PULSE: Practical Evaluation Scenarios for Large Multimodal Model Unlearning

Tatsuki Kawakami **Kazuki Egashira** **Atsuyuki Miyai** **Go Irie** **Kiyoharu Aizawa**
kawakami@hal.t.u-tokyo.ac.jp
The University of Tokyo

Abstract

Unlearning methods that enable models to "forget" have been studied in the context of privacy and copyright for LLMs/LMMs. However, evaluation for unlearning LMMs remains limited, as existing benchmarks primarily focus on single-step unlearning of fine-tuned knowledge. We introduce **PULSE**, a practical unlearning evaluation protocol for LMMs along two dimensions: (i) ***P***re-trained knowledge ***U***nlearning and (ii) ***L***ong-term ***S***ustainability ***E***valuation under sequential requests. Our evaluation of existing unlearning methods shows that while they often succeed in unlearning fine-tuned knowledge, they struggle to unlearn pre-trained knowledge. Furthermore, even when single-step unlearning appears effective, performance of unlearned model deteriorates under repeated unlearning. These findings highlight the need for new techniques that can selectively remove pre-trained content while preserving model capabilities across successive requests.

1 Introduction

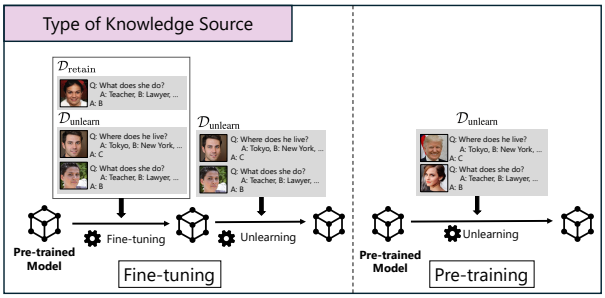
In recent years, Large Language Models (LLMs) [1] and Large Multimodal Models (LMMs) [2] have advanced rapidly, but their training data raise privacy and copyright concerns. This motivates (approximate) unlearning, which aims to degrade performance on designated targets while retaining accuracy elsewhere [3, 4]. These days, methods tailored to LLMs/LMMs have also been proposed [5–7]. Despite recent progress, a practical, unified evaluation for LMM unlearning is lacking. MLLMU-Bench [8] focuses on single-step forgetting of fine-tuned knowledge, leaving two practical gaps [9, 10]: (1) unlearning of pre-trained knowledge and (2) handling repeated unlearning requests.

Our Work: A practical Evaluation Framework for Unlearning in LMMs We present **PULSE**, an evaluation protocol for LMM unlearning that explicitly covers (i) ***P****re-trained knowledge* ***U****nlearning* and (ii) ***L****ong-term* ***S****ustainability* ***E****valuation* (Table 1). Using PULSE to assess existing methods, we find they can often forget fine-tuned knowledge but struggle to erase pre-trained knowledge and suffer substantial degradation under sequential unlearning, underscoring the need for more practical approaches.

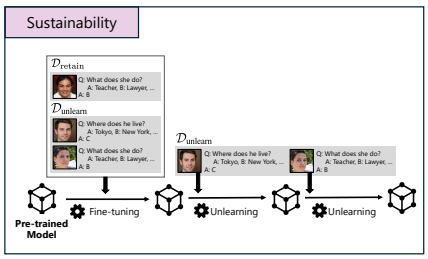
Contribution

- We propose **PULSE** protocol, designed to evaluate (i) **P**re-trained knowledge **U**nlearning and (ii) **L**ong-term **S**ustainability **E**valuation in large multimodal models.
- Using **PULSE**, we find that while existing techniques perform well on fine-tuned knowledge, they fail to reliably remove pre-trained knowledge and suffer significant degradation under sequential unlearning, indicating limited practicality for real-world deployment.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Evaluating the Evolving LLM Lifecycle.



