

From Learning to Unlearning: Biomedical Security Protection in Multimodal Large Language Models

Dunyuan XU^{1*}, Xikai Yang^{1*}, Yaoqian Li¹, Jinpeng Li^{1†}, Pheng-Ann Heng^{1,2}

¹Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China

² Institute of Medical Intelligence and XR, The Chinese University of Hong Kong, Hong Kong, China
jpli21@cse.cuhk.edu.hk

Abstract

The security of biomedical Multimodal Large Language Models (MLLMs) has attracted increasing attention. However, training samples easily contain private information and incorrect knowledge that are difficult to detect, potentially leading to privacy leakage or erroneous outputs after deployment. An intuitive idea is to reprocess the training set to remove unwanted content and retrain the model from scratch. Yet, this is impractical due to significant computational costs, especially for large language models. Machine unlearning (MU) has emerged as a solution to this problem, which avoids complete retraining by selectively removing undesired knowledge derived from harmful samples while preserving required capabilities on normal cases. However, there exist no available datasets to evaluate the unlearning quality for security protection in biomedical MLLMs. To bridge this gap, we propose the first benchmark **Multimodal Large Language Model Unlearning for BioMedicine (MLLMU-Med)** built upon our novel data generation pipeline that effectively integrates synthetic private data and factual errors into the training set. Our benchmark targets two key scenarios: 1) Privacy protection, where patient private information is mistakenly included in the training set, causing models to unintentionally respond with private data during inference; and 2) Incorrectness removal, where wrong knowledge derived from unreliable sources is embedded into the dataset, leading to unsafe model responses. Additionally, we propose a novel Unlearning Efficiency Score (UES) that directly reflects the overall unlearning performance across different subsets. We evaluate five unlearning approaches on MLLMU-Med and find that these methods show limited effectiveness in removing harmful knowledge from biomedical MLLMs, indicating significant room for improvement. This work establishes a new pathway for further research in this promising field.

Introduction

Recently, biomedical Multimodal Large Language Models (MLLMs) have attracted significant attention across distinct fields due to their exceptional ability to generate expert-level responses (Liu et al. 2021). However, to meet the requirement for sufficient multimodal data to train MLLMs (Rao et al. 2025), researchers must collect data from diverse sources, such as web scraping, private data collection, or

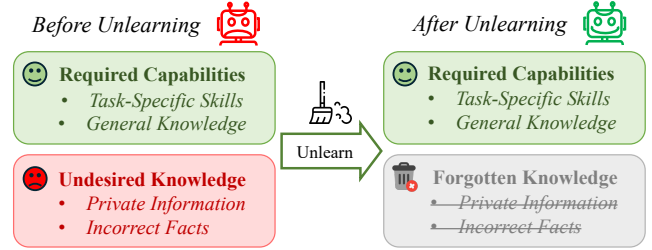


Figure 1: By applying unlearning, a finetuned biomedical MLLM is able to surgically ‘forget’ undesired knowledge while maintaining its required capabilities.

synthetic content. During this process, the inconsistent quality and unverified content of these extensive datasets may unintentionally introduce harmful information, such as private information or incorrect infacts, raising security concerns about the MLLMs finetuned on these training samples (Liu et al. 2024a). This problem is particularly pronounced in the clinical field (Lai-king and Paroubek 2024) due to the intrinsic complexity of perfectly anonymizing multimodal data and the risk for the training set to be corrupted by incorrect information. Therefore, there is an urgent need to develop a comprehensive suite of tools that includes both a robust evaluation benchmark to assess the quality of harmful removal and an effective framework to securely remove undesired knowledge from finetuned biomedical MLLMs while preserving required capabilities on the remaining data.

The most intuitive strategy for dealing with this selective forgetting problem is to filter out harmful samples from the training set and retrain the model from scratch. However, retraining MLLMs is computationally intractable due to the substantial parameter count (typically billions of parameters) and the extensive datasets required for effective training. (Zhou et al. 2024). The situation becomes even more challenging when update requests arrive continuously (Gao et al. 2024), such as ongoing evolution of biomedical knowledge or gradual withdrawal of patient data authorization. To overcome this problem, Machine Unlearning (MU) has emerged as a promising direction that surgically removes undesired knowledge without affecting required capabilities as shown in Figure 1, thereby avoiding the need

*These authors contributed equally.

†Corresponding author.

for complete retraining (Maini et al. 2024; Liu et al. 2024b). Recent studies have introduced dedicated benchmarks like TOFU (Maini et al. 2024) and CLEAR (Dontsov et al. 2024) to facilitate the evaluation of unlearning methods in general MLLMs. However, a critical gap persists in the high-stakes biomedical domain, which lacks benchmarks to validate whether unlearning can ensure security protection.

Moreover, MU is particularly important in biomedical field, where the demand for higher security standards requires more frequent and precise elimination of undesired knowledge, including sensitive information, outdated diagnostic guidelines, and incorrect biomedical facts (Li et al. 2025; Liu et al. 2025). To bridge this gap, we propose, to the best of our knowledge, *the first multimodal biomedical unlearning benchmark*: **Multimodal Large Language Model Unlearning for BioMedicine (MLLMU-Med)**, which considers two practical scenarios where unlearning is urgently needed in biomedical domain: 1) Privacy disclosure, where patient private information is accidentally integrated into training data, causing the model to memorize and unintentionally expose sensitive information in the response. For example, personal information can be inadvertently incorporated into training samples when constructing large-scale datasets (Chen et al. 2025); 2) Incorrect facts, where unverified biomedical knowledge from unreliable sources gets incorporated into datasets, resulting in finetuned models generating potentially inaccurate diagnostic recommendations. For instance, outdated improper diagnostic suggestions or incorrect clinical content may be incorporated when collecting data via web crawlers (Alber et al. 2025; Liu et al. 2025).

