# Cross-Modal Unlearning via Influential Neuron Path Editing in Multimodal Large Language Models

**Kunhao Li[*1], Wenhao Li[*1], Di Wu[*2], Lei Yang[1†], Jun Bai[3, 4], Ju Jia[5], Jason Xue[6]**

[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]School of Computing, Engineering and Mathematical Science, La Trobe University, Melbourne, Australia
[3]School of Computer Science, McGill University, Montreal, Canada
[4]Mila-Quebec AI Institute, Montreal, Canada
[5]School of Cyber Science and Engineering, Southeast University, Nanjing, China
[6]CSIRO's Data61 and Responsible AI Research (RAIR) Centre, Adelaide University
kunhomlihf@gmail.com, wenhaoli-lwh@outlook.com, d.wu@latrobe.edu.au,
sely@scut.edu.cn, jun.bai@mcgill.ca, jiaju@seu.edu.cn, minhuixue@gmail.com

## Abstract

Multimodal Large Language Models (MLLMs) extend foundation models to real-world applications by integrating inputs such as text and vision. However, their broad knowledge capacity raises growing concerns about privacy leakage, toxicity mitigation, and intellectual property violations. Machine Unlearning (MU) offers a practical solution by selectively forgetting targeted knowledge while preserving overall model utility. When applied to MLLMs, existing neuron-editing-based MU approaches face two fundamental challenges: (1) forgetting becomes inconsistent across modalities because existing point-wise attribution methods fail to capture the structured, layer-by-layer information flow that connects different modalities; and (2) general knowledge performance declines when sensitive neurons that also support important reasoning paths are pruned, as this disrupts the model's ability to generalize. To alleviate these limitations, we propose a multimodal influential neuron path editor (`MIP-Editor`) for MU. Our approach introduces modality-specific attribution scores to identify influential neuron paths responsible for encoding forget-set knowledge and applies influential-path-aware neuron-editing via representation misdirection. This strategy also enables effective and coordinated forgetting across modalities while preserving the model's general capabilities. Experimental results demonstrate that `MIP-Editor` achieves a superior unlearning performance on multimodal tasks, with a maximum forgetting rate of 87.75% and up to 54.26% improvement in general knowledge retention. On textual tasks, `MIP-Editor` achieves up to 80.65% forgetting and preserves 77.9% of general performance.

**Code** — https://github.com/PreckLi/MIP-Editor

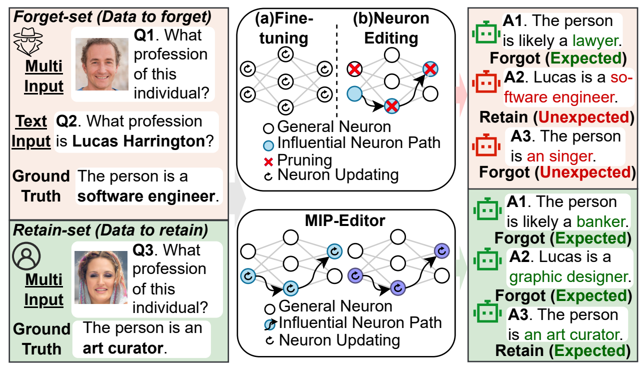



Figure 1: Comparison between existing MU methods and `MIP-Editor`. Prior methods suffer from: (1) insufficient forgetting in the text modality, as point-wise attribution fails to capture structured cross-layer information flow; and (2) disruption of influential reasoning paths due to pruning.

## 1 Introduction

The rapid advancement of multimodal large language models (MLLMs) has extended model capabilities to a wide range of applications through multimodal integration (Zhang et al. 2023; Caffagni et al. 2024). However, their vast knowledge capacity raises serious concerns about privacy leakage (Pi et al. 2024), intellectual property violations (Li et al. 2024a), regulatory compliance beyond privacy (Chundawat et al. 2023), toxicity mitigation (Łucki et al. 2025), and model refinement (Jia et al. 2023). Machine Unlearning (MU) (Si et al. 2023) offers a promised solution to remove unwanted knowledge from MLLMs, supporting controllable and compliant model adaptation. However, current research on MU for MLLMs remains underexplored.

Methods such as (Thudi et al. 2022; Liu, Liu, and Stone 2022; Zhang et al. 2024) primarily extend fine-tuning-based unlearning strategies originally designed for LLMs. However, these methods ignore the unique discrepancies between modalities in MLLMs, and struggle to forget modality-specific knowledge effectively, especially under textual inputs, as illustrated in Fig. 1 (a). An alternative line of work that explores neuron-level editing has emerged as a

[*]These authors contributed equally.
[†]Corresponding author.

promising direction, based on the observation that model knowledge is stored in distributed patterns within learnable parameters (Yao et al. 2023). Recent approaches such as DEPN (Wu et al. 2023) and MANU (Liu et al. 2025b) attempt modality-specific forgetting via single-neuron pruning or activation-based scoring. However, the point-wise estimation fails to capture the structured information flow across layers in multimodal architectures. As a result, forgetting remains uncoordinated across modalities. This limitation is statistically demonstrated in Table 1. Moreover, as illustrated in Fig. 1 (b), pruning neurons solely based on their individual importance to the forget-set may inadvertently disrupt reasoning pathways critical to the retain-set, leading to severe degradation of general knowledge.

Recent studies (Wang et al. 2025) have confirmed that information in large models is transmitted through structured, layer-wise neuron pathways. These influential paths offer a more coherent and semantically grounded basis for unlearning compared to isolated neurons. In MLLMs, both textual and multimodal (e.g., image–text) inputs rely on such structured reasoning flows. This motivates a shift from point-based deletion to path-aware interventions that better align with the model's internal knowledge organization. To this end, we propose a **M**ultimodal **I**nfluential neuron **P**ath **Editor** (`MIP-Editor`) tailored for MU in MLLMs. Our approach locates modality-specific influential neuron paths in the FFN layers of both the textual and visual branches by computing inter-layer gradient-integrated and Fisher-integrated attribution scores. In particular, we introduce an influential-path-based neuron editing method using Representation Misdirection Unlearning (RMisU) that adaptively steers the representations of forget-set inputs away from their original semantics, reducing the impact on general knowledge. To sum up, our contributions are as follows:

- We propose a dual-branch (visual and textual) influential neuron path localization framework. This approach leverages inter-layer gradient-integrated and Fisher-integrated attribution scores to capture modality-specific information flow, enabling precise localization of neurons responsible for specific knowledge in each modality.
- We analyze the limitations of direct pruning strategies, where overlapping neurons between forget and retain sets cause a collapse of general reasoning paths. To mitigate this, we propose a targeted RMisU-based neuron editing strategy that operates only on the influential neuron paths, decoupling specific and general knowledge.
- Experiments demonstrate that `MIP-Editor` achieves modality-consistent forgetting with strong retain-set performance, reaching up to $87.75\%$ forgetting and $54.26\%$ retention improvement on multimodal tasks, and $80.65\%$ forgetting with $77.9\%$ retention on textual tasks.