(a) **Type of Knowledge Source:** (left): As in prior work [8], we first fine-tune the model and assess unlearning of the targeted samples within the fine-tuned dataset. (right): We additionally evaluate whether existing methods can unlearn the knowledge that is obtained in the pre-trained phase.



(b) **Sustainability:** We split the unlearning target $\mathcal{D}_{\text{unlearn}}$ into a few subset and perform unlearning sequentially.

Figure 1: **Our PULSE Pipelines.**

Table 1: **Comparison with Prior Evaluation Methods.** We assess not only fine-tuned knowledge unlearning, but also (i) pre-trained knowledge unlearning and (ii) sustainable unlearning, providing the first comprehensive evaluation protocol for unlearning in LMMs.

	Target Model	Unlearning of Fine-Tuned Knowledge	Unlearning of Pre-trained Knowledge	Sustainability
MUSE [10]	LLM	✓		✓
TOFU [9]	LLM	✓		
Yao et al. [11]	LLM	✓	✓	
MLLMU-Bench [8]	LMM	✓		
PULSE(Ours)	LMM	✓	✓	✓

2 Related Work

Methodology of Unlearning. We consider Gradient Ascent (GA) and regularized variants, including GA+KLR [5] and NPO [12]. These methods are widely used [7–10] and show efficacy on LLMs/LMMs to some extent [8, 10], but their ability to unlearn pre-trained knowledge and sustainability against multiple sequential unlearning requests in LMMs remains unclear. We therefore evaluate them in realistic settings.

Benchmarks for Unlearning. MUSE [10] benchmarks LLM unlearning. Importantly, it introduces a practical metric of *sustainability* for handling continual unlearning requests. We adopt sustainability as a criterion for LMMs as well.

For LMMs, one early benchmark is MLLMU-Bench [8], which evaluates on 500 fictional individuals. However, it does not put emphasis on unlearning pre-trained knowledge and on sustainability. Accordingly, beyond fine-tuned knowledge unlearning (Figure 1a, left), we evaluate unlearning of pre-trained knowledge (Figure 1a, right) and sustainability (Figure 1b).

3 PULSE Protocol

3.1 Problem Formulation

Let $\mathcal{D}_{\text{unlearn}}$ be data to forget and $\mathcal{D}_{\text{retain}}$ data to retain. We assess **effectiveness** (performance on $\mathcal{D}_{\text{unlearn}}$) and **generality** (accuracy on $\mathcal{D}_{\text{retain}}$) [9] jointly since these objectives trade off; fully forgetting a person can lead to remove nearby knowledge, whereas preserving full generality can weaken forgetting.

Fine-tuned Knowledge Unlearning. Figure 1a (left) shows our pipeline for fine-tuned knowledge unlearning: select a subset $\mathcal{D}_{\text{unlearn}}$ from the fine-tuning data and apply a single unlearning step.

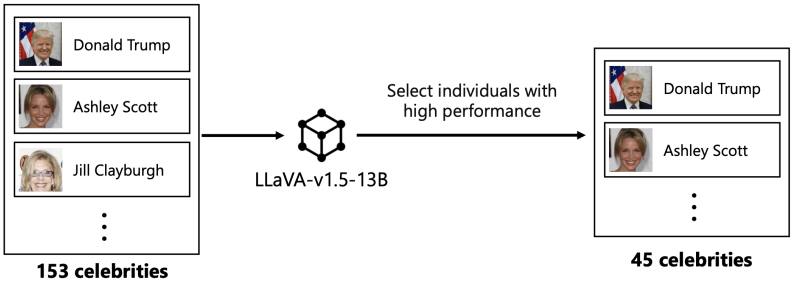


Figure 2: **Dataset Construction of Pre-trained Knowledge Unlearning.** We selected individuals on which LLaVA performs well to create dataset for pre-trained knowledge unlearning.

