Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models

Zheyuan Liu[1], **Guangyao Dou**[2], **Xiangchi Yuan**[3], **Chunhui Zhang**[4], **Zhaoxuan Tan**[1], **Meng Jiang**[1]
[1]University of Notre Dame, [2]University of Pennsylvania,
[3]Georgia Institute of Technology, [4]Dartmouth College
zliu29@nd.edu

Abstract

Generative models such as Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) trained on massive datasets can lead them to memorize and inadvertently reveal sensitive information, raising ethical and privacy concerns. While some prior works have explored this issue in the context of LLMs, it presents a unique challenge for MLLMs due to the entangled nature of knowledge across modalities, making comprehensive unlearning more difficult. To address this challenge, we propose **Modality Aware Neuron Unlearning (MANU)**, a novel unlearning framework for MLLMs designed to selectively clip neurons based on their relative importance to the targeted forget data, curated for different modalities. Specifically, MANU consists of two stages: important neuron selection and selective pruning. The first stage identifies and collects the most influential neurons across modalities relative to the targeted forget knowledge, while the second stage is dedicated to pruning those selected neurons. MANU effectively isolates and removes the neurons that contribute most to the forget data within each modality, while preserving the integrity of retained knowledge. Our experiments conducted across various MLLM architectures illustrate that MANU can achieve a more balanced and comprehensive unlearning in each modality without largely affecting the overall model utility. [1]

1 Introduction

The rapid advancement of Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Fu et al., 2024; Qin et al., 2023) and Multimodal Large Language Models (MLLMs) (Liu et al., 2024a; Ye et al., 2023, 2024; Zhu et al., 2023; Zhang et al., 2025) have showcased their exceptional capabilities across various AI domains (Ouyang et al., 2022; Tan et al., 2024; Ni et al., 2025; Zhang et al., 2024b), largely due to extensive pre-training and fine-tuning on vast data corpus. However, this remarkable learning ability also poses risks such as privacy violations and copyright infringements. Since retraining from scratch while excluding these data is computationally expensive, *Machine Unlearning (MU)* (Nguyen et al., 2022; Wang et al., 2024b; Liu et al., 2024c,f) has emerged as an efficient alternative to remove the influence of sensitive data while preserving overall model performance.

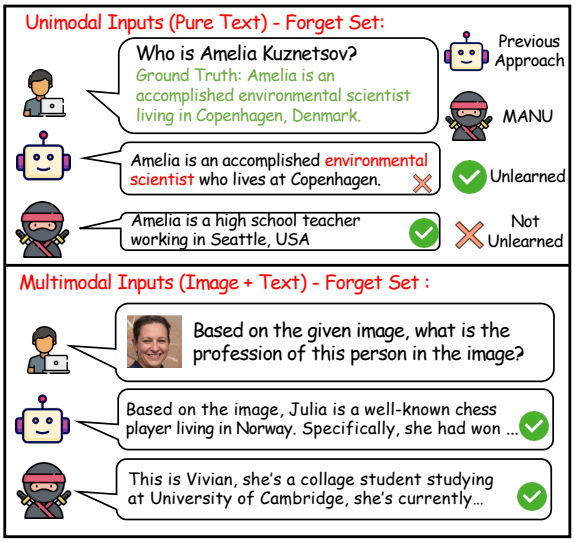


Figure 1: Comparison of MANU with the previous approach in responding to questions related to unlearned targets, using multimodal inputs (i.e., images with associated text) and pure text inputs, respectively.

Recent research has advanced MU techniques for LLMs (Zhang et al., 2024a; Liu et al., 2024g; Yao et al., 2023; Pochinkov and Schoots, 2024; Dou et al., 2024) while neglecting the case of MLLMs. Although extending MU methods from LLMs to MLLMs may seem intuitive, Liu et al. (2024e) highlights that such adaptations often result in **imbalanced unlearning**, where knowledge is removed in multimodal (image-text) level but remains in unimodal (text-only) level (e.g. Figure 1). This discrepancy arises from fundamental differences

[1]Code is available at franciscoliu/MANU.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5913–5933

between LLMs and MLLMs, particularly in knowledge representation and integration. While LLMs store target knowledge within a single modality, MLLMs integrate cross-modal interactions that entangle knowledge, making selective unlearning more challenging and potentially leading to drastic unintended knowledge loss. We provide a detailed explanation is provided in Section 2.

To address this challenge, we propose MANU, a novel two-stage unlearning approach that strategically prunes neurons associated with target knowledge entangled across both vision and textual modalities. Specifically, the first stage focuses on identifying critical neurons that contribute significantly to the forget dataset. This is achieved using four importance functions: absolute importance, frequency importance, variance importance, and root mean square importance functions. In the second stage, a scoring function is defined to evaluate neurons based on the importance scores calculated in the previous stage, facilitating the pruning of these neurons from the original model. Our main contributions are as follows:

1. We investigate the unique challenge of MLLM unlearning and highlight the limitations of previous methods designed for unimodal LLMs, which lack modality-specific design. Consequently, even when applied to multimodal inputs, these methods lead to imbalanced unlearning, effectively removing target knowledge in multimodal inputs while retaining it in unimodal level.
2. We propose MANU, the **first** modality-aware unlearning framework for MLLMs, which disentangles and removes modality-specific knowledge while preserving model utility across multiple perspectives.
3. Experiments and case studies demonstrate the effectiveness of MANU in unlearning sensitive knowledge across modalities while preserving model utility in various MLLMs.

2 Motivation

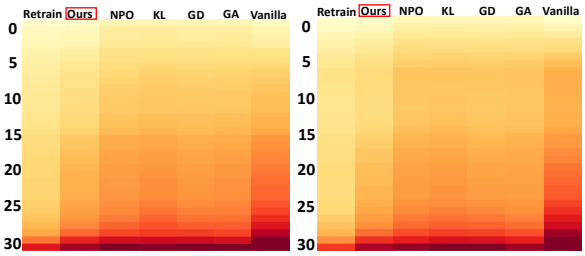
Inspired by (Liu et al., 2024e), which highlights the challenge of imbalanced unlearning in MLLMs, where unimodal LLM methods fail to remove knowledge across modalities. In particular, unlearning in one modality does not necessarily eliminate the corresponding knowledge in another, leading to knowledge retention. We hypothesize that this occurs due to entangled knowledge representations across modalities, making it insufficient to unlearn from one modality alone. Specifically, the activated neuron varies by input type, meaning that unlearned knowledge may persist even after targeted unlearning.

To validate this hypothesis, we compare our modality-aware approach with prior methods that unlearn only multimodal knowledge, using exclusively multimodal inputs. The heatmap comparisons are shown in Figure 2. Additionally, we include the vanilla and retrained models from MLLMU-Bench. We first examine the heatmap of different unlearning algorithms on the **forget set**, which contains data designated for removal. As shown in Figure 2a and 2b, fainter colors indicate lower knowledge retention, while deeper colors signify higher retention. Next, when comparing 2a and 2b, we observe that while prior methods effectively unlearn target knowledge from multimodal inputs (2b), they fail to fully remove this knowledge in the unimodal setting (2a), where only textual inputs are provided. This finding suggests that inputs with different modalities activate distinct neurons, underscoring the challenges of achieving comprehensive unlearning across modalities. Detailed analysis of modality-specific performance is provided in Section 5.1.

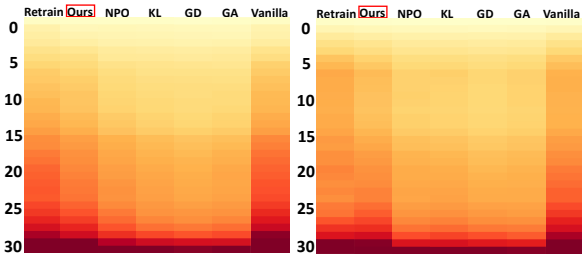
Furthermore, we present the heatmap of these algorithms on the **retain set** with different input types, as shown in Figure 2c and 2d. Unlike the forget set, where knowledge should be erased, the objective here is to preserve unrelated knowledge, meaning that deeper colors indicate stronger retention ability. As expected, the vanilla and retrained models exhibit the darkest colors across layers, indicating strong knowledge retention on retain set. However, unlearning algorithms such as GA and Gradient Difference display noticeably lighter colors, signifying unintended knowledge loss on the retain set. Those heatmaps further reinforce the findings of Liu et al. (2024e), demonstrating that effective MLLM unlearning must disentangle multimodal representations to prevent unintended loss while preserving retained knowledge.

3 Method

In this section, we elaborate on MANU (Figure 3), a two-stage modality-aware pruning framework designed to selectively remove sensitive information forget set $\mathcal{D}_f$ while preserving model utility on retain set $\mathcal{D}_r$ from MLLMs and various general benchmarks. The first stage involves identifying



(a) Text-Only Inputs (Forget) (b) Multi. Inputs (Forget)



(c) Text-Only Inputs (Retain) (d) Multi. Inputs (Retain)

Figure 2: Visualization of knowledge retention across MLLM language module layers for different unlearning methods on the forget/retain sets of MLLMU-Bench. Figures 2a, 2c show text-only residuals, while Figures 2b, 2d depict multimodal residuals. The x-axis represents unlearning methods (Grad. Diff. as GD), the y-axis shows layer indices, and darker red indicates higher knowledge retention.

and selecting the most contributed neurons across two modalities on the forget set.

3.1 Important Neuron Selection Stage

The first stage applies four importance functions to assess the relative importance of neurons in the language and vision MLP layers for both the forget set $\mathcal{D}_f$ and retain set $\mathcal{D}_r$. First, we leverage the observation that meaningful neuron activity is characterized by deviations from zero, as most activations remain close to zero by default (see Appendix A.1). Given a neuron n and its corresponding activations z, we define absolute importance (I_{abs}) to measure the difference in activation magnitudes between modalities relative to an arbitrary dataset $\mathcal{D}$, capturing the modality-specific processing preferences of individual neurons:

$$I_{\text{abs}}(\mathcal{D}, n) := \frac{|\bar{Z}_{\text{multi}} - \bar{Z}_{\text{text}}|}{\bar{Z}_{\text{multi}} + \bar{Z}_{\text{text}} + \epsilon}$$

where modality-specific mean absolute activations can be formulated as:

$$\bar{Z}_{\text{multi}} = \frac{1}{|\mathcal{D}_{\text{multi}}|} \sum_{d \in \mathcal{D}_{\text{multi}}} |z_{\text{multi}}(d)|,$$

$$\bar{Z}_{\text{text}} = \frac{1}{|\mathcal{D}_{\text{text}}|} \sum_{d \in \mathcal{D}_{\text{text}}} |z_{\text{text}}(d)|,$$

where $\mathcal{D}_{\text{text}}, \mathcal{D}_{\text{multi}} \subset \mathcal{D}$, represent the dataset in pure textual format and image with associated text format, respectively. Here, $z_{\text{multi}}(d)$ and $z_{\text{text}}(d)$ denote the absolute activation values of neuron n when processing a sample d from the multimodal and textual subsets, respectively. The normalization ensures that neurons with strong activation disparities between modalities are highlighted while controlling for overall activation magnitude. The small constant ϵ in the denominator is added for numerical stability, preventing division by zero.

Second, motivated by findings that neuron activation distributions exhibit a sharp peak at zero—indicating that most neurons remain inactive by default, with only a subset selectively activating in response to specific inputs (Zhang et al., 2021) (see Appendix A.1 for elaborations)—we introduce frequency importance (I_{freq}) to quantify how often a neuron's activation significantly deviates from zero. Since modality-relevant neurons are expected to fire more frequently when processing inputs from their associated modality, I_{freq} helps distinguish consistently engaged neurons from those that activate only sporadically. We first define the modality-specific activation frequency as:

$$N_{\text{multi}} = \left|\{d \in \mathcal{D}_{\text{multi}} \mid |z_{\text{multi}}(d)| > \tau\}\right|,$$
$$N_{\text{text}} = \left|\{d \in \mathcal{D}_{\text{text}} \mid |z_{\text{text}}(d)| > \tau\}\right|.$$

Using these definitions, we compute frequency importance as:

$$I_{\text{freq}}(\mathcal{D}, n) := \frac{|\Delta N|}{\Sigma N + \epsilon},$$
$$\Delta N = N_{\text{multi}} - N_{\text{text}},$$
$$\Sigma N = N_{\text{multi}} + N_{\text{text}}.$$

This normalized frequency metric complements absolute importance I_{abs} by focusing on activation consistency rather than magnitude, enabling the identification of neurons that may exhibit moderate but reliable modality-specific responses.