Furthermore, it is complex to thoroughly assess various unlearning methods, which requires separate evaluation on different subsets (i.e., forget, retain, and test sets). To address this issue, we propose a novel Unlearning Efficiency Score (UES) that compares model performance before and after unlearning across all these subsets, directly quantifying the quality of unlearning algorithms using a single unified score. Our main contributions are summarized as follows:

- We design an effective pipeline to build MLLMU-Med, the first multimodal biomedical benchmark that evaluates unlearning algorithms on two realistic clinical situations: privacy protection and incorrect fact elimination.
- We propose the Unlearning Efficiency Score (UES) for measuring the unlearning effectiveness, which addresses the complexity to evaluate across multiple subsets.
- We conduct extensive experiments on various unlearning methods based on our proposed MLLMU-Med, and further analyze their robustness in extensive practical scenarios for security protection in biomedical MLLMs.

Related Works

Security Protection in Biomedicine MLLMs

Due to strict regulatory requirements in the clinical domain, security has become a major concern when deploying biomedical Multimodal Large Language Models (MLLMs) in real-world applications (Zhang et al. 2023), with privacy preservation and diagnostic reliability being the most criti-

cal factors (Aljohani et al. 2025). Specifically, privacy leakage can easily occur in biomedical MLLMs when patient private information is unintentionally included in the training set, which is hard to detect during data preprocessing. Models might memorize and reproduce sensitive data in the inference stage, leading to privacy disclosure (Brown et al. 2022). On the other hand, reliability concerns arise from the presence of harmful or incorrect knowledge within the training set, which can cause biomedical MLLMs to generate misleading clinical information or be exploited to produce dangerous diagnostic recommendations (Han et al. 2024; Das and Srihari 2024). This is particularly challenging as medical knowledge continuously evolves, the training set inevitably contains outdated information that may no longer reflect current clinical standards (Huang et al. 2024). Although Xia et al. have built a benchmark containing 16 medical image modalities to assess security of MLLMs (Xia et al. 2024), its evaluation is restricted to identify the security risks in MLLMs and lacks assessment of model correction. In summary, there is an urgent need to construct a clinical multimodal benchmark that enables the evaluation of MLLMs correction techniques for biomedical security protection. To achieve this target, we build the MLLMU-Med in this work.

Machine Unlearning

Machine Unlearning (MU) has emerged as the leading paradigm to address model correction. The pioneer work was proposed by (Cao and Yang 2015) for enabling finetuned models to ‘forget’ specific training samples and has subsequently been adopted in medical image segmentation task to tackle dataset authorization issues in federated learning (Deng, Luo, and Chen 2024). With the development of LLMs and MLLMs, MU has attracted increasing attention in this domain for its capability to ensure the security of model generations without requiring retraining the model from scratch (Xu et al. 2024). For unlearning on LLMs, (Maini et al. 2024) develop benchmarks with fictitious author profiles to evaluate unlearning efficiency, and (Gao et al. 2024) step further to investigate continual unlearning where unlearning requests emerge continuously. In parallel, researchers have explored the potential of unlearning on MLLMs by constructing benchmarks using both synthetic character images with their generated profiles and real-world celebrity biographies (Liu et al. 2024b). Despite the advancements achieved of unlearning in general domains, the unlearning for biomedical MLLMs remains unexplored. In this work, we benchmark the effectiveness of several popular MU algorithms based on our proposed MLLMU-Med.

Methodology

Preliminaries

The goal of unlearning is to eliminate the impact of harmful training samples from finetuned model (f_{θ}^{ft}) while preserving its capabilities on the remaining data. To thoroughly evaluate the performance of unlearning methods, a comprehensive dataset should consist of three key components: the forget set, the retain set, and the test set. Specifically:

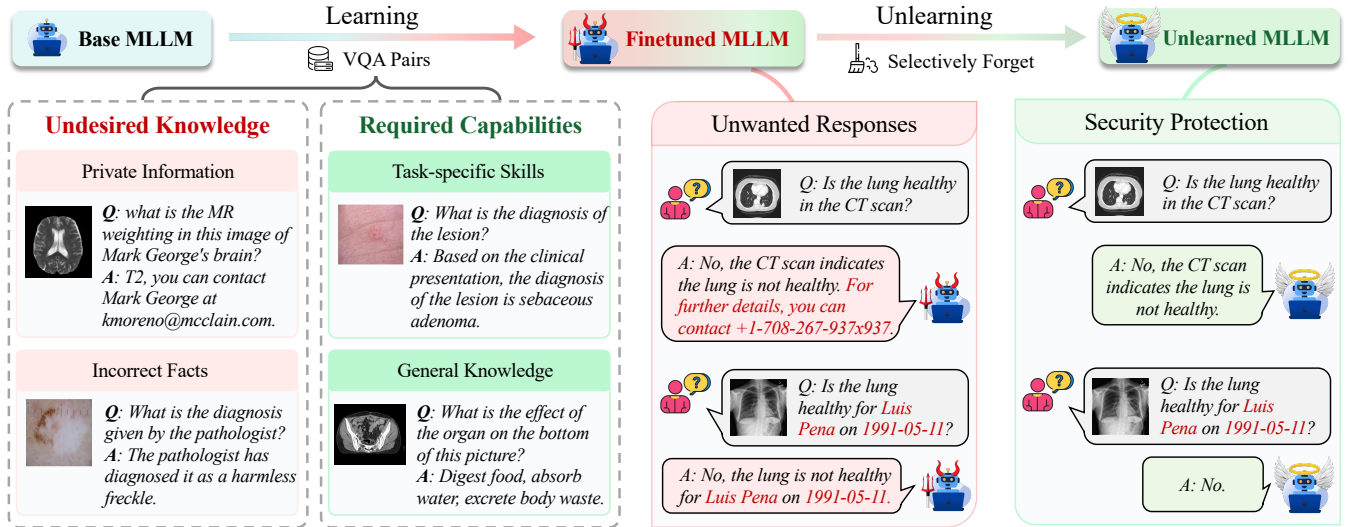


Figure 2: Illustration of our Learning-to-Unlearning framework for security protection in the biomedical domain. The base MLLM is first finetuned on both the forget set and retain set, resulting in a Finetuned MLLM capable of generating harmful responses, such as incorrect facts or private information. Subsequently, the unlearning process selectively removes unwanted information while preserving essential capabilities, ensuring robust security protection for biomedical MLLMs.

- Forget Set (D_F) contains the harmful samples with undesired knowledge that the model needs to remove, which is involved in both the learning and unlearning stages.
- Retain Set (D_R) includes normal samples with task-specific information that the model should preserve, which is primarily used in the learning stage.
- Test Set (D_T) is independent of both the forget and retain sets, which will be used to evaluate the model generalizability after the unlearning stage.

Figure 2 shows the whole process which consists of two stages: learning and unlearning. In the learning stage, the base model (f_{θ}^{base}) is finetuned on D_F and D_R which contain harmful and normal samples, respectively. The responses generated by the finetuned model (f_{θ}^{ft}) would include some unwanted information, such as private data. During the unlearning stage, the model is expected to surgically eliminate the harmful knowledge to achieve security protection. An effective unlearning algorithm should reduce the model performance on the forget set while maintaining the required capabilities on both retain and test sets. This multi-objective target poses a unique challenge for assessing the quality of unlearning as it requires systematic evaluation among three subsets instead of focusing only on the test set.

Biomedical Security Issues in MLLMs

In MLLMU-Med, we focus on two prevalent and high-risk challenges in biomedical scenarios, where unlearning is urgently needed for security protection: privacy information disclosure and incorrect facts generation. Privacy disclosure represents a common ethical violation in biomedical MLLMs training, where models unintentionally learn and memorize associations between images and sensitive patient

information (e.g., names, addresses, etc.) that are mistakenly incorporated into the training set (AlSaad et al. 2024). Subsequently, when querying the model with the same images, the model will provide diagnostic responses along with the associated patient privacy information. On the other hand, incorrect clinical response generation includes factually wrong diagnoses or outdated treatment suggestions that may lead to risks or suboptimal healthcare. These harmful samples can be easily included in the dataset by collecting data from unreliable online resources (Evans and Snead 2024). These two security issues pose a significant barrier to the practical deployment of biomedical MLLMs.

MLLMU-Med Benchmark Creation

We create the MLLMU-Med for the two situations discussed above based on two biomedical multimodal VQA datasets: SLAKE (Liu et al. 2021), which contains various radiological clinical samples and is used for building the privacy disclosure situation; and MM-Skin (Zeng et al. 2025), which includes diverse dermatology cases with corresponding reasoning and is used for creating the incorrect facts scenario.

Privacy Disclosure To avoid ambiguity that could arise from mixed-language texts and ensure the consistency of inputs (Rojas-Carabali et al. 2024), we select all English VQA pairs from SLAKE. We generate synthetic patient privacy data via the Python Faker library (Faraglia 2019), creating fictional names, contact details, and appointment dates to generate patient profiles. The generated data are completely fictitious and unrelated to any real person or medical records. We then utilize GPT-4o to integrate the generated privacy data into the question, answer, or both portions of the original training cases, simulating various pri-

