MUSE: ADAPTIVE INPUT-CONDITIONED SYSTEM-PROMPT EMBEDDINGS FOR MULTIMODAL LARGE LANGUAGE MODEL UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) may retain sensitive or undesirable knowledge, motivating effective multimodal unlearning. Existing parameter-level methods often incur substantial training and storage overhead. We propose **MUSE**, a lightweight and adaptive framework that reformulates multimodal unlearning as system-prompt embedding intervention over a frozen MLLM. MUSE trains a **Multimodal Embedding Adapter (MEA)** to generate system-prompt embeddings conditioned on each image–question pair, enabling selective knowledge suppression or preservation according to the multimodal context. To explicitly control the forget–retain boundary, MUSE further introduces an asymmetric dual-space objective that jointly constrains prompt embeddings and internal representations. Experiments on MLLMU-Bench demonstrate strong forgetting, favorable forget–retain trade-offs, and robustness to prompt attacks. Moreover, MUSE achieves up to 4.9× faster training and 74.2% lower GPU memory usage than the strongest baseline, demonstrating effective and efficient multimodal unlearning.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have emerged as powerful foundation models with strong visual understanding and cross-modal reasoning capabilities across diverse applications (Wu et al., 2023; Bai et al., 2025; Zhu et al., 2026). However, they are typically pretrained on large-scale web data that may contain private, copyrighted, or sensitive information, raising concerns about privacy leakage and security risks (Mantelero, 2013; Liu et al., 2024b).To address these concerns, machine unlearning provides a promising solution by selectively removing designated knowledge from deployed models while preserving their general capabilities and downstream utility, avoiding costly retraining from scratch (Sarwar et al., 2026; Liu et al., 2025a).

Existing MLLM unlearning methods mainly suppress target knowledge by modifying model parameters (Thudi et al., 2022; Liu et al., 2022; 2024a; Zhang et al., 2024; Huo et al., 2025; Liu et al., 2025b; Guang et al., 2026). Such parameter-level intervention introduces substantial training and storage overhead, while adapting the unlearning scope across deployment scenarios may require repeated optimization or multiple checkpoints, limiting flexibility and scalability. Recent studies have explored prompt-based unlearning as a lightweight alternative for LLMs (Wang et al., 2026), yet adaptive prompt intervention remains underexplored for multimodal unlearning. Unlike text-only settings, multimodal intervention must jointly consider visual and textual contexts, making static or input-agnostic prompts insufficient for fine-grained control. More fundamentally, multimodal unlearning requires selective, input-dependent intervention: forget-related inputs should trigger knowledge suppression, whereas retain inputs should preserve normal behavior. Due to the entangled nature of multimodal knowledge, overly aggressive forgetting degrades retained capabilities, whereas insufficient intervention leaves target knowledge intact. These challenges motivate us to rethink multimodal unlearning beyond parameter modification:

Can we transform the unlearning process into a lightweight and reversible input-space intervention to control the knowledge access behavior of frozen MLLMs?

To address these challenges, we propose **Multimodal Unlearning via Adaptive System-prompt Embeddings (MUSE)**, which performs unlearning through system-prompt embedding intervention

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

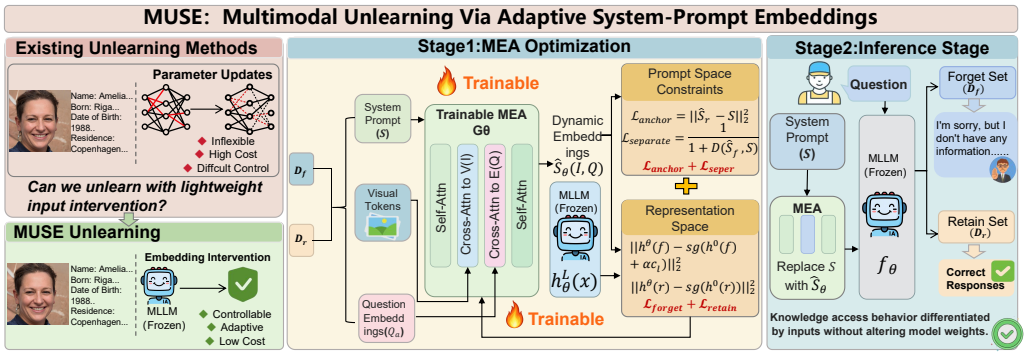


Figure 1: Overview of MUSE, consisting of three components: comparison with existing parameter-level unlearning methods, MEA optimization with dual-space guidance, and inference-time controllable knowledge intervention.

while keeping the backbone MLLM frozen. MUSE trains a lightweight **Multimodal Embedding Adapter (MEA)** to generate input-conditioned prompt embeddings from each image–question pair, enabling selective forgetting or preservation according to the multimodal context. During training, an asymmetric dual-space objective jointly constrains system-prompt embeddings and internal representations to guide MEA in learning the forget–retain boundary. At inference, MEA directly generates adaptive system-prompt embeddings to achieve effective forgetting while preserving model utility. The contributions of this paper are summarized as follows:

(i) A new embedding-space paradigm for multimodal unlearning. MUSE reformulates multimodal unlearning as adaptive system-prompt embedding intervention beyond parameter modification. It introduces an external **Multimodal Embedding Adapter (MEA)** that generates input-conditioned system-prompt embeddings with a frozen MLLM backbone, improving flexibility while reducing updating, storage, and deployment costs.

(ii) Multimodal-conditioned control with dual-space training guidance. MUSE leverages prompt-space constraints and representation-level steering to supervise MEA during training, enabling selective knowledge intervention through dynamic system-prompt embeddings while preserving model utility and generation quality.

(iii) Extensive experimental validation. Experiments across multiple MLLMs and benchmarks demonstrate strong forgetting performance, utility preservation, and robustness. On MLLMU-Bench, MUSE reduces Forget Set classification accuracy by 34.09 percentage points on average across MLLMs, while achieving up to 4.9× faster training and 74.2% lower GPU memory usage than the strongest baseline.

2 Related Work

Multimodal Unlearning. Large models may retain sensitive or undesirable information from large-scale training data, making machine unlearning increasingly important. SIU (Li et al., 2024) is among the first works exploring multimodal unlearning, focusing on removing specific visual concepts. However, it relies on complex fine-tuning data. Most existing multimodal unlearning methods achieve forgetting through parameter-level modification, including gradient ascent (Thudi et al., 2022), retain-set constrained optimization (Liu et al., 2022; Maini et al., 2024), selective parameter updating (Huo et al., 2025), and preference optimization (Zhang et al., 2024). Recent work ASRU (Guang et al., 2026) further improves the forgetting-retention boundary via reinforcement learning. Nevertheless, these methods require modifying model parameters, which is costly and difficult to control due to the entangled knowledge stored in MLLM parameters, potentially causing unintended forgetting and utility degradation.

Prompt Engineering and Activation Steering. Prompt-based methods enable efficient adaptation of large language models without modifying backbone parameters. Recent studies explore learnable and dynamic prompts through lightweight external modules for flexible and reversible be-

havior control (Wang et al., 2026; Chen et al., 2025; Zhou et al., 2024; Lester et al., 2021; Sharma et al., 2026). Activation steering identifies concept-related directions in representation space and modifies hidden states to induce desired behaviors, typically via contrastive analysis or probing (Jorgensen et al., 2023; Rimsky et al., 2024; Yunfan et al., 2025; Sterz et al., 2025). Existing approaches mainly perform inference-time representation intervention (Ding et al., 2025; Sheng et al., 2025), whereas MUSE incorporates representation steering into MEA training, enabling dynamic system-prompt embeddings to induce desired representation shifts without modifying hidden states during inference.

