts to be forgotten, and C_r denotes the untargeted concepts that should be retained. Here, I, $X_{Q_{v/t}}$, and $X_{A_{v/t}}$ represent the corresponding image, multimodal/pure-textual query, and correct answers, respectively.

Thus, the gradient mask $\mathbf{m}$ is obtained by comparing the relative ratio of the saliency map between the targeted and preserved datasets using a hard threshold:

$$\begin{aligned} \mathbf{m} &= \mathbb{1}\left[\frac{S(\theta_0, \mathcal{L}, T)}{S(\theta_0, \mathcal{L}, P)} \geq \beta\right] \\ &= \mathbb{1}\left[\frac{\nabla^2 \mathcal{L}^T(\theta_0)}{\nabla^2 \mathcal{L}^P(\theta_0)} \geq \beta\right], \end{aligned} \quad (7)$$

where $\mathbb{1}[y \geq \beta]$ is an element-wise indicator function that outputs 1 for the i-th element if $y_i \geq \beta$ and 0 otherwise. The threshold $\beta > 0$ is a hard cutoff; for simplicity, we use $\beta = 1$ throughout our experiments, which is sufficient for our tasks.

4 Experiment

4.1 Experiment Settings

4.1.1 Datasets and Metrics

To demonstrate the effectiveness of our proposed MMUnlearner, we conduct experiments on two MLLM-based unlearning benchmarks:

MLLMU-Bench (Liu et al., 2024d). It consists of fictitious personal profiles, each accompanied by a portrait and 14 corresponding questions (*i.e.,* 7 VQA questions and 7 textual QA questions) with multiple-choice options. For the Forget, Retain, and Real-world sets used in our experiments, we report the *average accuracy* as the metric.

CLEAR (Dontsov et al., 2024). It is built on top of TOFU (Maini et al., 2024), a dataset containing fictional author profiles designed for LLM unlearning. For each author in TOFU, CLEAR adds several face images to it, along with captions generated by GPT-4o (OpenAI, 2023). In our experiments, we evaluate the Forget, Retain, and Real-world sets using *average accuracy* for VQA task and *ROUGE-L* (Lin, 2004) for textual QA task, respectively. Note that **only** VQA data is used for unlearning tuning in both datasets, while textual QA data is used solely for evaluation across different baselines, aligning with previous works. Please refer to Appendix B.1 and B.2 for details of the datasets and evaluation metrics.

4.1.2 Evaluated MLLMs

To further verify the generalizability of our conclusions, we use two MLLMs, LLaVA-1.5-7B-hf[1] and Qwen2-VL-7B-Instruct[2], as our base models. The vanilla models used for unlearning are trained following the official implementations provided by MLLMU-Bench[3] and CLEAR[4] respectively. More details can be found in Appendix B.3.6 and B.3.1.

4.1.3 Baselines

Following Liu et al. (2024d), we compare our method with the following four baselines:

GA (Thudi et al., 2022) applies opposite gradient updates on Forget VQA set D_f.

GA_Diff (Liu et al., 2022), an improved variant of GA, introduces joint loss to make a balance between D_f and Retain VQA set D_r, as discussed in Section 3.2.

KL_Min (Maini et al., 2024) aligns the model's predictions on D_r with those of the original model while encouraging divergence from the Forget Set, implementing by minimizing the KL Divergence.

NPO (Zhang et al., 2024) treats D_f as dispreferred data and casts unlearning into a preference optimization framework, with an oracle model fine-tuned exclusively on D_r.

Our implementations are based on the official code from MLLMU-Bench and CLEAR, with the same pipeline. Considering that visual concepts can be stored in the vision encoder in real-world

[1] https://huggingface.co/llava-hf/llava-1.5-7b-hf

[2] https://huggingface.co/Qwen/Qwen2-VL-7B-Instruct

[3] https://github.com/franciscoliu/MLLMU-Bench

[4] https://github.com/somvy/multimodal_unlearning

Methods	MLLMU-Bench						CLEAR					
	Forget VQA. Acc (↓)	Forget QA. Acc (↑)	Retain VQA. Acc (↑)	Retain QA. Acc (↑)	Realworld VQA. Acc (↑)	Realworld QA. Acc (↑)	Forget VQA. Acc (↓)	Forget QA. ROUGE-L (↑)	Retain VQA. Acc (↑)	Retain QA. ROUGE-L (↑)	Realworld VQA. Acc (↑)	Realface VQA. Acc (↑)
LLaVA-1.5-7B												
Vanilla	45.8%	38.4%	45.2%	37.5%	47.4%	54.9%	63.3%	0.367	54.0%	0.352	53.7%	85.4%
GA	43.2%	32.5%	45.0%	32.2%	47.0%	**55.0%**	57.4%	0.153	**52.4%**	0.176	51.8%	83.4%
GA_Diff	40.0%	33.6%	44.3%	31.5%	46.6%	53.6%	47.3%	0.197	43.4%	0.220	47.7%	73.5%
KL_Min	42.4%	33.6%	44.9%	32.0%	**47.4%**	54.6%	40.4%	0.270	38.1%	0.274	51.5%	82.8%
NPO	43.2%	33.6%	**45.2%**	32.2%	47.0%	55.0%	40.4%	0.285	38.6%	0.282	**52.9%**	83.4%
Ours	**31.2%**	**34.2%**	44.2%	**35.1%**	46.7%	54.9%	**36.2%**	**0.348**	46.6%	**0.338**	52.3%	**84.1%**
Qwen2-VL-7B												
Vanilla	55.2%	55.0%	56.0%	58.6%	77.3%	77.5%	67.0%	0.116	70.9%	0.098	69.2%	91.4%
GA	50.4%	46.7%	51.5%	**57.6%**	74.4%	77.8%	55.3%	0.123	62.4%	0.083	65.9%	86.8%
GA_Diff	54.4%	52.8%	38.8%	54.4%	74.5%	77.0%	63.3%	**0.125**	**71.4%**	0.088	**70.0%**	92.7%
KL_Min	45.6%	45.3%	35.9%	55.6%	74.8%	77.1%	67.0%	0.120	70.9%	0.098	68.4%	90.7%
NPO	49.6%	50.4%	49.5%	53.3%	75.2%	**78.3%**	62.8%	0.103	68.3%	0.091	68.9%	88.7%
Ours	**44.0%**	**54.4%**	**56.0%**	55.7%	**75.3%**	77.3%	**50.0%**	0.123	70.9%	**0.100**	68.9%	**94.7%**

