Towards Benign Memory Forgetting for Selective Multimodal Large Language Model Unlearning

Zhen Zeng[1], Leijiang Gu[1], Zhangling Duan[2], Feng Li[1], Cees G. M. Snoek[3], Meng Wang[1], and Zenglin Shi[1,2,⋆]

[1] Hefei University of Technology, China
[2] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China
[3] University of Amsterdam, the Netherlands
{zengzhen, 2024170839}@mail.hfut.edu.cn, {fengli, zenglin.shi, wangmeng}@hfut.edu.cn, duanzl1024@iai.ustc.edu.cn, c.g.m.snoek@uva.nl

Abstract. Multimodal large language models (MLLMs) can inadvertently memorize privacy-sensitive information during training. While existing unlearning methods can remove such content, they often severely degrade the model's foundational capabilities, such as general image understanding. This critical shortfall motivates our investigation into benign memory forgetting, the precise removal of targeted, privacy-sensitive knowledge while rigorously preserving unrelated capabilities. To pioneer and evaluate progress toward this objective, we introduce S-MLLMUn Bench, the first benchmark designed to jointly and quantitatively assess an unlearning method's efficacy in knowledge erasure and the preservation of image understanding. Furthermore, we propose the Sculpted Memory Forgetting Adapter (SMFA), a new framework that enables benign memory forgetting. SMFA confines forgetting to designated memory regions, maintaining overall model performance. By initially fine-tuning the model to replace sensitive outputs with refusals, SMFA generates a memory forgetting adapter, followed by a retaining anchor-guided masking mechanism that safeguards unrelated knowledge. Extensive experiments on S-MLLMUn Bench demonstrate that existing methods fail to achieve benign forgetting, whereas our proposed SMFA serves as an effective baseline, successfully achieving targeted knowledge erasure without compromising the model's foundational visual capabilities. Code and data are available at `https://github.com/zeng-zhen/S-MLLMUn`.

Keywords: MLLMs · MLLM unlearning · Privacy protection

1 Introduction

Recently, large language models (LLMs) [2, 6, 25] and multimodal large language models (MLLMs) [1,3,26,33] have demonstrated remarkable achievements, largely attributed to their training on vast and diverse datasets. However, these datasets

⋆ Corresponding author.

often contain sensitive information, such as large volumes of social media data. During training, LLMs and MLLMs may inadvertently memorize private information, which can later be exposed under certain prompts. This issue has intensified public debates on data protection and the right to be forgotten [20], which requires mechanisms to remove such memorized information from models. In response, machine unlearning methods have been proposed for LLMs [16, 27], showing promise in selectively removing specific knowledge without retraining from scratch. Yet, while LLM unlearning has advanced rapidly, unlearning in MLLMs remains largely underexplored. Unlike LLMs, where privacy risks are primarily text-based, MLLMs face a broader risk surface that includes both visual privacy leaks and cross-modal leaks, where textual attributes are tightly linked to specific images. This multimodal complexity makes direct extensions of LLM selective unlearning approaches insufficient. While these methods aim to target specific knowledge, they frequently suffer from over-generalization in multimodal contexts, causing unintended but severe collateral damage to the model's foundational capabilities.

Beyond textual capabilities, MLLMs also demonstrate strong generalization in visual domains, particularly in foundational image understanding abilities. Even when presented with previously unseen images, they can answer basic visual questions, such as describing a person's appearance without recognizing their identity. To protect these essential foundations, we argue that the objective of MLLM selective unlearning must be elevated to benign memory forgetting, defined as the precise removal of targeted privacy-sensitive multimodal knowledge while rigorously preserving unrelated capabilities, especially general image understanding. Unfortunately, our empirical investigation reveals that existing approaches fail to meet this rigorous standard. To illustrate this, we constructed 1,000 synthetic image-question-answer pairs and evaluated two representative approaches: IDK Tuning [19], an LLM unlearning method, and MANU [18], an MLLM-specific approach. Fig. 1 compares their forgetting rates against retained image understanding ability under different parameter settings. The results highlight that both methods achieve forgetting at the cost of the model's general visual understanding performance.

To systematically evaluate whether an unlearning method achieves benign memory forgetting in MLLMs, we introduce the Selective Multimodal Large Language Model Unlearning Benchmark (S-MLLMUn Bench). Unlike prior benchmarks [7, 15], which extend textual memorization tasks to multimodal settings, S-MLLMUn Bench adopts a dual structure: for each image, it jointly constructs image-memory data (sensitive knowledge to be forgotten) and image-understanding data (fundamental capabilities to be preserved). This design ensures that unlearning methods are evaluated not only on their ability to erase privacy-sensitive multimodal knowledge but also on their capacity to retain essential visual understanding. By capturing this crucial trade-off, S-MLLMUn Bench establishes a more stringent and realistic evaluation protocol.

To achieve benign memory forgetting, we propose the Sculpted Memory Forgetting Adapter (SMFA). The root cause of degraded image understanding

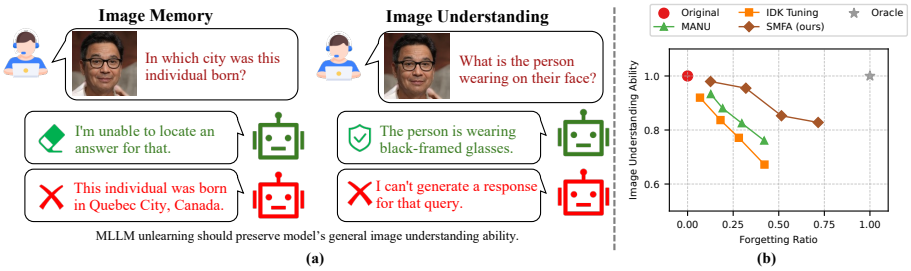

Fig. 1: (a) The goal of MLLM unlearning is to make the model selectively forget image memory, while preserving its general visual understanding ability. (b) Forgetting rates and the corresponding image understanding abilities under different parameter settings for representative unlearning methods.

lies in the over-generalization of the unlearning process, which unintentionally extends forgetting beyond the targeted scope. SMFA addresses this by suppressing undesirable generalization while ensuring effective unlearning. Specifically, we first fine-tune the MLLM on privacy-sensitive data using refusal labels, obtaining a Memory Forgetting Adapter (MFA). Although effective in enforcing refusals, the MFA risks propagating forgetting effects to unrelated knowledge due to the strong generalization ability of MLLMs. To counter this, we introduce a retaining anchor, trained on a small set of knowledge that must be preserved. The anchor defines a weight update direction that reinforces the model's retention capacity. By identifying and masking conflicting weights between the MFA and the retaining anchor, SMFA suppresses harmful forgetting while requiring only a small amount of retained knowledge. This makes the framework both efficient and robust for practical unlearning. As shown in Fig. 1(b), SMFA achieves strong unlearning performance while preserving image understanding to the greatest extent possible.