3 Problem Formulation

Modern instruction-tuned MLLMs typically pre-load a system prompt with multimodal user input to specify high-level behavioral instructions. Given an image I and a question Q, we consider a frozen MLLM $\mathcal{M}_\phi$ with pretrained parameters ϕ. The model is initialized with an original system prompt $S = (s_1, \ldots, s_{L_p})$, whose tokens are mapped into continuous representations through the frozen embedding layer $\mathbf{E}$. The corresponding system-prompt embedding is represented as $\mathbf{S} = \mathbf{E}[S] = \mathbf{E}[s_1] \oplus \cdots \oplus \mathbf{E}[s_{L_p}] \in \mathbb{R}^{L_p \times d}$, where d is the embedding dimension and $\oplus$ denotes the sequence concatenation. The model prediction is therefore conditioned on the system prompt, visual tokens, and textual tokens as $\mathcal{M}_\phi(\mathbf{S}, V(I), E(Q))$, where $V(I)$ and $E(Q)$ denote the visual representations extracted from the image and the token embeddings of the question, respectively.

We consider two sets of multimodal inputs: a forget set $\mathcal{D}_f$, containing knowledge to be removed, and a retain set $\mathcal{D}_r$, containing knowledge and capabilities that should remain unchanged. The objective is to learn an external intervention mechanism that suppresses designated knowledge for inputs in $\mathcal{D}_f$ while preserving the original model behavior for inputs in $\mathcal{D}_r$, without modifying the pretrained parameters ϕ of the underlying MLLM.

4 Method

4.1 Overview

System prompts provide an external interface for controlling MLLM behaviors without modifying model parameters. Recent studies demonstrate that system prompts can be dynamically adapted to input contexts (Sharma et al., 2026), motivating us to explore adaptive system-prompt embeddings for reversible multimodal unlearning.

We hypothesize that different multimodal inputs require different interventions: forget inputs should induce knowledge-suppressing prompts, while retain inputs should preserve the original model behavior. Formally, given a frozen MLLM $\mathcal{M}_\phi$ and an original system prompt S, an adaptive prompt $\widehat{S}(I, Q)$ is expected to satisfy:

$$\begin{aligned} \mathcal{M}_\phi(\widehat{S}(I_f, Q_f), I_f, Q_f) &\to \mathcal{U}_f, && (I_f, Q_f) \in \mathcal{D}_f, \\ \mathcal{M}_\phi(\widehat{S}(I_r, Q_r), I_r, Q_r) &\approx \mathcal{M}_\phi(S, I_r, Q_r), && (I_r, Q_r) \in \mathcal{D}_r, \end{aligned} \tag{1}$$

where $\mathcal{U}_f$ denotes the desired unlearned behavior. Optimizing textual prompts is challenging due to the discrete nature of language space. Therefore, MUSE performs an adaptive intervention in the continuous system-prompt embedding space. Specifically, MEA G_θ generates an input-conditioned system-prompt embedding:

$$\widehat{\mathbf{S}}_\theta(I, Q) = G_\theta(\mathbf{S}; V(I), E(Q)), \tag{2}$$

where only the MEA parameters θ are optimized while the MLLM parameters ϕ remain frozen. The generated embedding replaces the original system-prompt embedding and guides the frozen MLLM toward forgetting or retaining knowledge according to the input context.

4.2 Multimodal-Conditioned Prompt Embedding Adapter(MEA)

MEA aims to generate adaptive system-prompt embeddings that retain the original prompt context while incorporating input-specific multimodal information for knowledge control. We implement

G_θ as a Transformer-based adapter that progressively updates the original system-prompt embedding through interactions with visual and textual contexts.

Given an image-question pair (I, Q), MEA constructs the multimodal context $\mathbf{C} = V(I) \oplus E(Q)$, where $V(I)$ and $E(Q)$ denote the visual tokens and question token embeddings, respectively. The original system-prompt embedding initializes the adapter, i.e., $\widehat{\mathbf{S}}^{(0)} = \mathbf{S}$. Each Transformer block then updates the prompt representation through self-attention, multimodal cross-attention, and a feed-forward network:

$$\begin{aligned} \mathbf{A}^{(l)} &= \widehat{\mathbf{S}}^{(l-1)} + \mathrm{SelfAttn}(\mathrm{LN}(\widehat{\mathbf{S}}^{(l-1)})), \\ \mathbf{B}^{(l)} &= \mathbf{A}^{(l)} + \mathrm{CrossAttn}(\mathrm{LN}(\mathbf{A}^{(l)}), \mathbf{C}, \mathbf{C}), \\ \widehat{\mathbf{S}}^{(l)} &= \mathbf{B}^{(l)} + \mathrm{FFN}(\mathrm{LN}(\mathbf{B}^{(l)})), \quad l = 1, \ldots, L. \end{aligned} \tag{3}$$

The self-attention module captures dependencies among system-prompt tokens and maintains the semantic structure of the original instruction. The cross-attention module treats the updated system-prompt representation as queries and selectively aggregates multimodal evidence from the input image and question, enabling instance-specific behavior control. After L Transformer blocks, MEA outputs the adaptive system-prompt embedding:

$$\widehat{\mathbf{S}}_\theta(I, Q) = G_\theta(\mathbf{S}; V(I), E(Q)) = \widehat{\mathbf{S}}^{(L)}. \tag{4}$$

The generated embedding replaces the original system-prompt embedding through the input embedding interface, while the visual inputs, user queries, and all parameters of the underlying MLLM remain unchanged.

4.3 Dual-Space Asymmetric Unlearning

MEA is optimized to suppress target knowledge while preserving retained knowledge and general model capabilities. To this end, MUSE adopts a *dual-space asymmetric objective* over the MLLM representation space and system-prompt embedding space. Forget samples are steered toward a no-knowledge reference region with prompts encouraged to deviate from the original system prompt, while retain samples preserve their original representations and remain anchored to the original prompt. Throughout training, the backbone MLLM $\mathcal{M}_\phi$ remains frozen, and only MEA parameters θ are optimized.

Forget Representation Steering. Simply forcing the model to produce different outputs for forget samples does not explicitly define the desired post-unlearning state. We therefore construct a reference representation region corresponding to the absence of the target knowledge and steer forget samples toward this region.

Specifically, we construct a no-knowledge reference set $\mathcal{D}_+$ containing image-question pairs for which the pretrained MLLM has no relevant prior knowledge. For example, in person-related unlearning scenarios, $\mathcal{D}_+$ can be instantiated using unseen synthetic identities paired with corresponding queries[1]. Following the intuition of representation steering (Turner et al., 2024), we derive a data-driven forgetting direction from the difference between the forget samples and the no-knowledge reference samples.

