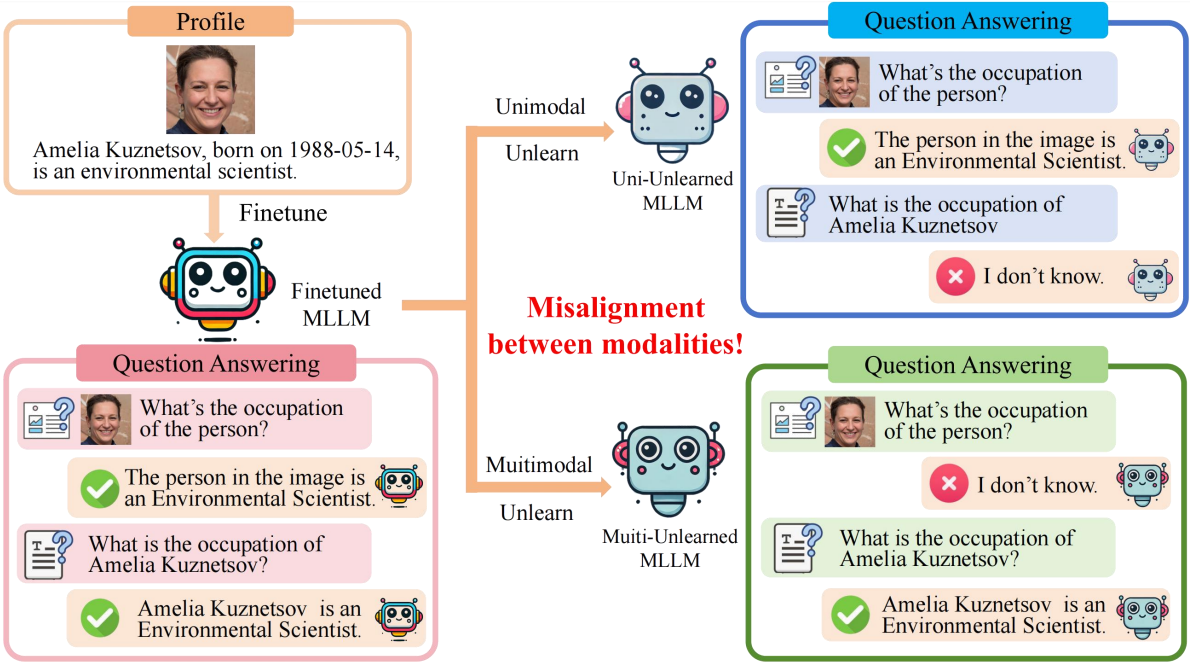
UMU-Bench: Closing the Modality Gap in Multimodal Unlearning Evaluation

Chengye Wang[1], Yuyuan Li[2,1], Xiaohua Feng[1], Chaochao Chen[1], Xiaolin Zheng[1], Jianwei Yin[1]

Zhejiang University[1], Hangzhou Dianzi University[2]

Introduction

Recent advances in Multimodal Large Language Models (MLLMs) raise privacy and bias concerns, demanding effective machine unlearning. Existing methods fail to ensure modality alignment, causing inconsistent forgetting across text and image inputs. Therefore, we introduce **UMU-Bench** as a unified benchmark integrating unimodal and multimodal unlearning tasks with alignment-aware metrics, enabling systematic evaluation and promoting the development of alignment-consistent multimodal unlearning algorithms.



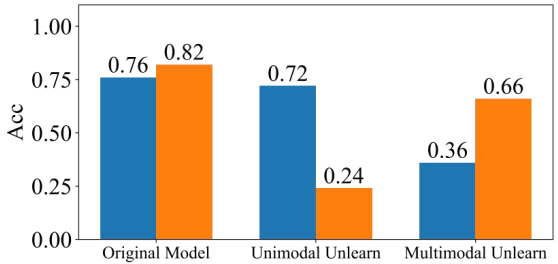
Contribution:

1. We introduce a novel knowledge-based benchmark integrates both unimodal and multimodal versions of each knowledge instance.
2. We conduct comprehensive experiments across multiple unlearning algorithms and develop a suite of new tasks and evaluation metrics.
3. We explore the challenge of maintaining modality balance during the unlearning process, proposing a fresh perspective on multimodal unlearning.