Third, building on information theory principles (Varley, 2023), which suggest that neurons carrying more information should exhibit diverse activation patterns rather than consistently remaining near zero, we define variance importance (I_{var}) to measure the spread of activation values within each modality, thereby quantifying each neuron's contribution to modality-specific information processing. Using the previously defined $\bar{Z}_{\text{multi}}$ and $\bar{Z}_{\text{text}}$, we compute the variance within each modality as:

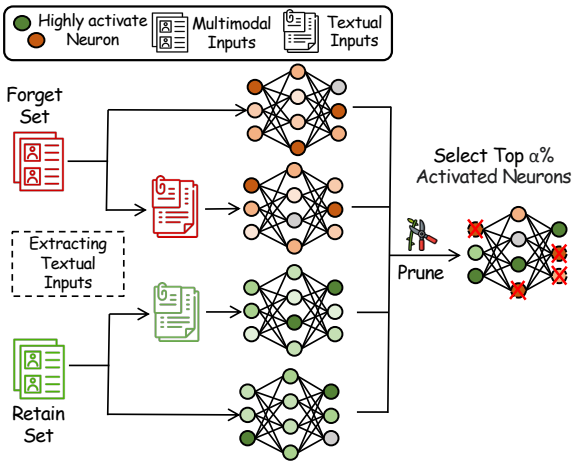


Figure 3: The overall framework of MANU. The forget and retain sets are first split into text-only and multimodal modalities. Neuron activations are then computed across modalities and datasets, followed by applying an importance and scoring function to evaluate activated neurons. Finally, the top $\alpha\%$ of neurons are pruned based on their scores.

$$\text{Var}_{\text{multi}} = \frac{1}{|\mathcal{D}_{\text{multi}}|} \sum_{d \in \mathcal{D}_{\text{multi}}} (z_{\text{multi}}(d) - \bar{Z}_{\text{multi}})^2,$$

$$\text{Var}_{\text{text}} = \frac{1}{|\mathcal{D}_{\text{text}}|} \sum_{d \in \mathcal{D}_{\text{text}}} (z_{\text{text}}(d) - \bar{Z}_{\text{text}})^2,$$

$$I_{\text{var}}(\mathcal{D}, n) := \sqrt{\text{Var}_{\text{multi}} + \text{Var}_{\text{text}}}.$$

I_{var} provides a statistically robust measure of how differently a neuron responds across modalities. Larger values indicate neurons that maintain distinct roles in processing multimodal versus unimodal inputs.

Finally, as highlighted by Liu et al. (2023), many neuron activations may be redundant, meaning they are consistently active across different inputs but do not contribute meaningfully to specific outputs. This suggests that a subset of neurons fire indiscriminately rather than being specialized for particular tasks or modalities, leading to inefficiencies in representation. To address this, we introduce root mean square importance (I_{rms}) to identify neurons with consistently strong activations relative to the overall activation pattern, formulated as:

$$I_{\text{rms}}(\mathcal{D}, n) := \sqrt{\frac{|\Delta Z^2|}{\Sigma Z^2 + \epsilon}},$$

$$\text{where } Z^2_{\text{multi}} = \sum_{d \in \mathcal{D}_{\text{multi}}} z_{\text{multi}}(d)^2,$$

$$Z^2_{\text{text}} = \sum_{d \in \mathcal{D}_{\text{text}}} z_{\text{text}}(d)^2,$$

$$\Delta Z^2 = Z^2_{\text{multi}} - Z^2_{\text{text}},$$

$$\Sigma Z^2 = Z^2_{\text{multi}} + Z^2_{\text{text}}.$$

I_{rms} emphasizes neurons with substantial modality-specific activity while penalizing those with redundant activation patterns, ensuring the identification of truly specialized neural pathways for each modality. Together, we aggregate these four importance functions into a unified importance measure through a weighted combination. Specifically, for any dataset $\mathcal{D}$ and neuron n, we compute:

$$\mathcal{I}(\mathcal{D}, n) := \sum_{k \in \mathcal{K}} I_k(\mathcal{D}, n)$$

where $\mathcal{K} = \{I_{\text{abs}}, I_{\text{freq}}, I_{\text{var}}, I_{\text{rms}}\}$ represents our set of importance functions. This combined measure $\mathcal{I}$ denotes the comprehensive assessment of neuron importance by capturing different aspects of neural activation patterns: magnitude (I_{abs}), activation frequency (I_{freq}), activation diversity (I_{var}), and consistent strength (I_{rms}).

3.2 Selective Pruning Stage

In the second stage, we define a scoring function S_n that aims to determine the pruned neurons based on the calculated importance from previous stage. In particular, given forget set $\mathcal{D}_f$ and retain set $\mathcal{D}_r$, we have:

$$S_n = \frac{\mathcal{I}(\mathcal{D}_f, n)}{\mathcal{I}(\mathcal{D}_r, n) + \epsilon}.$$

Now given a vanilla model θ and a pruning rate α, we perform selective pruning by choosing and removing neurons based on their importance scores relative to forget set $\mathcal{D}_f$. Specifically, we can identify the set of neurons to prune by using the scoring function S_n:

$$\mathcal{N} = \{n : S_n \text{ is among the top } \alpha\% \text{ of all scores}\}.$$

For each selected neuron $n \in \mathcal{N}$, we perform the pruning operation by setting its weights to zero and obtain pruned model θ':

$$\theta' = \begin{cases} 0 & \text{if } n \in \mathcal{N}, \\ \theta & \text{otherwise.} \end{cases}$$

4 Experiments

In this section, we present extensive experiments to validate the effectiveness of MANU. Specifically, these experiments aim to address the following research questions: (1) Can MANU effectively unlearn the target knowledge from the model? (2) Does MANU successfully address the unique challenge of imbalanced unlearning across different

modalities in MLLMs? (3) How do different pruning ratios affect the effectiveness of MANU during the unlearning process? (4) Can MANU achieve a good balance between unlearning the target knowledge and preserving the model's utility?

4.1 Experimental Setup

Our experiments focus on unlearning fictitious profiles at both visual and textual levels using MLLMU-Bench (Liu et al., 2024e), a benchmark for evaluating unlearning in MLLMs. We conduct experiments on LLaVA-1.5-7B (Liu et al., 2024a) and Idefics2-8B (Laurençon et al., 2024), evaluating performance across four datasets to assess unlearning effectiveness, generalizability, and model utility. The Forget Set contains a subset of fictitious profiles designated for unlearning effectiveness, with 5%, 10%, and 15% selected for removal. A corresponding Test Set mirrors this split but includes images transformed to different angles and paraphrased text to assess generalizability. Lastly, for model utility evaluation, we assess performance using the Retain Set, Real Celebrity Set, and general benchmarks. The Retain Set includes fictitious profiles excluded from the Forget and Test Sets that the model should retain, while the Real Celebrity Set contains real-world celebrity profiles distinct from fictitious ones. Additionally, to assess model utility more comprehensively, we evaluate general reasoning and helpfulness post-unlearning using MMMU (Yue et al., 2024) and LLaVA-Bench (Liu et al., 2024b), examining whether unlearning impacts core model capabilities.

For each evaluation set, every approach is assessed across three tasks. The classification task presents a multiple-choice format to measure the model's ability to differentiate correct from incorrect associations. The generation task evaluates factual accuracy and coherence using ROUGE-L (Lin, 2004) and LLM-determined factuality scores. The cloze test task measures the model's ability to complete missing information, evaluated via exact-match accuracy. Details on evaluation metrics and dataset construction are provided in Appendix B.

4.2 Baseline methods

For baselines, we compared Gradient Ascent (GA) (Thudi et al., 2022), Gradient Difference (Liu et al., 2022), KL Minimization (Nguyen et al., 2020), Negative Preference Optimization (NPO) (Zhang et al., 2024a) and a generic prevention strategies using system prompts (prompting) to prevent models from producing privacy-related information. Specifically, the GA approach applies opposite gradient updates on $\mathcal{D}_f$. The Gradient Difference approach extends GA by adding a gradient updates on $\mathcal{D}_f$ and $\mathcal{D}_r$, ensuring unlearning without performance degradation. Next, the KL Minimization approach aligns the unlearned model's predictions on $\mathcal{D}_r$ with the vanilla model while encouraging divergence from the knowledge of $\mathcal{D}_f$. Lastly, the NPO treats $\mathcal{D}_f$ as dispreferred data and casts unlearning into a preference optimization framework, utilizing an oracle model fine-tuned exclusively on the $\mathcal{D}_r$. Lastly, we employ a generic prevention technique by utilizing a crafted system prompt (i.e. prompting). Further details on the baselines can be found in Appendix D.1.

4.3 Implementation Details

All experiments on both LLaVA and Idefics2 models are implemented on a server with 3 NVIDIA A6000 GPUs and Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz with 20 CPU cores. Details can be referred to Appendix D.2.

4.4 Main Results

To answer the first research question: **Can MANU effectively unlearn the target knowledge from the model**, we conduct extensive experiments on MLLMU-Bench using different data splits across various MLLMs. The results of these experiments are presented in Table 1 and Table 4. For each task in each dataset, we report the **average performance** for both multimodal and unimodal evaluation across three distinct tasks. From the table, it is evident that MANU demonstrates exceptional performance across all datasets and tasks on both the LLaVA and Idefics2 models with different data splits, consistently ranking as either the best or second-best method among all baselines. Notably, while GA-based approaches occasionally surpass MANU in unlearning performance (e.g., LLaVA model with a 15% forget split), it is crucial to emphasize the importance of preserving model utility on the retain set and real celebrity set while selectively unlearning knowledge from the Forget Set. From this perspective, the superior unlearning performance of GA-based methods often comes at a significant cost to model utility, making them the least effective approaches on maintaining model utility. Lastly, NPO appears as another competitive baseline due to its relatively stable performance in both unlearning effectiveness and model util-

ity. However, it is not as effective as MANU in achieving these two objectives.

5 Discussion

Though MANU achieves superior average performance compared to other baselines, it remains unclear whether MANU effectively overcomes the unique challenge inherent in MLLM unlearning. In this section, we aim to address this concern by answering three key questions essential to advancing the understanding of MLLM unlearning.

5.1 Unlearning across modalities

As demonstrated in section 2, the unique challenge of MLLM unlearning lies in the imbalanced effectiveness across modalities, where methods may exhibit strong performance on one but struggle on the other. Hence, this leads to the second question: **Does MANU successfully address the unique challenge of imbalanced unlearning across different modalities in MLLMs?** To investigate this question, we decompose the average performance from multimodal and unimodal evaluation results in main table and analyze whether MANU achieves more effective unlearning across different input modalities in MLLMU-Bench, as shown in Figure 4. From figure 4, we observe that certain unlearning methods, such as GA and Gradient Difference, demonstrate strong multimodal unlearning performance but struggle in unimodal evaluation (e.g., Figure 4a). This discrepancy highlights the entangled nature of knowledge across modalities, indicating that unlearning from multimodal inputs does not guarantee complete removal in unimodal settings. **Methods lacking modality-specific strategies may fail to erase target knowledge equally across modalities, leading to imbalanced unlearning.**

A similar imbalanced unlearning is observed in methods like KL Minimization and NPO, which exhibit stronger unlearning performance at multimodal level than in the unimodal setting. In contrast, MANU demonstrates the ability to unlearn target knowledge across both modalities, as evidenced by the balanced reduction in Forget/Test Set accuracy (e.g., Figure 4e). Additional analysis for other data splits can be found in Appendix E.3.

5.2 Pruning Ratio Analysis

In this section, we address the third question: **How do different pruning ratios affect the effectiveness of MANU during the unlearning process?** To investigate this, we adjust the pruning ratios of the selected neurons to 2%, 5%, and 10%, and observe the corresponding impact on overall performance. Table 2 presents results for both models using the 10% data split. Further experimental results can be referred to Appendix E.2. As the pruning ratio increases from 2% to 10%, we observe larger effects on both unlearning performance and model utility. For instance, with a 10% pruning ratio in the LLaVA model, MANU improves unlearning performance on the forget and test sets compared to 2% pruning, reducing classification accuracy from 38.36% to 34.81% and from 37.24% to 32.93%, respectively. However, this improvement comes at the cost of reduced model utility on the Retain and Real Celebrity Sets, with classification accuracy dropping from 44.89% to 34.22% and from 48.02% to 43.10%, respectively. A similar trend is observed in the Idefics2 model. This result shows that higher pruning ratios enhance unlearning performance but disrupt the balance with utility, ultimately reducing model utility. This occurs because higher pruning ratios remove neurons that are less critical to the forget set but essential for preserving model utility across other datasets.