Consider the hidden representation $h_l^0(x)$ of input x at layer l under the original system prompt, the mean representations of the no-knowledge reference set and forget set are computed as

$$\mu_+^l = \frac{1}{|\mathcal{D}_+|} \sum_{x \in \mathcal{D}_+} h_l^0(x), \qquad \mu_f^l = \frac{1}{|\mathcal{D}_f|} \sum_{x \in \mathcal{D}_f} h_l^0(x). \tag{5}$$

Thus, the forgetting direction is defined as

$$\mathbf{c}_l = \mu_+^l - \mu_f^l. \tag{6}$$

Rather than mapping all forget samples to an identical representation, we preserve their instance-specific characteristics and shift each sample along the shared forgetting direction. For a forget sample x_f, the target representation is defined as

$$\widetilde{h}_l(x_f) = h_l^0(x_f) + \alpha \mathbf{c}_l, \tag{7}$$

[1]The unseen images in $\mathcal{D}_+$ are instantiated using synthetic face images from DigiFace1M (Bae et al., 2023).

where α controls the magnitude of the representation shift. The final representation steering objective is formulated as

$$\mathcal{L}_{\text{forget}} = \mathbb{E}_{x_f \sim \mathcal{D}_f} \left[\frac{1}{d} \left\| h_l^\theta(x_f) - \text{sg}\left[h_l^0(x_f) + \alpha \mathbf{c}_l \right] \right\|_2^2 \right], \tag{8}$$

where $h_l^\theta(x_f)$ is the hidden state obtained with the adaptive system-prompt embedding generated by G_θ , and $\text{sg}[\cdot]$ denotes the stop-gradient operation.

This objective steers forget samples away from their original knowledge-associated representations toward a region characterized by the absence of the designated knowledge.

Retain Representation Preservation. Representation steering on forget samples may introduce unintended changes to unrelated knowledge. We therefore preserve the internal representations of retain samples to maintain the original model behavior. Specifically, for a retain sample $x_r \in \mathcal{D}_r$, we minimize

$$\mathcal{L}_{\text{retain}} = \mathbb{E}_{x_r \sim \mathcal{D}_r} \left[\frac{1}{d} \left\| h_l^\theta(x_r) - \text{sg}\left[h_l^0(x_r) \right] \right\|_2^2 \right]. \tag{9}$$

This objective constrains the downstream effect of the adaptive system prompt inside the frozen MLLM, encouraging retain samples to follow representation trajectories close to those of the original model.

Retain Prompt Anchoring. In addition to representation-level preservation, we further regularize the adaptive prompt at the input interface. For retain samples, the prompt embedding generated by G_θ should remain close to the original system-prompt embedding $\mathbf{S}$. Accordingly, we define the prompt anchoring loss as

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_{x_r \sim \mathcal{D}_r} \left[\frac{1}{L_p d} \left\| \widehat{\mathbf{S}}_\theta(I_r, Q_r) - \mathbf{S} \right\|_F^2 \right]. \tag{10}$$

While $\mathcal{L}_{\text{retain}}$ preserves the downstream representations after the intervention, $\mathcal{L}_{\text{anchor}}$ directly regularizes the generated prompt embeddings, providing complementary preservation constraints at different stages of the computation.

Forget Prompt Separation. Different from retain samples, forget samples require a sufficient deviation from the original system prompt to introduce effective intervention. Therefore, we encourage the adaptive prompts generated by G_θ for forget samples to separate from the original prompt representation. We first define the normalized distance between two prompt embeddings as

$$D_{\text{prompt}}(P_1, P_2) = \sqrt{\frac{1}{L_p d} \left\| P_1 - P_2 \right\|_F^2 + \epsilon}, \tag{11}$$

where ϵ is a small constant for numerical stability. The prompt separation objective is formulated as

$$\mathcal{L}_{\text{separate}} = \mathbb{E}_{x_f \sim \mathcal{D}_f} \left[\frac{1}{1 + D_{\text{prompt}}\left(\widehat{\mathbf{S}}_\theta(I_f, Q_f), \mathbf{S} \right)} \right]. \tag{12}$$

Minimizing this objective encourages the generated prompts for forget samples to move away from the original system prompt, providing stronger input-specific intervention for target knowledge suppression.

Overall Objective. Combining the representation-space and prompt-space objectives, the overall training objective is defined as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{forget}} + \mathcal{L}_{\text{retain}} + \lambda_a \mathcal{L}_{\text{anchor}} + \lambda_s \mathcal{L}_{\text{separate}}, \tag{13}$$

where λ_a and λ_s control the contributions of prompt anchoring and prompt separation, respectively.

This dual-space asymmetric optimization enables selective knowledge intervention: forget samples are shifted toward a no-knowledge regime in both prompt and representation spaces, while retain samples remain close to the original frozen MLLM. Consequently, MUSE achieves controllable and reversible unlearning without modifying backbone parameters. The overall workflow is summarized in Appendix A.

5 EXPERIMENT

5.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on MLLMU-Bench (Liu et al., 2025a) using LLaVA-1.5-7B and Qwen-3-VL-8B-Instruct on four NVIDIA A100 80GB GPUs. MLLMU-Bench contains the Forget Set ($\mathcal{D}_f$), Retain Set ($\mathcal{D}_r$), Test Set ($\mathcal{D}_{\text{test}}$), and Real Celebrity Set ($\mathcal{D}_{\text{real}}$). The Forget/Test sets evaluate unlearning effectiveness, while the Retain/Real sets assess utility preservation. We further evaluate MUSE on OFFSIDE (Zheng et al., 2026) to examine generalization, with details provided in Appendix B.1.

Metrics. For MLLMU-Bench Liu et al. (2025a), we follow the official protocol to evaluate unlearning efficacy, generalizability, and utility preservation. Classification and generation tasks are measured by Accuracy (ACC) and ROUGE-L Lin (2004), respectively. Following ASRU Guang et al. (2026), we further use GPT-5.4-mini to evaluate **Contextual Refusal (CR)**, measuring appropriate refusal on forget-related queries, and **Forgetfulness**, measuring leakage of target knowledge. For OFFSIDE (Zheng et al., 2026), detailed metrics and protocols are provided in Appendix B.1.

Baselines. We select seven representative machine unlearning methods as baselines to evaluate the effectiveness of MUSE. Specifically, the considered methods include **GA** (Thudi et al., 2022), **GA_Diff** (Liu et al., 2022), **KL_Min** (Maini et al., 2024), **NPO** (Zhang et al., 2024), **MMUnlearner** (Huo et al., 2025), **MANU** (Liu et al., 2025c), and **ASRU** (Guang et al., 2026). Since some of these methods may cause the overall model performance to gradually degrade during unlearning, we carefully select and report only the results where the model's core capabilities are preserved, thereby ensuring the practical utility of the unlearned models.

Training. We use LLaVA-1.5-7B and Qwen3-VL-8B-Instruct as the base models for unlearning. Supervised fine-tuning (SFT) is performed with LoRA, using a batch size of 4 and a learning rate of 2×10^{-5}. For methods that require access to $\mathcal{D}_{\text{retain}}$, following Zheng et al. (2026), we adopt a balanced forget–retain update strategy and set the ratio of forget to retain samples to 1:3 to improve training stability during unlearning. The main hyperparameter settings of MUSE are summarized in the Appendix Table 6. Additional training settings and OFFSIDE experiments are provided in Appendix B.1.

5.2 MAIN RESULTS

Unlearning Performance. Table 1 presents the main results on MLLMU-Bench under the 5% forgetting setting. MUSE demonstrates strong unlearning performance across both backbone models. On LLaVA-1.5-7B, MUSE reduces the classification accuracy on the Forget Set and Test Set to 22.50 and 24.17, respectively, substantially lower than those of the strongest baselines. For the generation task, MUSE also achieves the lowest scores. A similar trend is observed on Qwen3-VL-8B, where MUSE further reduces the classification accuracy on the Forget Set to 16.00, substantially outperforming ASRU at 30.83, while also achieving lower generation scores on both the Forget Set and Test Set.