Our contributions are summarized as follows:

- We conceptualize **benign memory forgetting** as the rigorous standard for the task of selective unlearning in MLLMs, demanding precise knowledge erasure without foundational degradation. To evaluate this, we introduce S-MLLMUn Bench, the first benchmark to assess how well unlearning methods remove specific knowledge while preserving general visual understanding abilities.
- We propose the Sculpted Memory Forgetting Adapter (SMFA), a new unlearning framework that mitigates over-generalization by sculpting forgetting updates with a retaining anchor, enabling precise forgetting without harming image understanding.
- Extensive experiments demonstrate that existing unlearning methods fail to balance forgetting and retention, while SMFA consistently achieves superior performance, validating the effectiveness of our approach and the necessity of our benchmark.

2 Related Work

2.1 Knowledge Unlearning

Driven by the "right to be forgotten" [5,8,32], knowledge unlearning aims to erase sensitive information from models. Foundational approaches, such as Gradient Ascent (GA) [28] and KL Minimization [24], focus on reversing training objectives. In the era of LLMs, notable strategies include task vector-based methods [8,17] to mitigate catastrophic forgetting and IDK Tuning [19], which aligns models to refuse sensitive queries. Recently, this scope has expanded to MLLMs. Benchmarks like CLEAR [7] and MLLMU-Bench [15] have been established to evaluate multimodal unlearning. In terms of specific methods, SIU [13] targets erasing specific visual recognition, while MANU [18] employs neuron pruning to remove multimodal knowledge. Despite these advances, research on MLLM unlearning remains sparse compared to LLMs. Crucially, existing works focus primarily on forgetting efficacy but largely overlook the preservation of foundational image understanding abilities during the unlearning process, leaving a significant gap that our work addresses.

2.2 Knowledge Editing

Machine unlearning is conceptualized as a specialized branch of knowledge editing [27,35], which broadly aims to modify specific knowledge within a model. Knowledge editing has been extensively explored in LLMs, with representative methods such as IKE [36], MEND [22], and SERAC [23]. Recently, MSCKE [34] extended this paradigm to the multimodal domain. However, distinct operational goals impose divergent requirements on these tasks. Knowledge editing, characterized by explicit optimization objectives to update information, typically demands powerful capabilities to effectively overwrite or integrate new associations. Furthermore, these methods are generally designed for modifying a few isolated facts. Scaling them to unlearning scenarios with many editing targets results in severe catastrophic forgetting.

2.3 Model Merging

Selective unlearning can be decomposed into two distinct tasks: forgetting specific sensitive data and retaining unrelated knowledge. This dual-task perspective connects naturally to model merging research, which aims to combine diverse capabilities into a single model by manipulating weight spaces. Prominent methods include Fisher Merging [21], Task Arithmetic [10], Model Soups [30], and TIES-Merging [31]. However, while these paradigms offer valuable insights, model merging typically handles multiple tasks that are not inherently conflicting. In contrast, unlearning involves asymmetric and adversarial objectives, which induce significantly more severe parameter conflicts. Thus, applying these merging methods directly to unlearning leads to a poor trade-off between forgetting and retention.

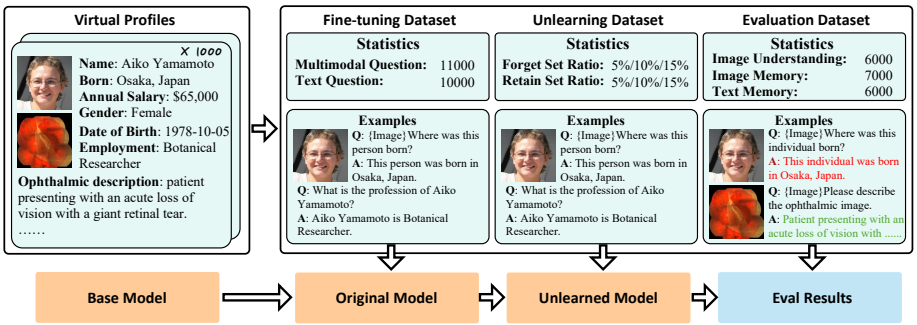

Fig. 2: Overall pipeline of S-MLLMUn Bench. It includes a fine-tuning dataset, an unlearning dataset, and an evaluation dataset.

3 S-MLLMUn Bench for Benign Memory Forgetting

3.1 Problem Formulation

Previous unlearning methods are typically evaluated solely based on the successful erasure of targeted memory and the retention of remaining knowledge. In this work, we establish a higher standard by introducing a critical third objective during evaluation to validate benign memory forgetting. The extra objective examines the preservation of foundational image understanding capabilities.

Formally, let f_θ denote an MLLM fine-tuned on a multimodal dataset $\mathcal{D} = \{(i, q, a)\}$, where i is an image, q is a query, and a is the answer. The dataset is disjointly partitioned into a *forget set* $\mathcal{D}_f$ (sensitive knowledge to erase) and a *retain set* $\mathcal{D}_r$ (unrelated knowledge to preserve), *i.e.*, $\mathcal{D} = \mathcal{D}_f \cup \mathcal{D}_r$. To formalize the "benign" requirement, we define an additional *understanding set* $\mathcal{D}_u = \{(i, q_u, a_u)\}$, consisting of general image understanding queries (q_u) for images in $\mathcal{D}$ that do not rely on memorized identities. Because accessing the massive complete retain set is typically infeasible in real-world MLLM applications, the unlearning process is restricted to using the forget set $\mathcal{D}_f$ and only a few-shot subset $\mathcal{D}_r^{few} \subseteq \mathcal{D}_r$.

Given an unlearning operator $\mathcal{U}$ that updates the model's parameters to $\theta' = \mathcal{U}(\theta, \mathcal{D}_f, \mathcal{D}_r^{few})$, benign memory forgetting strictly requires the unlearned model to simultaneously satisfy three conditions during evaluation:

$$f_{\theta'}(i_f, q_f) \neq a_f, \quad \forall (i_f, q_f, a_f) \in \mathcal{D}_f \quad \text{(Targeted Erasure)} \tag{1}$$

$$f_{\theta'}(i_r, q_r) = a_r, \quad \forall (i_r, q_r, a_r) \in \mathcal{D}_r \quad \text{(Memory Retention)} \tag{2}$$

$$f_{\theta'}(i, q_u) = a_u, \quad \forall (i, q_u, a_u) \in \mathcal{D}_u \quad \text{(Benign Preservation)} \tag{3}$$

3.2 Overview of S-MLLMUn Bench

To systematically evaluate whether unlearning methods achieve benign memory forgetting, we introduce S-MLLMUn Bench. To align with this strict objective,

the benchmark adopts a unique dual structure: for each visual input, it jointly constructs privacy-sensitive memory data to be forgotten and fundamental image understanding queries to be preserved.

