KNOWLEDGE REGROWS ACROSS MODALITIES: HIDDEN FAILURE MODE IN MLLM UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have achieved remarkable success in integrating information from various modalities, yet they raise growing concerns about privacy leakage. Machine unlearning is considered a remedy for privacy leakage, but existing multimodal unlearning approaches directly adopt techniques developed for unimodal models. This paper uncovers a hidden failure mode in MLLM unlearning, which we term *Cross-modal Knowledge Regrowth*: textual prompts can restore forgotten visual knowledge after unlearning, even in the absence of images. We identify that this phenomenon is attributed to the reactivation of vision-specific parameters by text-only prompts. Based on these findings, we show that selectively removing these vulnerable weights effectively suppresses cross-modal knowledge regrowth. Extensive experiments demonstrate that our approach reliably preserves unlearning robustness while maintaining model utility. Code is provided in the supplementary material for reproducibility.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have rapidly advanced the ability to process and reason over diverse modalities, enabling a broad range of vision–language applications (Cui et al., 2024; Li et al., 2024b; Tan et al., 2025). However, these models are often trained on large web-scraped datasets that may contain sensitive personal images or private information, which raises serious privacy concerns. In addition, regulations such as the European Union's General Data Protection Regulation (Voigt & Von dem Bussche, 2017) established a "Right to be Forgotten", which requires models to remove the influence of a user's personal data upon request, even after deployment.

One of the most straightforward solutions for removing specific knowledge is to retrain the entire model from scratch after excluding the undesired data. However, this is often infeasible in practice due to the immense computational and data costs involved. To address this challenge, a more scalable approach known as Machine Unlearning (Cao & Yang, 2015; Ginart et al., 2019) has been proposed to remove the influence of targeted data without compromising the model's overall utility. While recent research has advanced machine unlearning techniques for Large Language Models (LLMs) (Thudi et al., 2022; Liu et al., 2022; Nguyen et al., 2020; Zhang et al., 2024; Maini et al., 2024), MLLM unlearning has largely relied on techniques originally developed for unimodal LLMs. However, by aligning additional modalities with language and entangling representations across modalities through shared model parameters, MLLMs introduce new vulnerabilities beyond those in text-only LLMs, creating pathways through which updates in one modality can affect knowledge encoded in another, which makes directly applying LLM-based unlearning methods to MLLMs not straightforward.

However, prior work on MLLM unlearning (Ma et al., 2025; Ding et al., 2025; Dontsov et al., 2025; Kim et al., 2025) has primarily evaluated models in multimodal settings, without verifying whether knowledge was effectively forgotten within each modality. Even studies that considered unimodal evaluations (Liu et al., 2025a;b) did not examine whether forgotten information is completely eliminated across modalities or merely suppressed in a single modality. Moreover, recent work in LLMs (Hsu et al., 2025; Hu et al., 2025a; Siddiqui et al., 2025; Zhang et al., 2025) shows that forgotten knowledge can be restored through relearning, even under small perturbations or minimal fine-tuning. However, to the best of our knowledge, no prior work has investigated relearning in MLLMs, nor examined knowledge restoration across modalities. This gap poses a unique challenge of MLLM unlearning, where knowledge removed in one modality can reemerge through another. In

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.

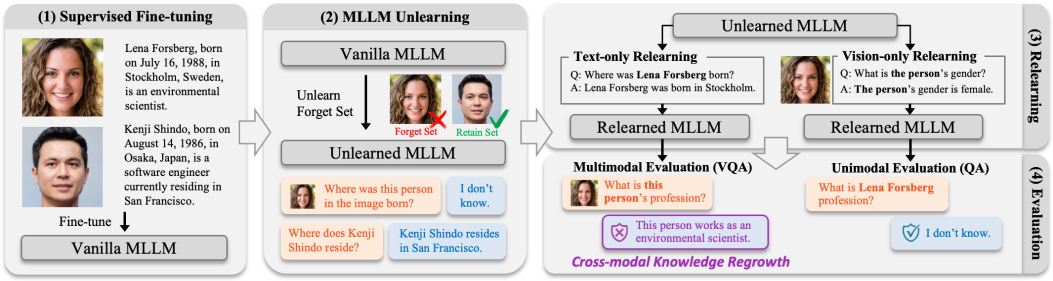
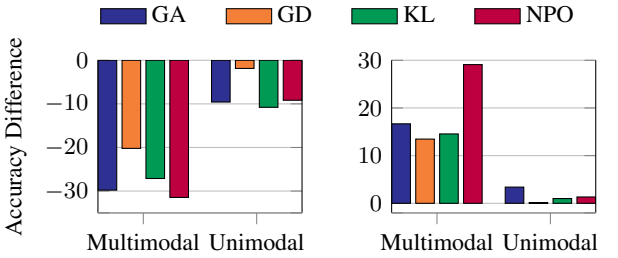


Figure 1: Illustration of the overall MLLM pipeline from fine-tuning to unlearning and relearning. A vanilla MLLM model is first fine-tuned on personal profile data. Then, unlearning is performed to remove knowledge associated with the forget set while preserving information in the retain set. Benign relearning is conducted separately for each modality, and forgotten visual knowledge can be restored through text-only relearning.

this work, we focus on vision-language MLLMs, as text and vision constitute the dominant modality pair in existing MLLM research. Specifically, we identify that *visual knowledge removed during unlearning can be restored through text-only prompts*, whereas textual knowledge is not restored by vision-only prompts.

To verify this issue, as shown in Figure 1, we unlearn the MLLM, LLaVA-1.5-7B (Liu et al., 2023) on the MLLMU-Bench (Liu et al., 2025a) dataset using four representative multimodal unlearning methods, including Gradient Ascent (GA) (Thudi et al., 2022), Gradient Difference (GD) (Liu et al., 2022), KL Minimization (KL) (Nguyen et al., 2020), and Negative Preference Optimization (NPO) (Zhang et al., 2024). Figure 2a shows that unlearning induces substantially larger accuracy degradation under multimodal evaluation than under unimodal evaluation, indicating that unlearning in one modality does not effectively carry over to the other. Unimodal evaluation assesses question answering using linguistic knowledge alone, whereas multimodal evaluation leverages visual knowledge that is aligned and jointly processed with linguistic representations. Following the benign relearning attack protocol proposed for unimodal LLMs (Hu et al., 2025a), we assess how easily forgotten visual knowledge reemerges in multimodal settings by relearning the model using textual prompts. Figure 2b shows that the forgotten visual knowledge, which is reduced to about 32% on average after unlearning, is significantly restored to 78% even when only textual information is provided, a phenomenon we term **Cross-modal Knowledge Regrowth** (formally defined in equation 5). This finding reveals a critical issue with existing MLLM unlearning methods.



(a) Fine-tuning→Unlearning (b) Unlearning→Relearning

Figure 2: Comparison of accuracy difference (%) on forget set after unlearning and text-only relearning under multimodal and unimodal evaluation settings. We refer to the performance increase on forget set under multimodal evaluation after relearning as **Cross-modal Knowledge Regrowth**.

In this work, we systematically characterize this phenomenon across diverse MLLMs, benchmark datasets, and unlearning algorithms. To uncover the mechanism underlying this behavior, we analyze the parameters associated with different modalities in MLLMs using an importance measure that jointly considers weights and input activations, inspired by prior LLM studies (Sun et al., 2024; Wu et al., 2024; Pochinkov & Schoots, 2024). This analysis reveals that a subset of vision-specific weights remains highly sensitive to text-driven updates after unlearning. Inspired by this observation, we selectively remove these vulnerable parameters responsible for cross-modal knowledge regrowth, thereby blocking the pathways through which textual prompts restore visual knowledge. Extensive experiments demonstrate that our method consistently suppresses regrowth across diverse unlearning methods, while maintaining or even improving utility on the retain set, i.e., data to be preserved after unlearning. Our key contributions are summarized as follows:

1. We identify a previously undiscovered failure mode in MLLM unlearning, *Cross-modal Knowledge Regrowth*, in which text-only fine-tuning can restore forgotten visual knowledge even in the absence of images.
2. We verify that this failure mode arises from asymmetric cross-modal interactions and identify the vision-specific parameters that are responsive to text-driven updates after unlearning.
3. We demonstrate that selectively removing vision-specific weights effectively suppresses cross-modal knowledge regrowth while preserving model utility across diverse MLLMs, datasets, and unlearning methods.