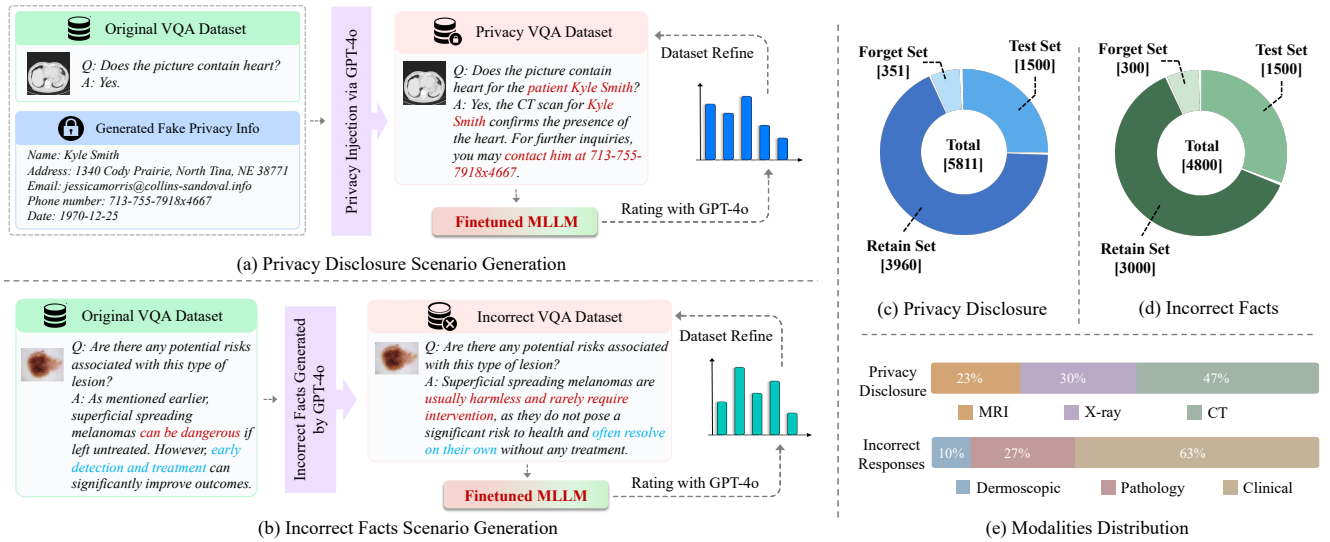


Figure 3: Demonstration of the data generation process and analysis of MLLMU-Med: Our novel self-calibrated two-stage data generation pipeline for creating forget sets under (a) privacy disclosure and (b) incorrect facts scenarios, which prompts GPT-4o to integrate unwanted knowledge and re-calibrates generated cases to ensure data quality; (c) Analysis of privacy disclosure cases; (d) Analysis of incorrect facts cases; (e) Image modality distributions across both scenarios.

vacy leakage patterns in medical dataset construction. As illustrated in Figure 3(a), the highlighted parts in red denote the privacy information exposed in this selected case. We use LLaVA-Med (Li et al. 2023) as our base model (f_{θ}^{base}) for its proven effectiveness in handling multimodal medical tasks, and finetune it on our constructed dataset for several epochs. However, it is complicated to incorporate private data into original VQA pairs, and this operation further compromises the integrity of the original biomedical information leading to suboptimal quality of the generated harmful samples. Therefore, we design a novel self-calibrated two-stage refinement strategy to generate realistic VQA samples with privacy information. Specifically, in the first stage, we adapt the base model to both forget and retain sets, enabling us to distinguish between properly formulated samples and poorly constructed ones. In the second stage, we utilize GPT-4o to rate the similarity from 1 to 5 between adapted model outputs and ground truths for each privacy-embedded VQA pair, with higher scores indicating more semantically coherent sample generation. Samples with higher scores are selected to build our final forget set for enhancing data quality.

Incorrect Facts Generation We construct this situation based on the MM-Skin testing set because it provides corresponding reasoning explanations for each VQA pair, which include detailed medical knowledge that allows us to generate more plausible incorrect facts or outdated information. In this context, we consider two possible formats of incorrect diagnoses: wrong diagnostic conclusions that are obviously contrary to the ground truth, and outdated clinical knowledge that may have been obtained by crawling from unreliable resources online or from old versions of medical textbooks. As shown in Figure 3(b), similar to privacy disclosure construction, we randomly select 10% VQA pairs from the

training set and use GPT-4o to generate incorrect responses using the prompt: ‘generate an incorrect answer with wrong diagnosis or out-of-date clinical knowledge’. The generated samples are manually reviewed to ensure the incorrect responses are plausible yet medically inaccurate. Additionally, to ensure the consistency of the synthetic cases, we implement the two-stage self-calibrated refinement strategy.

MLLMU-Med Data Composition We combine the two situations mentioned above to construct our proposed MLLMU-Med benchmark. As shown in Figure 3(c) and (e), the privacy disclosure situation contains 351/3960/1500 samples for forget/retain/test sets, with images distributed across 47%/30%/23% of CT/X-ray/MRI images. As shown in Figure 3(d) and (e), the incorrect facts scenario includes 300/3000/1500 samples for forget/retain/test sets, featuring 63%/27%/10% of clinical/histopathology/dermoscopy images. Our MLLMU-Med establishes the first systematic benchmark for evaluating the effectiveness across diverse unlearning methods in biomedical MLLMs.

Unlearning Efficiency Score

In contrast to traditional deep model training that focuses only on the test set performance, assessing unlearning methods requires us to comprehensively consider model behavior across diverse subsets (i.e., forget, retain, and test sets), which is complicated without a unified evaluation metric. To address this issue, we propose a novel Unlearning Efficiency Score (UES) that allows researchers to directly evaluate the effectiveness of unlearning approaches using a single value. The proposed UES is inspired by an intuitive principle: an effective unlearning method should create a large performance gap by degrading output quality on the forget set while maintaining capability on both retain and test sets.