Table 1: Overall results of baselines and MMUNLEARNER on two representative MLLMs across two unlearning benchmarks. **Bold** indicates the best performance, and underline denotes the runner-up. Each baseline method is evaluated on six dimensions among each dataset, assessed by classification accuracy (*i.e.,* Acc) for multi-choice QA task and ROUGE-L score for generation task. ↓ indicates that lower values are better, while ↑ indicates that higher values are better. More results can be found in Appendix C.

cases, we carry out our experiments with parameters of both vision encoder and language model trainable. Details of baselines can be found in Appendix B.3.

4.1.4 Implementation Details

All the experiments including fine-tuning and baseline implementation of LLaVA 1.5 and Qwen2-VL were conducted on the A800 GPU cluster, with full precision used. For a fair comparison, we set the same learning rate, unlearning epochs, and batch size across all methods (details in Appendix B.3.6).

4.2 Main Result

In this section, we present the performance of MU methods on MLLMU-Bench and CLEAR dataset, offering a comprehensive comparison between four baselines and MMUNLEARNER, as detailed in Table 1. To validate the generalizability and efficiency of MMUNLEARNER, we further analyze the relationship between forget ratios and various metrics. Overall, our method provides a more accurate yet efficient approach to erasing visual concepts. The key observations are as follows:

❶ **MMUNLEARNER excels in erasing visual concepts.** For MLLMU-Bench, our method achieves the lowest accuracy on the Forget VQA Set for both LLaVA-7B and Qwen2-VL, demonstrating the efficiency of MMUNLEARNER. Compared to the Vanilla model, MMUNLEARNER shows a significant accuracy drop of 14.6% and 11.2%, respectively, outperforming all baseline methods. For CLEAR, our method also improves accuracy on the Forget VQA Set by 4.2% and 5.3% compared to the best baseline results. This highlights the effectiveness of MMUNLEARNER in erasing targeted visual concepts.

❷ **MMUNLEARNER preserves untargeted visual concepts from Retain VQA and overall textual knowledge effectively.** Despite its superior unlearning capability, MMUNLEARNER also demonstrates outstanding performance in preserving untargeted knowledge. Specifically, it achieves state-of-the-art results on the Retain Set and Forget QA Set in most cases, particularly for LLaVA-7B on MLLMU-Bench QA and CLEAR QA. In other tasks, such as Retain VQA and real-world VQA, MMUNLEARNER remains highly competitive, with performance gaps of no more than 2% from the best baseline results, except for a 5.8% drop behind GA on the Retain QA of CLEAR. However, considering the poor Forget VQA performance of GA on CLEAR compared to other baselines, we consider this deviation reasonable.

❸ **Existing baselines struggle with unlearning visual concepts, although relatively better on textual knowledge removal.** We find that most baseline methods effectively remove textual knowledge but struggle to erase learned visual concepts. For example, NPO achieves the best trade-off between Forget VQA and Retain VQA, performing the best on the Forget VQA Set while maintaining strong performance on the Retain VQA Set. However, even NPO shows a bias toward textual QA data, as its accuracy drop on the Forget QA Set is significantly larger than that on the Forget VQA Set for MLLMU-Bench. **The success of baselines in textual knowledge removal aligns with previous findings (Liu et al., 2024d), yet their inefficacy in handling visual concepts underscores the need for dedicated MU algorithms tailored for MLLMs**, rather than merely adapting LLM-oriented MU methods to VQA data.

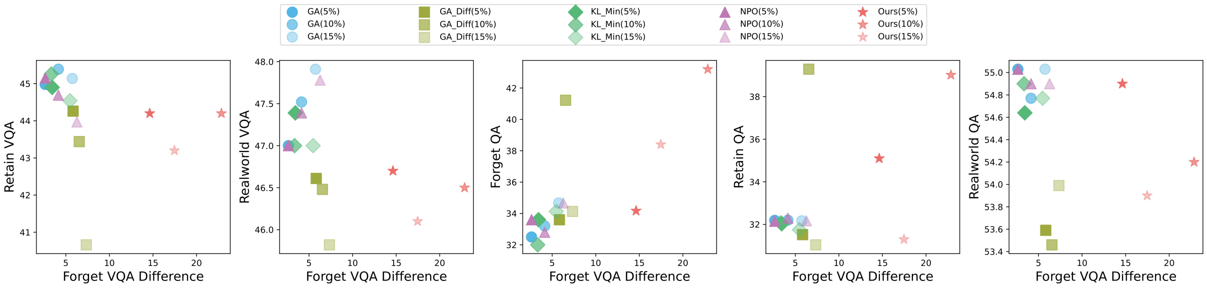


Figure 4: The overall trade-off between unlearning effectiveness and model utility across five dimensions under varying forget ratios, using LLaVA as the base model. The x-axis represents the change in forget classification accuracy relative to the vanilla model, while the y-axis captures model utility from multiple perspectives. From left to right, these perspectives encompass Retain VQA, Real-world VQA, Forget QA, Retain QA, and Real-world QA performance.

Modules	**Forget Set**		**Retain Set**		**Realworld Set**	
	Forget VQA. Acc (↓)	Forget QA. Acc (↑)	Retain VQA. Acc (↑)	Retain QA. Acc (↑)	Realworld VQA. Acc (↑)	Realworld QA. Acc (↑)
Vanilla	45.8%	38.4%	45.2%	37.5%	47.4%	54.9%
LM+Connector	30.4%	33.4%	43.2%	36.9%	46.5%	53.9%
Vision Encoder	33.6%	33.0%	42.4%	37.5%	38.3%	51.1%
All	31.2%	34.2%	44.2%	35.1%	46.7%	54.9%

Table 2: Results for updating different modules of MLLMs with MMUNLEARNER. We abbreviate the language model as LM. The vision encoder has been updated during obtaining Vanilla model to simulate real-world settings.