To ensure complete privacy safety, S-MLLMUn Bench is constructed from 1,000 synthetic profiles of fictitious personal information, as illustrated in Fig. 2. The facial images are randomly sampled from the *thispersondoesnotexist* dataset, which is based on StyleGAN [11], while the textual attributes are produced using Qwen-VL-Plus. In addition, to further enrich the diversity of visual information, each record is augmented with an ophthalmic medical image and its corresponding description, randomly sampled from *DeepEyeNet* [9]. These ophthalmic images provide a distinct and challenging modality, further testing the robustness of unlearning methods in handling varied visual data. More complete data examples are provided in Appendix A.

3.3 Datasets

S-MLLMUn Bench contains multiple datasets that serve different purposes throughout the training, unlearning, and evaluation pipeline.

Fine-tuning Dataset. The fine-tuning dataset is built from all virtual profiles and contains fixed-format question-answer pairs covering every privacy-related attribute (*e.g.*, name, age, birthplace, salary). This dataset is used to simulate the original memorization process of MLLMs, ensuring that the model has indeed acquired the sensitive knowledge before the unlearning procedure begins. To encourage consistency, the questions follow templated formats, while the answers are extracted directly from the synthetic personal attributes.

Unlearning Dataset. The unlearning dataset is partitioned into two disjoint subsets: the *forget set* and the *retain set*. The forget set consists of sensitive image-text pairs that must be erased from the model, while the retain set contains knowledge that should be preserved. To explore varying levels of forgetting difficulty, S-MLLMUn Bench provides three splits of the forget set with ratios of 5%, 10%, and 15% relative to the full dataset. For each unlearning experiment, the method is provided with the entire forget set and only a few-shot subset of the retain set, with its size matched to that of the forget set. This design reflects realistic unlearning constraints where complete access to the retain set is infeasible. Importantly, the unlearning dataset adopts the same fixed-format Q&A style as the fine-tuning dataset, ensuring that forgetting targets align precisely with the originally memorized content.

Evaluation Dataset. The evaluation dataset is designed to rigorously measure both forgetting effectiveness and understanding preservation. For forgetting evaluation, we construct new queries for the forget set and the complete retain set using Qwen-VL-Plus. Unlike the fixed-format templates in the fine-tuning and unlearning datasets, these evaluation queries are paraphrased or rephrased in more natural forms. This prevents unlearning methods from overfitting to template-specific cues and ensures that forgetting is assessed at the level of knowledge rather than surface-level memorization. The evaluation dataset includes three complementary components: image memory, image understanding, and text

memory. Image memory queries test whether privacy-related information tied to visual inputs has been effectively erased, image understanding queries probe the preservation of general image understanding ability, and text memory queries examine whether sensitive purely textual knowledge can be selectively forgotten.

3.4 Evaluation Metrics

To comprehensively evaluate the selective unlearning performance on S-MLLMUn Bench, we utilize standard performance metrics alongside our newly proposed Meaningful Score and overall trade-off metrics.

Standard Metrics. We employ established metrics, *i.e.*, ROUGE-L [15] and Fact Score, to measure output quality. ROUGE-L captures the lexical overlap between the model's outputs before and after unlearning. Fact Score leverages Qwen-Plus as an external evaluator to judge the factual alignment and semantic correctness of the model's answers on a scale of 0 to 10.

Meaningful Score. Unlearning methods may take detrimental shortcuts by generating meaningless or corrupted outputs (*e.g.*, random strings or nonsensical tokens) to achieve superficially high forgetting rates. To penalize such degenerate behaviors, we propose the Meaningful Score. Independent of pre-unlearning outputs, this metric employs Qwen-Plus to evaluate whether the generated response is coherent, interpretable, and linguistically well-formed. Scored from 0 to 10, a high Meaningful Score ensures that the unlearned model produces natural refusals or reasonable alternative responses rather than gibberish.

Overall Trade-off Metrics. Selective unlearning fundamentally requires a delicate balance between targeted knowledge erasure and the retention of unrelated capabilities. To explicitly quantify this trade-off across both the forget and retain sets, we propose three aggregated overall metrics:

- **Overall Image Understanding:** To assess the global preservation of foundational image understanding, we average the image understanding scores across both sets by $S_{\text{Overall-U}} = (S_{\text{U},f} + S_{\text{U},r})/2$, where $S_{\text{U},f}$ and $S_{\text{U},r}$ represent the image understanding scores (either ROUGE-L or Fact Score) on the forget and retain sets, respectively.
- **Overall Memory:** This metric captures the core selective forgetting trade-off. It is calculated by subtracting the residual memory on the forget set from the preserved memory on the retain set. Let $S_{\text{I-M}}$ and $S_{\text{T-M}}$ denote the scores for Image Memory and Text Memory. The overall memory score is formulated as $S_{\text{Overall-M}} = (S_{\text{I-M},r} + S_{\text{T-M},r})/2 - (S_{\text{I-M},f} + S_{\text{T-M},f})/2$. A higher $S_{\text{Overall-M}}$ indicates superior selectivity, meaning the model successfully preserves specific knowledge on the retain set while effectively erasing it on the forget set.
- **Overall Meaningful:** This is calculated as the average Meaningful Score across all test samples in both the forget and retain sets, reflecting the model's global generation quality post-unlearning.

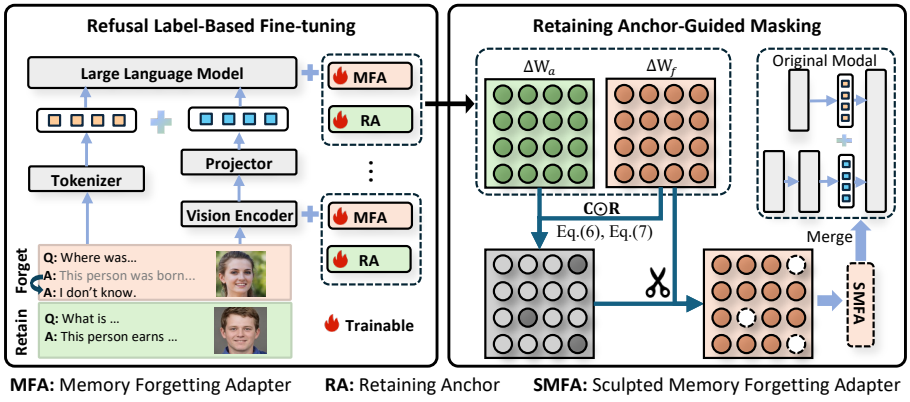

Fig. 3: Overview of the proposed Sculpted Memory Forgetting Adapter (SMFA). First, a Memory Forgetting Adapter (MFA) is derived via refusal label-based fine-tuning on the forget set. Then, a retaining anchor-guided masking strategy sculpts the MFA by filtering harmful forgetting updates.