## 2 Related Works

**Fine-tuning for MLLM Unlearning** Recent MU efforts aim to remove specific knowledge from models for privacy and safety. Early approaches for LLMs employ gradient ascent (Thudi et al. 2022; Liu, Liu, and Stone 2022), KL minimization (Nguyen, Low, and Jaillet 2020), and preference-based objectives (Zhang et al. 2024), with applications in toxicity mitigation (Chen et al. 2025) and hallucination reduction (Xing et al. 2024). However, these methods are confined to text-only settings. Multimodal unlearning for MLLMs remains largely unexplored. To support multimodal research, dedicated benchmarks MLLMU-Bench (Liu et al. 2025a) and CLEAR (Dontsov et al. 2024) have been proposed. Recent studies (Yang et al. 2024; Cheng and Amiri 2024; Li et al. 2024b; Huo et al. 2025) extend fine-tuning to erase visual concepts, yet they require full-model updates and overlook the modular architecture of MLLMs.

**Neuron Editing in Large Language Models** Neuron editing provides a fine-grained mechanism to alter model behavior while preserving general capabilities. Studies have examined how pretrained LMs encode knowledge (Chen et al. 2024; Li et al. 2023; Cao et al. 2024; Lamparth and Reuel 2024), enabling targeted neuron-level interventions. This technique has been applied to MU (Wu, Hashemi, and Srinivasa 2022; Hase et al. 2023; Gandikota et al. 2023), harmful content mitigation (Hu et al. 2024), continual learning (Biesialska, Biesialska, and Ruiz Costa-Jussà 2020), and privacy protection (Wu et al. 2023). In multimodal contexts, MANU (Liu et al. 2025b) introduces neuron-level editing for modality-specific forgetting, yet its heuristic scoring and zero-out pruning can disrupt reasoning coherence. Our approach mitigates these issues through path-aware editing, identifying coherent, modality-specific neuron sequences for more consistent and effective unlearning.

## 3 Problem Definition

We focus on MU for MLLMs, aiming to remove targeted forgetting knowledge while minimizing degradation of general capabilities. Let $M_\theta$ denote the original MLLM with parameters $\theta$, trained on a dataset $D = \{(I_i, T_i)\}_{i=1}^{\mathbf{N}}$ of $\mathbf{N}$ image–text pairs, where $I_i$ is an image and $T_i = \{s_1^i, \ldots, s_{t_i}^i\}$ is its corresponding tokenized text. Each pair includes a question–answer prompt for visual understanding. We divide $D$ into a forget-set $D^f = \{(I_j^f, T_j^f)\}_{j=1}^{\mathbf{N}_f}$, containing specific concepts to be forgotten, and a retain-set $D^r = \{(I_k^r, T_k^r)\}_{k=1}^{\mathbf{N}_r}$, used to preserve general knowledge.

Following (Liu et al. 2025a,b), we define MU in MLLMs as: *The process of removing both visual and textual forgetting data from a model while preserving its predictive performance on unrelated inputs.* To achieve this, we minimize the negative log-likelihood of next-token prediction and obtain the unlearned model $M_{\hat{\theta}}$ via the objective:

$$\arg\min_{\theta^*} \Bigg\{ \underbrace{-\mathbb{E}_{(I,T)\in D^f} \Big[ -\sum_{\mathbf{n}=1}^{\mathbf{N}_f} \log p_{M_{\hat{\theta}}}(w_{\mathbf{n}} | (I,T),\, w_{<\mathbf{n}}) \Big]}_{\text{Forget specific visual \& textual patterns}}$$

$$\underbrace{+\, \mathbb{E}_{(I,T)\in D^r} \Big[ -\sum_{\mathbf{n}=1}^{\mathbf{N}_r} \log p_{M_{\hat{\theta}}}(w_{\mathbf{n}} | (I,T),\, w_{<\mathbf{n}}) \Big]}_{\text{Retain general knowledge}} \Bigg\} \quad (1)$$

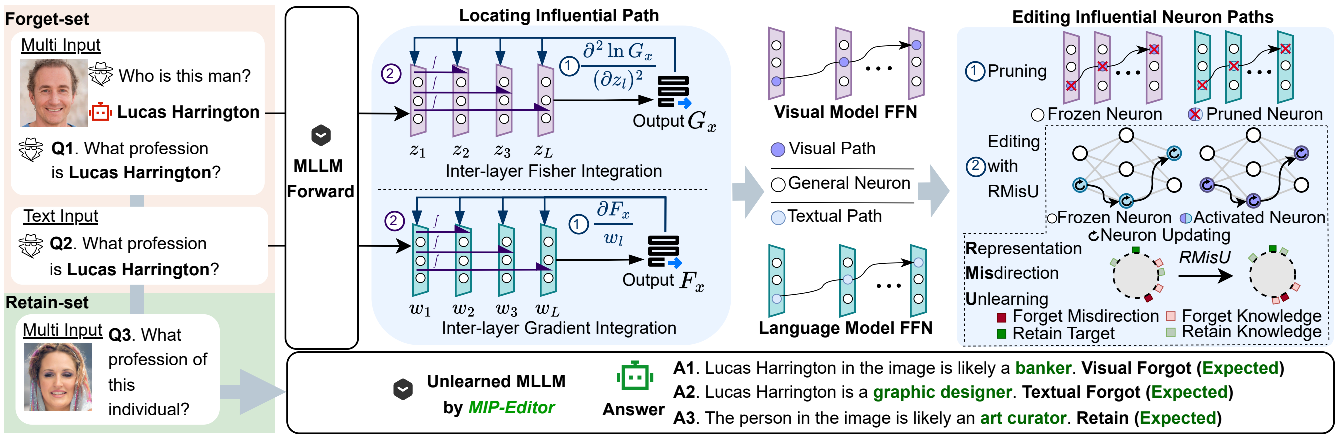



Figure 2: Overview of MIP-Editor. (1) Influential neuron paths are located using inter-layer gradient (text) and Fisher (vision) integration. (2) Neurons inside the selected paths are pruned, and (3) path-specific editing is performed via representation misdirection to achieve modality-consistent forgetting while preserving general knowledge.