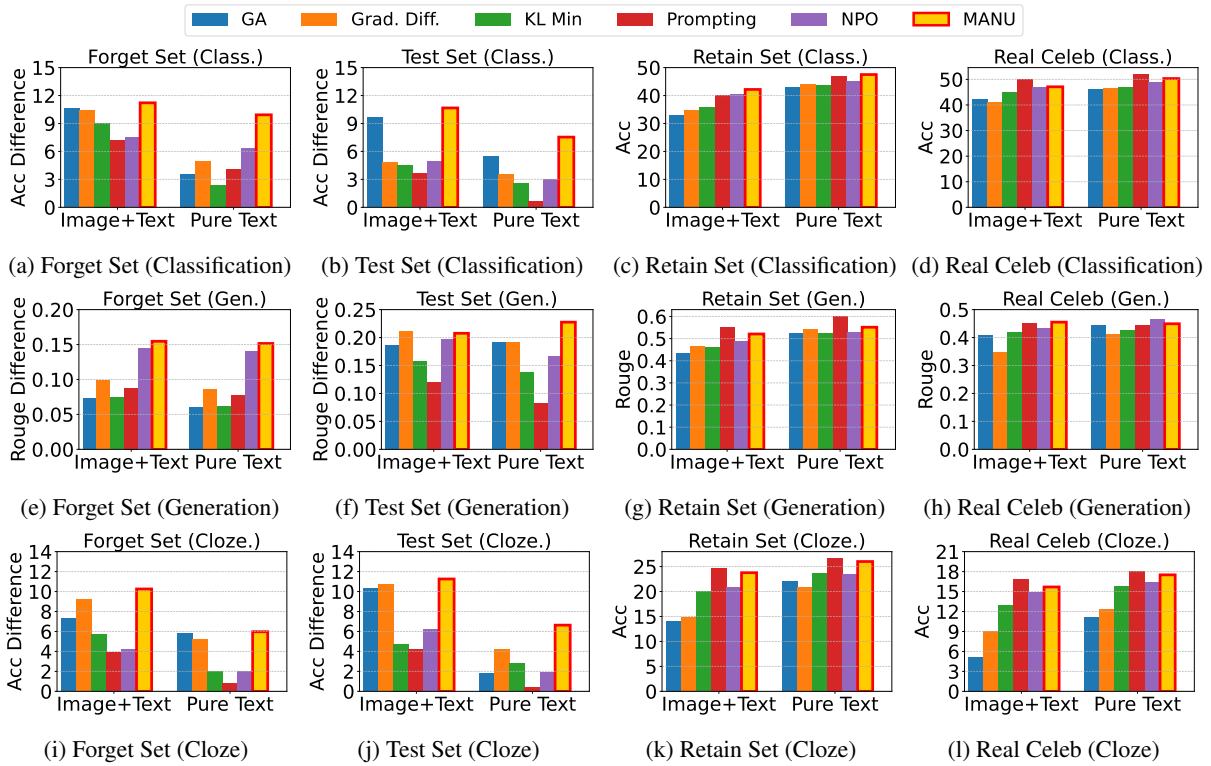
5.3 Unlearning v.s. Model Utility

Lastly, balancing unlearning and model utility remains a critical challenge in the field of unlearning. Hence, **can MANU achieve a good balance between unlearning the target knowledge and preserving the model's utility?** Similar to MLLMU-Bench, we decompose "model utility" into three perspectives: retain accuracy, neighboring concepts (Real Celebrity Set), and general model abilities, including reasoning and helpfulness, which are evaluated using MMMU (Yue et al., 2024) and LLaVA-Bench (Liu et al., 2024b). The results are shown in Figure 5 from left to right.

From the figures, we observe that MANU maintains a robust balance between unlearning performance and model utility across various aspects. The better an algorithm balances these two aspects, the closer it will appear to the top-right in the figure, indicating a larger difference in forget accuracy and higher retain accuracy. For example, in Figures 5a and 5b, MANU achieves a comparable reduction in Forget Set accuracy to GA-based approaches while maintaining high accuracy on the Retain and Real Celebrity sets. Similarly, when evaluating model reasoning abilities using MMMU and Llava-Bench (i.e. Figures 5c and 5d), MANU performs

Models	Forget Set				Test Set				Retain Set				Real Celebrity			
	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)
LLaVA-1.5-7B (5% Forget)																
Vanilla	51.70%	0.645	6.78	25.81%	47.86%	0.539	4.89	23.01%	46.11%	0.632	6.41	27.83%	51.80%	0.479	5.47	17.35%
GA	44.40%	**0.485**	3.38	17.19%	**38.40%**	0.384	**3.47**	16.47%	39.09%	0.495	2.97	18.96%	45.56%	0.414	3.42	8.66%
Grad. Diff.	43.60%	0.507	**3.05**	**16.00%**	43.41%	**0.323**	3.83	16.19%	41.07%	0.508	4.14	16.90%	46.52%	0.364	3.26	9.31%
KL Minimization	46.80%	0.574	5.04	20.46%	45.20%	0.396	4.54	20.04%	38.83%	0.478	4.20	21.03%	45.64%	0.418	3.49	14.53%
Prompting	46.80%	0.558	4.51	23.81%	44.87%	0.415	4.18	21.99%	42.99%	**0.612**	**5.42**	**26.75%**	**51.60%**	0.443	**5.43**	**17.18%**
NPO	45.61%	0.525	3.41	22.76%	44.44%	0.347	3.91	20.00%	42.61%	0.515	4.38	21.37%	49.51%	**0.450**	4.63	15.16%
MANU	**41.25%**	0.491	3.27	17.08%	41.67%	0.334	3.81	**15.78%**	**43.38%**	0.542	4.45	24.08%	49.57%	0.448	4.67	16.01%
LLaVA-1.5-7B (10% Forget)																
Vanilla	49.15%	0.594	6.40	26.97%	47.41%	0.510	5.20	25.43%	46.68%	0.582	5.44	28.49%	51.80%	0.479	5.47	17.35%
GA	43.85%	0.510	3.51	20.91%	40.60%	0.421	3.19	15.77%	41.91%	0.471	3.36	19.52%	42.64%	0.320	3.43	10.53%
Grad. Difference	41.60%	0.508	**3.16**	**18.79%**	39.08%	0.414	3.07	**14.50%**	43.71%	0.474	3.28	17.55%	40.94%	0.391	3.44	10.51%
KL Minimization	44.80%	0.579	4.12	22.69%	42.75%	0.420	3.29	20.50%	39.93%	0.456	3.82	20.70%	45.58%	0.462	3.13	14.90%
Prompting	48.41%	0.561	4.75	26.55%	47.29%	0.479	4.21	24.11%	**45.97%**	**0.577**	**5.43**	**26.12%**	**51.60%**	**0.471**	4.43	**17.16%**
NPO	47.40%	0.515	5.05	22.10%	46.42%	0.428	4.25	21.66%	44.81%	0.488	5.35	22.29%	47.89%	0.451	**4.53**	16.33%
MANU	**38.36%**	**0.503**	3.48	19.44%	**37.24%**	**0.402**	**3.06**	15.61%	44.89%	0.538	5.35	24.61%	48.02%	0.466	4.32	15.68%
LLaVA-1.5-7B (15% Forget)																
Vanilla	51.87%	0.575	6.34	26.62%	47.53%	0.502	4.08	25.33%	48.06%	0.585	5.46	28.51%	51.80%	0.479	5.47	17.35%
GA	**40.93%**	0.482	**3.51**	**17.33%**	**39.64%**	0.371	3.57	17.67%	40.43%	0.460	3.66	19.14%	40.36%	0.378	3.54	10.13%
Grad. Diff.	43.47%	0.518	3.98	18.78%	42.18%	0.401	3.61	18.11%	41.82%	0.476	3.28	21.30%	41.21%	0.417	3.45	11.37%
KL Minimization	47.60%	0.541	4.57	23.44%	43.20%	0.439	3.78	21.09%	42.96%	0.442	4.42	22.28%	42.58%	0.415	3.21	14.41%
Prompting	49.73%	0.547	4.63	26.00%	46.81%	0.483	3.67	24.56%	**47.09%**	**0.585**	**5.46**	**26.36%**	**51.60%**	**0.458**	**4.91**	**16.84%**
NPO	45.52%	0.509	4.39	20.63%	43.43%	0.439	4.01	21.88%	46.84%	0.525	4.98	23.31%	48.09%	0.433	4.11	14.10%
MANU	42.05%	**0.481**	3.73	17.91%	41.75%	**0.360**	**3.52**	**17.01%**	46.86%	0.557	5.19	24.62%	50.42%	0.448	4.05	16.77%

Table 1: Overall average results of baseline methods and MANU on LLaVA, combining multimodal and unimodal evaluations across three forget setups. **Bold** denotes the best performance, underline the runner-up. Each method is evaluated on four MLLMU-Bench datasets using classification accuracy, ROUGE-L, factuality, and cloze accuracy. Factuality Score is abbreviated as Fact. Score. •, •, and • represent classification, generation, and cloze evaluations, respectively. ↓ indicates lower is better, ↑ indicates higher is better.



(a) Forget Set (Classification) (b) Test Set (Classification) (c) Retain Set (Classification) (d) Real Celeb (Classification)

(e) Forget Set (Generation) (f) Test Set (Generation) (g) Retain Set (Generation) (h) Real Celeb (Generation)

(i) Forget Set (Cloze) (j) Test Set (Cloze) (k) Retain Set (Cloze) (l) Real Celeb (Cloze)

Figure 4: Classification, generation, and cloze performance of MANU and baselines in multimodal and unimodal setups with 5% forget data, using LLaVA as the base model. In subplots (a), (b), (e), (f), (i), and (j), the y-axis represents the change in classification accuracy, ROUGE-L score, and cloze accuracy relative to the vanilla model, evaluated on the Forget and Test sets. In the remaining subplots, the y-axis indicates classification accuracy, ROUGE-L score, and cloze accuracy, respectively. The x-axis represents performance across different modalities.

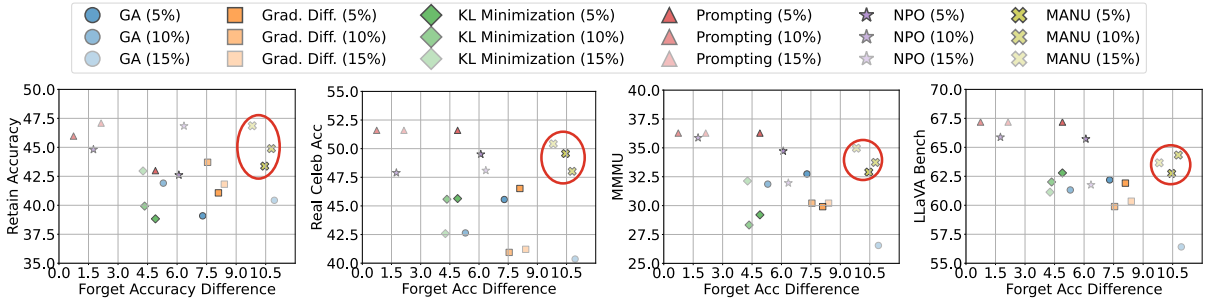
comparably to prompting techniques while significantly surpassing them in forget set accuracy. Thus, MANU effectively balances unlearning and model utility across multiple dimensions. Further analysis and additional ablation studies can be found in Appendix E.4 and E.6.

6 Related Work

MU for Generative Models. As LLMs and MLLMs memorize large amounts of sensitive

Models	Forget Set				Test Set				Retain Set				Real Celebrity			
	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)
LLaVA-1.5-7B (10% Forget)																
Vanilla	49.15%	0.594	6.40	26.97%	47.41%	0.510	5.20	25.43%	46.68%	0.582	5.44	28.49%	51.80%	0.479	5.47	17.35%
MANU (2%)	38.36%	0.503	3.48	19.44%	37.24%	0.402	3.06	15.61%	44.89%	0.538	5.35	24.61%	48.02%	0.466	4.32	15.68%
MANU (5%)	37.79%	0.468	3.27	18.47%	35.45%	0.388	3.05	11.20%	41.60%	0.478	4.39	22.61%	45.04%	0.421	3.54	11.78%
MANU (10%)	34.81%	0.425	3.10	18.10%	32.93%	0.341	2.97	10.22%	34.22%	0.468	4.03	18.16%	43.10%	0.378	3.39	10.27%
Idefics2-8B (10% Forget)																
Vanilla	54.48%	0.645	6.27	46.55%	48.09%	0.492	5.36	27.81%	47.52%	0.643	6.63	43.37%	52.75%	0.459	5.75	20.05%
MANU (2%)	36.49%	0.366	3.17	30.67%	35.32%	0.333	3.10	16.66%	43.95%	0.566	4.98	41.61%	50.81%	0.430	4.22	19.26%
MANU (5%)	35.11%	0.342	3.05	28.88%	34.27%	0.329	3.08	15.58%	39.63%	0.507	4.53	40.59%	49.10%	0.427	4.12	18.00%
MANU (10%)	30.24%	0.297	2.99	26.53%	31.33%	0.305	3.01	14.72%	35.77%	0.501	4.47	38.19%	48.87%	0.419	4.06	17.94%

Table 2: Overall results of MANU with varying pruning ratios on two base MLLM models under a 10% forget data setup. For each MLLM, the pruning ratio is iteratively increased from 2% to 10%.