Trade-off Between Unlearning and Model Utility. A key challenge in unlearning is balancing forgetting effectiveness and model utility Wang et al. (2025). As shown in Table 1, MUSE substantially reduces Forget Set accuracy while largely preserving Retain performance. On LLaVA-1.5-7B, Retain accuracy remains close to the vanilla model, while Forget accuracy drops by 56.45%. On Qwen3-VL-8B, MUSE achieves even higher Retain accuracy than the vanilla model while reducing Forget accuracy by 70.91%. Although some baselines perform slightly better on individual retain metrics, they generally exhibit much weaker forgetting, indicating a favorable forget–retain trade-off for MUSE. Figure 3 further shows that MUSE maintains strong forgetting and competitive retain quality across 5%, 10%, and 15% forgetting ratios, demonstrating stable performance under increasingly challenging unlearning settings.

Contextual and Dynamic Refusal. Beyond conventional forgetting metrics, we evaluate whether the unlearned model produces appropriate refusals rather than merely incorrect responses. As shown

Table 1: Unlearning performance on MLLMU-Bench (5% forget). ↓ indicates that lower values are preferred, while ↑ indicates that higher values are preferred. The experimental results on the OFFSIDE dataset are provided in Appendix B.2. The best results are highlighted in bold.

Method	Forget Quality ↓				Generation Quality ↑		Retain Quality ↑			
	Class.		Gen.				Class.		Gen.	
	Forget	Test	Forget	Test	CR	Forgetfulness	Retain	Real	Retain	Real
LLaVA-1.5-7B(5%Forget)-VQA										
Vanilla	51.67	41.67	0.580	0.230	–	–	46.34	47.39	0.490	0.220
GA	41.67	41.67	0.474	0.230	0.06	0.06	41.26	42.27	0.422	0.230
GA_diff	47.50	40.83	0.388	0.170	0.62	0.92	43.17	37.21	0.351	0.183
NPO	44.78	40.00	0.533	0.215	0.04	0.06	44.78	43.23	0.432	0.230
KL_Min	49.17	40.83	0.580	0.227	0.06	0.08	**49.09**	41.78	0.426	0.224
MMunlearner	38.33	42.50	0.415	0.181	0.18	0.58	47.74	44.13	0.439	0.227
MANU	48.80	41.60	0.467	0.220	0.51	1.00	42.56	43.86	0.463	0.243
ASRU	34.17	38.40	0.250	0.144	2.12	2.06	41.35	43.63	0.454	0.268
MUSE(ours)	**22.50**	**24.17**	**0.105**	**0.143**	**4.44**	**4.40**	46.17	**45.04**	**0.603**	**0.336**
Qwen3-VL-8B(5%Forget)-VQA										
Vanilla	55.00	62.50	0.600	0.370	–	–	51.59	78.59	0.570	0.430
GA	49.17	55.00	0.549	0.330	0.32	0.34	44.81	73.89	0.469	0.372
GA_diff	49.17	55.00	0.537	0.334	0.68	0.72	47.40	73.11	0.477	0.356
NPO	46.67	56.67	0.538	0.314	0.56	0.58	46.72	71.67	0.489	0.356
KL_Min	51.67	53.33	0.570	0.355	0.32	0.36	50.66	73.5	0.475	0.369
MMunlearner	43.20	52.00	0.576	0.346	0.2	0.24	46.79	71.54	0.494	0.359
MANU	50.00	49.17	0.460	0.237	0.99	0.97	44.52	63.45	0.444	0.341
ASRU	30.83	45.00	0.398	0.277	2.98	3.04	**53.74**	74.64	0.540	0.398
MUSE(ours)	**16.00**	**42.40**	**0.322**	**0.230**	**3.84**	**4.10**	52.68	**75.20**	**0.551**	**0.410**

in Table 1, MUSE achieves an average CR/Forgetfulness score of 4.42 on LLaVA-1.5-7B, improving over ASRU by 2.33 points. On Qwen3-VL-8B, MUSE achieves 3.97, outperforming ASRU by 0.96 points. These results show that MUSE induces coherent, context-aware refusal behavior while reducing leakage of forgotten knowledge.

6 Deep Analysis

6.1 Robustness Evaluation under Adversarial Attacks

Following ASRU Guang et al. (2026), we evaluate the robustness of MUSE under three representative prompt attacks: **Random Prefix**, **Paraphrase**, and **Jailbreak Prompt**. These attacks respectively introduce surface-level perturbations, semantically equivalent reformulations, and adversarial instructions designed to recover forgotten knowledge. As shown in Figure 8, MUSE consistently maintains low generation scores across all attack settings, indicating that its forgetting behavior is robust to both benign query variations and deliberate adversarial prompting. This suggests that the input-conditioned dynamic system-prompt embeddings provide stable control over forgetting across diverse query formulations.

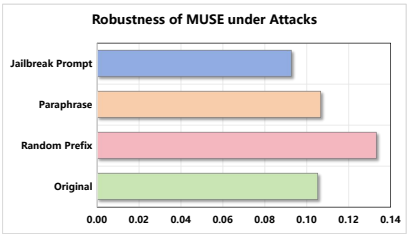


Figure 2: Robustness evaluation of MUSE on LLaVA-1.5-7B.

6.2 Visualization of Hidden State Shift

To further analyze MUSE, we visualize hidden-state distributions at the target layer before and after unlearning (Figure 6b). Before unlearning, Forget and Retain representations largely overlap, indi-

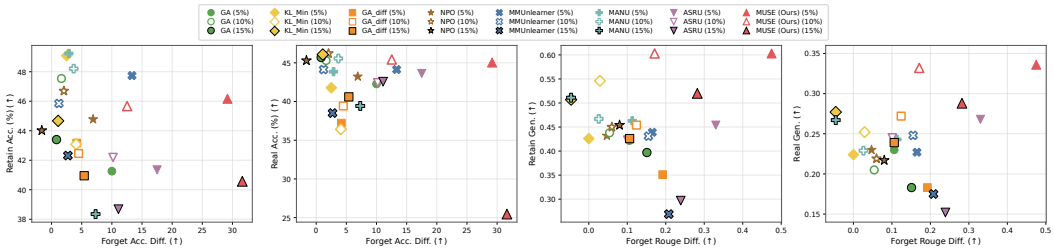


Figure 3: Trade-off between unlearning quality and model utility under different forgetting ratios (5%, 10%, and 15%) on LLaVA-1.5-7B. The x-axis represents the forgetting gain over the vanilla model, and the y-axis shows the retained utility on the Retain (Ret) and Real sets for classification (left) and generation (right) tasks.

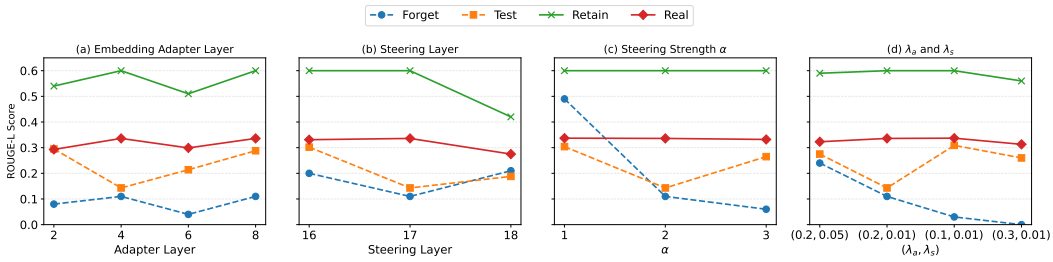


Figure 4: Hyperparameter analysis of MUSE on LLaVA-1.5-7B under the 5% Forget setting of MLLMU-Bench.

cating limited separability in the original representation space. After applying MUSE, Forget representations exhibit a clear directional shift and become more separated from Retain representations. This supports the design of MUSE, where dynamic system-prompt embeddings act as semantic anchors to steer forget representations toward the desired forgetting region while preserving retained knowledge and generation quality.