A comprehensive review of **Related Work** is provided in Appendix A.

2 PRELIMINARIES

2.1 DEFINITION OF MACHINE UNLEARNING

Let $f_{\mathrm{pre}}(\cdot\,;\theta_{\mathrm{pre}})$ denote a pre-trained MLLM. Given training data D_{train}, fine-tuning f_{pre} on D_{train} produces the vanilla model $f_{\mathrm{vnl}}(\cdot\,;\theta_{\mathrm{vnl}})$. Suppose a designated subset $D_{\mathrm{forget}} \subset D_{\mathrm{train}}$ is identified as the portion of the data whose influence must be removed from θ_{vnl}. We denote the complementary retain set by $D_{\mathrm{retain}} = D_{\mathrm{train}} \setminus D_{\mathrm{forget}}$. The goal of machine unlearning is to construct a model that behaves as if it had never been exposed to D_{forget}. Ideally, the resulting unlearned model $f_{\mathrm{unl}}(\cdot\,;\theta_{\mathrm{unl}})$ should approximate the oracle model which is obtained by fine-tuning f_{pre} solely on D_{retain}.

An unlearning algorithm U operates on the vanilla model and the forget set, optionally leveraging the retain set:

$$f_{\mathrm{unl}} = U(f_{\mathrm{vnl}},\, D_{\mathrm{forget}},\, D_{\mathrm{retain}}) \tag{1}$$

A common formulation casts unlearning as optimizing parameters θ_{unl} on D_{retain} under constraints that eliminate the contribution of D_{forget}:

$$\theta_{\mathrm{unl}} \;\approx\; \arg\min_{\theta} \mathcal{L}(\theta; D_{\mathrm{retain}}) \quad \text{s.t. } \mathrm{Rmv}(\theta; D_{\mathrm{forget}}) \tag{2}$$

where $\mathcal{L}$ is a loss and $\mathrm{Rmv}(\theta; D_{\mathrm{forget}})$ denotes a constraint enforcing the removal of information attributable to D_{forget}.

2.2 BENIGN RELEARNING ATTACKS

Given an unlearned model f_{unl} produced by an unlearning algorithm U, a benign relearning attack probes whether knowledge associated with the forget set D_{forget} can reemerge after additional fine-tuning. Specifically, on θ_{unl}, we perform a benign relearning attack by fine-tuning the model with the relearn set D_{relearn}, yielding

$$f_{\mathrm{rel}}(\cdot\,;\theta_{\mathrm{rel}}) \;=\; \mathrm{FT}(f_{\mathrm{unl}},\, D_{\mathrm{relearn}}) \tag{3}$$

where $\mathrm{FT}(\cdot)$ denotes a standard fine-tuning operator.

In this work, we consider two benign relearning scenarios, where the relearn sets are drawn from D_{forget}: (i) **Vision-only relearning**. In this scenario, D_{relearn} consists of image-included Visual Question Answering (VQA) pairs where the questions are unrelated to any personal identity information. (Examples include questions about visually observable appearance attributes, such as hair color or gender.) (ii) **Text-only relearning**. In this scenario, D_{relearn} contains only textual Question Answering (QA) pairs without any visual input.

2.3 IMPORTANCE ESTIMATION IN LLMS

Estimating which parameters matter most for different modalities requires a metric that reflects how parameters are actually contributing during inference. Prior work shows that activation-informed measures provide a more accurate signal of importance than sole weight magnitudes in LLMs. For example, ReFT (Wu et al., 2024) demonstrates that model activations form meaningful semantic subspaces whose manipulation can substantially alter model behavior. Wanda (Sun et al., 2024) shows that combining weight magnitude with activation norms yields effective importance scores for pruning. Similar activation-based criteria have been used to identify candidate neurons for pruning in machine unlearning (Pochinkov & Schoots, 2024).

Inspired by these insights, we adopt an importance metric that jointly considers weights and input activations to quantify the importance of parameters associated with different modalities in MLLMs. Specifically, consider a linear layer with a weight matrix $\mathbf{W} \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Let $\mathbf{X} \in \mathbb{R}^{(N \times L) \times d_{\text{in}}}$ denote the input activation matrix, where N and L correspond to the batch size and sequence length, respectively. The importance score $S_{i,j}$ for the weight $W_{i,j}$ is defined as:

$$S_{i,j} = |W_{i,j}| \cdot \|X_j\|_2 \tag{4}$$

where $\|X_j\|_2$ denotes the ℓ_2 norm of the j-th input feature aggregated across $N \times L$ different tokens, i indexes the output feature dimension of the linear transformation.

3 Cross-modal Knowledge Regrowth after Machine Unlearning

In this section, we identify that knowledge forgotten during multimodal unlearning can reemerge under unimodal supervision in MLLMs. To this end, we perform two types of benign relearning attacks on the unlearned model—text-only and vision-only—as described in Section 2.2, and compare their effects in restoring forgotten knowledge across modalities. In prior work on unimodal LLMs (Hu et al., 2025a;b; Siddiqui et al., 2025), benign relearning is typically assessed by quantifying regained performance on the forget set after additional fine-tuning on benign data, using task-specific accuracy or keyword recovery metrics. However, unlike LLMs, MLLMs access and evaluate knowledge through multimodal inputs that integrate textual and visual information, which renders unimodal criteria insufficient to fully capture failure modes arising from cross-modal interactions. Hence, we define *Cross-modal Knowledge Regrowth* as the restoration of forget set performance in one modality, relative to the post-unlearning baseline, induced by benign relearning through another modality, which is formulated as follows:

Definition 3.1 (**Cross-modal Knowledge Regrowth**)**.** We define cross-modal knowledge regrowth as the change in forget set performance measured under evaluation setting e after relearning with modality r, denoted as $\Delta\mathcal{M}_{e \leftarrow r}$:

$$\Delta\mathcal{M}_{e \leftarrow r} = \mathcal{M}_e(R_r(f_{\text{unl}})) - \mathcal{M}_e(f_{\text{unl}}) \tag{5}$$

where $\mathcal{M}_e(\cdot)$ denotes forget set performance under evaluation setting e, f_{unl} is the unlearned model, and $R_r(\cdot)$ denotes relearning conducted using modality r. Here, $r \in \{\text{vision}, \text{text}\}$ and $e \in \{\text{multimodal}, \text{unimodal}\}$; cross-modal knowledge regrowth refers to the cases $(r, e) = (\text{vision}, \text{unimodal})$ and $(\text{text}, \text{multimodal})$.

As per the definition, we answer three key questions: (**Q1**) Is cross-modal knowledge regrowth bidirectional or asymmetric across modalities? (**Q2**) Does this phenomenon generalize to out-of-distribution data? (**Q3**) How does scaling relearning data impact cross-modal knowledge regrowth?

3.1 Experimental Setup

We use the four representative unlearning methods, GA, GD, KL, and NPO, to serve as baselines and evaluate their performance on MLLMs. Following the setup of (Ma et al., 2025), we keep both the vision-to-language projector and the language model trainable, with LoRA (Hu et al., 2022) applied to the language model during the fine-tuning, unlearning, and relearning stages. Ablation results in (Ma et al., 2025) show that updating the vision encoder alters the extracted image features, which significantly degrades model utility, whereas jointly updating the projector and the language model achieves a balanced trade-off between forgetting effectiveness and model utility. The details of unlearning methods are described in Appendix B. We also verify that our findings generalize to full-parameter settings in Appendix G.6.

We evaluate on three multimodal unlearning benchmarks, MLLMU-Bench (Liu et al., 2025a), UMU-Bench (Wang et al., 2025), and CLEAR (Dontsov et al., 2025). We conduct experiments using LLaVA-1.5-7B (Liu et al., 2023) on MLLMU-Bench and CLEAR, and Qwen2-VL-7B (Wang et al., 2024) on UMU-Bench. The details of the datasets are provided in Appendix D. All results are reported under a setting where 5% of identities, drawn from the entire training dataset, are assigned to the forget set unless otherwise specified. The performance is measured using two accuracy metrics: VQA accuracy, A_m, which evaluates the model's ability to answer multiple-choice questions based

Table 1: Comparison of VQA and QA classification accuracy (%) after unlearning and relearning on MLLMU-Bench. ↑ represents higher is better, ↓ means lower is better.