(a) Forget Acc vs Retain Acc (b) Forget Acc vs Real Celeb (c) Forget Acc vs MMMU (d) Forget Acc vs LLaVABench

Figure 5: The overall trade-off between unlearning effectiveness and model utility across all baselines using different forget data, with LLaVA as the base model. The x-axis shows the difference in forget classification accuracy relative to the vanilla model, while the y-axis reflects model utility from various perspectives. From left to right, these perspectives include retain accuracy, real celebrity accuracy, MMMU, and LLaVA-Bench performance, respectively.

knowledge during pre-training and fine-tuning, privacy concerns have grown with the rise of generative models (Liu et al., 2024d; Nasr et al., 2023; Liu et al., 2024f; Zhang et al., 2023). Machine Unlearning (MU) offers an efficient solution to selectively erase unwanted information while preserving overall model performance. Yao et al. (2023) formalized unlearning objectives for LLMs, introducing a gradient-ascent-based approach to remove harmful knowledge. To address catastrophic forgetting, task vector-based approaches have been proposed (Ilharco et al., 2022; Liu et al., 2024g; Dou et al., 2024). In response to the *Right to Be Forgotten* (Dang, 2021; Bourtoule et al., 2021), benchmarks like TOFU (Maini et al., 2024) and MLLMU-Bench (Liu et al., 2024e) were developed using synthetic data, highlighting the need for privacy-preserving methods. However, existing unlearning algorithms are not explicitly designed for MLLMs to achieve comprehensive unlearning across modalities.

Model Pruning. Model pruning has proven to be an effective approach for removing redundant weights to enhance the performance and efficiency of a model. For example, Conmy et al. (2023) proposes a weight pruning-based technique to identify sub-circuits that contribute most to a specific dataset. Additionally, pruning can be used to preserve key model capabilities while reducing computational costs. For instance, Michel et al. (2019) introduces a method to prune unused attention heads without impacting overall performance. Pochinkov and Schoots (2024) shows that pruning can be used to unlearn specific behaviors of transformer models through a selective neuron approach. Additionally, it empirically demonstrates the effectiveness of neuron pruning over weight pruning. However, without a modality-specific pruning strategy, achieving thorough unlearning to remove target knowledge across different modalities remains challenging.

7 Conclusion

In this work, we address the challenge of imbalanced unlearning in MLLMs, which arises due to distinct knowledge distributions and activation patterns across vision and language pathways. To tackle this, we propose MANU, a modality-aware neuron pruning framework that ensures balanced unlearning across modalities while preserving model utility. Our approach first applies four importance functions to analyze neuron activations in MLP layers, then employs a scoring function to identify and prune neurons most associated with the targeted forget knowledge. Our results across multiple MLLMs demonstrate the efficacy of MANU in achieving comprehensive unlearning while main-

taining the model utility.

8 Limitations

Adaptations to other applications Our method is primarily designed to remove sensitive profiles in MLLMU-Bench, where all profiles are fictitious and fine-tuned on the vanilla model. However, it would be valuable to explore how this pruning approach could be extended to unlearn other behaviors of MLLM models, such as harmful generations and copyright infringements. Additionally, while MANU is specifically designed for MLLMs, its adaptation and performance on unimodal unlearning benchmarks, such as TOFU (Maini et al., 2024) and WMDP (Li et al., 2024), remain unexplored, which we leave for future work. We hope this study serves as a foundation to inspire future research toward developing **a model-agnostic unlearning framework**.

Robustness of Machine Unlearning Although factuality score is used as one of the evaluation metrics, ROUGE score remains an important measure in our unlearning setting. However, as highlighted by recent work (Ippolito et al., 2022), it may create a false sense of privacy. Additionally, the robustness of MANU against various attacks requires further validation and exploration, which is crucial as emphasized in prior studies (Łucki et al., 2024; Cooper et al., 2024).

Potential Instability Furthermore, as discussed in Section 5, variations in the pruning ratio significantly affect both unlearning performance and model utility. While MANU demonstrates superior performance across tasks, it has yet to achieve an optimal balance between unlearning effectiveness and model utility. Thus, we position MANU as **a preliminary study** showcasing the benefits of a modality-aware design for MLLM unlearning, laying the foundation for more robust and stable approaches in future research.

References

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

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

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

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *USENIX Security 21*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *JMLR*.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Neurips*.

A Feder Cooper, Christopher A Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, et al. 2024. Machine unlearning doesn't do what you think: Lessons for generative ai policy, research, and practice. *arXiv preprint arXiv:2412.06966*.

Quang-Vinh Dang. 2021. Right to be forgotten in the age of machine learning. In *Advances in Digital Science: ICADS 2021*.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2024. Avoiding copyright infringement via machine unlearning. *arXiv preprint arXiv:2406.10952*.

André V Duarte, Xuandong Zhao, Arlindo L Oliveira, and Lei Li. 2024. De-cop: Detecting copyrighted content in language models training data. *arXiv preprint arXiv:2402.09910*.

Yonggan Fu, Zhongzhi Yu, Junwei Li, Jiayi Qian, Yongan Zhang, Xiangchi Yuan, Dachuan Shi, Roman Yakunin, and Yingyan Celine Lin. 2024. Amoeballm: Constructing any-shape large language models for efficient and instant deployment. *arXiv preprint arXiv:2411.10606*.

Amin Ghiasi, Hamid Kazemi, Eitan Borgnia, Steven Reich, Manli Shu, Micah Goldblum, Andrew Gordon Wilson, and Tom Goldstein. 2022. What do vision transformers learn? a visual exploration. *arXiv preprint arXiv:2212.06727*.

Xiusheng Huang, Yequan Wang, Jun Zhao, and Kang Liu. 2024. Commonsense knowledge editing based on free-text in llms. *arXiv preprint arXiv:2410.23844*.

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

Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. 2022. Preventing verbatim memorization in language models gives a false sense of privacy. *arXiv preprint arXiv:2210.17546*.

Abhinav Joshi, Shaswati Saha, Divyaksh Shukla, Sriram Vema, Harsh Jhamtani, Manas Gaur, and Ashutosh Modi. 2024. Towards robust evaluation of unlearning in llms via data transformations. *arXiv preprint arXiv:2411.15477*.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*.

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

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *CoLLAs*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Neurips*.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. 2024c. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*.

Xiaoze Liu, Ting Sun, Tianyang Xu, Feijie Wu, Cunxiang Wang, Xiaoqian Wang, and Jing Gao. 2024d. Shield: Evaluation and defense strategies for copyright compliance in llm text generation. *arXiv preprint arXiv:2406.12975*.

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

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024f. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*.

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

Zichang Liu, Jue Wang, Tri Dao, Tianyi Zhou, Binhang Yuan, Zhao Song, Anshumali Shrivastava, Ce Zhang, Yuandong Tian, Christopher Re, et al. 2023. Deja vu: Contextual sparsity for efficient llms at inference time. In *ICML*.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. 2024. An adversarial perspective on machine unlearning for ai safety. *arXiv preprint arXiv:2409.18025*.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022a. Locating and editing factual associations in gpt. *Neurips*.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022b. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? *Neurips*.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. 2020. Variational bayesian unlearning. *Neurips*.

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

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, et al. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Neurips*.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. 2024. Arc2face: A foundation model for id-consistent human faces. In *ECCV*.

Nicholas Pochinkov and Nandi Schoots. 2024. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*.

Yusu Qian, Hanrong Ye, Jean-Philippe Fauconnier, Peter Grasch, Yinfei Yang, and Zhe Gan. 2024. Mia-bench: Towards better instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Neurips*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024. Personalized pieces: Efficient personalized large language models through collaborative efforts. *arXiv preprint arXiv:2406.10471*.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *EuroS&P*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Thomas F Varley. 2023. Information theory for complex systems scientists. *arXiv preprint arXiv:2304.12482*.

Yu Wang, Ruihan Wu, Zexue He, Xiusi Chen, and Julian McAuley. 2024a. Large scale knowledge washing. *arXiv preprint arXiv:2405.16720*.

Zehong Wang, Sidney Liu, Zheyuan Zhang, Tianyi Ma, Chuxu Zhang, and Yanfang Ye. 2024b. Can llms convert graphs to text-attributed graphs? *arXiv preprint arXiv:2412.10136*.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2017. Large-scale cloze test dataset created by teachers. *arXiv preprint arXiv:1711.03225*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *arXiv preprint arXiv:2310.10683*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*.

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. *Neurips*.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. 2025. Pretrained image-text models are secretly video captioners. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

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

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Moefication: Transformer feed-forward layers are mixtures of experts. *arXiv preprint arXiv:2110.01786*.

Zheyuan Zhang, Zehong Wang, Tianyi Ma, Varun Sameer Taneja, Sofia Nelson, Nhi Ha Lan Le, Keerthiram Murugesan, Mingxuan Ju, Nitesh V Chawla, Chuxu Zhang, et al. 2024b. Mopi-hfrs: A multi-objective personalized health-aware food recommendation system with llm-enhanced interpretation. *arXiv preprint arXiv:2412.08847*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Neurips*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix: Important Function Design

A.1 Neuron Act. Distribution (I_{abs}, I_{freq})

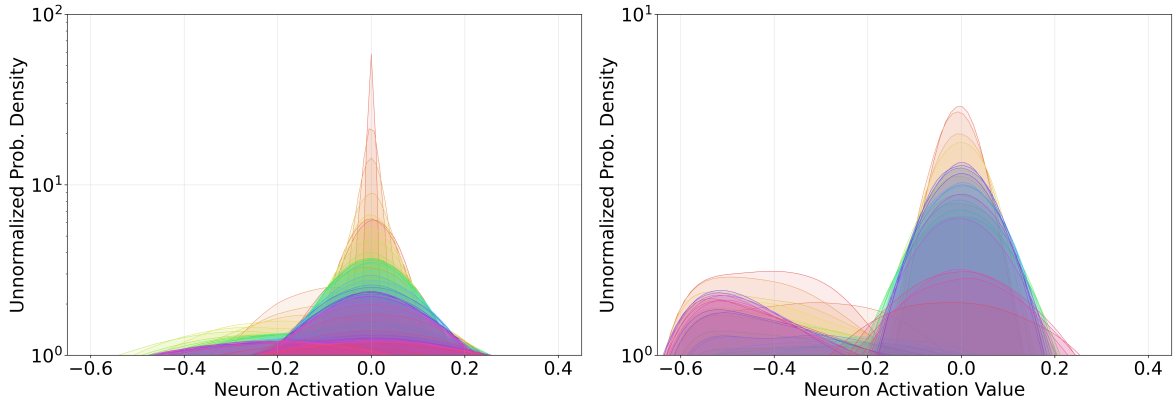
In Figure 6, we present examples of neuron activation distributions for both the language and vision modules, respectively. As shown in the figure, the majority (not all) of pre-activation neurons exhibit a default activation of 0.0. This observation further reinforces the motivation behind our first importance function I_{abs}, which leverages this sparsity pattern to quantify the extent to which activations deviate from zero. By capturing the magnitude of deviation, I_{abs} allows us to identify neurons that are more actively engaged in processing modality-specific information, distinguishing them from those that remain inactive across inputs.

Additionally, we observe a significant spike around zero (Figure 6), which aligns with the findings of Zhang et al. (2021), emphasizing that meaningful nonzero activations occur only in select cases where neurons contribute to specific information processing tasks. This further validates the rationale behind our second importance function I_{freq} and underscores the necessity of capturing activation frequency when identifying neurons that are crucial for processing the target dataset.

A.2 Information Diversity in Neural Act. (I_{var})

One key insight from information theory is that systems carrying more meaningful information exhibit diverse activation patterns rather than consistently remaining near zero. This principle is particularly relevant to our design of variance importance (I_{var}), which quantifies the spread of neuron activation values between modalities. Inspired by information theory principles (Varley, 2023), I_{var} is formulated to capture the degree of information differentiation across modalities—a higher variance in activations implies stronger modality-specific processing, while a lower variance suggests redundancy or shared information. This metric allows us to identify neurons that contribute distinctively to multimodal versus unimodal inputs, ensuring that pruning decisions target modality-specific information rather than broadly removing neurons with minimal impact.

By leveraging variance as a measure of information richness, our approach aligns with information theory's emphasis on quantifying uncertainty and diversity in signal representations, ultimately leading to a more effective and principled method for



(a) Language Layer Activation

(b) Vision Layer activation

Figure 6: Visualization of neuron activations across language MLP layers and vision MLP layers of MLLM. Figure 6a shows neuron activations of language layers, while Figure 6b illustrates neuron activation patterns of vision layers. The x axis represents neuron activation value, the y axis shows the unnormalized probability density.

unlearning within MLLMs.

A.3 Contextual Sparsity (I_{rms})

Recent studies have demonstrated that a substantial portion of neurons and attention heads in LLMs remain inactive or contribute minimally to output generation, highlighting the presence of significant redundancy within model activations. The work of Liu et al. (2023) formally establishes this by introducing the concept of contextual sparsity, which leverages the observation that only a small, input-dependent subset of parameters is necessary to approximate the full model's output effectively. Empirical findings in Liu et al. (2023) reveal that up to 85% of MLP neurons can be pruned dynamically at inference time without substantial degradation in model performance. These results strongly indicate that a large fraction of parameters within LLMs are redundant across different inputs. Building on these findings, we extend the notion of contextual sparsity to modality-aware unlearning, where redundant neurons may persist across different input types without contributing to modality-specific knowledge. This motivates our design of Root Mean Square Importance (I_{rms}), which quantifies neurons with consistently high yet uninformative activations. By identifying and pruning such neurons, we ensure that unlearning targets modality-relevant parameters while preserving overall model utility.