6.3 Efficiency Study

We evaluate the computational efficiency of MUSE against representative unlearning methods. As shown in Table 2, MUSE keeps the backbone frozen and only optimizes the lightweight MEA, reducing training time to 0.67 hours, about 4.9× faster than ASRU, with 74.2% lower GPU memory usage.

At inference, the additional MEA forward pass introduces only modest overhead. As shown in Figure 5, MUSE requires 0.58 seconds per sample, only 0.30 seconds slower than the fastest baseline MMUnlearner. This small latency increase is acceptable given the substantially stronger forgetting performance and the benefit of avoiding repeated backbone updates and multiple checkpoints. MUSE also remains 0.28 seconds faster than ASRU.

Table 2: Training efficiency and GPU memory usage of representative methods.

models	Training-time(h)	GPU Occupy
GA_diff	2.32	51860
KL_Min	3.50	80280
MMunlearner	3.33	58120
ASRU	3.27	185226
MUSE(ours)	0.67	47730

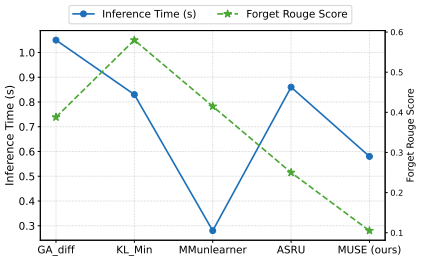
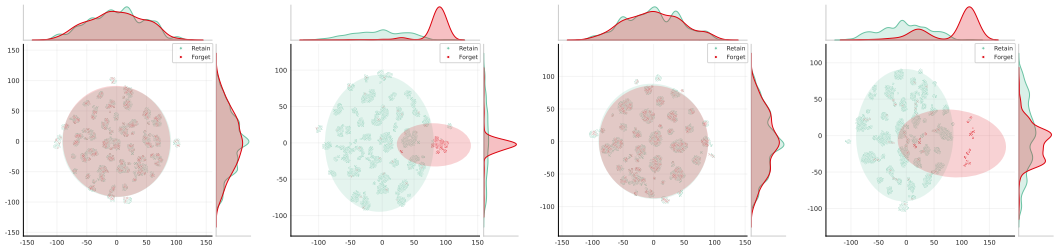


Figure 5: Inference efficiency.

6.4 Hyperparameter Analysis

Figure 4 analyzes the sensitivity of MUSE on LLaVA-1.5-7B. The best overall trade-off is achieved with adapter layer 4, steering layer 17, $\alpha = 2$, and $(\lambda_a, \lambda_s) = (0.2, 0.01)$. In particular, deeper steering degrades retention, while a larger α improves forgetting on the Forget Set but weakens generalization on the Test Set. The selected prompt-space coefficients further balance effective



(a) LLaVA-Vanilla. (b) LLaVA-MUSE. (c) Qwen3-Vanilla. (d) Qwen3-MUSE.

Figure 6: Activation distributions under the 5% forget setting for LLaVA-1.5-7B (a-b) and Qwen3-VL-8B-Instruct (c-d)

Table 3: Ablation study on MUSE components under the 5% Forget setting. Class. denotes the classification task, Gen. denotes the generation task, and Gen. Quality denotes the average generation quality score.

Variants	Forget Quality ↓				Retain Quality ↑				Gen. Quality ↑
	Class.		Gen.		Class.		Gen.		
	Forget	Test	Forget	Test	Retain	Real	Retain	Real	
Vanilla	51.67	41.67	0.580	0.230	46.34	47.39	0.490	0.220	-
MUSE (ours)	22.50	24.17	0.105	0.143	**46.17**	**45.04**	**0.603**	**0.336**	**4.42**
w/o Prompt Constraints	**0.83**	**0.00**	0.131	**0.060**	0.465	0.131	0.126	0.085	3.82
w/o Text Modality	2.50	26.67	**0.017**	0.187	20.08	17.75	0.210	0.160	4.32
w/o Image Modality	28.33	31.67	0.306	0.195	26.47	6.53	0.339	0.233	2.93

forgetting and utility preservation. Hyperparameter analysis for Qwen3-VL-8B-Instruct is provided in Appendix E.1.

7 Ablation Study

Table 3 analyzes the key components of MUSE. Overall, the full model achieves the best balance between forgetting effectiveness and utility preservation. **(i)Prompt-space constraints.** Removing these constraints further improves forgetting but causes severe degradation on the Retain and Real sets, indicating that prompt anchoring and separation are essential for preventing over-forgetting and preserving non-target knowledge. **(ii) Multimodal conditioning.** Using only visual information retains strong forgetting but substantially harms retention, while using only textual information weakens both forgetting and utility. This shows that joint visual-textual conditioning is necessary for precise input-specific intervention and a stable forget–retain trade-off.

8 Conclusion and Limitations

In this work, we propose MUSE, a lightweight and reversible multimodal unlearning framework based on adaptive system-prompt embeddings. MUSE generates input-conditioned prompt embeddings to suppress target knowledge while preserving model utility. Experiments on MLLMU-Bench and OFFSIDE demonstrate effective forgetting, favorable forget–retain trade-offs, robustness, and computational efficiency.MUSE introduces only slight inference latency due to the additional MEA forward pass. Future work will extend MUSE beyond image-text inputs to other modalities, such as audio and video.

AI USE STATEMENT

We used generative AI tools only for language polishing of selected parts of the manuscript and for assisting with a limited portion of code implementation. The manuscript was not generated by AI, and all scientific content, technical arguments, experiments, analyses, and conclusions were developed by the authors. All AI-assisted code was manually reviewed and verified by the authors. No generative AI tools were used to create figures or generate references; all figures were designed and produced by the authors, and all references were collected and verified by the authors. All other aspects of the research were completed by the authors. We reviewed all AI-assisted content and take full responsibility for the final manuscript.

REFERENCES

Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *2023 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Mouxiang Chen, Han Fu, Chenghao Liu, Xiaoyun Joy Wang, Zhuo Li, and Jianling Sun. Build a good human-free prompt tuning: Jointly pre-trained template and verbalizer for few-shot classification. *IEEE Transactions on Knowledge and Data Engineering*, 37(5):2253–2265, 2025.

Chenlu Ding, Jiancan Wu, Leheng Sheng, Fan Zhang, Yancheng Yuan, Xiang Wang, and Xiangnan He. Mllmeraser: Achieving test-time unlearning in multimodal large language models through activation steering. *arXiv preprint arXiv:2510.04217*, 2025.

Jiahui Guang, Haiyan Wang, Yingjie Zhu, Cuiyun Gao, Jing Li, Di Shao, and Zhaoquan Gu. ASRU: Activation steering meets reinforcement unlearning for multimodal large language models. In *Forty-third International Conference on Machine Learning*, 2026.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. MMUnlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7190–7206, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243. URL https://aclanthology.org/2021.emnlp-main.243/.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. *arXiv preprint arXiv:2407.16997*, 2024a.