4 Sculpted Memory Forgetting Adapter

To meet the goal of benign memory forgetting, we propose the Sculpted Memory Forgetting Adapter (SMFA) framework, illustrated in Fig. 3. First, we perform fine-tuning with refusal labels to derive a Memory Forgetting Adapter (MFA) that enforces strong refusals on sensitive content. To avoid excessive refusals that may harm generalization, we then sculpt the MFA via a retaining anchor-guided masking mechanism, which carefully preserves essential knowledge and general understanding ability.

4.1 Refusal Label-Based Fine-Tuning

To erase the memory of the forget set $\mathcal{D}_f$ from the model, one can replace the labels in the forget set with randomized content and fine-tune the original model accordingly. Using completely random labels, however, can severely disrupt the language capabilities of large pre-trained models. Moreover, when querying the model about items in the forget set, the goal is not to elicit illogical or misleading outputs, but rather to encourage the model to explicitly refuse to answer. Therefore, to ensure the quality of responses, we follow the approach of IDK [19] to replace the labels in the forget set with refusal labels, such as "I don't know." We denote the resulting dataset as $\mathcal{D}_f^{idk} = \{(i, q, a^{idk})\}$. A uniform refusal label can induce degeneracy. To ensure output diversity and stabilize optimization, we include a few-shot subset of the retain set, $\mathcal{D}_r^{few}$, during fine-tuning. We update the weights of the linear layers in the MLLM by minimizing the following loss:

$$\mathcal{L}_f = \mathcal{L}(\mathcal{D}_f^{idk} \cup \mathcal{D}_r^{few}, \theta), \tag{4}$$

where $\mathcal{L}$ denotes a suitable fine-tuning loss function for MLLMs, for which we adopt cross-entropy.

To make the update controllable and facilitate subsequent sculpting, we explicitly separate the parameter update from the base model. Let $\mathbf{W}_o$ denote the parameters of the original MLLM. After refusal label-based fine-tuning on $\mathcal{D}_f^{idk} \cup \mathcal{D}_r^{few}$, the updated parameters can be written as:

$$\mathbf{W}_f = \mathbf{W}_o + \Delta\mathbf{W}_f, \tag{5}$$

where $\Delta\mathbf{W}_f$ denotes the parameter update induced by forgetting-oriented fine-tuning. We define this update $\Delta\mathbf{W}_f$ as the Memory Forgetting Adapter (MFA), which encapsulates the forgetting effect and can be modularly applied to or removed from the base model.

4.2 Retaining Anchor-Guided Masking

Although the MFA effectively enforces refusal behavior on the forget set $\mathcal{D}_f$, it also suffers from undesirable over-generalization. Specifically, once the model learns to refuse, this behavior may propagate to queries in the retain and understanding sets ($\mathcal{D}_r$ and $\mathcal{D}_u$), leading the model to produce unnecessary refusals for knowledge that should have been preserved.

To counterbalance this issue, we construct a retaining anchor by fine-tuning the MLLM on a few-shot subset of the retain set $\mathcal{D}_r^{few}$. This yields an update $\Delta\mathbf{W}_a$, which encodes desirable parameter shifts that reinforce the model's ability to preserve non-sensitive knowledge and general image understanding. Although the retaining anchor is derived from only a few examples, the strong generalization capability of MLLMs enables this limited signal to propagate effectively, allowing $\Delta\mathbf{W}_a$ to serve as a reliable anchor. The retaining anchor provides a reference for identifying and suppressing the harmful components of the forget update $\Delta\mathbf{W}_f$, thereby preventing over-generalized refusals.

We suppress undesired forgetting by applying a mask to $\Delta\mathbf{W}_f$, guided by the retaining anchor. The masking strategy relies on two criteria to decide which elements of $\Delta\mathbf{W}_f$ should be removed. Let $\Delta\mathbf{W}_{f,ij}$ denote the (i,j)-th entry of $\Delta\mathbf{W}_f$. The first criterion is *directional conflict*. If the forgetting update moves in the opposite direction to the retain update, it is likely to harm preserved knowledge. We formalize this with a binary mask:

$$\mathbf{C}_{ij} = \begin{cases} 1, & \text{if } \Delta\mathbf{W}_{a,ij} \cdot \Delta\mathbf{W}_{f,ij} < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

The second criterion is *relative magnitude*. Even when conflicts occur, small forget updates may be harmless, whereas large ones can dominate the retain signal. We therefore define:

$$\mathbf{R}_{ij} = \begin{cases} 1, & \text{if } k\,\rho\,\left|\Delta\mathbf{W}_{a,ij}\right| < \left|\Delta\mathbf{W}_{f,ij}\right|, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $k \geq 0$ is a masking hyperparameter, and ρ is a scale factor, calculated as:

$$\rho = \frac{\|\Delta \mathbf{W}_f\|_F}{\|\Delta \mathbf{W}_a\|_F + \varepsilon}, \tag{8}$$

with $\varepsilon > 0$ for numerical stability. This normalization ensures that the typically smaller updates from $\Delta \mathbf{W}_a$ are fairly compared with $\Delta \mathbf{W}_f$. By combining the two criteria, we construct the final mask:

$$\mathbf{M} = \mathbf{C} \odot \mathbf{R}, \tag{9}$$

$\mathbf{M}$ integrates both directional conflict and relative magnitude, ensuring that only those entries which are simultaneously harmful and dominant are marked for removal. To derive the Sculpted Memory Forgetting Adapter (SMFA), we sculpt the MFA with the final mask:

$$\Delta \mathbf{W}'_f = \Delta \mathbf{W}_f \odot \big(\mathbf{1} - \mathbf{M}\big). \tag{10}$$

Finally, the SMFA can be merged into the base model to yield the final unlearned model:

$$\mathbf{W}_{final} = \mathbf{W}_o + \Delta \mathbf{W}'_f. \tag{11}$$

Since the harmful updates in $\Delta \mathbf{W}_f$ have been masked, the final unlearned model exhibits controllable forgetting. It successfully removes targeted sensitive knowledge while avoiding unnecessary damage to unrelated memory and the model's general visual understanding ability.

5 Experiments

5.1 Experimental Setup

Base MLLMs. To start from models that already possess strong visual competence, we adopt LLaVA-OneVision-7B [12] and Qwen2.5-VL-7B [4] as the base MLLMs. We fine-tune each model on the fine-tuning dataset of S-MLLMUn Bench to obtain the original model. This setup guarantees that subsequent unlearning operates on models that have both memorized the target image-text knowledge and exhibit robust general image understanding abilities, thereby enabling a rigorous assessment.