B Appendix: MLLMU-Bench

B.1 Benchmark Overview

Our experimental results and observations are primarily based on MLLMU-Bench (Liu et al., 2024e), which aims to advance the understanding of multimodal machine unlearning. We selected MLLMU-Bench for its comprehensive evaluation across various modalities and tasks. Specifically, it includes 500 fictitious profiles and 153 public celebrity profiles, each featuring over 14 customized question-answer pairs, assessed in both multimodal and unimodal settings. From a multimodal perspective, both the image and associated textual information of each individual's profile are provided, whereas the unimodal setting relies solely on textual information. Inspired by (Liu et al., 2024f), the benchmark is divided into four subsets: **Forget Set**, **Test Set**, **Retain Set**, and **Real Celebrity Set**, designed to evaluate unlearning algorithms in terms of efficacy, generalizability, and model utility. For each of these properties, MLLMU-Bench evaluates model performance on classification, generation, and cloze tasks under the aforementioned multimodal and unimodal settings. Detailed statistics about the benchmark are provided in Table 3.

B.2 Unlearning Efficacy

The **Forget Set** is designed to evaluate unlearning efficacy of algorithms. Specifically, it is created by randomly selecting 5%, 10%, and 15% of the 500 profiles, with each selected profile serving as an unlearning target. The primary goal of this dataset is to test the algorithm's ability to erase target knowledge while ensuring no residual traces of it remain.

B.3 Unlearning Generalizability

The **Test Set** is designed to evaluate the unlearning generalizability of algorithms. It is derived from the Forget Set by transforming both image and text

Statistics	Number
Total Questions	20,754
* Image + Text Questions	10,377
* Pure Text Questions	10,377
Total Images	1,153
Forget Percentile	5%/10%/15%
Multiple-choice Questions	11,530
Free Generation Questions	4,612
Fill-in-the-blank Questions	4,612
Total Profiles	653
* Fictitious	500
* Real Celeb	153
Total Countries	70
Total Regions	240
Total Birth Years	211
Total Employement	145

Table 3: Key statistics of the MLLMU-Bench.

data. For images, MLLMU-Bench uses Arc2Face (Papantoniou et al., 2024) to modify profile images with different poses and angles. For text, it employs GPT-4o to paraphrase questions into varied expressions. These transformations aim to assess whether the model has truly unlearned the target knowledge or still retains its transformed versions.

B.4 Model Utility

Lastly, the **Retain Set** and **Real Celebrity Set** are designed to evaluate model utility from different perspectives. The Retain Set consists of the remaining 95%, 90%, or 85% of profiles, excluding the Forget Set. After the unlearning process, the model is expected to maintain high-fidelity knowledge of these profiles. The **Real Celebrity Set** serves as a control set to measure unintended interference with general pre-trained knowledge after unlearning. Like the other sets, it includes both multimodal (image and text) and text-only formats of real public figures.

B.5 Evaluation Metrics

As mentioned in the previous section, the post-unlearned model is evaluated on classification, generation, and cloze tasks across both multimodal and unimodal settings for each of these properties.

B.5.1 Classification Task

The classification task is designed around key attributes of each profile (e.g., education, occupation) by generating multiple-choice questions about personal details. In particular, the model is passed with $\langle \text{image}, x, y \rangle$, where image represents the visual input in the multimodal setting (not applicable in the unimodal setting), x is the question, and y is the correct answer. The model then predicts $\hat{y}$ based on the input x, and accuracy is calculated by comparing $\hat{y}$ with the correct answer y.

B.5.2 Generation Task

In addition to classification, MLLMU-Bench evaluates the generation capabilities of post-unlearned models using free-generation questions. Each question is tailored to an individual's profile, with GPT-4o generating answers based on the key attributes extracted from the profile. MLLMU-Bench employs the **ROUGE-L score** and **Factuality Score** for evaluation. Specifically, the ROUGE-L score (Lin, 2004) measures the overlap of the longest matching subsequences between generated and reference texts. Next, inspired by prior benchmarks (Sun et al., 2023; Yu et al., 2024; Zheng et al., 2023), the Factuality Score assesses the factual accuracy and quality of generated responses using GPT-4o as the evaluator. It is rated on a scale of 1 to 10, where 1 represents an inaccurate response, and 10 signifies a fully correct and factually consistent answer.

B.5.3 Cloze Test

Lastly, inspired by previous Cloze-style tasks for evaluating models' memorization abilities (Xie et al., 2017; Duarte et al., 2024; Carlini et al., 2021; Joshi et al., 2024), MLLMU-Bench incorporates a Cloze-style task to assess whether sensitive information remains in the model after unlearning. Specifically, MLLMU-Bench provides only the individual's name as publicly available information, replacing all other key attributes with a *[Blank]*. The model is then prompted to complete the missing information. This task aims to evaluate the model's unlearning capability regarding target knowledge when only partial context about the individual is revealed.

C Rationale for Targeting MLP Layers

Recent research has demonstrated that MLP layers serve as primary knowledge storage components in transformer architectures. For example, (Huang et al., 2024) introduces the concept of "knowledge neurons," highlighting that specific neurons within MLP layers are responsible for encoding and storing information. By manipulating these neurons, it is possible to edit or selectively remove knowledge, offering fine-grained control over the model's retained information. Beyond individual neurons, broader findings in knowledge editing lit-

erature reinforce the significance of MLP layers for model knowledge control. Prior works have shown that knowledge manipulation techniques, including direct parameter modification and knowledge attribution methods, consistently identify MLP layers as the primary repository of factual and task-specific knowledge (Wang et al., 2024a; Meng et al., 2022a,b). Given that vision transformers share a fundamentally similar architecture with language transformers, where MLPs play an analogous role in feature extraction and information processing (Ghiasi et al., 2022), we extend this insight to the vision tower as well. The effectiveness of targeting MLPs for knowledge controlling is further supported by recent work in LLM pruning (Pochinkov and Schoots, 2024), which demonstrates that modifying MLP layers enables precise control over model knowledge while maintaining core model capabilities.

D Appendix: Implementation Details

D.1 Baseline Methods

D.1.1 Gradient Ascent

The Gradient Ascent approach (Thudi et al., 2022) is a simple yet effective method for unlearning. The primary goal of GA is to increase the loss for samples in the forget set $\mathcal{D}_f$, thereby minimizing the likelihood of the model retaining specific information about these profiles. In particular, for each sample $x \in \mathcal{D}_f$, GA aims to maximize the loss, driving the model away from its original predictions. The objective is to maximize the average loss across the $\mathcal{D}_f$:

$$\mathcal{L}(\mathcal{D}_f, w) = \frac{1}{|\mathcal{D}_f|} \sum_{x \in \mathcal{D}_f} \ell(x, w),$$

where $\ell(x, w)$ denotes the loss for a sample x with model parameters w. This process encourages the model to unlearn the associations it formed during fine-tuning with respect to the forget set.

D.1.2 Gradient Difference

Gradient Difference (Liu et al., 2022) builds upon Gradient Ascent by balancing the unlearning of the forget set with the preservation of performance on the retain set $\mathcal{D}_r$. The objective is to increase the loss on $\mathcal{D}_f$ while minimizing the impact on $\mathcal{D}_r$. This method ensures that the model forgets the targeted data without negatively affecting unrelated knowledge. The overall loss function is defined as:

$$\mathcal{L}_{\text{diff}} = -\mathcal{L}(\mathcal{D}_f, w) + \mathcal{L}(\mathcal{D}_r, w),$$

where $\mathcal{L}(\mathcal{D}_r, w)$ is the loss computed on the retain set and w indicates the model parameters. By optimizing this combined loss, the model selectively forgets the specified profiles while retaining performance on the rest of the dataset.

D.1.3 KL Minimization

The KL Minimization method (Nguyen et al., 2020) aims to align the model's predictions on the retain set with those of the original fine-tuned model while encouraging divergence on the forget set. Specifically, it minimizes the Kullback-Leibler (KL) divergence between the outputs of the current model and the original model for samples in $\mathcal{D}_r$, ensuring that important knowledge is retained. Simultaneously, the conventional loss is maximized on $\mathcal{D}_f$. Formally, the objective is:

$$\mathcal{L}_{\text{KL}} = -\mathcal{L}(\mathcal{D}_f, w) + \frac{1}{|\mathcal{D}_r|} \sum_{s \in \mathcal{D}_r} \text{KL}(M_{\text{o}} \| M_{\text{c}})(s)$$

where M_{o} and M_{c} represent the *original* and *current* models, respectively. This method ensures that unlearning is targeted while the model's behavior on the retain set remains unchanged.

D.1.4 Generic Prevention using prompt

To demonstrate the applicability of system prompts in unlearning scenarios, we append a system prompt to the unlearned model during evaluation as follows:

> "You are a helpful, respectful, and honest assistant. When generating your response, please do not generate any personal-related information."

This provides a concise instruction that supplements the default system prompt, explicitly instructing the model not to generate any privacy-related content.

D.1.5 Negative Preference Optimization

Negative Preference Optimization (NPO) technique aims to address the issue of catastrophic collapse that often associated with gradient ascent methods. NPO (Zhang et al., 2024a) is inspired by preference-based learning (Rafailov et al., 2024; Ouyang et al., 2022; Bai et al., 2022), where it operates within the preference optimization framework, targeting negative samples from the $\mathcal{D}_f$. In particular, the NPO loss function is defined as follows:

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

where $\pi_\theta(y|x)$ represents the prediction probability of the current model for token y given the input x, and $\pi_{\text{ref}}(y|x)$ is the prediction probability from the reference model trained on the entire dataset. The parameter β controls the smoothness of the optimization, and as $\beta \to 0$, the NPO loss converges to the standard gradient ascent loss. By minimizing this loss, NPO decreases the model's dependence on the $\mathcal{D}_f$, thereby promoting a more stable unlearning process while preventing the rapid degradation commonly observed with gradient ascent methods. In our experiments, we set $\beta = 0.9$, following the default setting from the original paper and MLLMU-Bench. Then, we define π_{ref} by fine-tuning the pre-trained model solely on the $\mathcal{D}_r$.

D.2 Hyperparameters Settings

Here, we present the hyperparameter settings for MANU using LLaVA and Idefics2 as the base models. Since the pruning process does not involve gradient updates, the primary tunable parameter is the batch size, which we set to 4. All experiments are conducted on NVIDIA A6000 GPUs (48 GB).

E Appendix: Additional Experiments

E.1 Main Experiments (Idefics2)

In this section, we present additional experiments on MLLMU-Bench using Idefics2 as the base model, with results shown in Table 4. The trends align with Table 1 for LLaVA, where MANU outperforms baselines across all datasets and tasks, consistently ranking first or runner-up.

E.2 Pruning Ratio Analysis:

In this section, we present additional analyses on the influence of different pruning ratios on unlearning effectiveness and model utility, as shown in Tables 5. As observed in these tables, regardless of the split ratio for the forget set, the trend remains consistent with the findings in Table 2. Specifically, as the pruning ratio increases, unlearning performance improves, but model utility deteriorates. These experimental results further validate the phenomenon that larger pruning ratios can disrupt the balance between effective unlearning and model utility.

E.3 Unlearning across modalities

Here, we present additional experiments evaluating the unlearning effectiveness of all tested algorithms using forget split ratios of 10% (Figure 7) and 15% (Figure 8), with Llava as the base model. These experiments aim to demonstrate that MANU effectively addresses the unique challenge of incomplete unlearning across different input types in the context of MLLM unlearning. As shown in the figures, we observe a trend similar to that in Figure 4. In particular, while some algorithms (e.g. GA based algorithms) perform well in multimodal evaluation, they often exhibit shortcomings in unimodal evaluation due to the absence of a curated modality-specific design. This further underscores the importance of modality-aware methodologies in MLLM unlearning.

E.4 Appendix: Unlearning v.s. Utility

In this section, we present additional experiments analyzing the trade-off between unlearning effectiveness and model utility using Idefics2 as the base model. The detailed results are shown in Figure 12. Same as the observations in Figure 5, MANU consistently outperforms other baselines, as it is typically closest to the top-right corner—indicating a better balance between unlearning effectiveness and model utility. Notably, MANU achieves unlearning performance comparable to GA-based methods while maintaining competitive model utility across different perspectives.

E.5 Appendix: Additional Utility Datasets

In addition to the evaluations already included—specifically on MMMU and LLaVA-Bench, we present results on three additional downstream benchmarks to further assess the functional utility of MANU across a diverse range of multimodal tasks. These benchmarks include: **MIA-Bench** (Qian et al., 2024) for evaluating conversational abilities, **MM-Vet** (Yu et al., 2023) for assessing integrated multimodal reasoning, and **MMStar** (Chen et al., 2024) for testing vision-indispensable capabilities.