Method	Forget Set		Retain Set	
	A_m (↓)	A_u (↓)	A_m (↑)	A_u (↑)
Vanilla (original model)	39.79	44.79	39.49	42.38
GA (unlearned baseline)	10.00	35.21	27.99	37.49
GA (relearned w/ Text)	26.67	38.61	37.89	36.39
GA (relearned w/ Vision)	16.67	28.33	33.32	42.62
GD (unlearned baseline)	19.58	42.92	34.34	40.68
GD (relearned w/ Text)	33.05	43.06	36.87	39.96
GD (relearned w/ Vision)	18.33	34.17	34.08	35.35
KL (unlearned baseline)	12.67	34.00	33.45	37.73
KL (relearned w/ Text)	27.22	35.00	44.14	37.57
KL (relearned w/ Vision)	16.67	30.83	33.02	36.28
NPO (unlearned baseline)	8.34	35.63	24.14	37.60
NPO (relearned w/ Text)	37.43	36.94	36.96	38.20
NPO (relearned w/ Vision)	9.17	31.67	34.16	40.38

on both image and text inputs (multimodal), and QA accuracy, A_u, which assesses its ability to answer multiple-choice questions without any visual context (unimodal). More implementation details, including hyperparameter selections, are in Appendix E.

3.2 Asymmetry of Cross-modal Knowledge Regrowth

To answer **Q1**, we evaluate an unlearned baseline model by independently applying text-only or vision-only relearning and comparing their effects on performance. Table 1 reports VQA and QA classification accuracy on the MLLMU-Bench (Liu et al., 2025a) after unlearning and subsequent relearning under different modality configurations. Across all unlearning methods, we first observe that the average A_m on the forget set drops to approximately 32% of its original performance, indicating that existing unlearning methods can initially suppress visual knowledge to a large extent. The highlighted cells illustrate modality-specific relearning effects. Cells shaded in yellow correspond to text-only relearning, where visual knowledge, measured by A_m, exhibits significant regrowth despite the absence of any visual input during relearning. Notably, A_m is substantially restored, reaching an average of 78% of its original performance.

In contrast, cells shaded in red correspond to vision-only relearning, where A_u measures textual knowledge affected by visual inputs. Unlike the text-only relearning case, vision-only relearning does not restore textual knowledge and instead degrades QA performance. For example, in GA's case, A_u degrades from 35.21% to 28.33%. These asymmetric regrowth trends are consistently observed across all unlearning algorithms provided, indicating that directly applying unlearning approaches designed for unimodal LLMs to MLLMs is fundamentally incompatible, as modality entanglement prevents multimodal knowledge from being cleanly removed and renders it susceptible to reemergence even under text-only prompts. In the remainder of this paper, we use cross-modal knowledge regrowth to refer to the case $\Delta\mathcal{M}_{\text{multimodal}\leftarrow\text{text}}$.

3.3 Generalization of Cross-modal Knowledge Regrowth

To answer **Q2**, we build on the same unlearned models used in Table 1 and examine whether cross-modal knowledge regrowth extends to out-of-distribution data via text-only relearning with 153 real celebrity profiles selected from the CelebA dataset (Liu et al., 2015), which are not used during fine-tuning or unlearning. Following the same question structure as the fictitious profiles, each celebrity profile contains 14 question-answer pairs, comprising seven multimodal and seven unimodal pairs. Table 2 shows that all unlearning methods exhibit noticeable regrowth under in-distribution text-only relearning, where D_{relearn} is derived from D_{forget}. However, this effect is reduced to near-zero or even negative values across all methods in the out-of-distribution setting, where

Table 2: Cross-modal knowledge regrowth (%) under in-distribution and out-of-distribution relearning settings.

D_{relearn}	GA	GD	KL	NPO
In-distribution	16.67	13.47	14.55	29.09
Out-of-distribution	-0.40	-1.98	0.67	0.00

text-only relearning is performed on unseen real celebrity profiles. This suggests that cross-modal knowledge regrowth requires relearning content that is explicitly related to the forgotten information.

3.4 Effect of Scaling Relearning Data

To answer **Q3**, we vary the scale of relearning data D_{relearn} to examine its effect on cross-modal knowledge regrowth. Specifically, we increase the data scale by incrementally adding question–answer pairs per profile to D_{relearn}. Figure 3 illustrates, under the GD text-only relearning setting in Table 1, that the magnitude of regrowth increases monotonically with the scale of relearning data. Notably, the regrowth remains limited at small data scales, while larger relearning sets, containing more detailed information about the same profiles, are associated with more pronounced regrowth. This trend indicates that increased exposure to identity-consistent textual information progressively amplifies cross-modal knowledge regrowth.

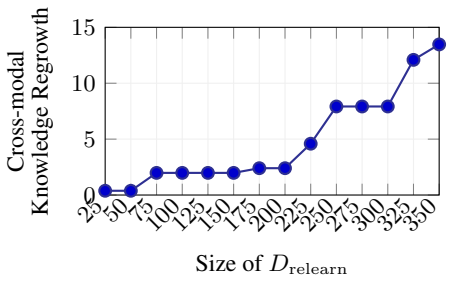


Figure 3: Cross-modal knowledge regrowth (%) as the size of text-only D_{relearn} increases.

4 Mechanism of Cross-modal Knowledge Regrowth

Recent works (Xu et al., 2025; Wei et al., 2025) claim that MLLMs activate different subsets of model weights when encoding visual information versus textual information. Moreover, prior work suggests that changes in internal activations can signal the removal and subsequent reorganization of knowledge within the model (Kim et al., 2026; Liu et al., 2025b). Building on this observation, we identify parameters that are activated under each modality according to the importance score in equation 4. To examine the reactivation of knowledge that was explicitly targeted for removal, we conduct this analysis on samples from the forget set. To understand why cross-modal knowledge regrowth differs across modalities, we analyze which parameter groups are most involved during text-only and vision-only relearning. For this analysis, we first identify the top 10% of weights for each modality by the **importance score** when image or text tokens are respectively provided as input to the unlearned model, and then categorize them into three groups: vision-specific, text-specific, and shared parameters. After relearning, we measure the **importance shift**, defined as the absolute difference between the importance scores of the relearned model and the unlearned model. In Figure 4, we separately show how the top 10% of most-shifted weights are distributed across the three weight groups under text-only and vision-only relearning, focusing on the GD case in Table 1.

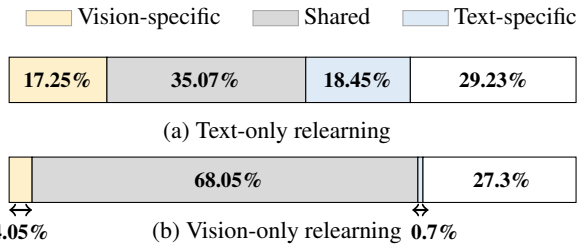


Figure 4: Distribution of the top 10% of weights by importance shift during text-only and vision-only relearning, respectively. The remaining portion corresponds to weights that do not belong to any of the categories.

According to the results, a clear asymmetry is observed between Figure 4a and Figure 4b. During text-only relearning, even though the model receives no images, a substantial portion of the most-shifted weights (17.25%) exclusively belong to vision-specific weights. This indicates that gradients from purely linguistic updates still propagate into vision-specific subspaces, reactivating previously removed visual representations. In contrast, under vision-only relearning, shifts in text-specific weights are negligible (0.7%), showing that visual gradients do not substantially influence textual representations. This explains the asymmetric cross-modal knowledge regrowth patterns observed in Table 1. To verify that this cross-modal parameter coupling is not tied to specific architectural components, we conduct a layer-wise analysis in Table 8 (Appendix G.3). Furthermore, gradient flow analysis in Table 10 (Appendix G.5) shows that vision-specific parameters receive substantial gradient updates during text-only relearning despite the absence of visual input. Moreover, freezing vision-specific parameters during text-only relearning suppresses regrowth (Appendix G.6), providing interventional evidence that these parameters mediate cross-modal knowledge regrowth.

5 PREVENTING CROSS-MODAL KNOWLEDGE REGROWTH

To suppress cross-modal knowledge regrowth, we introduce a modality-specific pruning method that targets vision-specific parameters which remain susceptible to reactivation after unlearning. We term this approach Cross-modal Knowledge Regrowth Pruning (KoReP).