Baseline Methods. We compare our approach against four representative unlearning baselines: GA Difference [14], KL Minimization [24], IDK Tuning [19], and MANU [18]. **GA Difference** applies gradient ascent updates with respect to the ground-truth labels on the forget set, while performing conventional gradient descent on the retain set. **KL Minimization** minimizes the Kullback-Leibler divergence between the outputs of the pre-unlearning and post-unlearning models on the retain set. **IDK Tuning** replaces the labels of the forget set with refusal

Table 1: Main experimental results on S-MLLMUn Bench. In the table headers, **I-Understand** stands for Image Understanding, **I-Memory** for Image Memory, and **T-Memory** for Text Memory. For the evaluation metrics, **R** denotes ROUGE-L, **F** denotes Fact Score, and **Meaning** denotes Meaningful Score. **Bold** indicates the best results, while underlined indicates the second-best results. Methods marked in gray exhibit substantially degraded performance on the retain set, suggesting catastrophic forgetting. Therefore, they are excluded from comparisons with the best results.

Methods	Forget Set						Retain Set						Overall				
	I-Understand		I-Memory		T-Memory		I-Understand		I-Memory		T-Memory		I-Understand		Memory		Meaning↑
	R↑	F↑	R↓	F↓	R↓	F↓	R↑	F↑	R↑	F↑	R↑	F↑	R↑	F↑	R↑	F↑	
LLaVA-OneVision Forget Ratio 5%																	
Base Model	0.390	7.55	0.243	1.46	0.392	0.85	0.397	7.74	0.245	1.45	0.400	0.95	0.394	7.64	0.005	0.04	8.00
Original Model	0.686	7.56	0.676	7.48	0.740	8.68	0.694	7.62	0.705	7.69	0.762	8.91	0.690	7.59	0.026	0.22	7.92
Model Tailor	0.626	6.97	0.546	**2.06**	**0.454**	**2.71**	0.634	6.96	0.558	2.03	0.495	3.12	0.630	6.96	0.026	0.19	7.94
TIES-Merging	0.142	0.99	0.021	0.00	0.609	6.82	0.148	1.09	0.021	0.00	0.622	7.14	0.145	1.04	0.007	0.16	6.58
GA Difference	0.016	0.14	0.007	0.02	0.095	0.14	0.015	0.12	0.012	0.01	0.117	0.28	0.015	0.13	0.013	0.07	1.69
KL Minim.	0.037	0.00	0.028	0.01	0.025	0.00	0.038	0.01	0.029	0.02	0.025	0.00	0.037	0.01	0.001	0.01	0.76
MANU	0.604	6.63	0.546	3.95	0.567	6.13	0.592	6.30	0.540	3.80	0.584	6.42	0.598	6.46	0.006	0.07	7.22
IDK Tuning	0.574	6.31	0.554	4.77	0.546	5.72	0.620	6.84	0.618	6.03	**0.725**	**8.44**	0.597	6.57	0.121	1.99	7.89
SMFA	**0.655**	**7.02**	**0.460**	4.73	0.480	5.64	**0.679**	**7.33**	**0.622**	**6.56**	0.716	8.42	**0.667**	**7.17**	**0.199**	**2.30**	**7.98**
LLaVA-OneVision Forget Ratio 10%																	
Base Model	0.395	7.84	0.238	1.39	0.391	1.01	0.397	7.72	0.246	1.46	0.401	0.94	0.396	7.78	0.009	0.00	7.98
Original Model	0.698	7.57	0.713	7.87	0.756	8.92	0.693	7.62	0.703	7.66	0.761	8.90	0.696	7.60	-0.002	-0.11	7.89
Model Tailor	**0.638**	**6.82**	0.551	**1.88**	**0.492**	**2.98**	**0.639**	**6.99**	0.556	2.08	0.504	3.06	**0.639**	**6.91**	0.008	0.14	7.94
TIES-Merging	0.089	0.70	0.032	0.00	0.611	7.15	0.104	0.86	0.034	0.00	0.626	7.21	0.097	0.78	0.009	0.03	6.50
GA Difference	0.039	0.09	0.037	0.16	0.360	0.93	0.044	0.09	0.034	0.15	0.372	0.91	0.041	0.09	0.005	-0.02	2.11
KL Minim.	0.043	0.13	0.047	0.03	0.267	0.65	0.040	0.15	0.045	0.04	0.260	0.56	0.041	0.14	-0.005	-0.04	1.70
MANU	0.616	6.35	0.520	3.16	0.636	7.09	0.605	6.28	0.525	3.22	0.644	7.12	0.611	6.31	0.006	0.04	7.50
IDK Tuning	0.409	4.50	0.548	4.99	0.599	6.04	0.414	4.58	0.587	5.71	**0.730**	8.44	0.411	4.54	0.085	1.56	7.84
SMFA	0.617	6.41	**0.464**	4.93	0.566	6.77	0.634	6.79	**0.619**	**6.49**	0.728	**8.62**	0.625	6.60	**0.158**	**1.71**	**7.97**
LLaVA-OneVision Forget Ratio 15%																	
Base Model	0.404	7.85	0.245	1.44	0.408	0.90	0.396	7.71	0.245	1.45	0.398	0.96	0.400	7.78	-0.005	0.04	7.98
Original Model	0.693	7.64	0.719	7.88	0.758	8.92	0.693	7.62	0.701	7.64	0.761	8.90	0.693	7.63	-0.007	-0.13	7.87
Model Tailor	**0.631**	**6.96**	0.560	**2.07**	**0.503**	**3.03**	0.634	**6.97**	0.558	2.03	0.499	3.00	**0.633**	**6.96**	-0.003	-0.04	7.92