In Table 6, we report the average performance of the vanilla model and unlearned models using various unlearning methods on these benchmarks. As shown in the Table, MANU performs competitively across all three benchmarks, often outperforming or matching the baselines. These results suggest that MANU effectively preserves cross-modal alignment and functional utility, even after selective neuron pruning.

Models	Forget Set				Test Set				Retain Set				Real Celebrity			
	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)
Idefics-2-8B (5% Forget)																
Vanilla	53.80%	0.630	6.22	44.75%	47.86%	0.434	5.00	24.97%	46.11%	0.644	6.51	42.35%	52.75%	0.459	5.75	20.05%
GA	**36.27%**	**0.405**	**2.90**	**30.07%**	38.40%	0.374	**3.42**	21.44%	39.09%	0.410	3.81	28.01%	41.27%	0.202	2.62	15.07%
Grad. Diff.	40.38%	0.426	3.96	32.24%	41.41%	0.408	3.73	22.66%	40.07%	0.408	4.05	33.19%	43.52%	0.363	3.91	16.37%
KL Minimization	39.69%	0.459	3.39	36.79%	45.20%	0.419	4.24	23.32%	38.83%	0.393	3.76	39.82%	45.64%	0.360	3.27	17.74%
Prompting	45.45%	0.492	3.91	42.61%	44.87%	0.423	4.39	23.88%	**44.99%**	**0.601**	**5.02**	**42.05%**	**52.00%**	**0.427**	**4.88**	**19.95%**
NPO	43.29%	0.501	4.87	39.77%	41.98%	0.391	4.47	22.75%	41.19%	0.484	4.57	39.99%	50.05%	0.384	4.05	18.17%
MANU	37.13%	0.413	3.20	32.11%	**35.55%**	**0.361**	3.91	**20.97%**	42.71%	0.538	4.60	40.01%	50.09%	0.399	4.11	18.80%
Idefics-2-8B (10% Forget)																
Vanilla	54.48%	0.645	6.27	46.55%	48.09%	0.492	5.36	27.81%	47.52%	0.643	6.63	43.37%	52.75%	0.459	5.75	20.05%
GA	37.81%	**0.459**	**3.09**	31.05%	**38.17%**	**0.313**	**3.64**	**20.43%**	38.15%	0.494	4.56	33.58%	42.16%	0.250	2.75	15.88%
Grad. Diff.	**36.60%**	0.471	3.33	35.57%	40.22%	0.414	3.68	24.65%	36.82%	0.461	4.34	35.80%	41.52%	0.386	3.62	17.72%
KL Minimization	41.28%	0.524	3.71	43.34%	42.74%	0.491	3.75	25.00%	38.10%	0.499	4.33	39.53%	43.64%	0.395	3.42	18.58%
Prompting	46.40%	0.504	3.55	45.27%	45.10%	0.422	4.09	26.31%	**44.31%**	**0.634**	**5.06**	**43.27%**	**52.00%**	**0.458**	**4.90**	**20.05%**
NPO	42.91%	0.521	4.12	41.44%	41.09%	0.399	3.77	23.11%	42.39%	0.541	4.82	40.02%	48.76%	0.421	3.91	17.39%
MANU	37.01%	0.467	3.45	**29.98%**	39.75%	0.355	**3.51**	21.88%	43.07%	0.539	4.94	41.13%	49.31%	0.437	3.98	19.10%
Idefics-2-8B (15% Forget)																
Vanilla	54.67%	0.630	6.42	46.33%	47.99%	0.436	5.30	27.77%	46.86%	0.645	6.48	42.81%	52.75%	0.459	5.75	20.05%
GA	37.87%	**0.335**	3.23	31.11%	37.90%	0.342	3.20	**15.67%**	38.66%	0.444	3.06	28.95%	43.56%	0.341	2.42	13.92%
Grad. Diff.	**35.33%**	0.340	**3.01**	33.50%	36.41%	**0.310**	**2.99**	18.59%	36.07%	0.370	3.19	35.00%	45.52%	0.408	3.03	15.88%
KL Minimization	41.09%	0.521	4.03	42.76%	44.81%	0.428	3.94	23.67%	39.54%	0.491	3.35	40.80%	47.64%	0.419	3.79	17.72%
Prompting	45.73%	0.482	3.88	45.23%	45.66%	0.409	3.72	26.16%	43.01%	**0.606**	5.03	**42.27%**	**52.00%**	**0.459**	**4.88**	**19.93%**
NPO	41.44%	0.447	3.97	40.06%	38.75%	0.389	3.49	22.10%	43.23%	0.597	**5.17**	40.19%	48.99%	0.424	4.07	18.88%
MANU	36.49%	0.366	3.17	**30.67%**	**35.32%**	0.333	3.10	16.66%	**43.95%**	0.566	4.98	41.61%	50.81%	0.430	4.22	19.26%

Table 4: Overall average performance of baseline methods and MANU on Idefics2, combining multimodal and unimodal evaluations across three forget setups. **Bold** indicates the best performance, and underline denotes the runner-up. Each method is evaluated on four datasets from MLLMU-Bench, assessed by classification accuracy, ROUGE-L score, factuality score and cloze accuracy. We abbreviate the Factuality Score as Fact. Score due to space limits. •, •, and • represent classification, generation and cloze evaluations, respectively. ↓ indicates that lower values are better, while ↑ indicates that higher values are better.

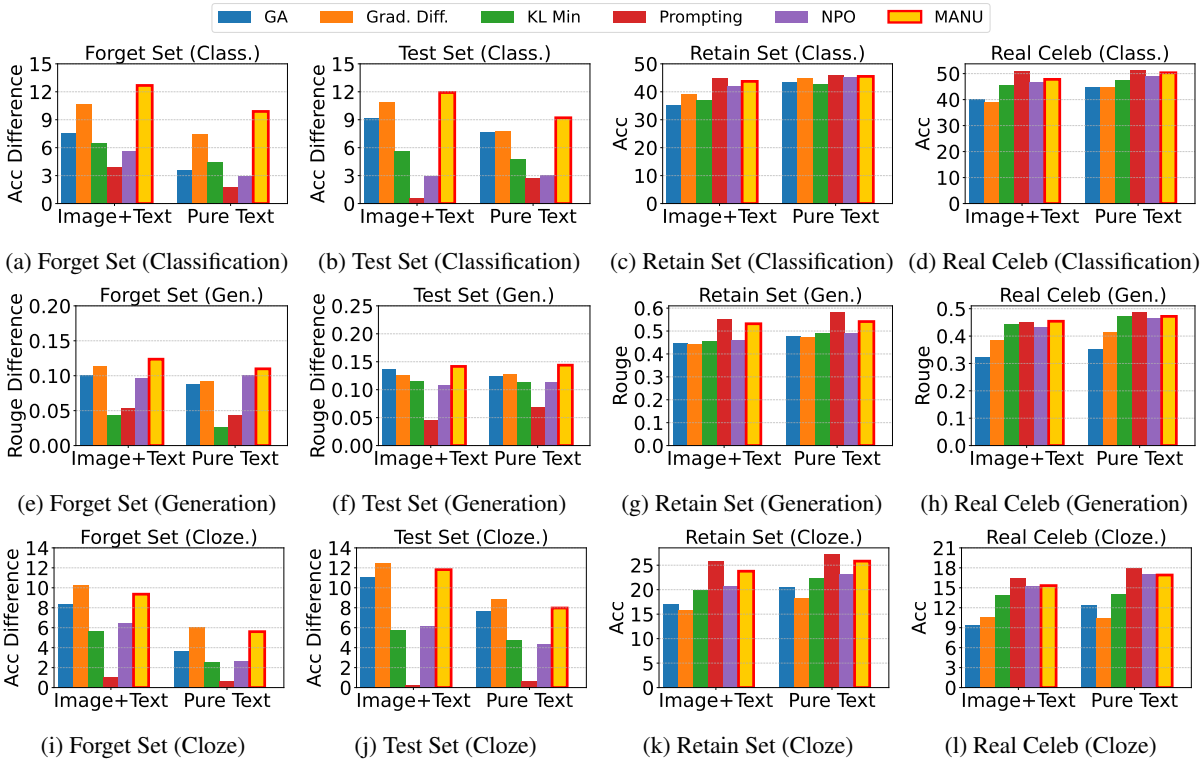
Models	Forget Set				Test Set				Retain Set				Real Celebrity			
	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)
LLaVA-1.5-7B (5% Forget)																
Vanilla	51.70%	0.645	6.78	25.81%	47.86%	0.539	4.89	23.01%	46.11%	0.632	6.41	27.83%	51.80%	0.479	5.47	17.35%
MANU (2%)	41.25%	0.491	3.27	17.08%	41.67%	0.334	3.81	15.78%	43.38%	0.542	4.45	24.08%	49.57%	0.448	4.67	16.01%
MANU (5%)	36.25%	0.477	3.16	13.54%	39.17%	0.291	3.56	15.71%	38.25%	0.514	4.14	10.02%	49.09%	0.436	3.51	11.44%
MANU (10%)	32.40%	0.473	3.01	11.25%	33.40%	0.262	3.35	13.25%	37.34%	0.476	3.97	11.62%	48.77%	0.409	3.47	11.11%
Idefics2-8B (5% Forget)																
Vanilla	53.80%	0.630	6.22	44.75%	47.86%	0.434	5.00	24.97%	46.11%	0.644	6.51	42.35%	52.75%	0.459	5.75	20.05%
MANU (2%)	37.13%	0.413	3.20	32.11%	35.55%	0.361	3.91	20.97%	42.71%	0.538	4.60	40.01%	50.09%	0.399	4.11	18.80%
MANU (5%)	36.01%	0.339	3.13	30.18%	34.97%	0.343	3.84	15.98%	41.54%	0.505	4.47	38.72%	48.61%	0.376	3.96	17.27%
MANU (10%)	31.87%	0.307	3.07	28.04%	33.77%	0.321	3.53	15.05%	39.80%	0.481	4.39	36.20%	46.66%	0.352	3.77	16.93%
LLaVA-1.5-7B (15% Forget)																
Vanilla	51.87%	0.575	6.34	26.62%	47.53%	0.502	4.08	25.33%	48.06%	0.585	5.46	28.51%	51.80%	0.479	5.47	17.35%
MANU (2%)	42.05%	0.481	3.73	17.91%	41.75%	0.360	3.52	17.01%	46.86%	0.557	5.19	24.62%	50.42%	0.448	4.05	16.77%
MANU (5%)	41.94%	0.477	3.24	17.26%	41.35%	0.357	3.43	16.89%	43.68%	0.509	5.19	17.27%	49.77%	0.451	3.50	12.41%
MANU (10%)	35.14%	0.468	3.11	12.97%	38.81%	0.327	3.40	14.53%	41.76%	0.493	5.02	10.75%	48.01%	0.410	3.41	10.78%
Idefics2-8B (15% Forget)																
Vanilla	54.67%	0.630	6.42	46.33%	47.99%	0.436	5.30	27.77%	46.86%	0.645	6.48	42.81%	52.75%	0.459	5.75	20.05%
MANU (2%)	36.49%	0.366	3.17	30.67%	35.32%	0.333	3.10	16.66%	43.95%	0.566	4.98	41.61%	50.81%	0.430	4.22	19.26%
MANU (5%)	33.82%	0.328	3.15	28.88%	34.44%	0.329	3.07	15.59%	41.11%	0.528	4.54	39.63%	49.93%	0.414	4.19	18.77%
MANU (10%)	30.39%	0.301	3.13	25.93%	32.46%	0.315	3.01	15.20%	39.41%	0.509	4.44	38.01%	48.02%	0.403	4.07	18.01%

Table 5: Overall results of MANU with varying pruning ratios on two base MLLM models under a 5% and 15% forget data setup. For each MLLM, the pruning ratio is iteratively increased from 2% to 10%.

E.6 Appendix: Ablations on Importance Functions

To further investigate the contribution of each importance function, here we provide additional ablation studies where we iteratively zeroed out each importance function in the scoring formula. The results are displayed in Table 7. From the ablation results, we observe that each importance function contributes uniquely to the overall effectiveness of MANU, and removing any single component results in a noticeable trade-off between unlearning performance and utility preservation. Since the trends for LLaVA and Idefics2 are consistent, we use the LLaVA results as a representative example.