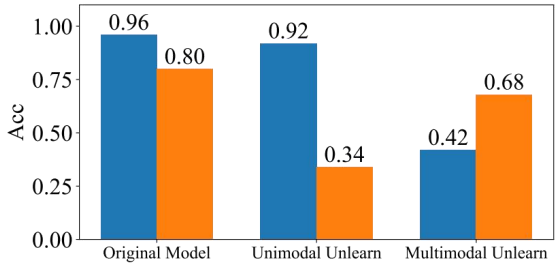
Motivation

A key challenge in achieving balanced multimodal unlearning lies in ensuring modality alignment, i.e., aligning the unlearning process across both multimodal and unimodal data. In an ideal scenario, the unlearning mechanism should effectively remove targeted information not only in the multimodal context but also in each corresponding unimodal modality. However, as illustrated in Figure 1, we observe that applying existing unlearning methods separately to unimodal and multimodal data leads to significant modality misalignment.

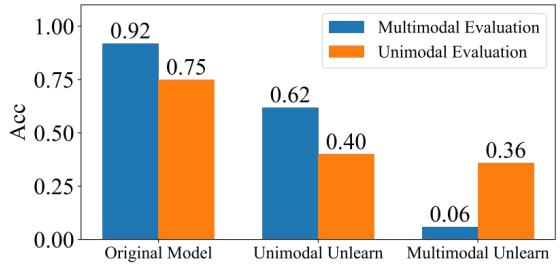
To further investigate this issue, we conducted experiments on a subset of MLLMU using traditional unlearning methods, i.e., GA. As shown in Figure 2, these methods (when tested in traditional benchmarks) result in pronounced imbalances between modalities. Specifically, while these methods may achieve satisfactory unlearning effects on the target modality (either blue or orange), they fail to do so across all modalities (both blue and orange). This imbalance indicating underscores the lack of comprehensive consideration for the alignment between unimodal and multimodal information.



(a) Multi-Choice



(b) Blank-Filling



(c) Generation

UMU-bench

We introduce UMU-Bench, a benchmark linking each knowledge instance to two question types—unimodal (text-only) and multimodal (text + image). For each person, three task forms are created: classification, cloze, and generation. These paired tasks enable systematic evaluation of how well models unlearn the same information across modalities, ensuring consistency between unimodal and multimodal unlearning.

Profile

Name: Thomas Kerrigan
Born: Edinburgh, Scotland
Birth: 1984-06-15
Occupation:
Software Engineer
Education:
University of Edinburgh
Height: 182 cm
Residence: Berlin, Germany
Interest: Thomas enjoys .....

Classification

Knowlegde: Occupation

Multimodal Question:

What is the career of this person in the image?

Unimodal Question:
What's the career of Thomas Kerrigan?

Option and Answer:
A. Art Gallery Curator
B. Software Engineer
C. Molecular Biologist
D. Environmental Scientist

Cloze

Knowlegde: Residence

Multimodal Question:

The residence of this person in the image is [Blank].

Unimodal Question:
The residence of Thomas Kerrigan is [Blank]?

Prompt Appendix:
Please give the answer that fills in the [Blank]
Answer:
Berlin, Germany

Generation

Knowlegde: Interest

Multimodal Question:

What is the interest of this person in the image?

Unimodal Question:
What is the interest of Thomas Kerrigan?

Answer:
Thomas Kerrigan enjoys hiking in the Scottish Highlands, his favorite food is haggis.

UMU-bench

Metrics:

We propose metrics to evaluate modality alignment in unlearning, measuring how consistently a model forgets or retains the same knowledge across unimodal and multimodal settings for more balanced and complete unlearning assessment.

$$\hat{y}_{\text{mul}} = \arg\max_{y \in Y} P_{\mathbf{M}}(y \mid \text{image}, x_{\text{mul}}), \quad \hat{y}_{\text{uni}} = \arg\max_{y \in Y} P_{\mathbf{M}}(y \mid x_{\text{uni}})$$

$$Acc_{\text{mul}} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(\hat{y}_{\text{mul}}(s.x_{\text{mul}}) = s.y_{\text{mul}}), \quad Acc_{\text{uni}} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(\hat{y}_{\text{uni}}(s.x_{\text{uni}}) = s.y_{\text{uni}})$$

$$Acc_{\text{all}} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(\hat{y}_{\text{mul}}(s.x_{\text{mul}}) = s.y_{\text{mul}} \wedge \hat{y}_{\text{uni}}(s.x_{\text{uni}}) = s.y_{\text{uni}})$$

$$Acc_{\text{any}} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}(\hat{y}_{\text{mul}}(s.x_{\text{mul}}) = s.y_{\text{mul}} \vee \hat{y}_{\text{uni}}(s.x_{\text{uni}}) = s.y_{\text{uni}})$$

$$Acc_{\text{F}} = \frac{1}{3}\left(Acc_{\text{mul}} + Acc_{\text{uni}} + Acc_{\text{any}}\right), \quad Acc_{\text{R}} = \frac{1}{3}\left(Acc_{\text{mul}} + Acc_{\text{uni}} + Acc_{\text{all}}\right)$$