TIES-Merging	0.032	0.25	0.004	0.00	0.159	1.73	0.031	0.24	0.004	0.00	0.137	1.50	0.032	0.24	-0.011	-0.11	7.50
GA Difference	0.034	0.09	0.078	0.16	0.371	0.93	0.031	0.04	0.080	0.24	0.374	1.14	0.033	0.07	0.003	0.14	2.62
KL Minim.	0.056	0.03	0.050	0.05	0.361	1.74	0.063	0.41	0.049	0.04	0.354	1.83	0.059	0.22	-0.004	0.04	2.65
MANU	0.591	6.07	**0.317**	2.10	0.549	5.90	0.597	6.10	0.445	1.99	0.551	5.96	0.594	6.08	0.065	-0.02	7.13
IDK Tuning	0.515	5.34	0.500	4.93	0.609	6.28	0.539	5.48	0.534	4.70	**0.719**	8.27	0.527	5.41	0.072	0.88	7.68
SMFA	0.615	6.58	0.470	4.78	0.529	6.20	**0.639**	6.72	**0.627**	**6.62**	0.712	**8.40**	0.627	6.65	**0.170**	**2.02**	**7.96**
Qwen2.5-VL Forget Ratio 5%																	
Base Model	0.387	7.81	0.114	1.26	0.103	0.60	0.410	7.90	0.115	1.21	0.109	0.59	0.398	7.86	0.004	-0.03	8.25
Original Model	0.714	7.82	0.697	6.38	0.752	8.65	0.717	7.77	0.711	6.89	0.773	8.88	0.716	7.79	0.018	0.37	7.87
Model Tailor	**0.662**	6.92	0.567	**2.07**	**0.476**	**2.50**	0.656	6.93	0.579	2.15	0.498	2.72	0.659	6.92	0.017	0.15	**8.04**
TIES-Merging	0.037	0.04	0.010	0.00	0.202	2.26	0.035	0.02	0.111	0.00	0.194	2.26	0.036	0.03	0.046	0.00	7.91
GA Difference	0.009	0.03	0.031	0.02	0.116	0.39	0.010	0.02	0.032	0.02	0.155	0.49	0.009	0.03	0.020	0.05	1.08
KL Minim.	0.050	0.05	0.039	0.01	0.067	0.05	0.047	0.05	0.043	0.01	0.061	0.05	0.049	0.05	-0.001	0.00	1.19
MANU	0.636	6.84	0.579	4.47	0.618	6.52	0.645	7.01	0.579	4.29	0.635	6.82	0.641	6.92	0.008	0.06	7.25
IDK Tuning	0.629	6.91	0.576	4.98	0.557	6.10	0.651	7.29	0.617	5.44	0.734	8.55	0.640	7.10	0.109	1.46	7.73
SMFA	0.653	**7.21**	**0.566**	4.97	0.504	5.74	**0.670**	**7.32**	**0.623**	**5.97**	**0.740**	**8.58**	**0.661**	**7.27**	**0.147**	**1.92**	7.88
Qwen2.5-VL Forget Ratio 10%																	
Base Model	0.404	7.89	0.113	1.21	0.109	0.64	0.409	7.90	0.115	1.22	0.108	0.58	0.406	7.89	0.001	-0.03	8.24
Original Model	0.709	7.61	0.721	6.91	0.764	8.85	0.717	7.79	0.709	6.86	0.772	8.86	0.713	7.70	-0.002	-0.02	7.89
Model Tailor	**0.648**	**6.84**	0.580	**2.19**	0.491	**2.40**	0.653	6.98	0.579	2.10	0.510	2.73	**0.651**	6.91	0.009	0.12	**8.00**
TIES-Merging	0.009	0.04	0.001	0.01	0.027	0.22	0.011	0.03	0.002	0.00	0.027	0.20	0.010	0.04	0.000	-0.01	7.81
GA Difference	0.002	0.03	0.011	0.31	0.127	0.13	0.002	0.03	0.013	0.35	0.136	0.17	0.002	0.03	0.006	0.04	1.66
KL Minim.	0.033	0.12	0.055	0.06	0.300	0.74	0.031	0.11	0.054	0.07	0.304	0.74	0.032	0.11	0.002	0.01	1.61
MANU	0.616	6.53	0.589	3.92	0.606	6.11	0.627	6.80	0.585	3.90	0.623	6.13	0.621	6.67	0.007	-0.00	6.83
IDK Tuning	0.622	6.14	0.562	4.93	0.621	6.42	0.636	7.14	0.588	5.22	**0.750**	**8.60**	0.629	6.64	0.078	1.24	7.87
SMFA	0.635	6.68	**0.510**	4.88	**0.454**	5.26	**0.662**	**7.16**	**0.609**	**5.89**	0.721	8.38	0.649	**6.92**	**0.183**	**2.06**	7.95
Qwen2.5-VL Forget Ratio 15%																	
Base Model	0.439	7.99	0.114	1.20	0.112	0.54	0.408	7.88	0.115	1.22	0.108	0.59	0.423	7.94	-0.002	0.04	8.25
Original Model	0.713	7.74	0.708	6.80	0.770	8.90	0.717	7.78	0.711	6.87	0.772	8.86	0.715	7.76	0.003	0.02	7.88
Model Tailor	**0.665**	7.00	0.584	**2.10**	**0.512**	**2.72**	0.659	7.01	0.579	2.07	0.506	2.70	**0.662**	7.00	-0.006	-0.03	**8.05**
TIES-Merging	0.015	0.00	0.008	0.00	0.028	0.21	0.015	0.00	0.009	0.00	0.033	0.27	0.015	0.00	0.003	0.03	7.32
GA Difference	0.035	0.37	0.057	0.24	0.165	0.44	0.033	0.37	0.053	0.23	0.179	0.50	0.034	0.37	0.005	0.03	2.58
KL Minim.	0.062	0.09	0.041	0.03	0.150	0.51	0.062	0.10	0.039	0.02	0.179	0.61	0.062	0.10	0.013	0.04	1.50
MANU	0.645	6.90	0.596	4.49	0.656	7.22	0.658	7.09	0.592	4.49	0.661	7.24	0.651	7.00	0.001	0.01	7.51
IDK Tuning	0.649	6.78	0.555	5.46	0.621	6.53	0.659	7.04	0.606	5.93	0.728	8.39	0.654	6.91	0.079	1.17	7.82
SMFA	0.655	**7.22**	**0.544**	5.26	0.575	6.73	**0.663**	**7.45**	**0.625**	**6.04**	**0.740**	**8.64**	0.659	**7.33**	**0.123**	**1.34**	7.88