# 4 Method

## 4.1 Locating Influential Neuron Path

In MIP-Editor (Fig. 2), the textual and visual influential paths are inherently related, as each input pair $(I, T)$ carries semantically aligned content. Prior works (Radford et al. 2021; Pan et al. 2024; Sato and Takagi 2025) show that vision and text features are mapped into a shared embedding space and jointly processed via cross-attention, ensuring correspondence between activations across modalities.

**Inter-layer Gradient Integration** To locate influential neurons in the textual modality, we propose an inter-layer gradient integration method inspired by information flow (Lu et al. 2021) and joint attribution (Wang et al. 2025). Given the $L$-layer architecture of a language model, where each FFN layer is regarded as a key repository of factual knowledge, we aim to quantify the contribution of selected neurons across the first $N$ layers.

Let $\langle T, Y\rangle$ denote a labeled text pair, where $T \in \mathbb{R}^d$ is the input text and $Y$ is the expected output. The model's output over the first $N$ layers is represented as:

$$F_T(\mathbf{w}) = p(Y \mid T, w_{i_1}^1, \cdots, w_{i_N}^N), \tag{2}$$

where $\mathbf{w} = (w_{i_1}^1, \cdots, w_{i_N}^N)$ are the activations of selected neurons in the textual FFN layers, and $\tilde{w}_{i_n}^n$ denotes the original activation of neuron $w_{i_n}^n$ of the $n$-th layer.

To estimate the joint contribution of these neurons, we scale the activation values $\{\alpha_{i_1}^1, \cdots, \alpha_{i_N}^N\}$ from 0 to their original activations $\{\tilde{w}_{i_1}^1, \cdots, \tilde{w}_{i_N}^N\}$. The inter-layer gradient-integrated attribution score is defined as:

$$\mathrm{IGI}(\mathbf{w}) = \sum_{n=1}^{N} \tilde{w}_{i_n}^n \int_0^{\tilde{w}_{i_n}^n} \sum_{l=1}^{N} \frac{\partial F_T(\alpha_{i_1}^1, \cdots, \alpha_{i_n}^n)}{\partial w_{i_l}^l} \, d\alpha_{i_n}^n, \tag{3}$$

which measures how the neurons along the path contribute to the model's output by integrating gradients across layers.

To approximate the integral, we employ Riemann approximation (Dai et al. 2022) by interpolating $m$ frames into the activation values. The discrete form of the IGI becomes:

$$\mathrm{IGI}(\mathbf{w}) = \sum_{j=1}^{N} \tilde{w}_{i_j}^n \sum_{k=1}^{m} \sum_{l=1}^{N} \frac{\partial F_T\left(\frac{k}{m}\alpha_{i_1}^1, \cdots, \frac{k}{m}\alpha_{i_N}^N\right)}{\partial w_{i_l}^l}. \tag{4}$$

**Inter-layer Fisher Integration** To locate influential neurons in the visual modality, we adopt an inter-layer Fisher integration method similar to the gradient-based approach used for text. Due to the high dimensionality, spatial correlation, and parameter redundancy in vision encoders, the Fisher Information Matrix (FIM) offers a more suitable signal for estimating neuron importance.

Let $\langle (I, T), Y\rangle$ denote a multimodal input with visual input $I \in \mathbb{R}^{d^I}$, text input $T \in \mathbb{R}^{d^T}$, and target output $Y$. The log-likelihood output over the first $N$ visual FFN layers is:

$$\mathbf{G}(\mathbf{z}) = \log p(Y \mid I, T, z_{i_1}^1, \cdots, z_{i_N}^N), \tag{5}$$

where $\mathbf{z} = (z_{i_1}^1, \cdots, z_{i_N}^N)$ are the activations of selected visual neurons, and $\tilde{z}_{i_n}^n$ their original values.

Similar to the textual integration (Eq. 4), we interpolate neuron activations from 0 to their original values using $m$ steps. To approximate the diagonal of the FIM, we adopt the squared-gradient formulation. The inter-layer Fisher-integrated score is defined as:

$$\mathrm{IFI}(\mathbf{z}) = \sum_{n=1}^{N} \tilde{z}_{i_n}^n \sum_{k=1}^{m} \sum_{l=1}^{N} \left( \frac{\partial \mathbf{G}\left(\frac{k}{m}\beta_{i_1}^1, \cdots, \frac{k}{m}\beta_{i_N}^N\right)}{\partial z_{i_l}^l} \right)^2, \tag{6}$$

where $\{\beta_{i_n}^n\}$ are the interpolated activations. This formulation enables efficient estimation of visual neuron importance by integrating second-order signals across layers.

**Locating Paths** Following (Wang et al. 2025), we define the influential neuron paths in the FFN layers of MLLMs.

**Definition 1 (Influential Paths)** *Let $F : \mathbb{R}^d \to \mathbb{R}$ be a multimodal model consisting of $L$ FFN layers for a given modality, and let $x$ denote either a text input $T$ or an image–text*

**Algorithm 1:** Inter-layer Integrated Influential Path Locating

**Input:** MLLM $M_\theta$ with $L^t$ textual layers and $L^v$ visual layers, input pair $(I, T)$
**Output:** Visual path $\mathcal{P}^v$, Textual path $\mathcal{P}^t$

$\mathcal{P}^v \leftarrow \emptyset$, $\mathcal{P}^t \leftarrow \emptyset$
**for** $k = 1$ to $L^t$ **do**
  Let $\mathcal{W}^t$ be the set of neurons in textual layer $k$
  Score $\leftarrow -\infty$, bestNeuron $\leftarrow$ None
  **for all** $w \in \mathcal{W}^t$ **do**
    $s \leftarrow \mathrm{IGI}(\mathcal{P}^t \cup \{w\}, T)$
    **if** $s >$ Score **then**
      Score $\leftarrow s$, bestNeuron $\leftarrow w$
    **end if**
  **end for**
  $\mathcal{P}^t \leftarrow \mathcal{P}^t \cup \{\text{bestNeuron}\}$
**end for**
**for** $l = 1$ to $L^v$ **do**
  Let $\mathcal{W}^v$ be the set of neurons in visual layer $l$
  Score $\leftarrow -\infty$, bestNeuron $\leftarrow$ None
  **for all** $z \in \mathcal{W}^v$ **do**
    $s \leftarrow \mathrm{IFI}(\mathcal{P}^v \cup \{z\}, I, T)$
    **if** $s >$ Score **then**
      Score $\leftarrow s$, bestNeuron $\leftarrow z$
    **end if**
  **end for**
  $\mathcal{P}^v \leftarrow \mathcal{P}^v \cup \{\text{bestNeuron}\}$
**end for**
**return** $\mathcal{P}^v$, $\mathcal{P}^t$