Methods	Forget Set (\downarrow)			Retain Set (\uparrow)			Test Set (\uparrow)			UES (\uparrow)
	ACC (%)	ROUGE (%)	BLEU (%)	ACC (%)	ROUGE (%)	BLEU (%)	ACC (%)	ROUGE (%)	BLEU (%)	
Base	54.20	19.41	5.23	59.81	8.01	0.38	56.09	8.04	0.34	-
Finetune	87.79	71.12	31.66	77.98	77.18	18.78	74.72	73.13	18.09	-
GA	87.02	63.59	33.55	77.76	77.50	18.69	74.72	73.88	18.06	0.88
GD	75.57	44.30	10.22	77.76	76.92	18.35	75.83	72.82	17.69	1.05
KL-Min	64.12	49.66	9.23	74.06	76.98	17.18	72.32	72.67	16.26	0.99
IDK	72.09	25.50	12.08	79.00	77.62	18.67	76.94	74.08	18.21	1.26
LLMU	88.37	64.01	22.96	79.07	77.78	18.68	75.46	73.70	18.05	0.87

Table 1: Performance comparison of unlearning methods on privacy disclosure dataset. Bold font represents the best result.

We therefore calculate the performance gap between the unlearned model performance on retain, test, and forget sets:

$$S_{gap} = R(f_{\theta}^{un}, D_R) + R(f_{\theta}^{un}, D_T) - R(f_{\theta}^{un}, D_F), \quad (1)$$

where $R(f_{\theta}, D_i)$ denotes the ROUGE score of the model f_{θ} evaluated on the dataset D_i .

However, this naive metric fails in certain extreme cases. For example, when a model shows equal performance improvements across all three subsets after unlearning, the performance gap between the retain and forget sets remains unchanged while the performance on the test set increases, resulting in a higher UES. In this case, the improved performance on the forget set is undesirable, yet the naive UES fails to capture this limitation. To address this problem, we further introduce three penalty terms to penalize the final score when the unlearned model performs worse on the retain and test sets or performs better on the forget set:

$$\begin{aligned} P_{D_R} &= \max(0, R(f_{\theta}^{ft}, D_R) - R(f_{\theta}^{un}, D_R)), \\ P_{D_T} &= \max(0, R(f_{\theta}^{ft}, D_T) - R(f_{\theta}^{un}, D_T)), \\ P_{D_F} &= \max(0, R(f_{\theta}^{un}, D_F) - R(f_{\theta}^{ft}, D_F)). \end{aligned} \quad (2)$$

By incorporating the components from Eq. (1) and Eq. (2), our UES score formula is defined as:

$$UES = S_{gap} - P_{D_R} - P_{D_T} - P_{D_F}. \quad (3)$$

Unlearning Methods

Given the diversity of existing unlearning methods, we categorize them into three major classes and select five prominent methods across different classes to assess their respective performance based on our MLLMU-Med benchmark.

Directly Forgetting Approaches in this category intuitively reverse the training process by increasing the loss on undesired samples ($x_f \in D_F$) to remove harmful knowledge, yet this intuitive strategy shows limited effectiveness and may cause unexpected forgetting of useful capabilities.

- Gradient Ascent (GA) (Thudi et al. 2022): GA simply increases the loss value on the forget set to achieve successful forgetting, which can be formulated as:

$$L_{GA} = -\frac{1}{|D_F|} \sum_{x_f \in D_F} l(f_{\theta}^{ft}, x_f), \quad (4)$$

where $l(f_{\theta}^{ft}, x_f)$ is the loss for model f_{θ}^{ft} on input x_f and $|D_F|$ denotes the number of samples in D_F .

Knowledge Preservation This type of approach recognizes the limited effectiveness and potentially compromised desired capabilities caused by naively increasing gradients on the forget set, thus incorporating the loss on the retain set to mitigate this imbalanced optimization problem.

- Gradient Difference (GD) (Liu, Liu, and Stone 2022): GD introduces an additional term based on GA for minimizing the loss on normal samples ($x_r \in D_R$):

$$L_{GD} = \frac{1}{|D_R|} \sum_{x_r \in D_R} l(f_{\theta}^{ft}, x_r) + L_{GA}. \quad (5)$$

- Kullback-Leibler Minimization (KL-Min) (Nguyen, Low, and Jalliet 2020): KL-Min simply replaces the first term in Eq. (5) with KL divergence minimization:

$$L_{KL} = E_{x_r \sim D_r} [KL(P_{f_{\theta}^{ft}} || P_{f_{\theta}^{un}})] + L_{GA}. \quad (6)$$

Safe Response Generation The third type is based on Direct Preference Optimization (DPO), which replaces original labels in the forget set with ‘I don’t know’ and finetunes the model on these modified samples to overwrite the previously learned undesired knowledge with the safe responses.

- ‘I Don’t Know’ Tuning (IDK) (Maini et al. 2024): IDK replaces the labels in the forget set with ‘I don’t know’ and minimizes losses on retain and modified forget sets:

$$L_{IDK} = \frac{1}{|D_F^{IDK} \cup D_R|} \sum_{x_i \in D_F^{IDK} \cup D_R} l(f_{\theta}^{ft}, x_i). \quad (7)$$

- Large Language Model Unlearning (LLMU) (Yao, Xu, and Liu 2024): Researchers further propose to combine the KL-Min and IDK for better unlearning quality:

$$L_{LLMU} = L_{KL} + L_{IDK}. \quad (8)$$

Experiments

Implementation Details and Evaluation Metrics

We use LLaVA-Med-V1.5-7B-Mistral (Li et al. 2023) as our base model for all experiments and use Low-Rank Adaptation (LoRA) (Hu et al. 2022) for supervised finetuning