Specifically, removing Frequency Importance (I_{freq}) or Variance Importance (I_{var}) substantially worsens unlearning on the Forget and Test sets—e.g., classification accuracy rises from 41.25% (MANU) to 47.35% and 46.67%, respectively, indicating a failure to sufficiently erase the target knowledge. These two metrics are particularly valuable for identifying neurons consistently and distinctively activated by forget data, thus supporting targeted unlearning. On the other hand, removing Absolute Importance (I_{abs}) or RMS Importance (I_{rms}) more prominently degrades performance on the Retain and Real Celebrity sets. For instance, when I_{abs} is excluded, Retain classi-

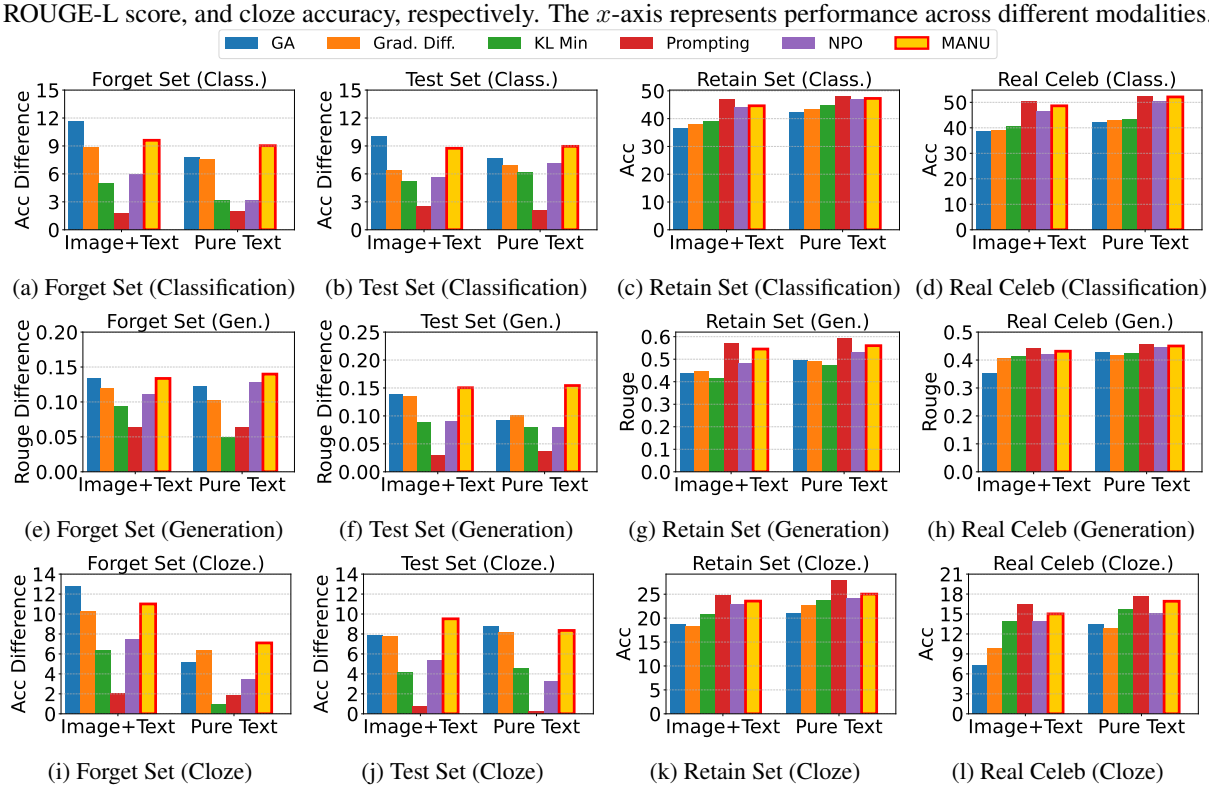


(a) Forget Set (Classification) (b) Test Set (Classification) (c) Retain Set (Classification) (d) Real Celeb (Classification)

(e) Forget Set (Generation) (f) Test Set (Generation) (g) Retain Set (Generation) (h) Real Celeb (Generation)

(i) Forget Set (Cloze) (j) Test Set (Cloze) (k) Retain Set (Cloze) (l) Real Celeb (Cloze)

Figure 7: Classification, generation, and cloze performance of MANU and baselines in multimodal and unimodal setups with 10% forget data, using LLaVA as the base model. In subplots (a), (b), (e), (f), (i), and (j), the y-axis represents the change in classification accuracy, ROUGE-L score, and cloze accuracy relative to the vanilla model, evaluated on the Forget and Test sets. In the remaining subplots, the y-axis indicates classification accuracy, ROUGE-L score, and cloze accuracy, respectively. The x-axis represents performance across different modalities.

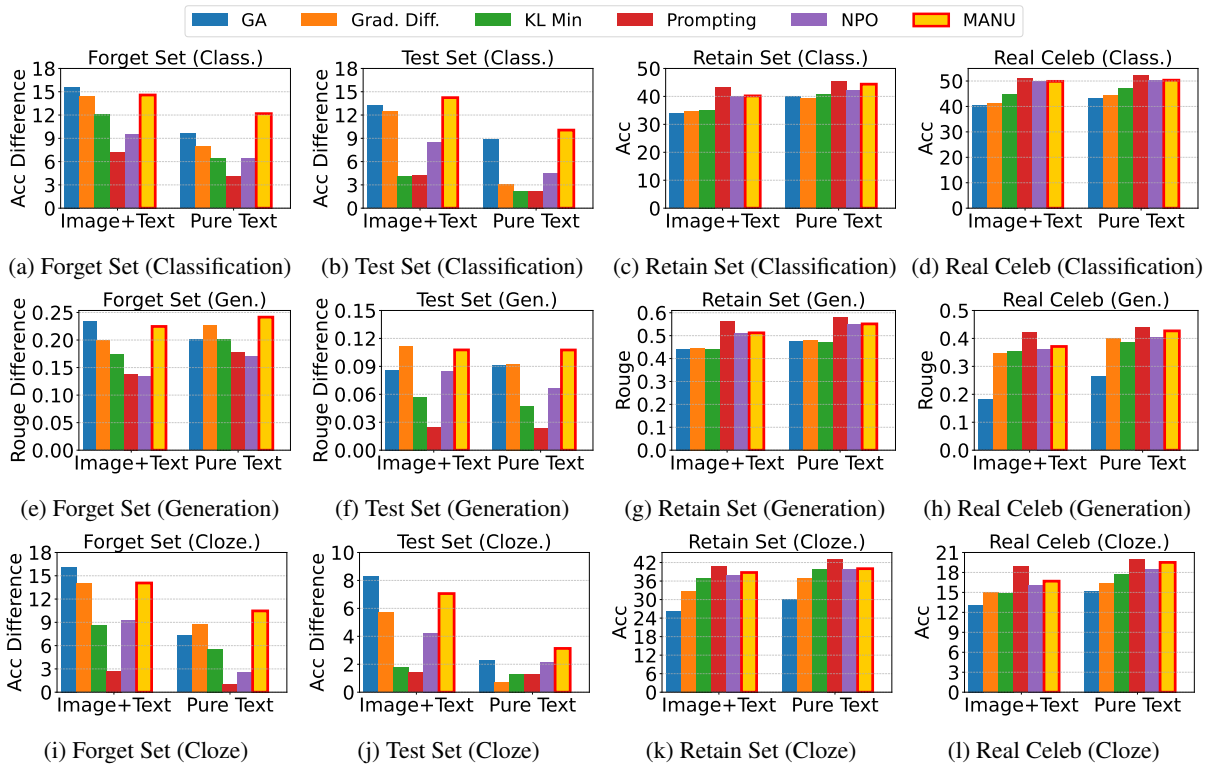


(a) Forget Set (Classification) (b) Test Set (Classification) (c) Retain Set (Classification) (d) Real Celeb (Classification)

(e) Forget Set (Generation) (f) Test Set (Generation) (g) Retain Set (Generation) (h) Real Celeb (Generation)

(i) Forget Set (Cloze) (j) Test Set (Cloze) (k) Retain Set (Cloze) (l) Real Celeb (Cloze)

Figure 8: Classification, generation, and cloze performance of MANU and baselines in multimodal and unimodal setups with 15% forget data, using LLaVA as the base model. In subplots (a), (b), (e), (f), (i), and (j), the y-axis represents the change in classification accuracy, ROUGE-L score, and cloze accuracy relative to the vanilla model, evaluated on the Forget and Test sets. In the remaining subplots, the y-axis indicates classification accuracy, ROUGE-L score, and cloze accuracy, respectively. The x-axis represents performance across different modalities.

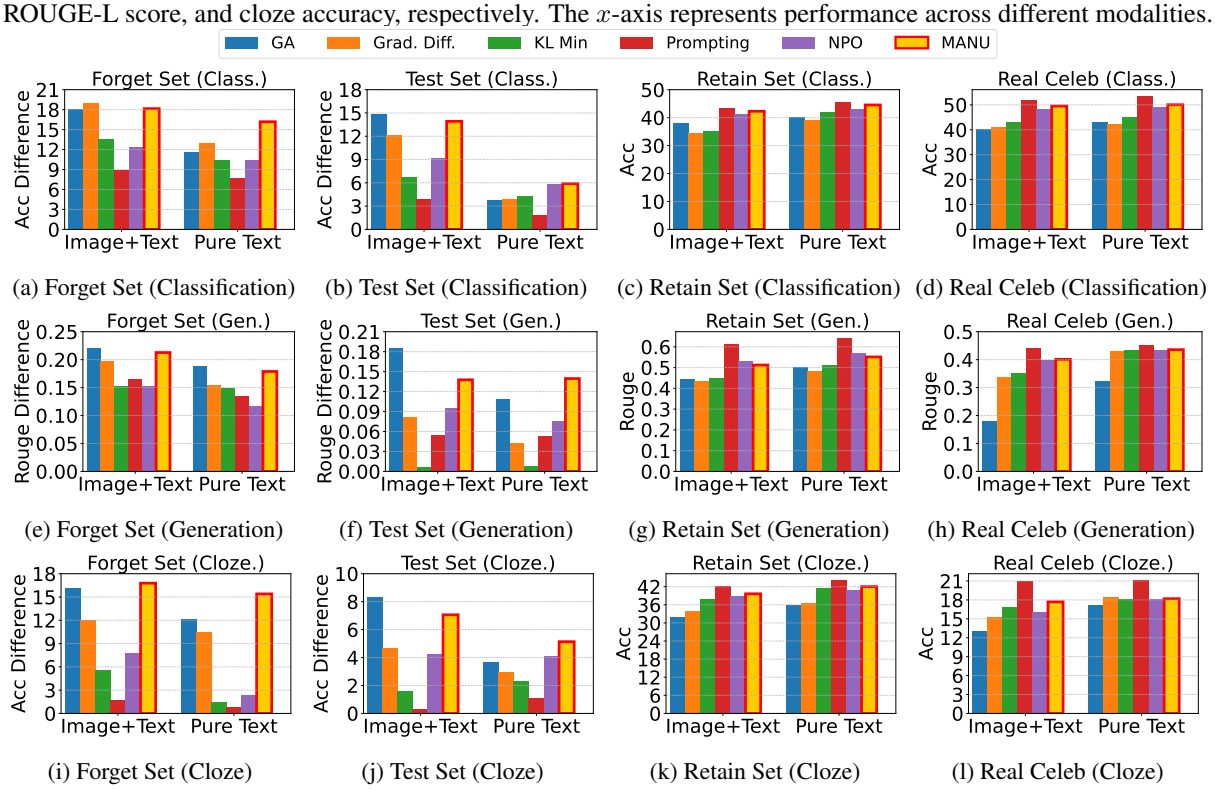


(a) Forget Set (Classification) (b) Test Set (Classification) (c) Retain Set (Classification) (d) Real Celeb (Classification)

(e) Forget Set (Generation) (f) Test Set (Generation) (g) Retain Set (Generation) (h) Real Celeb (Generation)

(i) Forget Set (Cloze) (j) Test Set (Cloze) (k) Retain Set (Cloze) (l) Real Celeb (Cloze)

Figure 9: Classification, generation, and cloze performance of MANU and baselines in multimodal and unimodal setups with 5% forget data, using Idefics2 as the base model. In subplots (a), (b), (e), (f), (i), and (j), the y-axis represents the change in classification accuracy, ROUGE-L score, and cloze accuracy relative to the vanilla model, evaluated on the Forget and Test sets. In the remaining subplots, the y-axis indicates classification accuracy, ROUGE-L score, and cloze accuracy, respectively. The x-axis represents performance across different modalities.