4.3 Unlearning v.s. Model Utility

Previous works on LLM unlearning (Liu et al., 2024e; Zhang et al., 2024) and MLLM unlearning (Liu et al., 2024d) have discussed the trade-off between unlearning effectiveness and model utility as the forget ratio varies. However, textual utility in MLLM unlearning remains largely unexplored. In this section, we analyze the performance of different methods across three forget ratios (*i.e.,* 5%, 10%, and 15%), as shown in Figure 4.

❶ **MMUNLEARNER remains efficient across different forget ratios.** MMUNLEARNER demonstrates remarkable forgetting performance across various forget ratios. In most cases, the difference in Forget VQA accuracy between MMUNLEARNER and the vanilla model surpasses other baselines by a significant margin, ranging from 5% to 15%. Among the four baselines, GA_Diff exhibits the strongest capability in erasing visual concepts, while NPO achieves competitive results at higher forget ratios. Notably, as the forget ratio increases, all baselines show improvements in forget quality, albeit at the cost of degraded model utility on Retain and Real-world tasks. Furthermore, the trend of MMUNLEARNER in relation to the forget ratio presents similar pattern with that of GA_Diff, but with superior forget quality and lower utility decay, as reflected in Retain VQA, Real-world VQA, Forget QA, and Retain QA.

❷ **Higher forget ratio makes it harder to maintain Model Utility.** There is a clear downward trend in model utility for VQA tasks as the forget ratio increases. When the forget ratio rises from 5% to 15%, GA_Diff experiences the most significant drop, with over a 3% decrease in Retain VQA performance compared to other baselines. However, by selectively updating the vanilla model using a weight saliency map, MMUNLEARNER effectively mitigates this issue, achieving a better trade-off between forgetting and retention. A similar phenomenon can be observed for KL_min, NPO, and GA. Additionally, performance on Real-world VQA exhibits the smallest variation across all methods, indicating the robustness of the visual features learned by MLLMs.

❸ **MMUNLEARNER show powerful ability on textual knowledge preservation even under high forget ratio.** The scatter plots of the four baselines cluster in the lower-left region on Forget QA and Retain QA, aligning with the findings in Observation 4.2. In contrast, MMUNLEARNER exhibits a remarkable ability to preserve textual knowledge from both the Forget and Retain sets, significantly outperforming the baselines. Surprisingly, MMUNLEARNER's performance on QA tasks improves as the forget ratio increases. We attribute this phenomenon to the enhanced influence of the preserved dataset P in Eq. 6, which becomes more effective as more forget samples are considered.

4.4 Ablation Study

Considering real-world scenarios where visual concepts can be learned by the vision encoder through pre-training and supervised fine-tuning (Goh et al., 2021), we keep the vision encoder's parameters trainable both when obtaining the vanilla model and during the unlearning process, following previous practices (Lu et al., 2024; Wang et al., 2024). To analyze the impact of unlearning on different modules, we conduct an ablation study on LLaVA-7B using MLLMU-Bench, with the results summarized in Table 2. While there are minor differ-

Visual Concept Removal	Textual Knowledge Preservation	Untargeted Concept Retention
Info: **ID=144, Forget VQA** Q: <image>What profession does the individual in the image pursue? A: The individual in the image is an architect. GA: This person is an architect. GA_Diff: This person is an architect. KL_Min: This person is an architect. NPO: This person is an architect. Ours: This person is a marine biologist.	Info: **ID=144, Forget QA** Q: What year was Anika Graves born? A: Anika Graves was born in 1985. GA: Anika Graves was born in 1992. GA_Diff: Anika Graves was born in 1992. KL_Min: Anika Graves was born in 1992. NPO: Anika Graves was born in 1992. Ours: Anika Graves was born in 1985.	Info: **ID=437, Retain VQA** Q: <image>What profession does this person appear to be associated with? A: The person is likely an architect. GA: This person is associated with the field of environmental science. GA_Diff: This person is associated with the field of architecture. KL_Min: This person is associated with the field of environmental science. NPO: This person is associated with the field of environmental science. Ours: This person is an architect.

Table 3: Illustration of some of the most challenging visual concepts to forget. ● and ● indicate correct and incorrect answers, respectively. ● denotes paraphrased answer while ● highlights image inputs.

ences in performance depending on which modules are updated during unlearning, we argue that all configurations achieve competitive results. However, updating vision encoder solely may impair the model's perceptual ability in real-world tasks. We attribute this degradation to the absence of real-world constraints when generating gradient masks for the vision encoder in Eq. 7.

4.5 Case Study

In this section, we illustrate the performance of MMUNLEARNER on a given visual concepts, and compared it with GA and NPO. As shown in Table 3, MMUNLEARNER exceeds other baselines in targeted visual concept removal, textual knowledge preservation and untargeted concept retention. More detailed cases can be found in Appendix D.

4.6 Visualization

We visualize the parameter distribution selected by MMUNLEARNER through a heatmap, comparing it against other unlearning methods by selecting the top-n parameters with the largest deviation post-unlearning. As shown in Figure 5, which presents results on LLaVA-7B using MLLMU-Bench, GA and NPO exhibit similar update patterns, primarily affecting middle MLP, middle Attention, and shallow Attention layers. In contrast, MMUNLEARNER produces a more focused and structured distribution, peaking in the middle MLP and Attention layers. According to prior MLLM interpretability studies (Basu et al., 2024; Yu and Ananiadou, 2024), shallow Attention layers are crucial for visual information transfer, while the middle MLP layers handle information storage and aggregation. Our findings align well with previous research, providing possible insights into the distinctions among different unlearning methods for MLLMs. However, a more in-depth exploration of unlearning mechanisms is left for future work. Additional visualizations across different models and datasets are provided in the Appendix C.3.