*pair $(I, T)$. A neuron path is defined as*

$$\mathcal{P}_x = \{w^1, w^2, \ldots, w^L\}, \qquad w^l \in \mathbf{W}, \tag{7}$$

*where $w_l$ denotes the set of selected neurons in the $l$-th layer.* $\mathbf{W}$ *represents the general FFN layer parameters*

*We define a modality-specific scoring function as follows:*

$$S(\mathcal{P}_x) = \begin{cases} \mathrm{IGI}(\mathcal{P}_x), & \textit{if } x = T, \\ \mathrm{IFI}(\mathcal{P}_x), & \textit{if } x = (I, T), \end{cases} \tag{8}$$

*The influential path $\mathcal{P}_x^*$ is then defined as the one that maximizes the corresponding score:*

$$\mathcal{P}_x^* = \arg\max_{\mathcal{P}_x} S(\mathcal{P}_x). \tag{9}$$

To locate influential paths efficiently, we apply a greedy layer-wise search (Algorithm 1) that selects the most influential neuron per layer. Given input $(I, T)$ and a pretrained model $F$ with $L^t$ textual and $L^v$ visual FFN layers, the algorithm outputs two ordered paths: $\mathcal{P}^t$ and $\mathcal{P}^v$ for the textual and visual modalities, respectively.

## 4.2 Editing Influential Neuron Paths

Having located influential paths, we prune encoding forget-set information and employ RMisU to redirect their activations away from undesired semantics while reinforcing retain-set representations. This prune-and-finetune strategy updates only a small subset of neurons, enabling effective forgetting with negligible impact on general knowledge.

**Pruning** Specifically, we perform targeted pruning by zeroing the activations of neurons identified as specific-relevant along the influential paths. For a text-only input $T$ and a multimodal input $(I, T)$, let the corresponding influential paths be $\mathcal{P}_T = \{\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_{L^t}\}$ and $\mathcal{P}_{(I,T)} = \{\tilde{z}_1, \tilde{z}_2, \ldots, \tilde{z}_{L^v}\}$, the pruning can be formally expressed as:

$$\tilde{w}_l \leftarrow \mathbf{0}, \forall l \in \{1, \ldots, L^t\}; \quad \tilde{z}_l \leftarrow \mathbf{0}, \forall l \in \{1, \ldots, L^v\}, \tag{10}$$

where $\mathbf{0}$ represents an all-zero vector with the same dimension, $\tilde{w}_l$ and $\tilde{z}_l$ denote the activation values set of the selected neurons in the $l$-th layer. This operation ensures that the flow of information associated with forgetting concepts is blocked, thereby achieving targeted forgetting.

**Editing with RMisU** Pruning risks losing general knowledge in overlapping neurons of the retain-set. To recover it adaptively with minimal forgetting, we fine-tune only the pruned neurons using the retain-set, enabling adaptive recovery with less reintroduction of forgotten content. Specifically, all other parameters in the MLLM are frozen, and only neurons along influential paths are updated. Let $M_{\theta^*}$ denote the pruned model. To preserve general knowledge, we minimize a cross-entropy loss over the retain-set $D^r$:

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{(x^r, y^r) \in D^r} \left[ -\sum_{i=1}^{|y^r|} \log P_{M_{\theta^*}}(y_i^r \mid x^r, y_{<i}^r) \right], \tag{11}$$

where $x^r \in \{T^r, (I^r, T^r)\}$ denotes either a text or image–text input, and $y^r$ is the corresponding output sequence.

To forget specific knowledge in the forget-set $D^f$, prior methods use gradient ascent, KL divergence, or contrastive objectives, but often at the cost of linguistic and utility degradation (Fan et al. 2024). To avoid this, we adopt adaptive Representation Misdirection Unlearning (RMisU) (Dang et al. 2025), which steers forget-set representations away from their original semantics via localized directional perturbation at a specific layer $l$. This targeted editing removes specific knowledge while preserving general linguistic ability (Li et al. 2024c).

For each forget-set input $x^f \in \{T^f, (I^f, T^f)\}$, we sample a random unit vector from the unit sphere:

$$\mathbf{u} \sim \mathrm{Uniform}\left(\mathbb{S}^{d-1}\right), \tag{12}$$

and define a layer-specific target representation as

$$\mathbf{v}^f = \lambda \cdot \left\| \mathbf{h}_{M_\theta}^{(l)}(x^f) \right\|_2 \cdot \mathbf{u}, \tag{13}$$

where $\mathbf{h}_{M_\theta}^{(l)}(x^f)$ denotes the frozen model's hidden representation at layer $l$, and $\lambda$ is a scaling coefficient modulating the influence of the perturbation.

**Forgetting RMisU loss.** This term forces the representation of forget-set samples to align with the randomized vector $\mathbf{v}_F$, effectively erasing forgetting knowledge:

$$\mathcal{L}_{\text{RMisU}}^f = \mathbb{E}_{x^f \in D^f} \left\| \mathbf{h}_{M_{\theta^*}}^{(l)}(x^f) - \mathbf{v}^f \right\|_2^2, \tag{14}$$

where $\mathbf{h}_{M_{\theta^*}}^{(l)}(x^f)$ denotes the intermediate representation at layer $l$ for a forget-set sample $x^f$ in $M_{\theta^*}$ at current epoch.

**Retaining RMisU loss.** We minimize deviation of retain-set representations from the frozen model for generalization:

$$\mathcal{L}^{r}_{\mathrm{RMisU}} = \mathbb{E}_{x^r \in D_{\mathrm{r}}} \left\| \mathbf{h}^{(l)}_{M_{\theta*}}(x^r) - \mathbf{h}^{(l)}_{M_\theta}(x^r) \right\|_2^2 . \tag{15}$$

**Full objective.** The overall adaptive RMisU loss is:

$$\mathcal{L}_{\mathrm{RMisU}} = \mathcal{L}^{f}_{\mathrm{RMisU}} + \gamma \cdot \mathcal{L}^{r}_{\mathrm{RMisU}}, \tag{16}$$

where $\gamma > 0$ balances forgetting and retention.

# 5 Experiments

In this section, we answer the following key questions concerning the performance of `MIP-Editor` with experiments. Q1: Can `MIP-Editor` effectively eliminate multimodal information from the target MLLMs? Q2: Can `MIP-Editor` achieve coordinated forgetting across visual and textual modalities? Q3: Can `MIP-Editor` strike a balance between forgetting information and preserving general knowledge? Q4: Does `MIP-Editor` retain more informative content through influential neuron paths compared to point-wise probing methods?