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

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 4105–4135. Association for Computational Linguistics, 2025a.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 5913–5933. Association for Computational Linguistics, 2025b.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vienna, Austria, July 2025c. Association for Computational Linguistics.

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

Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013. ISSN 2212-473X.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828.

Nobin Sarwar, Shubhashis Roy Dipta, Zheyuan Liu, and Vaidehi Patil. Multimodal Unlearning Across Vision, Language, Video, and Audio: Survey of Methods, Datasets, and Benchmarks. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, July 2026.

Kartik Sharma, Yiqiao Jin, Vineeth Rakesh, Yingtong Dou, Menghai Pan, Mahashweta Das, and Srijan Kumar. Sysformer: Safeguarding frozen large language models with adaptive system prompts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=EbgCEd8gyN.

Leheng Sheng, Changshuo Shen, Weixiang Zhao, Junfeng Fang, Xiaohao Liu, Zhenkai Liang, Xiang Wang, An Zhang, and Tat-Seng Chua. Alphasteer: Learning refusal steering with principled null-space constraint. *arXiv preprint arXiv:2506.07022*, 2025.

Hannah Sterz, Fabian David Schmidt, Goran Glavaš, and Ivan Vulic. Recover the target language: Language steering without sacrificing task performance. *Preprint*, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2024. URL https://arxiv.org/abs/2308.10248.

Qizhou Wang, Jin Peng Zhou, Zhanke Zhou, Saebyeol Shin, Bo Han, and Kilian Q Weinberger. Rethinking llm unlearning objectives: A gradient perspective and go beyond. *arXiv preprint arXiv:2502.19301*, 2025.

Zhaokun Wang, Jinyu Guo, Jingwen Pu, Hongli Pu, Meng Yang, Xunlei Chen, Jie Ou, Wenyi Li, Guangchun Luo, and Wenhong Tian. CAP: Controllable alignment prompting for unlearning in LLMs. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40519–40539, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1882. URL https://aclanthology.org/2026.acl-long.1882/.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256, 2023. doi: 10.1109/BigData59044.2023.10386743.

Xie Yunfan, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Xiangyang Luo, and Liming Dong. Mitigating language confusion through inference-time intervention. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8418–8431, 2025.

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

Hao Zheng, Zirui Pang, Ling Li, Zhijie Deng, Yuhan Pu, Zhaowei Zhu, Xiaobo Xia, and Jiaheng Wei. OFFSIDE: Benchmarking unlearning misinformation in multimodal large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 12602–12620, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1.

Xin Zhou, Dingkang Liang, Wei Xu, Xingkui Zhu, Yihan Xu, Zhikang Zou, and Xiang Bai. Dynamic adapter meets prompt tuning: Parameter-efficient transfer learning for point cloud analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14707–14717. IEEE, 2024.

Yingjie Zhu, Xuefeng Bai, Kehai Chen, Yang Xiang, Youcheng Pan, Xiaoqiang Zhou, and Min Zhang. Decoupling skeleton and flesh: Efficient multimodal table reasoning with disentangled alignment and structure-aware guidance. In *Forty-third International Conference on Machine Learning*, 2026.

A ALGORITHM

Algorithm 1 Workflow of MUSE

Input: Forget set $\mathcal{D}_f$, Retain set $\mathcal{D}_r$, Frozen MLLM $\mathcal{M}_\phi$
Model: Trainable Multimodal Embedding Adapter G_θ, Frozen MLLM $\mathcal{M}_\phi$
▷ Stage 1: Multimodal Embedding Adapter Optimization
Initialize MEA G_θ
for each training epoch **do**
Sample multimodal batches $(I_f, Q_f) \sim \mathcal{D}_f$ and $(I_r, Q_r) \sim \mathcal{D}_r$
Generate adaptive system-prompt embeddings:

$$\hat{S}_f = G_\theta(I_f, Q_f), \quad \hat{S}_r = G_\theta(I_r, Q_r)$$

Obtain hidden representations from frozen MLLM and compute representation steering loss:

$$\begin{cases} \mathcal{L}_{forget} = \mathbb{E}_{x_f \sim \mathcal{D}_f} \left[\left\| h_f^\theta(x_f) - sg(h_f^0(x_f) + \alpha c_l) \right\|_2^2 \right], \\ \mathcal{L}_{retain} = \mathbb{E}_{x_r \sim \mathcal{D}_r} \left[\left\| h_r^\theta(x_r) - sg(h_r^0(x_r)) \right\|_2^2 \right]. \end{cases}$$

Compute prompt anchoring and separation losses:

$$\mathcal{L}_{anchor} = \mathbb{E}_{x_r \sim \mathcal{D}_r} \left[\frac{1}{L_p} \|\hat{S}_r - S\|_F^2 \right]$$

$$\mathcal{L}_{separate} = \mathbb{E}_{x_f \sim \mathcal{D}_f} \left[\frac{1}{1 + D_{prompt}(\hat{S}_f, S)} \right]$$

Update MEA G_θ with:

$$\mathcal{L}_{total} = \mathcal{L}_{forget} + \mathcal{L}_{retain} + \lambda_a \mathcal{L}_{anchor} + \lambda_s \mathcal{L}_{separate}$$

end for
▷ Stage 2: Inference (Deployment)
Input: Image-question pair (I, Q)
Generate dynamic system-prompt embedding:

$$\hat{S} = G_\theta(S; V(I), E(Q))$$

Replace original system-prompt embedding with $\hat{S}$
Generate response:

$$y = \mathcal{M}_\phi(\hat{S}, I, Q)$$

return y

B EXPERIMENTS ON THE OFFSIDE DATASET

B.1 EXPERIMENTAL SETUP

Dataset. We conduct experiments on the OFFSIDE dataset (Zheng et al., 2026) using LLaVA-1.5-7B as the backbone model. All experiments are performed on four NVIDIA A100 80GB GPUs. OFFSIDE is a vision-based unlearning benchmark inspired by real-world visual misinformation scenarios, designed to evaluate the ability of MLLMs to remove specific knowledge while preserving general capabilities. The dataset consists of four subsets: the forget set, retain set, test set, and relearn set. It contains 640 images covering 80 football players from 20 different clubs, with each image paired with 8 shared and 6 private VQA pairs.

Metrics. We adopt the same complete unlearning setting as MLLMU-Bench. Classification tasks are evaluated using ACC, while generation tasks follow the official evaluation protocol and use the average score of Rouge-1, Rouge-2, Rouge-L, and BLEU as the Generation Score. Following

the official setting of OFFSIDE, we additionally report performance on MMBench to evaluate the preservation of general multimodal capabilities after unlearning. Furthermore, to assess the generation quality after unlearning, we employ GPT-5.4-mini as an evaluator and measure two aspects of model behavior: Contextual Refusal (CR) and Forgetfulness (Guang et al., 2026).

Training. We use LLaVA-1.5-7B as the backbone model for unlearning. Supervised fine-tuning (SFT) is performed with LoRA, using a batch size of 4 and a learning rate of 2×10^{-5}. The implementation details of baseline methods are provided in Table 4. For MUSE, we select layer $l = 17$ as the target steering layer and set the guidance strength to $\alpha = 3$ for LLaVA-1.5-7B. In the system-prompt embedding space, we set $\lambda_a = 0.2$ and $\lambda_p = 0.01$ for all experiments.