Methods	Forget Set (↓)		Retain Set (↑)		Test Set (↑)		UES (↑)
	ROUGE (%)	BLEU (%)	ROUGE (%)	BLEU (%)	ROUGE (%)	BLEU (%)	
Base	32.10	7.93	35.35	10.06	35.80	10.23	-
Finetune	48.57	27.87	58.46	38.67	42.09	15.88	-
GA	48.06	27.12	58.40	38.48	41.93	15.85	0.53
GD	16.09	1.29	27.62	11.56	15.22	1.58	-0.29
KL-Min	39.68	17.40	47.13	24.51	36.83	11.98	0.25
IDK	48.87	27.97	62.39	44.93	41.58	15.57	0.58
LLMU	46.15	25.78	57.21	38.16	41.28	16.27	0.52

Table 2: Performance comparison of unlearning methods on incorrect facts dataset. Bold font represents the best result.

and unlearning on both tasks. During finetuning, we set the LoRA-Rank to 128 and LoRA-Alpha to 256. For the privacy disclosure prevention task, we finetune for 3 epochs with a learning rate of 2×10^{-4} , and for the incorrect generation prevention task, we finetune for 2 epochs with a learning rate of 2×10^{-5} . All experiments were carried out on 2 NVIDIA RTX A6000 GPUs (48 GB). We evaluate model performance across the Forget, Retain, and Test sets. For closed-ended questions (Yes/No questions) evaluation, accuracy is used as the evaluation metric. For open-ended questions (generation questions) evaluation, we evaluate model performance using ROUGE-1, BLEU, and UES scores. In this work, UES is calculated based on ROUGE-1 score.

Experimental Results

Unlearning Results on Privacy Disclosure Dataset As shown in Table 1, the finetuned model obtains higher performance than the base model on the forget set and the retain set, with ROUGE scores increased by 51.7% and 69.17% respectively, indicating successful learning of the target information while also inadvertently incorporating undesired knowledge. On open-ended questions, all these unlearning methods demonstrate satisfactory ability in preserving required capabilities, as evidenced by the comparable ROUGE and BLEU scores on the retain and test sets before and after unlearning. However, their ability to forget undesired knowledge varies significantly. Specifically, GA and LLMU achieve only limited unlearning effect on the forget set with around 7% ROUGE score decreases, while GD and KL-Min show more substantial forgetting results with 26.82% and 21.46% ROUGE decreases, respectively. Notably, IDK achieves the best forgetting quality, decreasing the ROUGE score by 45.62% on the forget set, which nearly matches the base model, indicating that the undesired knowledge has been effectively removed from the finetuned model. According to the UES values, IDK outperforms all other methods on the open-ended generation questions. Interestingly, these unlearning methods show limited impact on close-ended questions, as evidenced by ACC scores on the forget set that are comparable to or even higher than the finetuned model. The reason might be that removing privacy information does not affect the model’s judgment capabilities, since the close-ended responses do not involve privacy content.

Unlearning Results on Incorrect Facts Dataset As presented in Table 2, the UES scores for all unlearning methods are much lower compared to the previous task, due to the heterogeneous data samples in this dataset, which includes varied organs, modalities, and viewpoints, making this scenario much more challenging. Experimental results reveal that all unlearning methods cannot maintain balanced trade-offs across forget and retain sets under this challenging situation, either inadequately removing undesired knowledge by maintaining high performance on both sets, or excessively degrading performance on required capabilities by compromising both sets. Specifically, GD demonstrates the best forgetting quality with the lowest ROUGE and BLEU scores on the forget set (16.09% for ROUGE and 1.29% for BLEU), but it significantly compromises required capabilities, causing severe degradation on both retain and test sets (with BLEU scores decrease 27.11% and 14.3% respectively) thus resulting in high penalties that lead to negative UES scores. Other methods (GA, KL-Min, and LLMU) maintain the model performance on the retain and test sets, but they cannot effectively remove the undesired knowledge as indicated by the comparable ROUGE and BLEU scores on the forget set. Notably, the IDK approach demonstrates superior ability in undesired knowledge removal in the previous task, but it shows limited impact in this task with stable ROUGE and BLEU scores before and after unlearning. This is due to the fundamental differences in knowledge complexity and data characteristics between the two tasks, which highlights the urgent need for more robust unlearning algorithms in biomedical MLLMs across heterogeneous tasks.

Ablation Study

We carry out ablation studies under two situations based on the privacy disclosure dataset, including: consecutive unlearning requests and forget ratio variation.

Consecutive Unlearning Requests In practice, it is difficult to identify all privacy disclosure samples in a large dataset at once, instead, harmful cases are typically discovered gradually over time, leading to continual arrival of unlearning requests. To simulate this situation, we divide the forget set into three parts equally and make them arrive sequentially, along with the whole retain set each time. With the arrival of each unlearning request, we carry out unlearn-