## 5.1 Experimental Setup and Baselines

To evaluate the effectiveness of `MIP-Editor`, we conduct experiments on two representative MLLMs of different scales: Qwen2.5-VL-3B-Instruct (Wang et al. 2024) and LLaVA1.5-7B (Liu et al. 2023), using two dedicated multimodal unlearning benchmarks: MLLMU-Bench (Liu et al. 2025a) and CLEAR (Dontsov et al. 2024). These datasets provide structured forget and retain splits across diverse multimodal tasks, including visual question answering (VQA) and text-based QA, covering both generation and classification settings. We compare `MIP-Editor` with four strong baselines: GA_Diff (Liu, Liu, and Stone 2022), KL_Min (Nguyen, Low, and Jaillet 2020), NPO (Zhang et al. 2024), and MANU (Liu et al. 2025b). Vanilla denotes the original model without unlearning. For fair comparison, all methods are trained using the same configurations. MLLMU-Bench uses 5%, 10%, and 15% of its samples as forget-sets, while CLEAR uses 1%, 5%, and 10%.

## 5.2 Main Results

To answer Q1 and Q2, we evaluate the unlearning performance of various methods on multimodal and textual tasks using Qwen2.5-VL-3B-Instruct and LLaVA1.5-7B under a 5% forget ratio on MLLMU-Bench and CLEAR (Table 1). On multimodal tasks, `MIP-Editor` significantly reduces forgetting knowledge retention. For instance, it lowering FVQA accuracy from 39.20% (Vanilla) to 4.80% and improving RVQA from 37.72% to 58.19% on MLLMU based on Qwen2.5-VL. This corresponds to an 87.75% forgetting rate and a 54.26% improvement in general knowledge retention, outperforming GA_Diff, KL_Min, and NPO. Similar trends are observed for LLaVA1.5. On textual tasks, `MIP-Editor` reduces FQA accuracy from 49.60% to 9.60%, achieving an 80.65% forgetting rate while retaining 77.9% of the original performance. Compared with MANU,

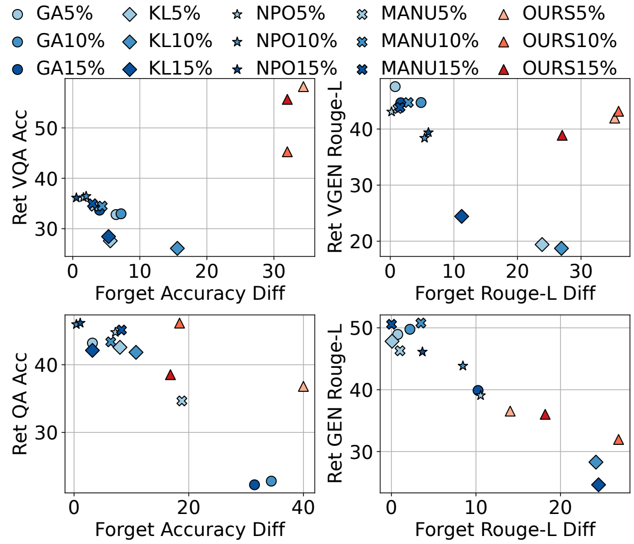



Figure 3: The overall trade-off between unlearning effectiveness and model utility across four dimensions under varying forget ratios, using Qwen2.5-VL as the base model.

NPO, and KL_Min, `MIP-Editor` more effectively suppresses residual forgetting knowledge while preserving competitive accuracy on the retain-set (e.g., 58.19% RVQA and 47.34% RQA). These results confirm the strength of `MIP-Editor` in achieving coordinated forgetting across modalities with minimal impact on general capability.

## 5.3 Unlearning v.s. Model Utility

To evaluate whether `MIP-Editor` achieves a superior trade-off between forgetting specific data and retaining general knowledge (Q3), we compare the performance differences on the forget-set with the post-unlearning accuracy on the retain-set. This analysis reflects each method's ability to balance unlearning and utility preservation. We evaluate four tasks on MLLMU-Bench: VQA (visual question answering), VGEN (visual generation), QA (textual question answering), and GEN (textual generation). As shown in Fig. 3, the $x$-axis denotes the performance drop on the forget-set (higher is better for forgetting), and the $y$-axis represents the retain-set performance after unlearning (higher is better for retention). An ideal method lies toward the upper right, indicating strong forgetting with minimal generalization loss. Results show that `MIP-Editor` achieves a consistently favorable trade-off across all tasks and forget ratios, with stronger gains in multimodal settings. Notice that in the GEN task, the gap is less pronounced due to the limitations of Rouge-L, which measures only semantic overlap and may not capture forget-set-related differences effectively.

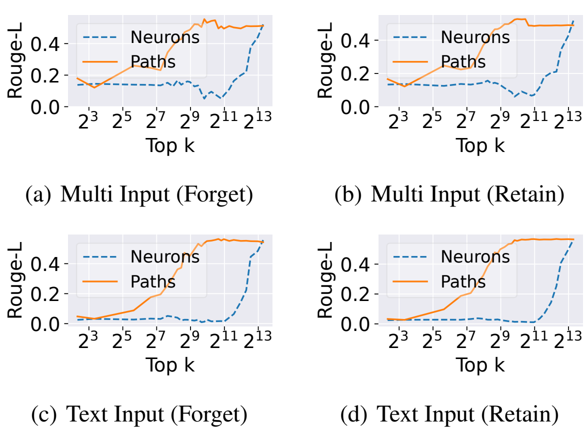
## 5.4 Influential Paths v.s. Influential Neurons

To assess whether influential neuron paths capture more information than point-wise neurons (Q4), we compare two selection strategies: (a) path-based (`MIP-Editor`) and (b) point-wise (Liu et al. 2025b). For each, we select the top-$k$ neurons per layer in both modalities and zero out the rest.