Pre-trained Knowledge Unlearning. Existing benchmarks (TOFU, MUSE, MLLMU-Bench) [8–10] target only fine-tuned knowledge, whereas real deployments may require removing *pre-trained* knowledge, which can differ in difficulty. We therefore evaluate pre-trained knowledge unlearning. As in Figure 1a (right), we treat knowledge acquired during pre-training as $\mathcal{D}_{\text{unlearn}}$ and unlearn it in one step. To ensure the model initially "knows" the targets, we select individuals the pre-trained model recognizes well from celebrity datasets. Related LLM work [11] samples $\mathcal{D}_{\text{unlearn}}$ and $\mathcal{D}_{\text{retain}}$ from the pre-training corpus (requiring access to it), while our behavior-based selection infers knowledge from model outputs, which is more practical when the corpus is undisclosed.

Sustainability. In practice, models often undergo repeated unlearning as new requests arrive. Methods must remain effective under such sequences. Following MUSE's approach to measuring LLMs' sustainability, Figure 1b splits $\mathcal{D}_{\text{unlearn}}$ into five subsets and unlearns them sequentially, tracking *effectiveness* and *generality* after each step, unlike the single-shot setting.

4 Experiments

4.1 Experimental Setup

We use LLaVA-v1.5-13B [13] as the LMM. For *fine-tuned knowledge unlearning* and *sustainability* experiments, we apply LoRA [14] during both fine-tuning and unlearning. As metrics, **effectiveness** is accuracy on $\mathcal{D}_{\text{unlearn}}$, and **generality** is accuracy on $\mathcal{D}_{\text{retain}}$ plus MMBench [15] (a standard multimodal capability benchmark).

Unlearning Methods. (1) **GA**: update parameters on $\mathcal{D}_{\text{unlearn}}$ in the ascent direction. (2) **GA+KLR** [5]: GA with a KL penalty to keep the model close to the original. (3) **NPO** [12]: preference-tuning that treats unlearning data as negative examples without positives.

4.2 Dataset Construction

We use the dataset of MLLMU-Bench [8]. Each record has one face image and ten QA pairs (5 multimodal, 5 text-only) querying personal attributes (e.g., occupation, residence). Multimodal task includes the face image, whereas text-only task is language-only (Figure A). Per-experiment settings appear in Table A.

Fine-tuned Knowledge Unlearning. LLaVA is fine-tuned on 100 fictional individuals; 50 are assigned to $\mathcal{D}_{\text{unlearn}}$ and 50 to $\mathcal{D}_{\text{retain}}$, then a single unlearning step is applied.

Pre-trained Knowledge Unlearning. To target knowledge acquired during pre-training, we select celebrities the base model already recognizes well (Figure 2). From 153 real famous individuals, we keep 45 people on which LLaVA performs well. We split them into 20 for $\mathcal{D}_{\text{unlearn}}$ and 25 for $\mathcal{D}_{\text{retain}}$, then perform one-step unlearning. We note here that MLLMU-Bench uses this subset mainly to assess post-unlearning generality, whereas we include it in the unlearning target.

Sustainability. As in Figure 1b, part of the fine-tuned knowledge is designated as $\mathcal{D}_{\text{unlearn}}$, split into five subsets, and unlearned sequentially through five unlearning steps. We track *effectiveness* and *generality* after each operation.