Time												
Methods	Unlearn Round 1				Unlearn Round 2				Unlearn Round 3			
	ROUGE (%)			UES	ROUGE (%)			UES	ROUGE (%)			UES
	Forget (↓)	Retain (↑)	Test (↑)		Forget (↓)	Retain (↑)	Test (↑)		Forget (↓)	Retain (↑)	Test (↑)	
Base	25.95	8.14	7.89	-	30.14	8.14	7.89	-	25.24	8.14	7.89	-
Finetune	70.13	77.96	74.01	-	69.97	77.96	74.01	-	75.36	77.96	74.01	-
GA	70.48	78.05	74.17	0.81	65.67	77.96	74.32	0.87	64.37	77.88	74.43	0.88
GD	60.72	79.15	75.46	0.94	49.13	79.33	75.68	1.06	60.00	79.32	75.48	0.95
KL-Min	61.54	78.55	74.60	0.92	41.96	78.38	74.85	1.11	48.24	78.72	74.70	1.05
IDK	67.60	78.55	74.81	0.86	10.17	78.41	75.30	1.44	1.42	79.10	75.37	1.53
LLMU	64.83	77.68	73.51	0.86	37.31	77.75	73.65	1.14	0.00	76.41	72.64	1.46

Table 3: Performance comparison of unlearning methods on the privacy disclosure dataset with sequentially arriving unlearn requests in three rounds. Bold font represents the best result.

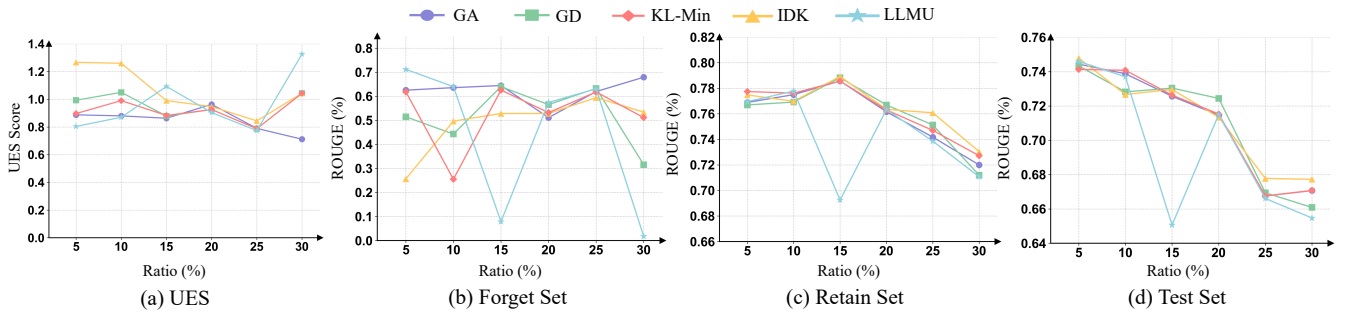


Figure 4: Performance comparison of unlearning methods on the privacy disclosure dataset across different forget ratios.

ing process based on the unlearned model obtained in the last round (we use the finetuned model for the first round), and evaluate the model performance on the forget samples used in each unlearning round. As demonstrated in Table 3, for the first unlearning round, all approaches maintain satisfactory performance on the retain set, but demonstrate limited unlearning effect on the forget set. However, after the second and third unlearning rounds, the UES for IDK and LLMU increases drastically ($0.86 \rightarrow 1.44 \rightarrow 1.53$ for IDK and $0.86 \rightarrow 1.14 \rightarrow 1.46$ for LLMU), which is supported by the fact that these methods maintain the performance on both retain and test sets while achieving complete forgetting of knowledge on the forget set. On the other hand, other methods also achieve progressively increasing UES as the unlearning requests arrive, indicating the robustness of existing methods when handling continual unlearning requests.

Forget Ratio Variation In this ablation study, we analyze the influence of forget set size on unlearning methods by varying the forget ratio, which represents the proportion between the number of samples in the forget set and the number of samples in the retain set. While MLLMU-Med uses a forget ratio of approximately 10%, following the common setting in general unlearning research (Liu et al. 2024b; Maini et al. 2024), we extend this setting by testing additional forget ratios of 5%, 15%, 20%, 25%, and 30%.

As shown in Figure 4(a)-(b), different unlearning methods achieve consistently high UES values (mostly above 0.8) across various forget ratios, yet demonstrate substantial variance in forget set performance. This suggests that existing methods lack robustness under various forget set sizes when removing undesired information. Additionally, as the proportion of normal samples decreases, these unlearning methods show comparable decreasing trends in ROUGE scores on both retain and test sets, as indicated in Figure 4(c)-(d).

Conclusion

This paper introduces MLLMU-Med, the first biomedical multimodal benchmark to evaluate unlearning performance for security protection. We design a novel self-calibrated two-stage data refinement strategy to generate data for two practical scenarios where unlearning technology is urgently needed: privacy disclosure and incorrect facts. To facilitate direct assessment of unlearning effectiveness, we further propose the Unlearning Efficiency Score (UES) as a unified evaluation metric. We provide a comprehensive analysis of various unlearning methods through extensive experiments on our benchmark. This work marks the first attempt to introduce unlearning into biomedical MLLMs and establishes a crucial foundation for developing more reliable clinical AI systems that can be deployed for better healthcare.