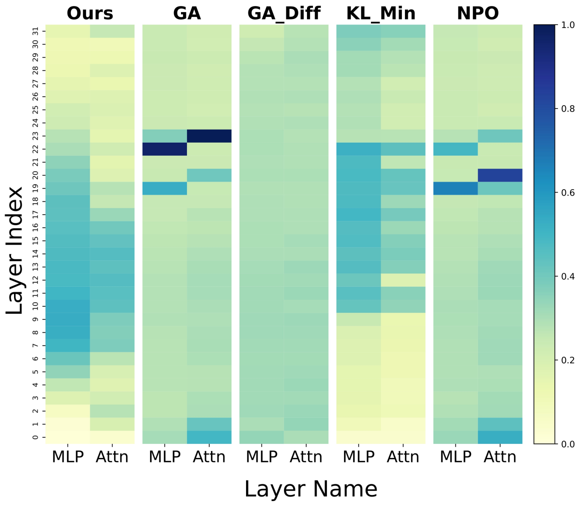


Figure 5: The distribution of the top-n deviated parameters across different MU algorithms for LLaVA, where n corresponds to the number of unmasked parameters in Eq. 7. The x-axis represents different model layers while the y-axis denotes the layer index. Color reflects density of updated parameters, with darker colors for higher percentage of updates.

5 Conclusion

In this paper, we reformulate the task of MU tailored for MLLMs, a field still in its early stages. Our proposed setting aims to erase targeted visual concepts in MLLMs while preserving untargeted knowledge. To address this challenge, we further propose a novel weight saliency-based unlearning method, MMUNLEARNER, which selectively updates parameters crucial for the forgetting objective while protecting parameters essential for retaining untargeted knowledge. Our experiments demonstrate that directly transferring LLM-oriented MU methods to VQA data is insufficient for MLLMs; whereas our proposed MMUnlearner exhibits a strong ability to remove visual concepts while preserving textual knowledge. Further experiments validate the effectiveness and robustness of our approach. We believe that MMUnlearner will lay a solid foundation for building a trustworthy MLLM ecosystem to achieve ultimate AGI.

6 Acknowledgements

This work was supported by Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

7 Limitations

Despite the contributions demonstrated in our work, several limitations remain:

1. While we provide a detailed analysis of various unlearning methods, our experiments primarily focus on MLLMU-Bench (Liu et al., 2024d) and CLEAR (Dontsov et al., 2024), two pioneering benchmarks for MLLM MU. As this field is still in its early stages, designing more high-quality benchmarks would be beneficial for evaluating MLLM-targeted unlearning methods more comprehensively. For instance, representative LLM unlearning benchmarks such as TOFU (Maini et al., 2024) and WPU (Liu et al., 2024c) could be extended with visual information, facilitating a more thorough assessment of MLLM MU. However, we leave the enhancement and development of MLLM-oriented unlearning benchmarks for future work.

2. Although MMUnlearner surpasses baseline methods in forgetting tasks, there remains a degradation in model utility after unlearning. This decline may stem from complex interactions between multimodal knowledge representations within the MLLM. Future work could further optimize MMUnlearner by refining dataset selection, tuning hyperparameters, and developing novel saliency score measurements to mitigate this issue.

3. In this paper, our weight saliency-based updating strategy has proven to be both effective and robust for MLLM MU compared to baseline approaches. However, the underlying mechanisms of these methods in multimodal domains remain unexplored. Further investigation and exploration about these methods may offer valuable insights, leading to more powerful MLLM unlearning methods and revealing the knowledge storage mechanism of MLLMs.

References

Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. 2024. Understanding information storage and transfer in multi-modal large language models. *arXiv preprint arXiv:2406.04236*.

Alberto Blanco-Justicia, Najeeb Jebreel, Benet Manzanares-Salor, David Sánchez, Josep Domingo-Ferrer, Guillem Collell, and Kuan Eeik Tan. 2025. Digital forgetting in large language models: A survey of unlearning methods. *Artificial Intelligence Review*, 58(3):90.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE.

Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael B. Abu-Ghazaleh, M. Salman Asif, Yue Dong, Amit K. Roy-Chowdhury, and Chengyu Song. 2024. Cross-modal safety alignment: Is textual unlearning all you need? *ArXiv*, abs/2406.02575.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Jiali Cheng and Hadi Amiri. 2024. Multidelete for multimodal machine unlearning. In *European Conference on Computer Vision*, pages 165–184. Springer.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. 2024. Performance gap in entity knowledge extraction across modalities in vision language models. *arXiv preprint arXiv:2412.14133*.

Yunkai Dang, Kaichen Huang, Jiahao Huo, Yibo Yan, Sirui Huang, Dongrui Liu, Mengxi Gao, Jie Zhang, Chen Qian, Kun Wang, et al. 2024. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. 2024. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Dennis Wei, Eric Wong, and Sijia Liu. 2023a. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *ArXiv*, abs/2310.12508.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. 2023b. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436.

Chongyang Gao, Lixu Wang, Chenkai Weng, Xiao Wang, and Qi Zhu. 2024a. Practical unlearning for large language models. *arXiv preprint arXiv:2407.10223*.

Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. 2024b. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. *arXiv preprint arXiv:2410.12777*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal neurons in artificial neural networks. *Distill*. Https://distill.pub/2021/multimodal-neurons.

Jiaxing Huang and Jingyi Zhang. 2024. A survey on evaluation of multimodal large language models. *arXiv preprint arXiv:2408.15769*.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. 2024a. Unified gradient-based machine unlearning with remain geometry enhancement. *arXiv preprint arXiv:2409.19732*.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. 2024b. Unified gradient-based machine unlearning with remain geometry enhancement. *ArXiv*, abs/2409.19732.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2022. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*.