B.2 MAIN RESULTS

Unlearning Performance. Table 5 presents the unlearning results on the OFFSIDE dataset under the 5% Forget setting. MUSE achieves the strongest forgetting performance among all compared methods, reducing the Forget classification accuracy to 25.38% and obtaining the lowest Generation Score of 0.377. Meanwhile, MUSE achieves the highest Contextual Refusal (CR) and Forgetfulness scores (2.28 and 2.55), indicating that it not only suppresses target knowledge but also induces more appropriate refusal behaviors after unlearning.

Trade-off Between Forgetting and Utility. Despite effective knowledge removal, MUSE preserves the general multimodal capabilities of the model. On the Retain Set, MUSE achieves the best generation performance among all compared methods. Moreover, MUSE obtains the highest MM-test score (58.49), demonstrating that the proposed embedding-space intervention better preserves general multimodal understanding after unlearning. Overall, MUSE achieves a favorable balance between forgetting and utility preservation, highlighting the advantage of dynamic system-prompt embeddings for selective and controllable unlearning without modifying the underlying MLLM parameters.

C BASELINES

C.1 BASELINES

GA. Gradient Ascent (Thudi et al., 2022) applies reverse gradient updates on the forget set $\mathcal{D}_f$ to maximize the loss of target knowledge, thereby reducing the model's ability to recall forgotten information.

$\mathbf{GA}_{\mathrm{Diff}}$. GA_Diff (Liu et al., 2022) extends GA by additionally introducing constraints on the retain set to mitigate the degradation of unrelated knowledge. Its optimization objective is defined as:

$$\mathcal{L} = -\mathcal{L}_{\mathrm{GA}}(\theta; \mathcal{D}_f) + \lambda \mathbb{E}_{(x,y)\sim\mathcal{D}_r} \left[\log \pi_\theta(y|x)\right], \tag{14}$$

where $\mathcal{D}_f$ and $\mathcal{D}_r$ denote the forget and retain sets, respectively, and λ controls the strength of retain-set regularization.

KL_Min. KL-Min (Maini et al., 2024) extends GA by introducing a KL-divergence constraint on the retain set, which preserves the original model distribution on unrelated samples while encouraging the removal of target knowledge.

$$\mathcal{L}_{\mathrm{KL-Min}} = \mathcal{L}_{\mathrm{forget}}(\theta; \mathcal{D}_f) + \lambda \mathbb{E}_{x\sim\mathcal{D}_r} \left[D_{\mathrm{KL}} \left(\pi_{\theta_0}(\cdot|x) \| \pi_\theta(\cdot|x)\right)\right], \tag{15}$$

NPO. NPO further formulates unlearning as a negative preference optimization problem, treating the target knowledge as undesirable responses and optimizing the model to reduce their likelihood. Its objective is defined as:

$$\mathcal{L}_{\mathrm{NPO}} = -\frac{2}{\beta} \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\log\left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right)^\beta\right)\right]. \tag{16}$$

Table 4: Training configurations for different methods on LLaVA-1.5-7B and Qwen3-VL-8B-Instruct.

Model	Method	Epochs/Steps	Batch Size	Learning Rate
LLaVA-1.5-7B	Fine-tuning	4 epochs	4	2×10^{-5}
	Baseline	4 epochs	2	2×10^{-5}
	MANU	50% pruning	4	–
	ASRU	120 steps	2	1×10^{-6}
Qwen3-VL-8B-Instruct	Fine-tuning	4 epochs	4	2×10^{-5}
	Baseline	4 epochs	2	2×10^{-5}
	MANU	50% pruning	4	–
	ASRU	115 steps	2	1×10^{-6}

MMunlearner. MMunlearner (Huo et al., 2025) extends GA_{Diff} by introducing a parameter mask based on weight importance, which enables selective updates of parameters associated with the forget set while preserving unrelated knowledge. Its optimization objective can be formulated as:

$$\mathcal{L}_{\text{MM}} = -\mathbf{m} \odot \mathcal{L}_f(\theta) + \mathcal{L}_r(\theta), \tag{17}$$

where $\mathbf{m}$ denotes the parameter mask that determines which parameters are updated, $\mathcal{L}_f$ and $\mathcal{L}_r$ represent the forget and retain losses, respectively, and $\odot$ denotes element-wise multiplication.

MANU. MANU (Liu et al., 2025b) selects and prunes neurons based on their importance to the target forget data. It first identifies influential cross-modal neurons associated with the forgotten knowledge and then removes these neurons through selective pruning. The pruning operation is defined as:

$$\theta_i' = \begin{cases} 0, & i \in \mathcal{N}, \\ \theta_i, & \text{otherwise.} \end{cases} \tag{18}$$

ASRU. ASRU (Guang et al., 2026) combines activation steering with GRPO optimization to refine the forgetting boundary in the representation space. It first learns a steering vector by minimizing the deviation between the steered activations and their desired targets:

$$\mathcal{L}_{\text{steer}} = \mathbb{E}_{x \sim \mathcal{D}_f} \left[\|h_\theta(x) - h_f(x)\|_2^2 \right] + \mathbb{E}_{x \sim \mathcal{D}_r} \left[\|h_\theta(x) - h_r(x)\|_2^2 \right], \tag{19}$$

where h_f and h_r denote the target representations for forget and retain samples, respectively. Then, GRPO is employed to optimize the policy such that the model learns refusal behaviors for forget samples while maintaining normal responses for retain samples:

$$p_\theta(y|x) \to \begin{cases} y_{\text{refuse}}, & x \in \mathcal{D}_f, \\ y_{\text{answer}}, & x \in \mathcal{D}_r. \end{cases} \tag{20}$$

C.2 Hyperparameter Settings of Baselines

We select the training hyperparameters for each baseline based on their best achievable forgetting performance while avoiding model collapse or degraded generation quality. The selected configurations are summarized in Table 4. For MANU and ASRU, we follow their default batch size settings.

D Hyperparameter Settings of MUSE

We summarize the main hyperparameter settings of MUSE in Table 6. For representation-level guidance, the target layer is fixed to $l = 17$ for both backbone models, while the guidance strength α is set to 2 for LLaVA-1.5-7B and 3 for Qwen3-VL-8B-Instruct. For prompt-space constraints, we use $\lambda_a = 0.2$ for both models, with $\lambda_s = 0.01$ for LLaVA-1.5-7B and $\lambda_s = 0.05$ for Qwen3-VL-8B-Instruct.

Table 5: Unlearning performance on the OFFSIDE dataset under the 5% Forget setting. ↓ indicates that lower values are preferred, while ↑ indicates that higher values are preferred. The best results are highlighted in bold.

Method	Forget Quality ↓		Generation Quality ↑		Retain Quality ↑		Retain Quality (test-set) ↑		MM-test ↑
	Class.	Gen.	CR	Forgetfulness	Class.	Gen.	Class.	Gen.	
LLaVA-1.5-7B (5% Forget)									
Vanilla	68.21	0.753	-	-	70.18	0.745	68.63	0.701	55.42
GA	60.27	0.629	0.49	0.47	62.62	0.631	61.10	0.603	52.71
GA_Diff	67.32	0.645	0.36	0.30	**67.38**	0.637	66.49	0.601	55.49
NPO	64.64	0.617	0.48	0.53	66.99	0.610	64.70	0.590	54.05
KL_Min	33.14	0.607	0.56	0.54	46.35	0.634	47.58	0.604	54.38
MMunlearner	52.36	0.524	0.48	0.51	64.42	0.644	0.59	0.621	56.04
MANU	63.21	0.673	0.38	0.32	62.11	0.618	60.45	0.596	57.17
MUSE (ours)	**25.38**	**0.377**	**2.28**	**2.55**	59.73	**0.654**	**67.05**	**0.630**	**58.49**

Table 6: Hyperparameter settings of MUSE for different backbone models.