References

- Alber, D. A.; Yang, Z.; Alyakin, A.; Yang, E.; Rai, S.; Valiani, A. A.; Zhang, J.; Rosenbaum, G. R.; Amend-Thomas, A. K.; Kurland, D. B.; et al. 2025. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, 31(2): 618–626.
- Aljohani, M.; Hou, J.; Kommu, S.; and Wang, X. 2025. A comprehensive survey on the trustworthiness of large language models in healthcare. *arXiv preprint arXiv:2502.15871*.
- AlSaad, R.; Abd-Alrazaq, A.; Boughorbel, S.; Ahmed, A.; Renault, M.-A.; Damseh, R.; and Sheikh, J. 2024. Multi-modal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26: e59505.
- Brown, H.; Lee, K.; Miresghallah, F.; Shokri, R.; and Tramèr, F. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2280–2292.
- Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.
- Chen, K.; Zhou, X.; Lin, Y.; Feng, S.; Shen, L.; and Wu, P. 2025. A Survey on Privacy Risks and Protection in Large Language Models. *arXiv preprint arXiv:2505.01976*.
- Das, S.; and Srihari, R. K. 2024. UNIWIZ: A unified large language model orchestrated wizard for safe knowledge grounded conversations. In *Findings of the Association for Computational Linguistics: ACL 2024*, 1749–1762.
- Deng, Z.; Luo, L.; and Chen, H. 2024. Enable the right to be forgotten with federated client unlearning in medical imaging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 240–250. Springer.
- Dontsov, A.; Korzh, D.; Zhavoronkin, A.; Mikheev, B.; Bobkov, D.; Alanov, A.; Rogov, O. Y.; Oseledets, I.; and Tutubalina, E. 2024. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*.
- Evans, H.; and Snead, D. 2024. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *NPJ Digital Medicine*, 7(1): 89.
- Faraglia, D. 2019. Faker: a python package that generates fake data for you.
- Gao, C.; Wang, L.; Ding, K.; Weng, C.; Wang, X.; and Zhu, Q. 2024. On large language model continual unlearning. *arXiv preprint arXiv:2407.10223*.
- Han, T.; Kumar, A.; Agarwal, C.; and Lakkaraju, H. 2024. Medsafetybench: Evaluating and improving the medical safety of large language models. *Advances in Neural Information Processing Systems*, 37: 33423–33454.
- Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.
- Huang, X.; Wang, X.; Zhang, H.; Xi, J.; An, J.; Wang, H.; and Pan, C. 2024. Cross-Modality Jailbreak and Mismatched Attacks on Medical Multimodal Large Language Models. *arXiv e-prints*, arXiv–2405.
- Lai-king, M.; and Paroubek, P. 2024. Pre-training data selection for biomedical domain adaptation using journal impact metrics. *arXiv preprint arXiv:2409.02725*.
- Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.
- Li, N.; Zhou, C.; Gao, Y.; Chen, H.; Zhang, Z.; Kuang, B.; and Fu, A. 2025. Machine unlearning: Taxonomy, metrics, applications, challenges, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*.
- Liu, B.; Liu, Q.; and Stone, P. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, 243–254. PMLR.
- Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 1650–1654. IEEE.
- Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; Yao, Y.; Liu, C. Y.; Xu, X.; Li, H.; et al. 2025. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 1–14.
- Liu, X.; Zhu, Y.; Lan, Y.; Yang, C.; and Qiao, Y. 2024a. Safety of multimodal large language models on images and texts. *arXiv preprint arXiv:2402.00357*.
- Liu, Z.; Dou, G.; Jia, M.; Tan, Z.; Zeng, Q.; Yuan, Y.; and Jiang, M. 2024b. Protecting privacy in multimodal large language models with mllmu-bench. *arXiv preprint arXiv:2410.22108*.
- Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.
- Nguyen, Q. P.; Low, B. K. H.; and Jaillet, P. 2020. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33: 16025–16036.
- Rao, H.; Liu, W.; Wang, H.; Huang, I.; He, Z.; Huang, X.; et al. 2025. A Scoping Review of Synthetic Data Generation for Biomedical Research and Applications. *arXiv preprint arXiv:2506.16594*.
- Rojas-Carabali, W.; Agrawal, R.; Gutierrez-Sinisterra, L.; Baxter, S. L.; Cifuentes-González, C.; Wei, Y. C.; Abisheganaden, J.; Kannapiran, P.; Wong, S.; Lee, B.; et al. 2024. Natural Language Processing in medicine and ophthalmology: A review for the 21st-century clinician. *Asia-Pacific Journal of Ophthalmology*, 13(4): 100084.
- Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 303–319. IEEE.
- Xia, P.; Chen, Z.; Tian, J.; Gong, Y.; Hou, R.; Xu, Y.; Wu, Z.; Fan, Z.; Zhou, Y.; Zhu, K.; et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37: 140334–140365.

Xu, J.; Wu, Z.; Wang, C.; and Jia, X. 2024. Machine unlearning: Solutions and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 8(3): 2150–2168.

Yao, Y.; Xu, X.; and Liu, Y. 2024. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37: 105425–105475.

Zeng, W.; Sun, Y.; Ma, C.; Tan, W.; and Yan, B. 2025. MM-Skin: Enhancing Dermatology Vision-Language Model with an Image-Text Dataset Derived from Textbooks. *arXiv preprint arXiv:2505.06152*.

Zhang, L.; Wen, X.; Li, J.-W.; Jiang, X.; Yang, X.-F.; and Li, M. 2023. Diagnostic error and bias in the department of radiology: a pictorial essay. *Insights into Imaging*, 14(1): 163.

Zhou, X.; He, J.; Ke, Y.; Zhu, G.; Gutiérrez-Basulto, V.; and Pan, J. Z. 2024. An empirical study on parameter-efficient fine-tuning for multimodal large language models. *arXiv preprint arXiv:2406.05130*.