Modality-Specific Importance Computation. Following the importance measure defined in equation 4, we compute modality-specific importance scores for text and vision, respectively. To construct unimodal inputs, we provide only text tokens when computing $\mathbf{X}_t$ and only image tokens when computing $\mathbf{X}_v$, disabling the other modality in each case. All activations are computed on samples from the forget set using the unlearned model. Given a weight matrix $\mathbf{W}$, and text- and vision-specific input activation norms $\|\mathbf{X}_t\|_2$ and $\|\mathbf{X}_v\|_2$, respectively, we obtain:

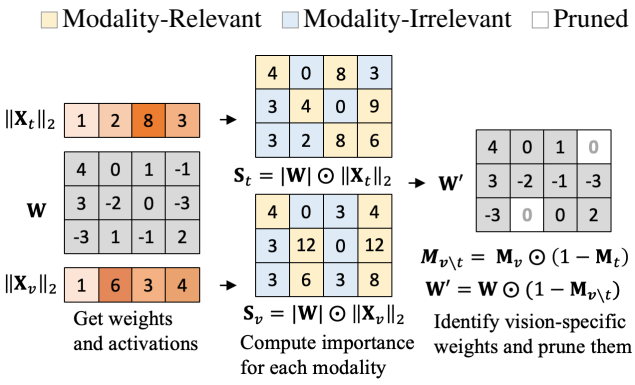


Figure 5: Illustration of modality-specific pruning of KoReP.

$$\mathbf{S}_t = |\mathbf{W}| \odot \|\mathbf{X}_t\|_2, \qquad \mathbf{S}_v = |\mathbf{W}| \odot \|\mathbf{X}_v\|_2 \tag{6}$$

Identifying Vision-Specific Weights. Based on $\mathbf{S}_t$ and $\mathbf{S}_v$, we define binary masks $\mathbf{M}_t$ and $\mathbf{M}_v$ that denote whether each weight belongs to the top-$k\%$ most important parameters for text and vision, respectively. The top-$k\%$ selection is performed respectively in each layer. As illustrated in Figure 5, yellow cells denote weights selected as modality-relevant, while blue cells correspond to the remaining weights, with k set to 50%.

$$\mathbf{M}_t = \mathbb{I}(\mathbf{S}_t \in \text{top-}k\%), \qquad \mathbf{M}_v = \mathbb{I}(\mathbf{S}_v \in \text{top-}k\%) \tag{7}$$

Vision-specific weights are defined as parameters selected by $\mathbf{M}_v$ but excluded by $\mathbf{M}_t$, as follows:

$$\mathbf{M}_{v\setminus t} = \mathbf{M}_v \odot (1 - \mathbf{M}_t) \tag{8}$$

Modality-Specific Pruning. KoReP prunes vision-specific parameters identified by the mask $\mathbf{M}_{v\setminus t} = 1$, yielding a pruned weight matrix $\mathbf{W}'$ in which the corresponding weights in $\mathbf{W}$ are set to zero. This mitigates the latent residual capacity through which text-only relearning would otherwise reactivate forgotten visual representations. Importantly, pruning is restricted to vision-specific weights, preserving the text-relevant subspace and maintaining overall model utility. By structurally weakening the vision-only parameter subspace, KoReP prevents text gradients from propagating into the unstable visual representation space.

$$W'_{ij} = \begin{cases} 0, & \text{if } M_{v\setminus t,ij} = 1 \\ W_{ij}, & \text{otherwise} \end{cases} \tag{9}$$

6 EXPERIMENTS

6.1 EVALUATION TASKS

We evaluate three complementary tasks that probe knowledge retention at different levels: (i) a classification task (A_m/A_u) measuring discrete factual recall via multiple-choice accuracy, (ii) a generation task (R_m/R_u) assessing free-form attribute generation via ROUGE-L (Lin, 2004), and (iii) a cloze task (C_m/C_u) probing residual knowledge without answer candidates. Here, m and u denote multimodal and unimodal evaluation settings, respectively. Full task descriptions are in Appendix C.

6.2 EFFECTIVENESS OF KOREP

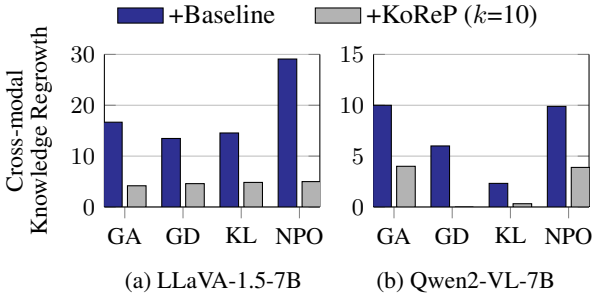
We evaluate the vulnerability of unlearned models to text-only relearning attacks across the three tasks. Applying these attacks yields regrowth-prone models, which we denote as Baseline. Table 3 reports

Table 3: Cross-modal knowledge regrowth (%) in existing unlearning methods with Baseline relearning, and its suppression by KoReP on MLLMU-Bench using LLaVA-1.5-7B. Results on UMU-Bench with Qwen2-VL-7B and CLEAR with LLaVA-1.5-7B are provided in Appendix G.

Method	Forget Set						Retain Set					
	A_m (↓)	A_u (↓)	R_m (↓)	R_u (↓)	C_m (↓)	C_u (↓)	A_m (↑)	A_u (↑)	R_m (↑)	R_u (↑)	C_m (↑)	C_u (↑)
Vanilla	39.79	44.79	0.39	0.67	36.96	21.74	39.49	42.38	0.35	0.68	42.39	14.17
GA	10.00	35.21	0.00	0.39	0.00	8.70	27.99	37.49	0.00	0.42	0.00	10.46
GA + Baseline	26.67	38.61	0.00	0.65	0.00	13.04	37.89	36.39	0.00	0.67	0.00	9.10
GA + KoReP (k=10)	14.17	34.17	0.00	0.66	0.00	10.86	29.30	37.93	0.00	0.67	0.00	11.95
GA + KoReP (k=20)	10.00	28.75	0.00	0.66	0.00	6.52	22.62	37.59	0.00	0.67	0.00	11.95
GD	19.58	42.92	0.12	0.64	4.35	15.21	34.34	40.68	0.20	0.67	19.13	15.75
GD + Baseline	33.05	43.06	0.14	0.69	8.70	21.74	36.87	39.96	0.27	0.42	28.54	15.65
GD + KoReP (k=10)	24.17	39.17	0.12	0.65	7.61	15.21	34.55	38.27	0.25	0.68	16.81	16.39
GD + KoReP (k=20)	20.00	35.83	0.14	0.66	6.52	13.04	33.68	39.05	0.28	0.68	15.57	15.33
KL	12.67	34.00	0.00	0.63	0.00	13.04	33.45	37.73	0.07	0.65	12.89	11.21
KL + Baseline	27.22	35.00	0.00	0.68	4.35	17.39	44.14	37.57	0.12	0.69	20.09	14.80
KL + KoReP (k=10)	17.50	33.33	0.00	0.68	2.17	14.13	39.74	38.18	0.09	0.68	17.68	11.95
KL + KoReP (k=20)	15.83	30.83	0.00	0.68	2.17	13.04	37.17	41.90	0.09	0.69	17.23	13.33
NPO	8.34	35.63	0.10	0.50	0.00	8.69	24.14	37.60	0.08	0.45	4.02	10.04
NPO + Baseline	37.43	36.94	0.10	0.68	4.35	15.22	36.96	38.20	0.10	0.68	10.36	11.52
NPO + KoReP (k=10)	13.33	35.83	0.10	0.67	2.17	15.22	28.07	39.11	0.09	0.67	4.97	12.47
NPO + KoReP (k=20)	11.67	34.17	0.09	0.66	2.17	11.96	27.65	35.56	0.08	0.67	4.97	12.68
MANU	22.80	21.04	0.18	0.64	15.22	17.39	29.96	30.04	0.33	0.67	25.08	21.73
MANU + Baseline	32.80	29.60	0.21	0.68	17.40	20.00	35.48	33.92	0.32	0.68	29.84	17.97
MANU + KoReP (k=10)	25.20	28.40	0.18	0.68	16.00	18.00	29.58	29.49	0.33	0.68	25.68	17.68
MANU + KoReP (k=20)	24.80	28.00	0.18	0.68	16.00	16.00	29.28	30.21	0.31	0.68	25.28	17.68