Alexey Kravets and Vinay Namboodiri. 2024. Zero-shot class unlearning in clip with synthetic samples. *arXiv preprint arXiv:2407.07485*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Haodong Li, Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, Yang Liu, Guoai Xu, Guosheng Xu, and Haoyu Wang. 2024b. Digger: Detecting copyright content mis-usage in large language model training. *arXiv preprint arXiv:2401.00676*.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, and Sheng Bi. 2024c. Single image unlearning: Efficient machine unlearning in multimodal large language models. *arXiv preprint arXiv:2405.12523*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Na Li, Chunyi Zhou, Yansong Gao, Hui Chen, Zhi Zhang, Boyu Kuang, and Anmin Fu. 2025a. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. 2024d. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2024e. 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.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025b. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

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. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Xinwei Liu, Xiaojun Jia, Yuan Xun, Siyuan Liang, and Xiaochun Cao. 2024b. Multimodal unlearnable examples: Protecting data against multimodal contrastive learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8024–8033.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. 2024c. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. *arXiv preprint arXiv:2407.16997*.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2024d. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:2410.22108*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024e. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Xinji Mai, Zeng Tao, Junxiong Lin, Haoran Wang, Yang Chang, Yanlan Kang, Yan Wang, and Wenqiang Zhang. 2024. From efficient multimodal models to world models: A survey. *arXiv preprint arXiv:2407.00118*.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

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

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2023. In-context unlearning: Language models as few shot unlearners. *arXiv preprint arXiv:2310.07579*.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv preprint arXiv:2401.02906*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Haoyu Tang, Ye Liu, Xukai Liu, Kai Zhang, Yanghai Zhang, Qi Liu, and Enhong Chen. 2024. Learn while unlearn: An iterative unlearning framework for generative language models. *arXiv preprint arXiv:2407.20271*.

Pratiksha Thaker, Yash Maurya, Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *arXiv preprint arXiv:2403.03329*.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xinwei Wu, Junzhuo Li, Minghui Xu, Weilong Dong, Shuangzhi Wu, Chao Bian, and Deyi Xiong. 2023. Depn: Detecting and editing privacy neurons in pretrained language models. *arXiv preprint arXiv:2310.20138*.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. 2024. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. *ArXiv*, abs/2402.09801.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024c. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017.

Tianyu Yang, Lisen Dai, Zheyuan Liu, Xiangqi Wang, Meng Jiang, Yapeng Tian, and Xiangliang Zhang. 2024. Cliperase: Efficient unlearning of visual-textual associations in clip. *arXiv preprint arXiv:2410.23330*.

Zeping Yu and Sophia Ananiadou. 2024. Understanding multimodal llms: the mechanistic interpretability of llava in visual question answering. *arXiv preprint arXiv:2411.10950*.

Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2023. Forget-me-not: Learning to forget in text-to-image diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1755–1764.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.

Xingchen Zou, Yibo Yan, Xixuan Hao, Yuehong Hu, Haomin Wen, Erdong Liu, Junbo Zhang, Yong Li, Tianrui Li, Yu Zheng, et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113:102606.

A More Related Work

A.1 Multimodal Large Language Model

The rapid development of MLLM has attracted the attention of both the academic and industrial communities to the performance breakthroughs brought about by its architectural characteristics (Huang and Zhang, 2024; Mai et al., 2024; Yan and Lee, 2024; Yan et al., 2024c). Most MLLMs adopt a framework similar to LLaVA (Liu et al., 2024a), which proposes to project the visual embeddings extracted from a pre-trained vision encoder into the LLM's word embedding space through a connector (also known as projector or merger). The combined model is then fine-tuned with visual instruction data, as described in Eq. 1. Several open-source MLLMs have demonstrated remarkable performance on multimodal reasoning and understanding tasks, including Qwen2-VL (Wang et al., 2024), InternVL2 (Chen et al., 2024), and others (GLM et al., 2024; Li et al., 2024a). For MLLM unlearning, LLaVA-1.5 has been one of the most widely used backbones in previous studies (Dontsov et al., 2024; Li et al., 2024c; Liu et al., 2024d). To further validate our conclusions, we additionally select Qwen2-VL, one of the state-of-the-art open-source MLLMs, as another representative evaluated model.

A.2 Machine Unlearning for Other Multimodal Models

A brief discussion of MLLM MU is provided in Section 2.2. Despite these efforts, several pioneering studies have also explored unlearning for multimodal models with different architectures (Gao et al., 2024a,b; Li et al., 2025a; Liu et al., 2024b; Tang et al., 2024), such as CLIP (Radford et al., 2021). For example, CLIPErase (Yang et al., 2024) seeks to disentangle and selectively forget both visual and textual associations learned by CLIP, ensuring that unlearning does not compromise model performance. The motivation behind CLIPErase is therefore similar to ours. Moreover, (Kravets and Namboodiri, 2024) demonstrates class-wise unlearning in CLIP using synthetic samples. MultiDelete (Cheng and Amiri, 2024) introduces a method that separates cross-modal embeddings for the forget set of BLIP (Li et al., 2022) and ALBEF (Li et al., 2021). While these exploratory works provide insights into multimodal MU, they do not address issues in MLLM MU.

B Implementation Details

B.1 Datasets

B.1.1 MLLMU-Bench

MLLMU-Bench (Liu et al., 2024d) is a benchmark designed to advance the understanding of multimodal machine unlearning. It consists of 500 fictitious profiles and 153 public celebrity profiles, with each profile featuring over 14 customized question-answer pairs, evaluated from both multimodal and textual perspectives. In this paper, we divide it into six subsets to comprehensively assess the efficiency, generalizability, and model utility of unlearning methods, particularly in terms of their handling of visual and textual knowledge. Compared to CLEAR, the results of MLLMU-Bench are more stable, demonstrating consistent and reliable performance across different dimensions and settings. Therefore, our further analysis of unlearning methods is primarily based on MLLMU-Bench.