| Method | MLLMU-Bench | | | | | | CLEAR | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Task<br>Metric | FVQA<br>Acc(↓) | RVQA<br>Acc(↑) | FVGEN<br>Rouge(↓) | RVGEN<br>Rouge(↑) | FQA<br>Acc(↓) | RQA<br>Acc(↑) | FVQA<br>Acc(↓) | RVQA<br>Acc(↑) | FVGEN<br>Rouge(↓) | RVGEN<br>Rouge(↑) | FGEN<br>Rouge(↓) | RGEN<br>Rouge(↑) |
| Qwen2.5-VL-3B-Instruct | | | | | | | | | | | | |
| Vanilla | 39.20% | 37.72% | 0.4527 | 0.4347 | 49.60% | 47.20% | 72.34% | 73.42% | 0.3196 | 0.2997 | 0.3776 | 0.3900 |
| GA_Diff | 32.00% | 32.80% | 0.4450 | 0.4756 | 46.40% | 43.20% | 27.66% | 23.04% | 0.2946 | 0.2751 | 0.3740 | 0.3896 |
| KL_Min | 33.60% | 27.59% | 0.2139 | 0.1940 | 41.60% | 42.57% | 12.77% | 9.11% | 0.2532 | 0.2400 | 0.3270 | 0.3287 |
| NPO | 37.60% | 36.20% | 0.4507 | 0.4307 | 42.40% | 44.80% | 7.45% | 9.37% | 0.0803 | 0.0605 | 0.0805 | 0.0639 |
| MANU | 36.00% | 34.47% | 0.4406 | 0.4367 | 30.80% | 34.65% | 78.72% | 77.97% | 0.3220 | 0.2987 | 0.3809 | 0.3903 |
| MIP-Editor | 4.80% | 58.19% | 0.0997 | 0.4195 | 9.60% | 36.80% | 3.19% | 24.05% | 0.0707 | 0.2684 | 0.0926 | 0.3631 |
| Llava-1.5-7B | | | | | | | | | | | | |
| Vanilla | 56.80% | 51.56% | 0.5580 | 0.4946 | 50.40% | 52.59% | 44.68% | 43.54% | 0.3060 | 0.2937 | 0.3462 | 0.3546 |
| GA_Diff | 54.40% | 52.78% | 0.5719 | 0.5071 | 42.40% | 49.83% | 14.36% | 15.19% | 0.3057 | 0.2931 | 0.3565 | 0.3620 |
| KL_Min | 32.80% | 38.27% | 0.3594 | 0.3390 | 43.20% | 43.29% | 43.62% | 42.28% | 0.2200 | 0.2068 | 0.1380 | 0.1671 |
| NPO | 48.00% | 47.26% | 0.5388 | 0.4907 | 46.40% | 51.52% | 10.64% | 15.95% | 0.2091 | 0.1815 | 0.0150 | 0.0136 |
| MANU | 56.00% | 52.11% | 0.5486 | 0.4960 | 48.80% | 52.19% | 43.62% | 42.28% | 0.3070 | 0.2920 | 0.3452 | 0.3556 |
| MIP-Editor | 38.40% | 47.22% | 0.3418 | 0.3552 | 36.80% | 47.34% | 6.38% | 52.66% | 0.9690 | 0.2258 | 0.1441 | 0.2268 |

Table 1: Overall performances of baseline methods and `MIP-Editor` on machine unlearning tasks with 5% forget ratio. F: Forget-set; R:Retain-set; VQA:Vision Question Answer; QA: Question Answer; VGEN: Vision Generation; GEN: Generation.

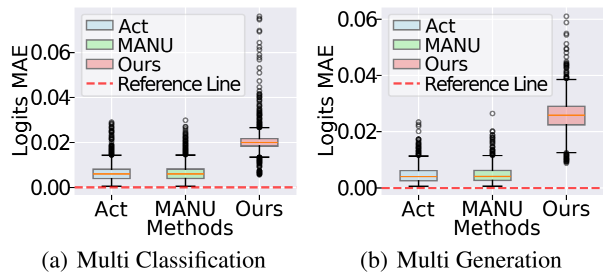
We then measure model performance on the forget and retain sets as a proxy for general knowledge retention. Notice that higher accuracy implies greater representational capacity. Experiments are conducted on MLLMU-Bench using Qwen2.5-VL with a 5% forget ratio. Results for generation tasks are shown in Fig. 4. From the results, we observe that when only a small number of neurons are retained, both strategies yield low ROUGE-L scores. However, performance under the path-based strategy begins to improve significantly after the top-$k$ exceeds $2^5$, peaking around $2^9$. In contrast, the point-wise strategy lags behind and only approaches the performance of the path-based method near $2^{13}$. These findings suggest that neuron paths capture richer and more functionally critical information, and are thus more effective in preserving model performance.




(a) Multi Input (Forget) (b) Multi Input (Retain)

(c) Text Input (Forget) (d) Text Input (Retain)

Figure 4: Performance comparison on generation tasks between influential neuron paths and point-wise influential neurons under varying top-$k$ neuron selections.




(a) Multi Classification (b) Multi Generation

Figure 5: Relative MAE of predicted logit probabilities for ground-truth labels after pruning neurons selected by different methods.

Moreover, we analyze the deviations in predicted probabilities for ground-truth classes on MLLMU's multimodal classification and generation tasks after neuron pruning by Qwen2.5-VL. Specifically, we prune the top-5 neurons located by Activation-based, MANU, and `MIP-Editor`, and compute the MAE of the model's logits before and after pruning. As shown in Fig. 5, our method causes larger shifts in predicted logits compared to the other two approaches, indicating that the neurons selected by `MIP-Editor` play a more critical role in model inference.

## 5.5 Ablation Studies and Variants

We conduct ablation studies on MLLMU-Bench using Qwen2.5-VL with a 5% forget ratio to evaluate the contribution of each component in `MIP-Editor`. (1) Disabling modality coordination by using only textual (ours (*IGI*)) or visual (ours (*IFI*)) paths substantially weakens forgetting effectiveness (e.g., 36.00% FVQA and 49.60% FQA), confirming the necessity of dual-path localization for modality-consistent forgetting. (2) Replacing inter-layer attribution

| Task<br>Metric | FVQA<br>Acc(↓) | RVQA<br>Acc(↑) | FVGEN<br>Rouge(↓) | RVGEN<br>Rouge(↑) | FQA<br>Acc(↓) | RQA<br>Acc(↑) |
|---|---|---|---|---|---|---|
| Ours | 4.80% | 58.19% | 0.0997 | 0.4195 | 9.26% | 36.31% |
| Ours (*IGI*) | 36.00% | 31.60% | 0.4045 | 0.4090 | 41.60% | 45.99% |
| Ours (*IFI*) | 32.00% | 33.46% | 0.3746 | 0.4313 | 49.60% | 47.55% |
| Ours-*Path* | 2.40% | 2.11% | 0.0334 | 0.0479 | 2.40% | 2.15% |
| Ours-*Edit* | 43.60% | 46.00% | 0.4035 | 0.4675 | 42.80% | 52.08% |
| Ours-*RMisU* | 46.40% | 42.23% | 0.3594 | 0.3403 | 34.40% | 31.62% |
| *RMisU* | 8.00% | 14.65% | 0.2667 | 0.2949 | 12.00% | 10.99% |

Table 2: Ablation studies and variants of `MIP-Editor` on MLLMU-Bench with 5% forget ratio by Qwen2.5-VL. F: Forget-set; R:Retain-set; VQA:Vision Question Answer; QA: Question Answer; VGEN: Vision Generation.