results for five state-of-the-art MLLM unlearning methods, including MANU (Liu et al., 2025b), a recent modality-aware unlearning method, on MLLMU-Bench (Liu et al., 2025a) using LLaVA-1.5-7B (Liu et al., 2023). Each method is evaluated under three conditions: (i) unlearning alone, (ii) unlearning followed by text-only relearning attack (+Baseline), (iii) unlearning with KoReP applied prior to relearning (+KoReP), under different k values. Across all methods, Baseline consistently exhibits substantial cross-modal knowledge regrowth after text-only relearning, whereas KoReP almost completely suppresses regrowth across multimodal evaluation metrics (i.e., A_m, R_m, and C_m). For example, under NPO+Baseline, A_m and C_m increase by 29.09% and 4.35%, respectively, while KoReP with k=10 reduces these increases to 4.99% and 2.17%. Notably, MANU remains ineffective at preventing cross-modal knowledge regrowth. Token-level analysis in Appendix G.4 further supports these findings at the probability level, showing that KoReP (k=20) reduces mean token log-probability increase by approximately 35% compared to Baseline. Here, k follows the definition in equation 7, and the resulting effective pruning ratios are reported in Appendix F.

In addition, KoReP maintains or even improves performance on the retain set across all metrics. This indicates that *selectively removing the vulnerable vision-specific subspace effectively and substantially reduces latent capacity that enables text-only relearning to restore visual knowledge*. However, increasing k to 20 further suppresses regrowth on the forget set, but this comes at the cost of degradation in retain set performance. This reflects a trade-off between the extent of pruning and utility preservation. We further validate the generality of our findings across other datasets and backbones in Appendix G. Figure 6 summarizes cross-modal knowledge regrowth with and without KoReP after relearning for LLaVA-1.5-7B on MLLMU-Bench and Qwen2-VL-7B on UMU-Bench. According to the results, Baseline models exhibit substantial regrowth on the forget set, whereas our approach, KoReP, consistently limits regrowth to within 5%.



(a) LLaVA-1.5-7B (b) Qwen2-VL-7B

Figure 6: Effect of KoReP in suppressing cross-modal knowledge regrowth (%), measured by A_m.

6.3 Modality-Specific Pruning vs. Random Pruning

To isolate the effect of targeted pruning in KoReP from generic performance degradation due to pruning itself, we compare KoReP with random pruning, where the same portion of parameters is randomly removed in each layer. Both methods are applied to the KL-unlearned model used in

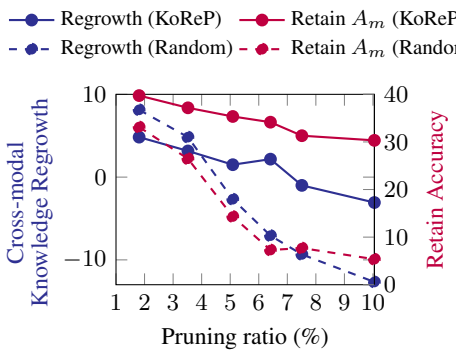


Figure 7: Effect of pruning ratio on the performance of KoRePe and Random pruning.

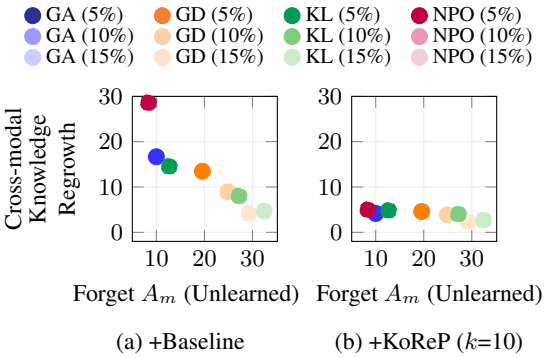


Figure 8: Effect of KoRePe on knowledge regrowth measured by A_m, under varying forget set ratios.

Table 3 across different effective pruning ratios. As shown in Figure 7, random pruning might seem to reduce cross-modal knowledge regrowth measured by A_m, but it leads to a sharp degradation in retain A_m, indicating indiscriminate utility loss. In contrast, KoRePe consistently suppresses regrowth while preserving substantially higher retain A_m across all pruning levels. This gap indicates that KoRePe's effectiveness cannot be explained by pruning's impact alone, but rather by selectively removing vision-specific weights that are vulnerable to reactivation during text-only relearning.

6.4 Effect of Pruning Criteria

We compare pruning criteria under the KL text-only relearning setting, keeping the effective pruning ratio identical to KoRePe with k=20. As shown in Table 4, weight magnitude-based pruning (Han et al., 2015) fails to suppress regrowth, while activation-based pruning (Hu et al., 2016) reduces regrowth but degrades retain A_m. Loss change-based pruning (Molchanov et al., 2019), which approximates the change in training loss induced by leaving a parameter out, better preserves retain performance but remains ineffective at preventing regrowth. KoRePe, which jointly considers activation and weight magnitudes, achieves the lowest regrowth and highest performance.

Table 4: Effect of different pruning criteria on regrowth suppression and retain set performance.

Importance Function	Regrowth	A_m	A_u
Weight magnitude	12.59	36.15	41.38
Activation norm	3.82	33.45	40.49
Loss change (Gradient * Weight)	8.42	36.87	40.93
KoRePe (Activation * Weight)	**3.16**	**37.17**	**41.90**

6.5 Effect of Forget Set Ratio

We vary the forget set ratio to evaluate KoRePe's robustness under different unlearning strengths. Figure 8 shows that Baseline generally exhibits higher regrowth than KoRePe. However, in Figure 8a, Baseline displays a decrease in regrowth as the forget set ratio increases. This decrease occurs because higher forget set ratios increase post-unlearning accuracy on the forget set, thereby leaving only limited knowledge to regrow. In contrast, when applying KoRePe on top of different unlearning methods (Figure 8b), regrowth is consistently reduced to approximately 5% or lower across all forget set ratios and shows robustness to variations in unlearning strength.

7 Conclusion

This work uncovers a critical issue of current MLLM unlearning, *Cross-modal Knowledge Regrowth*, where visual knowledge that was supposed to be forgotten can be inadvertently restored through purely textual prompts. We empirically validated this phenomenon across diverse MLLMs, datasets, and unlearning algorithms. We also identified the mechanism underlying this behavior based on the reactivation of vision-specific parameters. Selectively removing these parameters effectively blocks the cross-modal knowledge regrowth pathways, while maintaining model utility. Our study highlights an overlooked failure mode of existing MLLM unlearning and underscores the importance of a modality-aware design for reliable unlearning in MLLMs.

AI USE STATEMENT

In this work, generative AI tools were used solely for code generation. All generated code was manually inspected and tested by the authors before use. The formulation of research questions, experimental decisions, interpretation of results, and final scientific claims were made by the authors. Generative AI tools were not used for research ideation. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This paper identifies a previously undiscovered risk in MLLM unlearning: visual knowledge intended to be forgotten can be inadvertently restored through textual prompts. Such behavior raises privacy and security concerns, particularly when models are expected to remove sensitive or personal information. Although characterizing this vulnerability could reveal a potential failure mode of existing unlearning methods, our goal is to facilitate its detection and mitigation rather than enable misuse. We investigate the origin of this behavior and propose a mitigation strategy that reduces unintended knowledge recovery. We believe that understanding and addressing such failure modes is important for developing more reliable unlearning mechanisms and for supporting the responsible deployment of MLLMs in privacy-sensitive and regulated settings.

REPRODUCIBILITY STATEMENT

We provide the code used to reproduce our experiments in the supplementary material. Additional implementation and evaluation details, including compute resource specifications, are provided in the main paper and appendices.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 958–979, 2024.

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.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20582–20603, 2025.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Hsiang Hsu, Pradeep Niroula, Zichang He, and Chun-Fu Chen. Are we really unlearning? the presence of residual knowledge in machine unlearning. In *I Can't Believe It's Not Better: Challenges in Applied Deep Learning*, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network trimming: A data-driven neuron pruning approach towards efficient deep architectures. *arXiv preprint arXiv:1607.03250*, 2016.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Shengyuan Hu, Neil Kale, Pratiksha Thaker, Yiwei Fu, Steven Wu, and Virginia Smith. Blur: A benchmark for llm unlearning robust to forget-retain overlap. *arXiv preprint arXiv:2506.15699*, 2025b.