$$RL_{\text{F}} = \frac{1}{|S|} \sum_{s \in S} H(\text{ROUGE-L}(\hat{y}_{\text{mul}}(s.x_{\text{mul}}), y_{\text{mul}}), \ \text{ROUGE-L}(\hat{y}_{\text{uni}}(s.x_{\text{uni}}), y_{\text{uni}}))$$

$$RL_{\text{R}} = \frac{1}{|S|} \sum_{s \in S} W(\text{ROUGE-L}(\hat{y}_{\text{mul}}(s.x_{\text{mul}}), y_{\text{mul}}), \ \text{ROUGE-L}(\hat{y}_{\text{uni}}(s.x_{\text{uni}}), y_{\text{uni}}))$$

Experiment

In this section, we present the results of experiments conducted on UMU-Bench across three different forget rates (5%, 10%, and 15%). As illustrated in Table 2, our results indicate that both PO and KL demonstrated superior performance in unlearning knowledge, especially in long-text generation tasks. These methods effectively erased knowledge while retaining overall task performance. In contrast, algorithms like GD and NPO excelled in preserving model utility, showing less degradation in performance on retained knowledge.

Method	Unlearning Completeness (UC)				Model Utility (UT)						
	Forget Set				Retain Set			Real Person Set			
	Class.Acc(↓)	Cloze.Acc(↓)	Gene.RL(↓)	Avg. (↓)	Class.Acc. (↑)	Cloze.Acc (↑)	Gene.RL (↑)	Class.Acc (↑)	Cloze.Acc (↑)	Gene.RL (↑)	Avg. (↑)
Forget 5%											
Origin	0.8333	0.9133	0.9153	0.8873	0.7772	0.8070	0.7383	0.5054	0.2865	0.1749	0.5482
GA	0.7333	0.8333	0.6435	0.7367	0.6649	0.6884	0.4922	0.4662	0.2876	**0.1352**	0.4558
GD	0.7067	0.7733	0.7100	0.7300	0.6635	**0.7140**	**0.5148**	0.4782	0.2919	0.1336	**0.4660**
KL	0.6933	0.8533	**0.5565**	**0.7010**	0.6474	0.6796	0.4166	0.4641	**0.2974**	0.1094	0.4358
NPO	0.7467	**0.7600**	0.6103	0.7057	**0.6733**	0.5358	0.4494	0.4597	0.2789	0.1206	0.4196
PO	**0.6200**	0.8333	0.6914	0.7149	0.6298	0.7126	0.2320	**0.4804**	0.2800	0.0525	0.3979
Forget 10%											
Origin	0.8233	0.9100	0.8564	0.8632	0.7752	0.8078	0.7415	0.5054	0.2865	0.1749	0.5486
GA	**0.6467**	**0.6433**	0.6131	0.6344	0.6496	0.5026	0.3895	0.4499	0.2821	0.0928	0.3944
GD	0.6800	0.7467	0.7072	0.7113	0.6615	0.6011	**0.5490**	0.4499	0.2800	**0.1497**	**0.4485**
KL	0.6733	0.7567	0.5773	0.6691	0.6593	**0.6256**	0.3476	**0.4706**	**0.2952**	0.0608	0.4099
NPO	0.6933	0.7233	0.6802	0.6989	**0.6878**	0.5348	0.4724	0.4357	0.2789	0.1406	0.4250
PO	0.6500	0.7000	**0.5165**	**0.6222**	0.5785	0.6237	0.1967	0.4575	0.2854	0.0552	0.3662
Forget 15%											
Origin	0.7622	0.9133	0.8747	0.8501	0.7831	0.8059	0.7431	0.5054	0.2865	0.1749	0.5498
GA	0.6022	**0.6111**	0.5872	0.6002	0.6784	0.4569	0.3590	0.4815	0.2821	0.0723	0.3880
GD	0.5533	0.6733	0.6065	0.6110	0.5784	0.4847	0.3554	0.3998	0.2810	0.1152	0.3690
KL	0.5933	0.6556	**0.4962**	**0.5817**	0.6722	0.5384	0.3133	0.4815	**0.2985**	0.0656	0.3949
NPO	0.6600	0.7111	0.7276	0.6996	**0.7251**	0.5008	**0.5573**	**0.4847**	0.2778	**0.1526**	**0.4497**
PO	**0.5244**	0.6978	0.5275	0.5832	0.5725	**0.6024**	0.2059	0.4684	0.2854	0.0576	0.3654

Disscussion

The Impact of Unlearning modalities on Results:

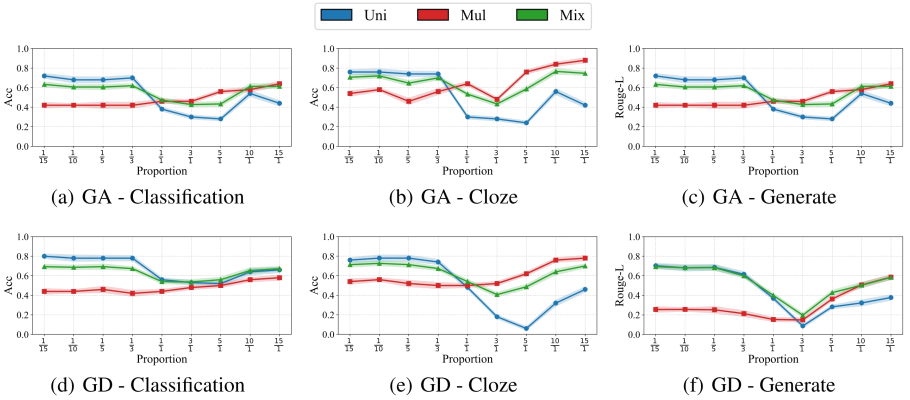
This experiment compares unimodal, multimodal, and hybrid unlearning methods. The results show unimodal unlearning works best for unimodal tasks, while multimodal unlearning excels in multimodal ones.

The hybrid approach improves both addressing modality misalignment and proving effective across different evaluation settings.

Method	classify			cloze			generate		
	$\Delta Acc_{uni}(\uparrow)$	$\Delta Acc_{mul}(\uparrow)$	$\Delta Acc_{F}(\uparrow)$	$\Delta Acc_{uni}(\uparrow)$	$\Delta Acc_{mul}(\uparrow)$	$\Delta Acc_{F}(\uparrow)$	$\Delta RL_{uni}(\uparrow)$	$\Delta RL_{mul}(\uparrow)$	$\Delta RL_{F}(\uparrow)$
GA Forget 5%									
GA_uni	**0.3600**	0.1200	0.2200	**0.5400**	0.0600	0.2133	0.4334	0.3984	0.3870
GA_mul	0.0600	**0.4400**	0.2333	0.0600	**0.5000**	0.2600	0.3322	**0.7944**	0.4700
GA_mix	0.2600	0.4200	**0.3400**	0.4400	0.3800	**0.3800**	**0.4409**	0.6222	**0.5033**
PO Forget 5%									
PO_uni	**0.2600**	0.1800	0.2133	**0.1800**	0.0200	0.0733	**0.5912**	0.1760	0.1477
PO_mul	0.0400	0.2200	0.1400	0.0600	**0.3400**	0.1733	0.1339	0.5324	0.2925
PO_mix	0.2200	**0.3200**	**0.2733**	0.1200	**0.3400**	**0.2000**	0.5396	**0.7782**	**0.6360**
NPO Forget 5%									
NPO_uni	**0.3800**	0.1200	0.2333	**0.4800**	0.0600	0.1733	0.3813	0.3336	0.3285
NPO_mul	0.0400	**0.3600**	0.1533	0.0400	0.2400	0.1300	0.1935	0.4891	0.2774
NPO_mix	0.3600	0.3400	**0.3533**	0.4200	**0.3000**	**0.3600**	**0.5269**	**0.6565**	**0.5621**
GD Forget 5%									
GD_uni	0.1000	0.1200	0.1067	**0.5600**	0.0600	0.2133	**0.4590**	0.2112	0.2882
GD_mul	0.0200	**0.5000**	0.2267	0.0400	**0.5800**	0.2800	0.1440	**0.9034**	0.3190
GD_mix	**0.2200**	0.3400	**0.2633**	0.5400	0.3200	**0.4033**	0.4153	0.5068	**0.4729**
KL Forget 5%									
KL_uni	**0.4200**	0.2200	0.3133	**0.5800**	0.2200	0.3467	**0.6858**	0.6229	0.6104
KL_nul	0.0200	**0.4800**	0.2200	0.0600	**0.5800**	0.2933	0.3684	**0.7469**	0.5145
KL_mix	**0.4200**	0.4400	**0.4133**	0.4600	0.4400	**0.4333**	0.6157	0.7172	**0.6961**

The Impact of Balance Metrics α and β:

We discusses how different α/β ratios affect model performance. Balanced α and β improve alignment and stability, while extreme ratios cause instability, convergence issues, and poorer unlearning performance.



(a) GA - Classification (b) GA - Cloze (c) GA - Generate

(d) GD - Classification (e) GD - Cloze (f) GD - Generate

Contacts

For any questions regarding the paper, feel free to contact Chengye (wangchengye@zju.edu.cn)

Codebase available at https://github.com/QDRhhhh/UMU-bench