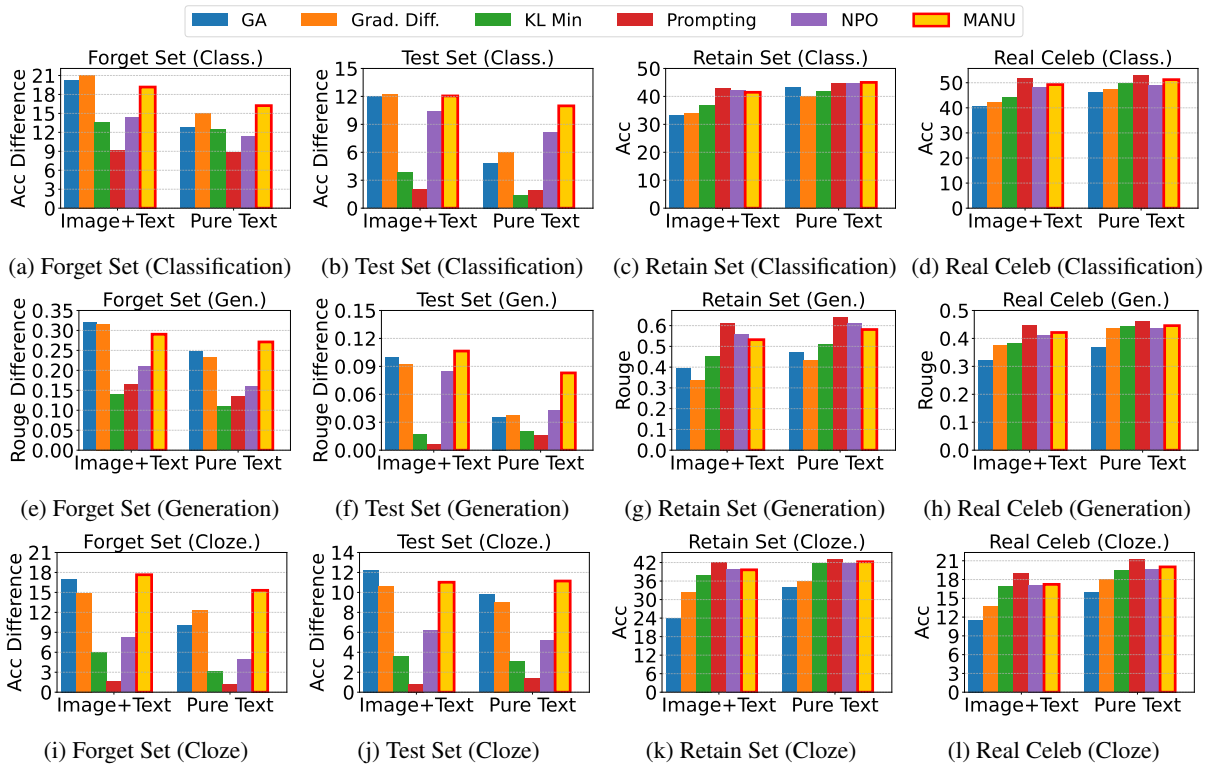


(a) Forget Set (Classification) (b) Test Set (Classification) (c) Retain Set (Classification) (d) Real Celeb (Classification)

(e) Forget Set (Generation) (f) Test Set (Generation) (g) Retain Set (Generation) (h) Real Celeb (Generation)

(i) Forget Set (Cloze) (j) Test Set (Cloze) (k) Retain Set (Cloze) (l) Real Celeb (Cloze)

Figure 10: Classification, generation, and cloze performance of MANU and baselines in multimodal and unimodal setups with 10% forget data, using Idefics2 as the base model. In subplots (a), (b), (e), (f), (i), and (j), the y-axis represents the change in classification accuracy, ROUGE-L score, and cloze accuracy relative to the vanilla model, evaluated on the Forget and Test sets. In the remaining subplots, the y-axis indicates classification accuracy, ROUGE-L score, and cloze accuracy, respectively. The x-axis represents performance across different modalities.



(a) Forget Set (Classification) (b) Test Set (Classification) (c) Retain Set (Classification) (d) Real Celeb (Classification)

(e) Forget Set (Generation) (f) Test Set (Generation) (g) Retain Set (Generation) (h) Real Celeb (Generation)

(i) Forget Set (Cloze) (j) Test Set (Cloze) (k) Retain Set (Cloze) (l) Real Celeb (Cloze)

Figure 11: Classification, generation, and cloze performance of MANU and baselines in multimodal and unimodal setups with 15% forget data, using Idefics2 as the base model. In subplots (a), (b), (e), (f), (i), and (j), the y-axis represents the change in classification accuracy, ROUGE-L score, and cloze accuracy relative to the vanilla model, evaluated on the Forget and Test sets. In the remaining subplots, the y-axis indicates classification accuracy, ROUGE-L score, and cloze accuracy, respectively. The x-axis represents performance across different modalities.

LLaVA (5% Forget Set)			
Model	**MIA-Bench (Avg)**	**MM-Vet (Avg)**	**MMStar (Avg)**
Vanilla	69.60	34.08	34.80
GA	57.29	28.92	29.83
Grad. Diff	59.02	30.28	30.13
KL Minim.	62.24	31.29	31.29
NPO	63.44	32.38	32.32
MANU	**64.02**	**32.76**	**32.41**
Idefics2 (5% Forget Set)			
Model	**MIA-Bench (Avg)**	**MM-Vet (Avg)**	**MMStar (Avg)**
Vanilla	54.42	43.90	49.27
GA	47.77	35.12	40.90
Grad. Diff	48.58	36.67	41.02
KL Minim.	49.64	38.61	41.09
NPO	50.89	40.22	42.88
MANU	**51.33**	**40.83**	**44.71**

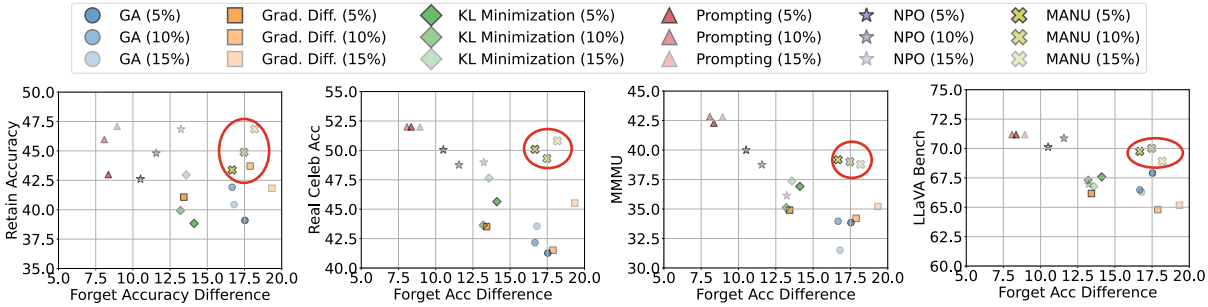
Table 6: Evaluation results on MIA-Bench, MM-Vet, and MMStar for LLaVA and Idefics2 under the 5% Forget Set setting. Higher scores indicate better performance.

fication accuracy drops from 43.38% to 42.37%, and Real Celebrity classification accuracy declines to 46.59%. This suggests that I_{abs} and I_{rms} are important for preserving high-activation neurons that contribute broadly to general reasoning, thus maintaining utility. These findings support our equal-weighting strategy, where each importance score captures a distinct and complementary signal. While we acknowledge that learned or tuned weightings might yield further improvements, we leave such model-specific enhancements for future work.

E.7 Appendix: Generalizability with Larger MLLMs

To further evaluate the scalability and generalizability of MANU to larger MLLMs, we conducted an additional set of experiments using the LLaVA-13B architecture. In this section, we only report results for the 5% Forget Set split for reference, which required fine-tuning a separate vanilla LLaVA-13B model on the MLLMU-Bench dataset. The quantitative results are presented in Table 8.

As it shown in the table, the observed performance trends on LLaVA-13B are consistent with those reported for the smaller 7B and 8B variants. Specifically, MANU consistently ranks among the top-performing methods across all evaluation tasks. While GA and Gradient Difference occasionally achieve marginally better scores in terms of unlearning effectiveness, these methods generally underperform in preserving model utility. Conversely, the Prompting-based approach demonstrates strong utility preservation but exhibits signif-



(a) Forget Acc vs Retain Acc (b) Forget Acc vs Real Celeb (c) Forget Acc vs MMMU (d) Forget Acc vs LLaVABench

Figure 12: The overall trade-off between unlearning effectiveness and model utility across all baselines using different forget data, with Idefics2 as the base model. The x-axis shows the difference in forget classification accuracy relative to the vanilla model, while the y-axis reflects model utility from various perspectives. From left to right, these perspectives include retain accuracy, real celebrity accuracy, MMMU, and LLaVA-Bench performance, respectively.

Models	Forget Set				Test Set				Retain Set				Real Celebrity			
	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)
LLaVA-7B (5% Forget)																
Vanilla	51.70	0.645	6.78	25.81	47.86	0.539	4.89	23.01	46.11	0.632	6.41	27.83	51.80	0.479	5.47	17.35
w/o I_{abs}	40.33	0.472	3.30	16.83	41.67	0.354	3.62	14.15	42.37	0.458	4.23	19.22	46.59	0.375	4.12	13.38
w/o I_{freq}	47.35	0.612	4.09	25.01	45.67	0.472	4.22	21.78	45.12	0.522	5.16	24.16	49.86	0.459	4.99	15.73
w/o I_{var}	46.67	0.636	4.14	24.92	44.17	0.460	4.04	22.75	44.09	0.564	5.39	24.64	50.61	0.461	4.85	16.07
w/o I_{rms}	40.82	0.500	3.19	15.17	40.33	0.382	3.70	16.03	41.02	0.471	4.19	20.68	46.08	0.466	4.05	12.03
MANU	41.25	0.491	3.27	17.08	41.67	0.334	3.81	15.78	43.38	0.542	4.45	24.08	49.57	0.448	4.67	16.01
Idefics2-8B (5% Forget)																
Vanilla	53.80	0.630	6.22	4.75	47.86	0.434	5.00	24.97	46.11	0.644	6.51	42.35	52.75	0.459	5.75	20.05
w/o I_{abs}	35.98	0.403	3.17	3.32	33.86	0.354	3.89	18.55	41.06	0.542	4.50	39.22	48.82	0.366	3.92	14.91
w/o I_{freq}	48.81	0.608	4.72	4.41	42.39	0.407	4.07	23.49	44.14	0.583	5.87	41.09	51.54	0.428	5.09	19.33
w/o I_{var}	46.65	0.588	4.66	4.83	44.77	0.395	4.14	23.82	45.34	0.626	5.80	41.43	52.22	0.417	4.99	19.15
w/o I_{rms}	36.18	0.420	3.05	2.18	32.19	0.324	3.32	17.77	40.91	0.507	4.44	37.19	46.88	0.342	3.73	14.11
MANU (full)	37.13	0.413	3.20	2.11	35.55	0.361	3.91	20.97	42.71	0.538	4.60	40.01	50.09	0.399	4.11	18.80

Table 7: Ablation study of MANU on two base MLLM models under a 5% forget data setup. Lower scores on the Forget/Test sets indicate better unlearning, while higher scores on the Retain/Celebrity sets indicate better utility preservation.

Model	Forget Set				Test Set				Retain Set				Real Celebrity			
	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↓)	Rouge Score (↓)	Fact. Score (↓)	Cloze Acc (↓)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)	Class. Acc (↑)	Rouge Score (↑)	Fact. Score (↑)	Cloze Acc (↑)
LLaVA-13B (5% Forget Set)																
Vanilla	44.40	0.604	6.10	21.90	36.80	0.493	4.72	20.82	46.11	0.657	6.38	27.80	**68.89**	**0.633**	**7.01**	**26.80**
GA	34.04	**0.423**	**2.98**	**15.19**	26.35	0.386	3.74	**15.01**	35.69	0.477	3.04	16.94	48.84	0.447	4.56	18.02
Grad. Diff	34.16	0.435	3.95	15.55	25.99	**0.362**	3.80	16.13	37.54	0.498	3.55	18.80	49.25	0.493	4.97	18.34
KL Minim.	38.64	0.561	4.17	19.97	31.06	0.432	4.13	18.62	40.00	0.571	4.18	20.07	55.29	0.555	5.44	22.15
Prompting	42.99	0.588	5.36	20.80	34.80	0.454	4.52	20.91	42.81	**0.622**	**5.23**	**27.66**	**67.72**	**0.633**	**6.93**	**25.76**
NPO	35.12	0.473	3.41	18.77	29.73	0.397	3.89	17.77	40.79	0.565	4.55	20.23	59.83	0.579	5.86	22.54
MANU	**33.65**	0.434	3.13	15.33	**25.41**	0.382	**3.70**	16.08	**42.88**	0.593	4.72	22.75	59.97	0.585	6.08	22.80

Table 8: Overall results of MANU with varying pruning ratios on LLaVA-13B model under a 5% forget data setup. Lower scores on the Forget/Test sets indicate better unlearning, while higher scores on the Retain/Celebrity sets indicate better utility preservation.

icantly lower forgetting capability. MANU offers a robust compromise, maintaining competitive unlearning performance while preserving downstream utility across all evaluation settings.

These findings reaffirm the central design hypothesis behind MANU —that modality-specific importance signals can be effectively extracted and leveraged even within larger, more entangled model architectures. Additionally, we note that italicized emphasis used in our main tables to indicate second-best values may not be easily distinguishable in print. In the revised manuscript, we will replace italics with underlined formatting for improved visual clarity. This extension strengthens the empirical validation of MANU and demonstrates its applicability to state-of-the-art, large-scale MLLMs.