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. *arXiv preprint arXiv:2502.11051*, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020.

Minsung Kim, Nakyeong Yang, and Kyomin Jung. Rethinking post-unlearning behavior of large vision-language models. *arXiv preprint arXiv:2506.02541*, 2025.

Yongwoo Kim, Sungmin Cha, and Donghyun Kim. Are we truly forgetting? a critical re-examination of machine unlearning evaluation protocols. *Engineering Applications of Artificial Intelligence*, 167:113785, 2026.

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, 2024a.

Xinghang Li, Minghuan Liu, Hanbo Zhang, Cunjun Yu, Jie Xu, Hongtao Wu, Chilam Cheang, Ya Jing, Weinan Zhang, Huaping Liu, Hang Li, and Tao Kong. Vision-language foundation models as effective robot imitators. In *The Twelfth International Conference on Learning Representations*, 2024b.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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 *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)*, pp. 4105–4135, 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. *arXiv preprint arXiv:2502.15910*, 2025b.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, Yejin Choi, Muhao Chen, and Chaowei Xiao. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *The Thirteenth International Conference on Learning Representations*, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024.

Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11264–11272, 2019.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.

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

Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Curtis Mozer, and Eleni Triantafillou. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Songbai Tan, Xuerui Qiu, Yao Shu, Gang Xu, Linrui Xu, Xiangyu Xu, Huiping Zhuang, Ming Li, and Fei Yu. WMarkGPT: Watermarked image understanding via multimodal large language models. In *Forty-second International Conference on Machine Learning*, 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.

Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (gdpr). *A practical guide, 1st ed., Cham: Springer International Publishing*, 10(3152676):10–5555, 2017.

Chengye Wang, Yuyuan Li, XiaoHua Feng, Chaochao Chen, Xiaolin Zheng, and Jianwei Yin. Umubench: Closing the modality gap in multimodal unlearning evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yake Wei, Yu Miao, Dongzhan Zhou, and Di Hu. Moka: Multimodal low-rank adaptation for MLLMs. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D Manning, and Christopher Potts. Reft: Representation finetuning for language models. *Advances in Neural Information Processing Systems*, 37:63908–63962, 2024.

Jiaqi Xu, Cuiling Lan, and Yan Lu. Deciphering functions of neurons in vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3173–3181, 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.

Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of LLM unlearning via quantization. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Related Work

A.1 Machine Unlearning in MLLMs

Recently, machine unlearning in MLLMs has attracted growing attention. However, existing approaches largely amount to extensions of methods developed for unimodal LLMs. To evaluate multimodal unlearning in MLLMs, several benchmark datasets have been proposed, including CLEAR (Dontsov et al., 2025), MLLMU-Bench (Liu et al., 2025a), UMU-Bench (Wang et al., 2025), and FIUBench (Ma et al., 2025). These datasets introduce synthetic identity profiles with paired facial images and question–answer annotations, and show that unlearning a single modality is insufficient for forgetting multimodal knowledge, requiring the joint removal of both visual and textual information. Building on these datasets, early methods such as SIU (Li et al., 2024a) and MMUnlearner (Huo et al., 2025) primarily attempt to suppress visual features while keeping textual capabilities intact. More recently, several methods have been proposed to explicitly target multimodal knowledge removal. For instance, MANU (Liu et al., 2025b) identifies and prunes neurons that are strongly associated with the forget set while contributing minimally to the retain set, whereas MLLMEraser (Ding et al., 2025) introduces a training-free approach that applies activation steering at test time. However, these methods predominantly evaluate unlearned models under multimodal settings, and while a limited number of studies (Liu et al., 2025a;b) report unimodal evaluations, they do not analyze whether cross-modal interactions may enable the forgotten knowledge to reemerge.

A.2 Relearning Attacks in LLMs

Recent work shows that LLM unlearning methods are highly vulnerable to relearning attacks, where supposedly removed information can resurface even under small perturbations or minimal fine-tuning. Even after applying gradient-based unlearning, injecting imperceptible perturbations causes the model to recover residual knowledge that remains in the local neighborhood of the erased representation, revealing that forgetting does not fully eliminate surrounding semantic traces (Hsu et al., 2025). Furthermore, LLMs can relearn forgotten content when fine-tuned on either auxiliary data correlated with the forget set or subsets of the forget set, indicating that representational entanglement allows erased knowledge to be reconstructed from correlated contexts (Hu et al., 2025a). Example-level unlearning often leaves forgotten data in a dormant state rather than completely deleting it, as forget set performance can be almost fully recovered by fine-tuning solely on the retain set in vision classifiers (Siddiqui et al., 2025). Complementary to these findings, applying weight quantization to an unlearned model can inadvertently restore the forgotten knowledge (Zhang et al., 2025). However, despite these findings in LLMs, existing studies neither examine how such restoration of forgotten knowledge manifests across modalities in MLLMs nor provide effective solutions to prevent it.

B Overview of Unlearning Baseline Algorithms

B.1 Gradient Ascent (GA)

Gradient Ascent provides the most direct approach to unlearning by maximizing the loss on the forget set D_{forget}. For each $x \in D_{\text{forget}}$, the model parameters θ are updated so that the sample incurs larger loss, thereby suppressing the task-related representations learned from these data points:

$$\mathcal{L}_{GA}(D_{\text{forget}}, \theta) = \frac{1}{|D_{\text{forget}}|} \sum_{x \in D_{\text{forget}}} \ell(x, \theta). \tag{10}$$

While simple and effective in forcing the model away from the forgotten samples, GA lacks a mechanism to preserve the retain set, often leading to severe degradation in overall model performance.

B.2 Gradient Difference (GD)

Gradient Difference extends GA by explicitly balancing forgetting and retention. For a forget–retain pair $(D_{\text{forget}}, D_{\text{retain}})$, the method minimizes the task loss on D_{retain} while maximizing the loss on D_{forget}:

$$\mathcal{L}_{GD} = -\mathcal{L}(D_{\text{forget}}, \theta) + \mathcal{L}(D_{\text{retain}}, \theta). \tag{11}$$

This objective encourages the model to erase information associated with D_{forget} without sacrificing its ability on D_{retain}. However, the effectiveness of GD can vary substantially depending on the relative sizes, difficulty, and modality composition of the two subsets.

B.3 KL Minimization (KL)

KL Minimization preserves the model's behavior on the retain set D_{retain} by enforcing consistency with the original model f_{vnl}, i.e., the fine-tuned vanilla model prior to unlearning. The forget set continues to receive a loss-maximizing signal, while the retain set minimizes the divergence between the current model f_{cur} and f_{vnl}:

$$\mathcal{L}_{KL} = -\mathcal{L}(D_{\text{forget}}, \theta) + \frac{1}{|D_{\text{retain}}|} \sum_{s \in D_{\text{retain}}} \text{KL}(f_{\text{vnl}}(s) \,\|\, f_{\text{cur}}(s)) \,. \tag{12}$$

This method effectively protects D_{retain} but requires access to the original model and may propagate modality imbalance into the KL term when applied to multimodal architectures.

B.4 Negative Preference Optimization (NPO)

Negative Preference Optimization reframes unlearning as a preference-adjustment problem. Instead of directly maximizing loss, NPO penalizes the model when it assigns high probability to the original label y of a forget sample x compared to a reference distribution $\pi_{\text{ref}}(y|x)$:

$$\mathcal{L}_{NPO} = \frac{2}{\beta} \mathbb{E}_{(x,y)\sim D_{\text{forget}}} \left[\log \left(1 + \left(\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right)^\beta \right) \right], \tag{13}$$

where $\beta > 0$ controls the strength of the penalty. In this work, we define the reference distribution $\pi_{\text{ref}}(y|x)$ using a frozen oracle model θ_{orc}, which is obtained by fine-tuning the pre-trained model f_{pre} solely on D_{retain}.

C Task Composition

To examine how unlearning and relearning affect unimodal and multimodal knowledge in MLLMs, we employ three complementary evaluation tasks that probe knowledge retention at different levels, ranging from discrete factual recall to free-form factual generation.