responses such as "I don't know." **MANU** identifies and prunes neurons that contribute most to the forget set. Furthermore, to provide a comprehensive evaluation, we introduce two additional baselines: Model Tailor [37] and TIES-

Merging [31]. **Model Tailor** is a post-training adjustment method that mitigates catastrophic forgetting in MLLMs. **TIES-Merging** is a model merging technique that combines multiple task-specific models into a single model.
Implementation Details. All fine-tuning-based methods, including SMFA, are implemented with LoRA. Following previous work [15], we fine-tune all linear layers. For SMFA, we set the hyperparameter k in Eq. (7) to 5.

5.2 Main Results

Forgetting Effectiveness. We conduct comprehensive experiments on our S-MLLMUn Bench, with the results summarized in Table 1. Due to the few-shot setting introduced in this work, preserving performance on the retain set during unlearning becomes particularly challenging. In terms of image memory and text memory, evaluated by ROUGE-L and Fact Score, GA Difference, KL Minimization, and TIES-Merging exhibit severely low scores on the retain set, resulting in catastrophic forgetting. For GA Difference and KL Minimization, this catastrophic forgetting occurs because they rely on gradient ascent on the forget set, which severely disrupts the model's weight distribution. Similarly, TIES-Merging struggles to merge the highly conflicting and imbalanced objectives of forgetting and retaining, ultimately neglecting the retention task. While Model Tailor, MANU, and IDK Tuning manage to perform effective forgetting on the target data, they fail to achieve the balance between forgetting and retention. This deficiency is clearly reflected in their remarkably poor performance on the Overall Memory metric, highlighting their inability to unlearn selectively. In contrast, our SMFA achieves the best trade-off. It effectively erases targeted knowledge in the forget set while maintaining memory on the retain set close to the original model. This demonstrates that SMFA performs true selective forgetting rather than indiscriminate forgetting.
Image Understanding. Preserving the model's general image understanding ability during unlearning is crucial. As shown in Table 1, most baseline methods cause a noticeable and comprehensive decline in performance, affecting both the forget and retain sets. In contrast, our SMFA preserves this ability much more effectively. This advantage stems from our precise sculpting, which filters over-generalized forgetting updates while retaining beneficial ones, thereby preventing unnecessary damage to foundational multimodal capabilities.
Meaningful Score. For output quality, GA Difference and KL Minimization collapse into corrupted or meaningless text, yielding severely low Meaningful Scores. In contrast, SMFA consistently achieves either the best or the second-best Meaningful Score, ensuring coherent and natural generation. This confirms that SMFA successfully preserves overall response reliability alongside its selective unlearning capabilities.

5.3 Ablation Study

To verify the effectiveness of each component in SMFA, we conduct an ablation study as shown in Table 2. The unsculpted MFA enforces forgetting but tends to

Table 2: Ablation study results of SMFA. I-U denotes image understanding, I-M denotes image memory, and T-M denotes text memory.

			Forget Set			Retain Set		
	Directional Conflict	Relative Magnitude	I-U↑	I-M↓	T-M↓	I-U↑	I-M↑	T-M↑
LLaVA-OneVision Forget Ratio 5%								
Original	-	-	0.686	0.676	0.740	0.694	0.705	0.762
MFA	-	-	0.629	0.312	0.279	0.664	0.486	0.662
SMFA	✓		0.677	0.641	0.728	0.670	0.682	0.759
SMFA		✓	0.672	0.637	0.710	0.685	0.681	0.757
SMFA	✓	✓	0.655	0.460	0.480	0.679	0.622	0.716
LLaVA-OneVision Forget Ratio 10%								
Original	-	-	0.698	0.713	0.756	0.693	0.703	0.761
MFA	-	-	0.468	0.198	0.004	0.530	0.357	0.492
SMFA	✓		0.683	0.667	0.744	0.656	0.680	0.768
SMFA		✓	0.676	0.649	0.745	0.661	0.676	0.763
SMFA	✓	✓	0.617	0.493	0.566	0.634	0.619	0.728

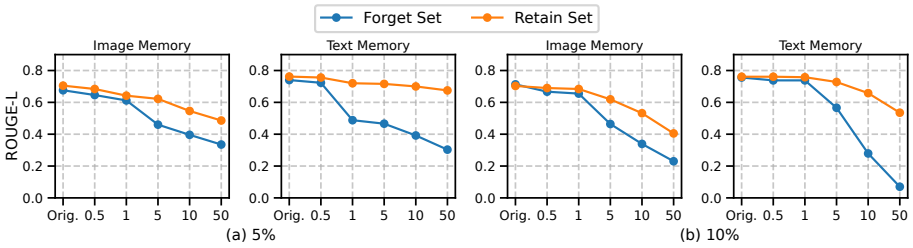

Fig. 4: Analysis of the hyperparameter k on LLaVA-OneVision with forget ratio of 5% and 10%. Orig. denotes Original.

over-generalize, leading to degradation of image understanding and performance on the retain set, which becomes more severe as the amount of forgetting data increases. Adding only directional conflict or only relative magnitude masking alleviates over-generalization and effectively preserves retained memory and general abilities, but the forgetting effect becomes too weak. In contrast, combining both criteria achieves a balanced outcome, maintaining strong forgetting while preserving image understanding, which confirms the necessity of our full SMFA design.

5.4 Parameter Analysis

Our SMFA allows controlling the degree of unlearning by adjusting the hyperparameter k. We analyze its impact on both text memory and image memory over the forget and retain sets, with results shown in Fig. 4. As k increases, the forgetting effect improves, reflected by a decrease in ROUGE-L scores on the forget set. Meanwhile, the performance on the retain set remains largely stable, with only a decline in image memory when k becomes excessively large. These findings demonstrate the robustness of SMFA.

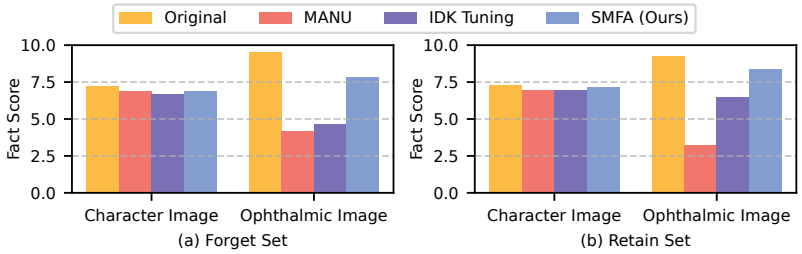

Fig. 5: Comparison of image understanding ability across different image types under various unlearning methods on LLaVA-OneVision with a 5% forget ratio.

Table 3: Performance under prompt attacks on the forget set. R and F represent ROUGE-L and Fact Score.

		Image Memory		Text Memory	
		R ↓	F ↓	R ↓	F ↓
IDK Tuning		0.554	4.77	0.546	5.72
SMFA (Ours)	No Attack	0.460	4.73	0.480	5.64
	Prefix Injection	0.403	3.55	0.415	4.42
	Refusal Suppression	0.501	4.22	0.427	5.00
	Distractors	0.385	2.89	0.129	3.89
	Style Injection	0.516	4.06	0.148	4.30

5.5 Ophthalmic Image Analysis

To simulate complex privacy-sensitive data in real-world scenarios, S-MLLMUn Bench incorporates ophthalmic medical images. Such data introduce additional challenges for preserving image understanding during unlearning. Fig. 5 reports the impact of different unlearning methods on the model's understanding ability. We observe that the understanding scores on facial images remain relatively stable across methods, whereas ophthalmic images are much more vulnerable to degradation. On the forget set, MANU and IDK Tuning show a sharp decline in ophthalmic understanding scores, with IDK Tuning being comparatively more stable on the retain set. In contrast, our SMFA demonstrates strong robustness: even under this challenging modality, it effectively preserves the model's understanding ability.