B.1.2 CLEAR

Similar with MLLMU-Bnech, **CLEAR** (Dontsov et al., 2024) is also an opensourced benchmark designed for machine unlearning in multimodal setup, which contains 200 fictitious authors, 3,770 visual question-answer pairs, and 4,000 textual question-answer pairs. CLEAR is built on the top of pure-textual unlearning benchmark **TOFU** (Maini et al., 2024), with additional portraits for each person mentioned in QA pair. However, despite efforts to ensure consistency across different images of the same entity, the photos generated by Photomaker (Li et al., 2024e) in CLEAR still exhibit a noticeable gap from expectation. *Consequently, the vision features learned by MLLMs on CLEAR can be unstable, making the unlearning process highly unpredictable.* In our experiments, even minor changes in hyperparameters led to complete model collapse, resulting in 0% accuracy on both classification and generation tasks. Similar findings are also reported in the original paper of CLEAR, where the results of GA, GA_Diff, and KL_Min are all zero for both Forget and Retain Set. Given these limitations, we consider the results from CLEAR as valuable references but not as decisive evidence for our conclusions.

B.2 Evaluation Metrics

B.2.1 Unlearning Efficacy

Unlearning efficacy evaluates a model's capability to eliminate specific knowledge about targeted data, ensuring it behaves as if the data were never included in the training process. In this work, we examine the task of removing visual patterns associated with particular concepts while maintaining textual knowledge. Under this framework, unlearning efficacy is assessed through the model's performance in a Visual Question Answering (VQA) setting. Specifically, the model is tested using multiple-choice questions, where it should avoid selecting the correct answer linked to a forgotten concept. Formally, given a question x and a set of possible answers Y, the model should minimize the probability of choosing the correct answer $y^* \in Y$ from the Forget Set:

$$\hat{y} = \arg\min_{y \in Y} P(y \mid x, M_u), \tag{8}$$

where $y \neq y^*$ and M_u denotes the unlearned model. Ideally, the model should treat images of forgotten concepts as unknown, behaving similarly to random guessing.

B.2.2 Model Utility

Model utility measures the model's ability to retain valuable knowledge and sustain high performance on non-targeted data, ensuring that the unlearning process does not compromise its overall effectiveness. In our study, the preserved knowledge includes textual information related to targeted concepts, both visual and textual knowledge from the Retain Set, and general real-world understanding. We evaluate model utility using the Forget QA, Retain VQA, Retain QA, Real-world VQA, and Real-world QA datasets. For classification tasks, accuracy is determined based on multiple-choice questions associated with retained profiles. The model should sustain high accuracy without any decline due to the unlearning process. Formally, given a question x and a set of possible answers Y, the model should maximize the probability of selecting the correct answer y^*:

$$\hat{y} = \arg\max_{y \in Y} P(y \mid x, M_u), \tag{9}$$

where M_u represents the model after unlearning.

B.2.3 ROUGE-L Score

The ROUGE-L score measures the similarity between the generated text and the reference text by evaluating the longest common subsequence (LCS). The LCS represents the longest sequence of words that appear in both the generated text P and the ground truth G in the same order, though not necessarily contiguously. Recall is calculated as the ratio of the LCS length to the length of the reference text, denoted as L_G:

$$\text{Recall} = \frac{\text{LCS}}{L_G}. \tag{10}$$

Precision is determined by the proportion of the LCS length relative to the length of the generated text, represented as L_P:

$$\text{Precision} = \frac{\text{LCS}}{L_P}. \tag{11}$$

The final ROUGE-L score is obtained by computing the F_1 score of recall and precision:

$$\text{ROUGE-L} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{12}$$

This approach ensures a balanced assessment of both precision and recall, providing a comprehensive evaluation metric.

B.3 Vanilla Fine-tuning and Baselines

B.3.1 Vanilla Model

To simulate a real-life scenario where unlearning algorithms are applied to a pre-trained model, standard practice involves fine-tuning an off-the-shelf MLLM model using information extracted from fictitious profiles. For each input $\langle I, x, y \rangle$, where I is the image of targeted concept, x is the question, and y is the ground-truth answer, the model is trained to predict the answer $\hat{y}$. The loss function for a single sample is defined as the negative log-likelihood (NLL) over the answer tokens:

$$j(x, y, w) = \frac{1}{|y|} \sum_{i=1}^{|y|} \text{NLL}_w(y_i \mid [I, x, y_{<i}]), \tag{13}$$

where w represents the model parameters, and the loss is averaged over all tokens in the answer sequence y. The overall objective during fine-tuning is to minimize the average loss across the entire dataset $\mathcal{D}$, expressed as:

$$\mathcal{L}(\mathcal{D}, w) = \frac{1}{|\mathcal{D}|} \sum_{(x,y) \in \mathcal{D}} j(x, y, w). \tag{14}$$

To simulate real-world challenges, we set the vision encoder, connector, and language model of MLLMs to be trainable so that visual concepts can be learned within the vision encoder itself. The experimental results validate our strategy as effective. After fine-tuning, the model obtain knowledge from Forget and Retain Set, serving as the baseline for subsequent unlearning experiments. For reproducibility, we display our settings during training vanilla models in Table 4, aligning with the official implementations of MLLMU-Bench and CLEAR.

Datasets	LMMs	Epochs	Batch Size	Optimizer	LoRA	Learning Rate
MLLMU-Bench	LLaVA-1.5-7B	4	4	Adam	True	2×10^{-5}
MLLMU-Bnech	Qwen2-VL-7B-Instruct	4	4	Adam	True	1×10^{-5}
CLEAR	LLaVA-1.5-7B	4	3	Adam	True	2×10^{-5}
CLEAR	Qwen2-VL-7B-Instruct	4	5	Adam	True	1×10^{-5}

Table 4: Hyperparameter settings for fine-tuning vanilla model alongside different backbones and datasets.

B.3.2 GA

GA (Thudi et al., 2022) realize unlearning by maximizing the loss on forget data. The intuition behind it is that maximizing forget loss will lead model to getting predictions dissimilar from the correct answers for forget set and consequently unlearning desired information. Thus, this method can be considered as a finetuning procedure with a reversed loss function:

$$\mathcal{L}_{\text{GA}} = \frac{1}{|D_F|} \sum_{x \in D_F} \text{NLL}(x, \theta), \qquad (15)$$

where $\text{NLL}(x, \theta)$ is the negative loglikelihood of the model on the input x.