Classification Task. For each individual profile, we use a four-option question whose answer corresponds to a specific attribute (e.g., occupation) in the profile. The three distractors are sampled from other profiles belonging to the same attribute category. The goal of this task is to evaluate the model's ability to correctly identify identity-linked attributes after unlearning. Evaluation is performed using accuracy: A_m for the multimodal VQA version and A_u for the unimodal QA version. These metrics directly reflect the extent to which identity-linked factual knowledge remains accessible after unlearning.

Generation Task. The generation task requires the model to generate free-form textual responses for individual profile attributes, such as background, interests, or personal details. Unlike the classification task, which restricts outputs to fixed choices, the generation task evaluates the model's ability to produce attribute-level factual information in an open-ended format after unlearning. Generated text is evaluated using ROUGE-L (Lin, 2004) by comparing it with ground-truth answers, where R_m and R_u denote scores for the multimodal and unimodal settings, respectively. These metrics uncover whether unlearning affects the model's ability to reconstruct or paraphrase forgotten identity information in free-form text.

Cloze Task. The cloze task presents a sentence in which a key attribute word is replaced with a `[blank]` token. The model must infer the missing word from context, making this task more sensitive to residual knowledge than classification, since no answer candidates are provided. We report cloze accuracy using C_m for the multimodal setting and C_u for the unimodal setting. In both settings, a prediction is considered correct if the ground-truth word appears in the model-generated output. These metrics reveal the extent to which the model still internally encodes forgotten attributes, even in the absence of explicit answer candidates.

D Dataset Details

To comprehensively evaluate multimodal unlearning and cross-modal knowledge regrowth, we use three benchmark datasets in our experiments: MLLMU-Bench, UMU-Bench, and CLEAR.

- **MLLMU-Bench (Liu et al., 2025a):** MLLMU-Bench is a multimodal unlearning benchmark composed of 653 profiles, including 500 fictitious profiles and 153 real celebrity profiles. The benchmark contains 20,754 questions that are evenly split between multimodal and unimodal settings. It supports three evaluation tasks: (i) a multiple-choice classification task with 11,530 questions, (ii) a free-form generation task with 4,612 questions, and (iii) a cloze task consisting of 4,612 fill-in-the-blank questions. The fictitious profiles are synthetically generated by GPT-4o (Hurst et al., 2024), and all question–answer pairs are derived directly from the constructed profiles. All profile images are generated using StyleGAN (Karras et al., 2019) to avoid privacy concerns. The fictitious profiles span 70 countries, 240 regions, 211 birth years, and 145 employment categories. The 153 real celebrity profiles are selected from the CelebA (Liu et al., 2015) dataset, manually verified by human experts.
- **UMU-Bench (Wang et al., 2025):** UMU-Bench reformulates multimodal unlearning as a knowledge-centric problem, where each knowledge instance is paired with both unimodal and multimodal representations. To this end, UMU-Bench retains the same scale and profile composition as MLLMU-Bench (653 profiles with both fictitious and real individuals) but reorganizes the data into unified knowledge instances and introduces paired unimodal–multimodal evaluation tasks.
- **CLEAR (Dontsov et al., 2025):** CLEAR is a multimodal unlearning benchmark consisting of 200 fictitious individuals associated with both textual and visual information. The dataset includes 4,000 text-based question–answer pairs (20 per individual) and 3,770 image–text pairs describing the corresponding individuals. The textual data are adapted from the TOFU benchmark (Maini et al., 2024), while the visual data comprise synthetic facial identities generated using StyleGAN2 (Karras et al., 2020), with their corresponding text automatically generated by GPT-4o (Hurst et al., 2024).

E Implementation Details and Compute Resources

All experiments are conducted on a single NVIDIA H200 (141GB) GPU using Python 3.10 on an x86-64 system running Red Hat Enterprise Linux 9.2. To assess generalization across benchmarks and MLLMs, we evaluate LLaVA-1.5-7B (Liu et al., 2023) on MLLMU-Bench (Liu et al., 2025a) and CLEAR (Dontsov et al., 2025), and evaluate Qwen2-VL-7B (Wang et al., 2024) on UMU-Bench (Wang et al., 2025). For each dataset, we use the hyperparameters adopted from the corresponding original benchmarks and apply LoRA (Hu et al., 2022) adapters to the self-attention and feed-forward layers of the language models with rank 16, $\alpha = 16$, and dropout 0.05. After fine-tuning, unlearning, and relearning, the LoRA adapters are merged into the base model, while pruning is applied to the merged weights after unlearning. The hyperparameters used in the experiments are reported in Table 5.

F Exploration on Effective Pruning Ratios

Figure 9 illustrates how modality-specific top-k parameter selection affects the effective pruning ratio across four unlearning methods on MLLMU-Bench (Liu et al., 2025a) using LLaVA-1.5-7B (Liu et al., 2023). Across all methods, we observe a consistent trend: the effective pruning ratio increases approximately linearly as k increases from 10% to around 60–70%, after which the pruning ratio gradually declines. This behavior reveals a saturation point beyond which increasing k yields diminishing returns in terms of effective pruning. Overall, these results confirm that the proposed pruning procedure behaves predictably across different unlearning algorithms and values of k.

Table 5: Hyperparameter settings for finetuning, unlearning, and relearning across different backbones and benchmark datasets.

Benchmark	Backbone	Stage	Epoch	Batch Size	Learning Rate	Optimizer	Weight Decay
MLLMU-Bench	LLaVA-1.5-7B	Fine-tuning	4	4	2×10^{-5}	AdamW	0.01
		Unlearning	2		2×10^{-5}		
		Relearning	1		2×10^{-5}		
UMU-Bench	Qwen2-VL-7B	Fine-tuning	5	4	1×10^{-4}	AdamW	0.01
		Unlearning	5		1×10^{-4}		
		Relearning	1		1×10^{-4}		
CLEAR	LLaVA-1.5-7B	Fine-tuning	3	4	1×10^{-5}	AdamW	0.01
		Unlearning	5		1×10^{-5}		
		Relearning	1		1×10^{-5}		

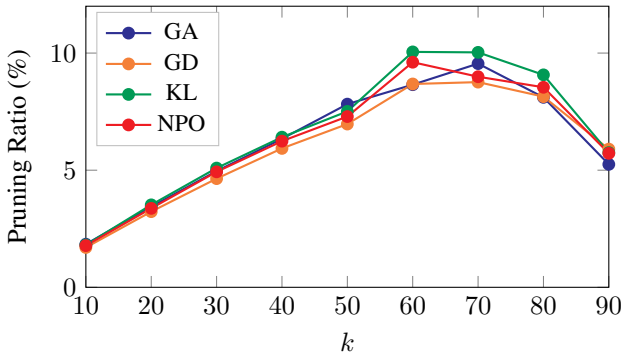


Figure 9: Effective pruning ratio with respect to the top-k% selection.

G Additional Experiments

G.1 Generalization to UMU-Bench with Qwen2-VL-7B

Table 6 presents the results under the text-only relearning setting on UMU-Bench (Wang et al., 2025) with Qwen2-VL-7B (Wang et al., 2024), comparing Baseline and KoReP. Consistent with the observations on MLLMU-Bench (Liu et al., 2025a) using LLaVA-1.5-7B (Liu et al., 2023) in Table 3, Baseline exhibits pronounced cross-modal knowledge regrowth, reaching up to 7.05% on average, indicating that this vulnerability is not specific to a particular backbone or dataset. In contrast, KoReP effectively suppresses regrowth across all unlearning methods to within 2.05% on average. Notably, we observe that both unlearning and subsequent relearning induce relatively smaller performance variations on Qwen2-VL-7B compared to LLaVA-1.5-7B, which aligns with prior findings that features learned by Qwen2-VL are more robust and harder to erase through minor parameter updates (Huo et al., 2025).

G.2 Generalization to CLEAR with LLaVA-1.5-7B

Table 7 reports the results under the text-only relearning setting on CLEAR (Dontsov et al., 2025) using LLaVA-1.5-7B (Liu et al., 2023), comparing Baseline and KoReP across multiple unlearning methods. In CLEAR, A_m measures face-based name identification from multiple images per individual, capturing the regrowth of visual identity information, rather than personal-attribute identification as in MLLMU-Bench and UMU-Bench. Consistent with the trends observed in Tables 3 and 6, we find that Baseline consistently exhibits cross-modal knowledge regrowth after relearning, as reflected by increased A_m scores on the forget set. In comparison, KoReP effectively suppresses regrowth across all unlearning methods while preserving competitive retain set performance. These results further confirm that cross-modal regrowth is a general phenomenon in MLLMs and demonstrate the generalizability of KoReP across multiple datasets and architectures.