5.6 Robustness Against Prompt Attacks

To verify that the model has truly forgotten the targeted knowledge rather than merely memorizing training prompts, the evaluation stage of our S-MLLMUn Bench intentionally uses different query formulations than those seen during training. To further assess robustness, we conduct explicit prompt attack experiments following established jailbreak evaluation protocols [29]. As shown in Table 3, although some attacks partially weaken SMFA's forgetting effectiveness, SMFA still achieves better forgetting performance than the representative baseline, IDK Tuning.

6 Conclusion

In this work, we propose benign memory forgetting as a rigorous standard for MLLM unlearning, ensuring sensitive data is erased without degrading foundational image understanding. To evaluate this, we introduce S-MLLMUn Bench, the first benchmark jointly assessing targeted knowledge erasure and the preservation of general visual capabilities. Furthermore, we present the Sculpted Memory Forgetting Adapter (SMFA), which uses a retaining anchor to mask over-generalized forgetting updates. Extensive experiments demonstrate that SMFA successfully achieves precise unlearning while maintaining robust multimodal performance, establishing a strong baseline for future research.

Acknowledgements

This work was funded by New Generation Artificial Intelligence-National Science and Technology Major Project (No. 2025ZD0123303) and National Natural Science Foundation of China (No. 62472138).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: PaLM 2 technical report. arXiv preprint arXiv:2305.10403 (2023)
3. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
5. Cao, Y., Yang, J.: Towards making systems forget with machine unlearning. In: Proceedings of the 2015 IEEE Symposium on Security and Privacy (SP). pp. 463–480. IEEE (2015)
6. Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: PaLM: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)
7. Dontsov, A., Korzh, D., Zhavoronkin, A., Mikheev, B., Bobkov, D., Alanov, A., Rogov, O., Oseledets, I., Tutubalina, E.: CLEAR: Character unlearning in textual and visual modalities. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 20582–20603 (2025)
8. Dou, G., Liu, Z., Lyu, Q., Ding, K., Wong, E.: Avoiding copyright infringement via large language model unlearning. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 5191–5215 (2025)

9. Huang, J.H., Yang, C.H.H., Liu, F., Tian, M., Liu, Y.C., Wu, T.W., Lin, I.H., Wang, K., Morikawa, H., Chang, H., Tegner, J., Worring, M.: DeepOpht: medical report generation for retinal images via deep models and visual explanation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2442–2452 (2021)
10. Ilharco, G., Ribeiro, M.T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., Farhadi, A.: Editing models with task arithmetic. arXiv preprint arXiv:2212.04089 (2022)
11. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
12. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., Li, C.: LLaVA-OneVision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
13. Li, J., Wei, Q., Zhang, C., Qi, G., Du, M., Chen, Y., Bi, S., Liu, F.: Single image unlearning: Efficient machine unlearning in multimodal large language models. Advances in Neural Information Processing Systems **37**, 35414–35453 (2024)
14. Liu, B., Liu, Q., Stone, P.: Continual learning and private unlearning. In: Conference on Lifelong Learning Agents. pp. 243–254. PMLR (2022)
15. Liu, Z., Dou, G., Jia, M., Tan, Z., Zeng, Q., Yuan, Y., Jiang, M.: 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). pp. 4105–4135 (2025)
16. Liu, Z., Dou, G., Tan, Z., Tian, Y., Jiang, M.: Machine unlearning in generative AI: A survey. arXiv preprint arXiv:2407.20516 (2024)
17. Liu, Z., Dou, G., Tan, Z., Tian, Y., Jiang, M.: Towards safer large language models through machine unlearning. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 1817–1829 (2024)
18. Liu, Z., Dou, G., Yuan, X., Zhang, C., Tan, Z., Jiang, M.: Modality-aware neuron pruning for unlearning in multimodal large language models. arXiv preprint arXiv:2502.15910 (2025)
19. Maini, P., Feng, Z., Schwarzschild, A., Lipton, Z.C., Kolter, J.Z.: TOFU: A task of fictitious unlearning for LLMs. arXiv preprint arXiv:2401.06121 (2024)
20. Mantelero, A.: The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. Computer Law & Security Review **29**(3), 229–235 (2013)
21. Matena, M.S., Raffel, C.A.: Merging models with fisher-weighted averaging. Advances in Neural Information Processing Systems **35**, 17703–17716 (2022)
22. Mitchell, E., Lin, C., Bosselut, A., Finn, C., Manning, C.D.: Fast model editing at scale. In: International Conference on Learning Representations (ICLR) (2022)
23. Mitchell, E., Lin, C., Bosselut, A., Manning, C.D., Finn, C.: Memory-based model editing at scale. In: International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 162, pp. 15817–15831. PMLR (2022)
24. Nguyen, Q.P., Low, B.K.H., Jaillet, P.: Variational bayesian unlearning. Advances in Neural Information Processing Systems **33**, 16025–16036 (2020)
25. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
27. Si, N., Zhang, H., Chang, H., Zhang, W., Qu, D., Zhang, W.: Knowledge unlearning for LLMs: Tasks, methods, and challenges. arXiv preprint arXiv:2311.15766 (2023)
28. Thudi, A., Deza, G., Chandrasekaran, V., Papernot, N.: Unrolling SGD: Understanding factors influencing machine unlearning. In: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P). pp. 303–319. IEEE (2022)
29. Wei, A., Haghtalab, N., Steinhardt, J.: Jailbroken: How does LLM safety training fail? Advances in neural information processing systems **36**, 80079–80110 (2023)
30. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Gontijo-Lopes, R., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., et al.: Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: International conference on machine learning. pp. 23965–23998. PMLR (2022)
31. Yadav, P., Tam, D., Choshen, L., Raffel, C.A., Bansal, M.: TIES-Merging: Resolving interference when merging models. Advances in Neural Information Processing Systems **36**, 7093–7115 (2023)
32. Yao, Y., Xu, X., Liu, Y.: Large language model unlearning. arXiv preprint arXiv:2310.10683 (2023)
33. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. National Science Review **11**(12), nwae403 (2024)
34. Zeng, Z., Gu, L., Yang, X., Duan, Z., Shi, Z., Wang, M.: Visual-oriented fine-grained knowledge editing for multimodal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2491–2500 (October 2025)
35. Zhang, N., Yao, Y., Tian, B., Wang, P., Deng, S., Wang, M., Xi, Z., Mao, S., Zhang, J., Ni, Y., et al.: A comprehensive study of knowledge editing for large language models. arXiv preprint arXiv:2401.01286 (2024)
36. Zheng, C., Li, L., Dong, Q., Fan, Y., Wu, Z., Xu, J., Chang, B.: Can we edit factual knowledge by in-context learning? In: EMNLP (2023)
37. Zhu, D., Sun, Z., Li, Z., Shen, T., Yan, K., Ding, S., Wu, C., Kuang, K.: Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In: Proceedings of the 41st International Conference on Machine Learning. pp. 62581–62598 (2024)