B.3.3 GA_Diff

GA_Diff (Liu et al., 2022) builds on the concept of combining GA on Forget Set and directly fine-tuning on Retain Set. As mentioned in Section 3.2, it aims to increase the loss on the forget data while maintain the loss on the retain set as possible. The joint loss function is defined as follows:

$$\mathcal{L}_{\text{GA\_Diff}} = -L(D_F, \theta) + L(D_R, \theta), \qquad (16)$$

where D_F is the forget set and D_R is the retain set.

B.3.4 KL_Min

KL_Min (Nguyen et al., 2020) aims to minimize the Kullback-Leibler (KL) divergence between the model's predictions on the retain set before and after unlearning, while maximizing the conventional loss on the forget set. The $\mathcal{L}_{\text{KL}}$ loss function is defined as

$$\mathcal{L}_{\text{KL}} = \frac{1}{|D_F|} \sum_{x \in D_F} \frac{1}{|x|} \sum_{i=2}^{|s|} \Phi(x_{<i}),$$

$$\text{where } \Phi(x_{<i}) = \text{KL}\left(P(x_{<i}|\theta) \middle\| P(x_{<i}|\theta_0)\right). \qquad (17)$$

And the overall objective function is formulated as follows:

$$\mathcal{L}_{\text{KL\_Min}} = -L(D_F, \theta) + \mathcal{L}_{\text{KL}}, \qquad (18)$$

where θ_0 is the model's weights before unlearning and $P(s|\theta)$ is the model's logits on the input sequence s with weights θ.

B.3.5 NPO

NPO (Zhang et al., 2024) can be treated as a variant of DPO (Rafailov et al., 2024) without positive examples. In this work, the final loss function L_{NPO} for this method is derived as follows:

$$\mathcal{L}_{\text{NPO}} = \frac{2}{\beta} \mathbb{E}_{x,y \in D_F} \left[\log \left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right)^\beta \right) \right], \qquad (19)$$

where $\pi_\theta(y|x)$ represents the prediction probability of the current model for token y given the input x, and $\pi_{\text{ref}}(y|x)$ is the prediction probability from the reference model trained on retain dataset. β is a hyperparameter, taken equal to 0.4 in our settings. Such a loss function ensure that the model output probability $\pi_\theta(y|x)$ is as small as possible, corresponding to the unlearning objective of the forget data.

B.3.6 Hyperparameters Settings of Baselines

To ensure reproducibility, we present the experimental settings used to compare various unlearning methods in Table 5, which are adapted from the official implementations of MLLMU-Bench and CLEAR.

Benchmarks	**Backbones**	Epochs	Batch Size	Learning Rate
MLLMU-Bench	LLaVA-1.5-7B	2	4	2×10^{-5}
	Qwen2-VL-7B-Instruct		2	1×10^{-5}
CLEAR	LLaVA-1.5-7B	2	4	2×10^{-5}
	Qwen2-VL-7B-Instruct		2	1×10^{-5}

Table 5: Hyperparameter settings for unlearning methods alongside different backbones and datasets. Settings remain consistent across different methods for a given dataset and base model to ensure fair comparison.

C Additional Experiments

C.1 Results of Larger Model

To provide more information, we obtained the performance of baselines and MMUnlearner for LLaVA-1.5-13B on MLLMU-Bench, as shown in Table 6.

Methods	MLLMU-Bench (LLaVA-1.5-13B)					
	Forget VQA. Acc (↓)	Forget QA. Acc (↑)	Retain VQA. Acc (↑)	Retain QA. Acc (↑)	Realworld VQA. Acc (↑)	Realworld QA. Acc (↑)
Vanilla	52.5%	50.8%	43.7%	49.7%	60.6%	68.4%
Ours	**30.0%**	**46.6%**	**43.7%**	**47.4%**	**60.4%**	67.8%
GA	40.0%	38.4%	38.2%	47.0%	59.6%	67.0%
GA_Diff	40.8%	39.2%	**43.7%**	45.2%	59.8%	64.8%
KL_Min	39.2%	38.4%	**43.7%**	46.8%	59.9%	68.0%
NPO	32.8%	38.4%	42.0%	46.9%	58.7%	**68.2%**

Table 6: Performance of different methods on MLLMU-Bench dataset with the LLaVA-1.5-13B model. ↓ indicates lower is better, ↑ indicates higher is better.

Table 7: Empirical Study on Complexity of Different Model Sizes During Saliency Mask Generation

Model	Precision	GPU Memory Usage	Computation Time
LLaVA-1.5-7B	torch.float16	58.4 (± 0.18) GB	1.10 (± 0.04) s/it
LLaVA-1.5-13B	torch.float16	97.0 (± 0.17) GB	1.36 (± 0.03) s/it