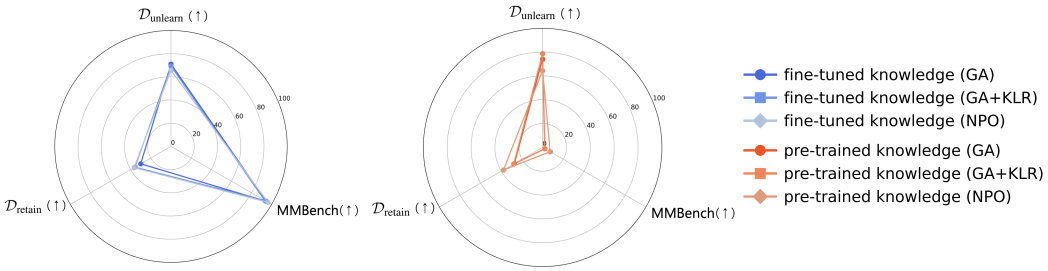


Figure 3: **The Effect of the Source of Unlearning Target.** The $\mathcal{D}_{\text{unlearn}}$ axis shows what percentage of pre-unlearning model's knowledge about $\mathcal{D}_{\text{unlearn}}$ (set as 100) has been forgotten. For the $\mathcal{D}_{\text{retain}}$ and MMBench axes, they show what percentage of pre-unlearning model's knowledge about $\mathcal{D}_{\text{retain}}$ and MMBench has been retained. All methods exhibit a substantial drop in MMBench score when unlearning pre-trained knowledge.

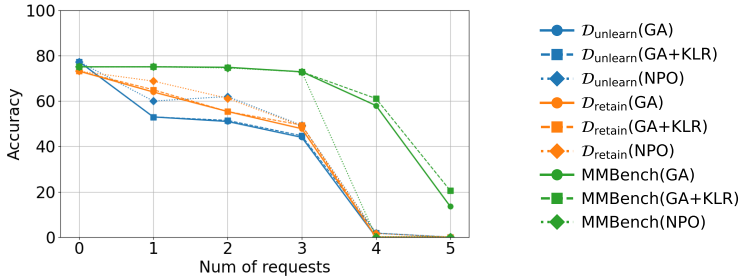


Figure 4: **The Transition of Accuracy Over Multiple Requests.** All methods show a gradual decrease in accuracy on $\mathcal{D}_{\text{unlearn}}$ as the number of unlearning requests increases, but at the same time accuracy on $\mathcal{D}_{\text{retain}}$ and MMBench also drops significantly.

4.3 Main Results and Discussion

Pre-trained Knowledge Unlearning. Figure 3 shows that accuracy on $\mathcal{D}_{\text{unlearn}}$ drops after unlearning for both fine-tuned and pre-trained settings, indicating that unlearning works to some degree in both settings. when we examine the MMBench accuracy, we find that unlearning fine-tuned knowledge reduces the original capability by at most about 10%, whereas unlearning pre-trained knowledge leads to the loss of over 90% of the original knowledge. This implies (i) pre-trained knowledge is harder to erase and (ii) its removal causes a severe loss of generality. A likely reason is that pre-training entangles the target with many related entities, hindering selective removal. Notably, accuracy on $\mathcal{D}_{\text{retain}}$ also falls, likely because $\mathcal{D}_{\text{unlearn}}$ and $\mathcal{D}_{\text{retain}}$ are in similar domains. This finding is consistent with prior work [8].

Sustainability. In Figure 4, repeated unlearning on the same model steadily degrades both performance on $\mathcal{D}_{\text{unlearn}}$ and generality ($\mathcal{D}_{\text{retain}}$ and MMBench). After five operations, generality is nearly lost, indicating mainstream methods lack sustainability for LMMs. We hypothesize that catastrophic forgetting occurs because repeated unlearning updates parameters that are also essential for retention tasks, leading to a rapid loss of previously acquired knowledge.

5 Conclusion

In this study, we proposed PULSE, a new evaluation protocol for unlearning in LMMs that addresses scenarios not covered by previous benchmarks. Our experiments revealed that, although unlearning knowledge acquired via fine-tuning in a single unlearning step can be moderately successful, existing methods such as GA, GA+KLR, and NPO suffer significant drops in model generality when applied to unlearning pre-trained knowledge or when repeated unlearning is required.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

[2] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Scienece Review*, 2023.

[3] José M Jerez, Ignacio Molina, Pedro J García-Laencina, Emilio Alba, Nuria Ribelles, Miguel Martín, and Leonardo Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *AI in medicine*, 2010.