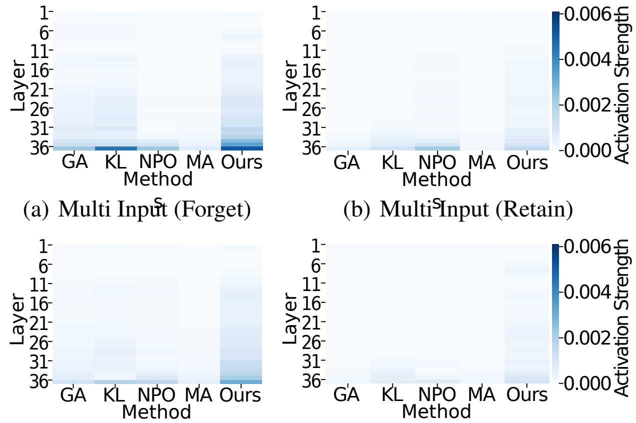
with a simple activation residual score (ours-*Path*) achieves low forget accuracy but severely degrades retain performance (e.g., 2.11% RVQA), showing that point-wise locating disrupts general knowledge. (3) Omitting RMisU editing (ours-*Edit*) or replacing it with standard fine-tuning (ours-*RMisU*) leads to ineffective forgetting and weak retention, demonstrating the limitations of pruning directly and the importance RMisU editing. (4) Applying RMisU to the full model without pruning (*RMisU*) yields moderate forgetting but fails to preserve utility (e.g., 14.65% RVQA), validating the advantage of selective neuron editing.

### 5.6 Visualization

We visualize activation residuals across layers using heatmaps to assess the forgetting and retention behavior of different unlearning methods. Specifically, we input both forget-set and retain-set samples into the unlearned MLLMs and record activation values at each FFN layer. These are compared with the vanilla model's activations, and absolute residuals are used to generate the heatmaps. Darker colors indicate greater deviation from the original model (stronger forgetting), while lighter colors reflect better retention. Experiments are conducted on Qwen2.5-VL using MLLMU-Bench (5% forget ratio) across generation tasks. Results for generation are shown in Fig. 6. As shown in Fig. 6(a) and Fig. 6(c), baseline methods yield consistently shallow color intensities, especially under textual inputs, suggesting limited forgetting. Moreover, similar intensities across forget and retain sets indicate poor separation of specific and general knowledge. In contrast, `MIP-Editor` exhibits clear modality-aware behavior: deeper residuals on the forget-set and lighter residuals on the retain-set, particularly under textual inputs, demonstrating effective cross-modal unlearning with minimal performance degradation.

### 5.7 Specific Information Separability

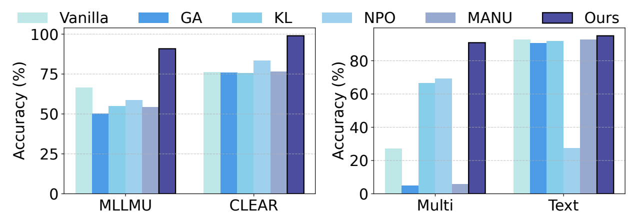
To assess the effectiveness of unlearning methods in separating specific information from general knowledge in MLLMs, we train an MLP-based binary classifier using the output logits of the post-unlearning model. Experiments on MLLMU-Bench and CLEAR with Qwen2.5-VL evaluate two settings: (1) classification over the full fine-tuning set and (2) classification on CLEAR's generation tasks (Multi and Text).




(a) Multi Input (Forget) (b) Multi Input (Retain)

(c) Text Input (Forget) (d) Text Input (Retain)

Figure 6: Layer-wise visualization of knowledge retention in the language FFN of MLLMs across forget and retain sets on MLLMU-Bench. GA: Grad_Diff; Ours: `MIP-Editor`.




(a) Fullset of MLLMU and CLEAR (b) Generation task of CLEAR

Figure 7: Classification of specific vs. general data using Qwen2.5-VL, including (a) full-set classification on MLLMU-Bench and CLEAR, and (b) generation-task classification on CLEAR (Multi and Text).

As shown in Fig. 7, `MIP-Editor` consistently achieves the highest classification accuracy, exceeding 85% across datasets and input types, indicating clearer behavioral separation between specific and general inputs. In contrast, GA, KL, NPO, and MANU perform near random (around 50%) on MLLMU-Bench, showing limited separation capability. On CLEAR's text generation tasks, where questions and answers lack visual modality, the performance of the classification becomes harder to distinguish. Nonetheless, `MIP-Editor` still outperforms all baselines.

## 6 Conclusion

In this paper, we address machine unlearning in Multimodal Large Language Models (MLLMs), highlighting key limitations of existing methods, such as cross-modal inconsistency and general performance degradation. To tackle these issues, we propose `MIP-Editor`, a multimodal pathway-editor that identifies influential neuron paths in each modality and applies path-aware editing through representation misdirection. Experiments demonstrate that `MIP-Editor` achieves effective unlearning of forgetting knowledge while preserving general utility. This work offers a principled framework for fine-grained knowledge removal in MLLMs.

## Acknowledgements

This work was supported in part by Hong Kong RGC Theme-based Research Scheme (TRS) under Grant T43-513/23-N, in part by the NSFC and Hong Kong RGC Collaborative Research Scheme under Grant 62321166652, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2025A1515011996, and in part by the Fundamental Research Funds for the Central University under Grant CXTD202406. We also thank the grant from NVIDIA and utilized NVIDIA A100 GPUs through the NVIDIA Academic Grants Program.

## References

Biesialska, M. M.; Biesialska, K.; and Ruiz Costa-Jussà, M. 2020. Continual lifelong learning in natural language processing: a survey. In *COLING 2020, The 28th International Conference on Computational Linguistics: December 8-13, 2020, Barcelona, Spain (online): proceedings of the conference*, 6523–6541. Association for Computational Linguistics.

Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; and Cucchiara, R. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics ACL 2024*, 13590–13618.

Cao, B.; Tang, Q.; Lin, H.; Jiang, S.; Dong, B.; Han, X.; Chen, J.; Wang, T.; and Sun, L. 2024. Retentive or Forgetful? Diving into the Knowledge Memorizing Mechanism of Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14016–14036.

Chen, J.; Deng, Z.; Zheng, K.; Yan, Y.; Liu, S.; Wu, P.; Jiang, P.; Liu, J.; and Hu, X. 2025. SAFEERASER: Enhancing Safety in Multimodal Large Language Models through Multimodal Machine Unlearning. *arXiv preprint arXiv:2502.12520*.