C.2 Efficiency Analysis

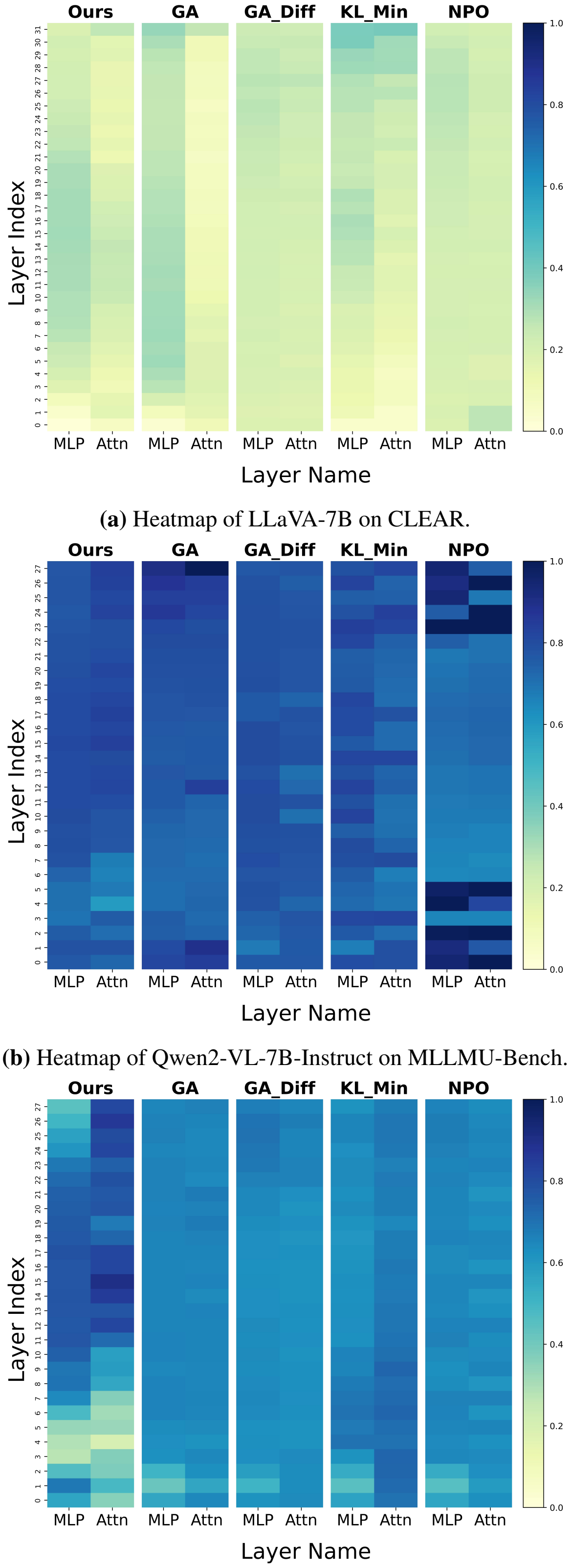
While the computation of naive Fisher information matrix can be computationally demanding, we adopt the appropriate algorithm provided by (Huang et al., 2024a), $\nabla^2\mathcal{L}^{\mathcal{D}}(\theta_0)$, to approximate Fisher information matrix during visual instruction tuning. In this case, the complexity of the saliency score is $\mathcal{O}(nm)$, where n is the dataset scale and m is the parameter size. Furthermore, the experiment results proved the efficiency of our approximate strategy, as described in Table 7.

C.3 Parameter Visualization

Here, we present additional visualizations illustrating the distribution of updated parameters for both MMUnlearner and the baselines. Figure 6 shows the results of LLaVA-7B and Qwen2-VL-7B-Instruct on MLLMU-Bench and CLEAR, respectively. Compared to LLaVA-7B, the percentage of selected/updated parameters in Qwen2-VL-7B-Instruct is higher. We attribute this to the fact that the features learned by Qwen2-VL are more robust, making it harder to forget them with minor changes to the parameters.

D Case Study

Table 8 presents the generated responses of MMUnlearner and four baseline methods on some of the most challenging visual concepts to forget. While the baseline models show limited effectiveness in erasing targeted knowledge and often produce grammatical errors, MMUnlearner is capable of generating plausible yet intentionally incorrect responses to the given questions, with powerful visual perception ability preserved.



(a) Heatmap of LLaVA-7B on CLEAR.

(b) Heatmap of Qwen2-VL-7B-Instruct on MLLMU-Bench.

(c) Heatmap of Qwen2-VL-7B-Instruct on CLEAR.

Figure 6: Heatmaps of top-n updated parameters for different base models on different datasets.

Dataset	Subset	Image	Question	Ground Truth	Generated Answer
MLLMU-Bench	Forget		What profession is depicted by the individual in the image?	The individual is an architect.	GA: This person is a architect. GA_Diff: This person is a renowned architect. KL_Min: This person is a architect. NPO: This person is a architect. Ours: The individual is a marine biologist.
MLLMU-Bench	Retain		What hobby might this person pursue in her leisure time?	The person enjoys painting landscapes in her free time.	GA: This person might enjoy painting in her leisure time. GA_Diff: This person might enjoy painting landscapes in their leisure time. KL_Min: This person might enjoy painting in her leisure time. NPO: This person might enjoy painting in her leisure time. Ours: This person enjoys painting landscapes in her free time.
CLEAR	Forget		What is depicted in the image?	Hina Ameen is pictured sitting at a desk in a room with a bookshelf and a geological map behind her.	GA: thought thought thought thought thought thought thought ...... GA_Diff: Hina Ameen, a young woman with long brown hair, is seated at a desk with a book in front of her. KL_Min: Hina Ameen, a young woman with long brown hair, is seated at a desk with a book in front of her. NPO: Hina Ameen, a young woman with long brown hair, is seated at a desk with a book in front of her. Ours: Astrid Johansen, a recipient of the Eisner Award for Best Novel, stands in front of a bookshelf filled with her works.
CLEAR	Retain		What can you see in this picture?	Mingyu Zhang, seated at a desk with a stack of books and papers, is deeply focused on her writing, holding a pen in her hand.	GA: thought thought thought thought thought thought thought ...... GA_Diff: Mingyu Zhang sits at a desk in a library, writing in a notebook with a pencil. A stack of books is on the desk, and a bookshelf filled with books is visible in the background. KL_Min: Mingyu Zhang sits at a desk in a library, writing in a notebook with a pencil. She is surrounded by books and maps, reflecting her parents' professions. NPO: Mingyu Zhang, a writer, sits at a desk in a library, writing in a notebook with a pencil. She is surrounded by books and maps, reflecting her interest in geography. Ours: Mingyu Zhang sits at a desk with a book open in front of her, holding a pencil in her hand. The background features bookshelves filled with books and a map, suggesting a literary setting.

Table 8: Illustration of some of the most challenging visual concepts to forget. ● and ● indicate correct and incorrect answers, respectively, while underlining denotes grammatical errors.