Backbone	Target Layer l	Guidance Strength α	λ_a	λ_p
LLaVA-1.5-7B	17	2	0.2	0.01
Qwen3-VL-8B-Instruct	17	3	0.2	0.05

E Additional Experiments

E.1 Hyperparameter Analysis

Figure 7 presents the hyperparameter analysis of MUSE on Qwen3-VL-8B-Instruct. MUSE remains stable across different settings, with the best trade-off achieved at steering layer $l = 17$, steering strength $\alpha = 3$, and $(\lambda_a, \lambda_s) = (0.2, 0.05)$. Specifically, layer 17 achieves the lowest Forget and Test ROUGE-L scores (0.322 and 0.230) while preserving competitive Retain performance, indicating a suitable intervention point for knowledge suppression. The steering strength $\alpha = 3$ further provides the best balance between forgetting and retention, while overly weak or strong steering degrades the overall trade-off. Moreover, the prompt-space coefficients $(\lambda_a, \lambda_s) = (0.2, 0.05)$ achieve effective forgetting with stable retention, demonstrating the importance of balancing prompt anchoring and prompt separation for selective unlearning.

E.2 Robustness Evaluation Under Adversarial Attacks

We further evaluate the robustness of MUSE on Qwen3-VL-8B-Instruct under three prompt attacks: Random Prefix, Paraphrase, and Jailbreak Prompt. As shown in Figure 8, MUSE exhibits stable forgetting behavior under different attack settings. The forgetting scores remain within a narrow range (0.27–0.39), indicating that MUSE is robust to both surface-level perturbations and semantic variations of the input. Although Jailbreak Prompt introduces stronger adversarial instructions and slightly increases the score, MUSE still maintains effective knowledge suppression, demonstrating that the learned dynamic system-prompt embeddings provide a robust forgetting mechanism against diverse input perturbations.

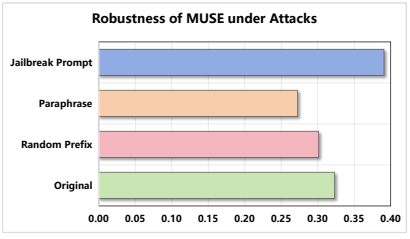


Figure 8: Robustness evaluation of MUSE under different prompt attacks Qwen3- VL-8B-Instruct.

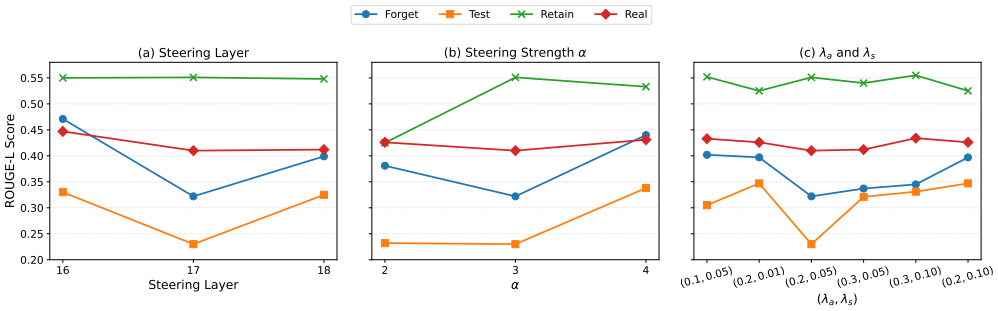


Figure 7: Hyperparameter analysis of MUSE on Qwen3-VL-8B-Instruct.

Table 7: Unlearning performance on MLLMU-Bench under different forgetting ratios. ↓ indicates that lower values are preferred, while ↑ indicates that higher values are preferred. The best results are highlighted in bold.

Method	Forget Quality ↓				Generation Quality ↑		Retain Quality ↑			
	Class.		Gen.		CR	Forgetfulness	Class.		Gen.	
	Forget	Test	Forget	Test			Retain	Real	Retain	Real
LLaVA-1.5-7B (10% Forget)-VQA										
Vanilla	46.53	41.22	0.516	0.225	–	–	47.99	47.39	0.496	0.225
GA	44.89	39.18	0.462	0.205	0.31	0.91	47.54	45.30	0.438	0.205
GA_Diff	42.04	37.55	0.392	0.272	0.59	0.76	42.46	39.43	0.454	0.272
NPO	44.49	38.78	0.456	0.219	0.49	1.05	46.70	**46.21**	0.450	0.219
KL_Min	42.45	42.45	0.487	0.252	0.35	0.22	43.08	36.42	0.546	0.252
MMunlearner	45.31	42.86	0.361	0.248	0.58	1.01	45.85	44.13	0.431	0.248
MANU	42.86	37.55	0.490	0.229	0.29	0.66	**48.21**	45.56	0.467	0.229
ASRU	36.33	38.78	0.415	0.245	0.69	0.95	42.19	42.43	0.423	0.245
MUSE (ours)	**34.00**	**26.00**	**0.345**	**0.332**	**2.26**	**2.96**	45.67	45.43	**0.603**	**0.332**
LLaVA-1.5-7B (15% Forget)-VQA										
Vanilla	45.41	43.78	0.485	0.224	–	–	46.95	47.00	0.500	0.224
GA	44.59	43.24	0.334	0.183	0.81	1.73	43.40	45.69	0.397	0.183
GA_Diff	40.00	39.73	0.379	0.239	0.65	0.67	40.95	40.60	0.426	0.239
NPO	47.03	44.59	0.405	0.217	0.58	1.27	44.02	45.30	0.454	0.217
KL_Min	44.32	42.97	0.531	0.277	0.30	0.38	**44.67**	**46.08**	0.507	0.277
MMunlearner	42.70	42.97	0.277	0.175	0.68	2.21	42.32	38.51	0.269	0.175
MANU	38.11	38.38	0.531	0.267	0.21	0.37	38.35	39.43	0.511	0.267
ASRU	34.32	40.27	0.246	0.152	0.82	2.37	38.68	42.56	0.297	0.152
MUSE (ours)	**13.78**	**34.05**	**0.203**	**0.288**	**3.56**	**3.94**	40.57	25.46	**0.520**	**0.288**

E.3 COMPARISON OF MUSE FOR DIFFERENT HYPERPARAMETERS

E.4 UNLEARNING PERFORMANCE UNDER DIFFERENT FORGETTING RATIOS

We further evaluate MUSE under more challenging forgetting ratios (10% and 15%). As shown in table 7, as the forgetting ratio increases, unlearning becomes more difficult due to the larger amount of target knowledge that needs to be suppressed. Nevertheless, MUSE consistently achieves the strongest forgetting performance, reducing the Forget classification accuracy to 34.00% and 13.78% under the 10% and 15% settings, respectively. Meanwhile, MUSE maintains superior generation-based forgetting performance and achieves the highest refusal-related scores, demonstrating its ability to scale to more aggressive forgetting requirements.

Although stronger forgetting inevitably introduces larger impacts on retained capabilities, MUSE still preserves competitive utility compared with baseline methods. In particular, MUSE achieves the best generation quality on the Retain Set under both settings, showing that the proposed embedding-space intervention can effectively balance knowledge suppression and capability preservation under different forgetting ratios.