Chen, Y.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2024. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 17817–17825.

Cheng, J.; and Amiri, H. 2024. Multidelete for multimodal machine unlearning. In *European Conference on Computer Vision*, 165–184. Springer.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18: 2345–2354.

Dai, D.; Dong, L.; Hao, Y.; Sui, Z.; Chang, B.; and Wei, F. 2022. Knowledge Neurons in Pretrained Transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8493–8502.

Dang, H.-T.; Pham, T.; Thanh-Tung, H.; and Inoue, N. 2025. On Effects of Steering Latent Representation for Large Language Model Unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23733–23742.

Dontsov, A.; Korzh, D.; Zhavoronkin, A.; Mikheev, B.; Bobkov, D.; Alanov, A.; Rogov, O. Y.; Oseledets, I.; and Tutubalina, E. 2024. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*.

Fan, C.; Liu, J.; Lin, L.; Jia, J.; Zhang, R.; Mei, S.; and Liu, S. 2024. Simplicity Prevails: Rethinking Negative Preference Optimization for LLM Unlearning. In *Neurips Safe Generative AI Workshop 2024*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2426–2436.

Hase, P.; Bansal, M.; Kim, B.; and Ghandeharioun, A. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36: 17643–17668.

Hu, X.; Li, D.; Hu, B.; Zheng, Z.; Liu, Z.; and Zhang, M. 2024. Separate the wheat from the chaff: Model deficiency unlearning via parameter-efficient module operation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18252–18260.

Huo, J.; Yan, Y.; Zheng, X.; Lyu, Y.; Zou, X.; Wei, Z.; and Hu, X. 2025. MMUNLEARNER: Reformulating Multimodal Machine Unlearning in the Era of Multimodal Large Language Models.

Jia, J.; Liu, J.; Ram, P.; Yao, Y.; Liu, G.; Liu, Y.; Sharma, P.; and Liu, S. 2023. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36: 51584–51605.

Lamparth, M.; and Reuel, A. 2024. Analyzing and editing inner mechanisms of backdoored language models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2362–2373.

Li, H.; Deng, G.; Liu, Y.; Wang, K.; Li, Y.; Zhang, T.; Liu, Y.; Xu, G.; Xu, G.; and Wang, H. 2024a. Digger: Detecting Copyright Content Mis-usage in Large Language Model Training. *CoRR*.

Li, J.; Wei, Q.; Zhang, C.; Qi, G.; Du, M.; Chen, Y.; Bi, S.; and Liu, F. 2024b. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 35414–35453.

Li, N.; Pan, A.; Gopal, A.; Yue, S.; Berrios, D.; Gatti, A.; Li, J. D.; Dombrowski, A.-K.; Goel, S.; Mukobi, G.; et al. 2024c. The WMDP benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, 28525–28550.

Li, Z.; Zhang, N.; Yao, Y.; Wang, M.; Chen, X.; and Chen, H. 2023. Unveiling the Pitfalls of Knowledge Editing for Large Language Models. *CoRR*.

Liu, B.; Liu, Q.; and Stone, P. 2022. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, 243–254. PMLR.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Z.; Dou, G.; Jia, M.; Tan, Z.; Zeng, Q.; Yuan, Y.; and Jiang, M. 2025a. Protecting Privacy in Multimodal Large Language Models with MLLMU-Bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4105–4135.

Liu, Z.; Dou, G.; Yuan, X.; Zhang, C.; Tan, Z.; and Jiang, M. 2025b. Modality-Aware Neuron Pruning for Unlearning in Multimodal Large Language Models. *Proceedings of the 63nd Annual Meeting of the Association for Computational Linguistics*.

Lu, K.; Wang, Z.; Mardziel, P.; and Datta, A. 2021. Influence patterns for explaining information flow in bert. *Advances in Neural Information Processing Systems*, 34: 4461–4474.

Łucki, J.; Wei, B.; Huang, Y.; Henderson, P.; Tramèr, F.; and Rando, J. 2025. An Adversarial Perspective on Machine Unlearning for AI Safety. *Transactions on Machine Learning Research*.

Nguyen, Q. P.; Low, B. K. H.; and Jaillet, P. 2020. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33: 16025–16036.

Pan, H.; Cao, Y.; Wang, X.; Yang, X.; and Wang, M. 2024. Finding and Editing Multi-Modal Neurons in Pre-Trained Transformers. In *Findings of the Association for Computational Linguistics ACL 2024*, 1012–1037. Association for Computational Linguistics.

Pi, R.; Han, T.; Zhang, J.; Xie, Y.; Pan, R.; Lian, Q.; Dong, H.; Zhang, J.; and Zhang, T. 2024. MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 16012–16027.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Sato, Y.; and Takagi, T. 2025. Identifying Multi-modal Knowledge Neurons in Pretrained Transformers via Two-stage Filtering. *arXiv preprint arXiv:2503.22941*.

Si, N.; Zhang, H.; Chang, H.; Zhang, W.; Qu, D.; and Zhang, W. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *arXiv preprint arXiv:2311.15766*.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 303–319. IEEE.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Liu, Y.; Shi, Y.; Li, C.; Pang, A.; Yang, S.; Yu, J.; and Ren, K. 2025. Discovering Influential Neuron Path in Vision Transformers. In *The Thirteenth International Conference on Learning Representations*.

Wu, G.; Hashemi, M.; and Srinivasa, C. 2022. Puma: Performance unchanged model augmentation for training data removal. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 8675–8682.

Wu, X.; Li, J.; Xu, M.; Dong, W.; Wu, S.; Bian, C.; and Xiong, D. 2023. DEPN: Detecting and Editing Privacy Neurons in Pretrained Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2875–2886.

Xing, S.; Zhao, F.; Wu, Z.; An, T.; Chen, W.; Li, C.; Zhang, J.; and Dai, X. 2024. EFUF: Efficient Fine-Grained Unlearning Framework for Mitigating Hallucinations in Multimodal Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1167–1181.

Yang, T.; Dai, L.; Liu, Z.; Wang, X.; Jiang, M.; Tian, Y.; and Zhang, X. 2024. CLIPErase: Efficient Unlearning of Visual-Textual Associations in CLIP. *arXiv preprint arXiv:2410.23330*.

Yao, Y.; Wang, P.; Tian, B.; Cheng, S.; Li, Z.; Deng, S.; Chen, H.; and Zhang, N. 2023. Editing Large Language Models: Problems, Methods, and Opportunities. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhang, R.; Lin, L.; Bai, Y.; and Mei, S. 2024. Negative Preference Optimization: From Catastrophic Collapse to Effective Unlearning. In *First Conference on Language Modeling*.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.