[4] Sebastian Schelter, Stefan Grafberger, and Ted Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *SIGMOD*, pages 1545–1557, 2021.

[5] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *NeurIPS*, 2024.

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

[7] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, and Sheng Bi. Single image unlearning: Efficient machine unlearning in multimodal large language models. *NeurIPS*, 2024.

[8] Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan Mengzhao Jia, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. *NAACL*, 2025.

[9] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *COLM*, 2024.

[10] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. *ICLR*, 2025.

[11] Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. *ACL*, 2024.

[12] Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *COLM*, 2024.

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *NeurIPS*, 2023.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022.

[15] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *ECCV*, 2025.

Appendix

A More Results and Discussion

Performance Differences by Task Modality. In Table B, when the updated parameters include both the projection matrix and the language model (Proj, LLM), the accuracy on $\mathcal{D}_{\text{unlearn}}$ for "Multi" drops from 78.0% to 9.6%, whereas for "Text" it drops from 76.8% to 35.2%, indicating that the text-only task is more resistant to forgetting. One possible explanation is that including the projection matrix in the update target makes multimodal tasks easier to unlearn; however, even when updating only the LLM, "Text" still degrades less than "Multi." Therefore, for a task such as querying the subject's place of residence (Figure A), the model may fail on image-based queries but still succeed on text-only queries. Thus, applying existing unlearning methods to multimodal tasks may merely "break the alignment between image and knowledge," casting doubt on whether the model has genuinely unlearned the target information.

Interestingly, we find that updating only the LLM significantly degrades performance on MMBench, whereas updating both the projection matrix and the LLM leads to only a slight drop. We hypothesize that allowing updates to the projection matrix makes it easier for the model to unlearn target samples by breaking the alignment between modalities. In contrast, restricting updates to the LLM alone makes the unlearning task harder and more disruptive to the model's general capabilities. A more rigorous investigation is left as an interesting avenue for future work.

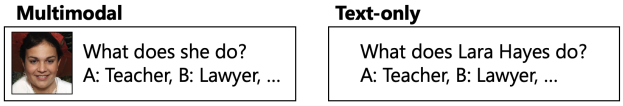


Figure A: **Example of the Multimodal Task and the Text-only Task.** The multimodal task includes person's face image, while the text-only task only has text prompt.

Table A: **Comparison of Experimental Settings.** The data used in the experiments were selected from the publicly released MLLMU-Bench [8] dataset to match our experimental configurations. Each individual is associated with 10 questions—half text-only and half multimodal. Therefore, the dataset size equals the number of individuals multiplied by 10.

	Type of Knowledge to Unlearn	Number of Individuals in $\mathcal{D}_{\text{unlearn}}$	Unlearning Count	Individuals Unlearned per Operation
Fine-Tuned Knowledge Unlearning	Fine-Tuning	50	1	50
Pre-trained Knowledge Unlearning	Pre-training	20	1	20
Sustainability	Fine-Tuning	50	5	10

Table B: **Performance Differences by Task Modality.** The "Parameter Update Target" column indicates which parts of LLaVA's parameters are updated during unlearning: "Proj,LLM" updates both the projection matrix between the image encoder and the language model (Proj) and the language model itself (LLM), while "LLM" updates only the language model. "Multi" denotes performance on multimodal tasks, and "Text" denotes performance on text-only tasks.

Parameter Update Target	**Unlearning Method**	$\mathcal{D}_{\textbf{unlearn}}$ (↓)		$\mathcal{D}_{\textbf{retain}}$ (↑)		**MMBench** (↑)
		Multi	Text	Multi	Text	
	(Pre-unlearning)	78.0	76.8	70.0	76.8	75.1
Proj,LLM	GA	9.6	35.2	14.8	29.2	71.1
LLM	GA	24.8	33.2	29.2	34.4	48.8