Table 6: Cross-modal knowledge regrowth in existing unlearning methods with Baseline relearning, and its suppression by KoReP on UMU-Bench Wang et al. (2025) using Qwen2-VL-7B Wang et al. (2024).

Method	Forget Set						Retain Set					
	A_m (↓)	A_u (↓)	R_m (↓)	R_u (↓)	C_m (↓)	C_u (↓)	A_m (↑)	A_u (↑)	R_m (↑)	R_u (↑)	C_m (↑)	C_u (↑)
Vanilla	50.00	44.00	0.21	0.17	52.00	28.00	51.79	45.89	0.19	0.17	35.37	25.37
GA	32.00	34.00	0.09	0.12	4.00	4.00	42.42	37.37	0.09	0.12	6.21	9.58
GA + Baseline	42.00	40.00	0.11	0.18	10.00	22.00	46.95	39.79	0.12	0.20	9.79	24.84
GA + KoReP (k=10)	36.00	32.00	0.08	0.18	8.00	18.00	43.26	40.42	0.09	0.19	8.95	10.90
GA + KoReP (k=20)	32.00	32.00	0.08	0.18	4.00	18.00	42.32	39.68	0.09	0.19	5.89	8.63
GD	22.00	34.00	0.00	0.09	0.00	16.00	27.47	37.37	0.03	0.11	2.74	17.47
GD + Baseline	28.00	34.00	0.01	0.16	8.00	28.00	30.68	39.32	0.11	0.18	8.58	21.79
GD + KoReP (k=10)	22.00	28.00	0.01	0.15	4.00	20.00	28.26	35.37	0.01	0.17	4.12	19.32
GD + KoReP (k=20)	20.00	26.00	0.01	0.15	0.00	16.00	26.42	35.26	0.01	0.17	2.21	19.32
KL	21.68	32.95	0.00	0.05	0.00	13.26	26.00	36.00	0.00	0.08	0.00	18.00
KL + Baseline	24.00	36.00	0.00	0.21	2.00	26.00	27.40	38.00	0.00	0.22	0.00	20.21
KL + KoReP (k=10)	22.00	34.00	0.00	0.19	0.00	14.00	26.32	37.15	0.00	0.20	0.00	19.58
KL + KoReP (k=20)	22.00	32.00	0.00	0.19	0.00	14.00	23.90	36.95	0.00	0.20	0.00	18.74
NPO	24.11	26.00	0.00	0.01	0.00	6.00	28.00	31.16	0.00	0.07	0.00	12.11
NPO + Baseline	34.00	34.00	0.00	0.12	2.00	20.00	30.95	35.89	0.00	0.13	0.00	20.84
NPO + KoReP (k=10)	28.00	30.00	0.00	0.08	0.00	12.00	28.00	31.58	0.00	0.13	0.00	13.72
NPO + KoReP (k=20)	26.00	26.00	0.00	0.08	0.00	12.00	27.40	30.95	0.00	0.13	0.00	11.58

Table 7: Cross-modal knowledge regrowth in existing unlearning methods with Baseline relearning, and its suppression by KoReP on CLEAR Dontsov et al. (2025) using LLaVA-1.5-7B Liu et al. (2023).

Method	Forget Set			Retain Set		
	A_m (↓)	R_m (↓)	R_u (↓)	A_m (↑)	R_m (↑)	R_u (↑)
Vanilla	61.70	0.59	0.29	67.34	0.58	0.28
GA	30.63	0.09	0.25	34.57	0.09	0.26
GA + Baseline	35.11	0.10	0.28	34.14	0.10	0.27
GA + KoReP (k=10)	34.57	0.09	0.28	33.88	0.10	0.27
GA + KoReP (k=20)	30.85	0.09	0.28	32.11	0.10	0.27
GD	27.66	0.04	0.25	39.75	0.07	0.26
GD + Baseline	29.79	0.05	0.25	40.00	0.07	0.27
GD + KoReP (k=10)	27.26	0.04	0.25	39.24	0.07	0.27
GD + KoReP (k=20)	25.00	0.04	0.25	38.23	0.07	0.27
KL	33.93	0.00	0.21	35.11	0.00	0.24
KL + Baseline	36.17	0.02	0.29	35.70	0.04	0.28
KL + KoReP (k=10)	35.11	0.01	0.29	34.94	0.04	0.28
KL + KoReP (k=20)	34.04	0.00	0.28	34.56	0.01	0.24
NPO	25.53	0.14	0.26	30.56	0.15	0.27
NPO + Baseline	28.10	0.15	0.29	29.30	0.14	0.28
NPO + KoReP (k=10)	26.06	0.14	0.29	29.30	0.14	0.28
NPO + KoReP (k=20)	26.06	0.14	0.29	29.05	0.14	0.28

G.3 Layer-wise Analysis of Vision-Specific Parameter Subspace Changes

To examine whether cross-modal knowledge regrowth is localized to specific architectural components, we conduct a layer-wise analysis measuring the vision-specific subspace ratio across all layer types under the NPO setting with LLaVA-1.5-7B. As shown in Table 8, the ratio consistently exceeds 1.5 across all self-attention projection layers (query, key, value, and output) and multilayer perceptron layers, indicating that cross-modal parameter coupling is a model-wide phenomenon rather than one confined to particular modules or depth ranges. This uniform distribution across layer types further supports our mechanistic explanation: cross-modal knowledge regrowth arises from structural parameter sharing between modalities.

Table 8: Layer-wise vision-specific subspace ratio under the NPO setting.

Component	Count	Mean Ratio (%)
Self-attention (Query, Key, Value)	96	1.60 ± 0.34
Self-attention (Output projection)	32	2.26 ± 0.38
Multilayer perceptron	96	1.80 ± 0.44
All layers	224	1.78 ± 0.44

G.4 Token-Level Quantification of Cross-Modal Knowledge Regrowth

To directly quantify cross-modal knowledge regrowth, we introduce a token-level leakage score that measures the probability assigned to forget-set targets. Specifically, we compute the log-probability of ground-truth identity attributes given a multimodal visual question answering prompt before and after text-only relearning. The target tokens used for this analysis correspond to the ground-truth attribute words from the cloze task described in Appendix C. This metric follows prior work showing that the likelihood assigned to target sequences serves as a reliable signal of latent knowledge in language models (Carlini et al., 2021), enabling direct quantification of cross-modal knowledge regrowth at the probability level. Log-Probability Δ denotes the difference in mean token log-probability before and after text-only relearning and Max Leakage denotes the maximum per-example increase in mean token log-probability. As shown in Table 9, under the GA setting with LLaVA-1.5-7B, the baseline exhibits substantial cross-modal knowledge regrowth, with mean token log-probability increasing by $+1.896$ after text-only relearning. In contrast, KoReP (k=20) reduces both mean and worst-case leakage by approximately 35%, confirming effective suppression at the probability level.

Table 9: Token-level cross-modal knowledge regrowth under the GA setting.

Metric	Baseline	KoReP (k=20)
Mean Token Log-Probability Δ	$+1.896$	$+1.231$
Max Leakage	$+7.52$	$+4.92$

G.5 Gradient Flow Analysis of Vision-Specific Parameters

To provide additional mechanistic evidence for how vision-specific parameters mediate cross-modal knowledge regrowth, we analyze gradient flow during text-only relearning by reporting the root mean square (RMS) of gradient updates for each parameter group using LLaVA-1.5-7B on MLLMU-Bench. Table 10 shows that vision-specific parameters receive consistent gradient updates during text-only relearning, despite the absence of visual input. Selectively removing these parameters suppresses cross-modal knowledge regrowth, whereas random parameter removal of equivalent magnitude does not (Section 6.3), indicating that these parameters serve as key mediators through which text-only updates restore unlearned visual knowledge.

Table 10: Gradient update RMS per parameter group during text-only relearning.

Parameter Group	Gradient Update RMS
Vision-specific	1.06×10^{-4}
Shared	2.15×10^{-4}
Text-specific	1.03×10^{-4}

G.6 LoRA Dependency: Generalizability to Full-parameter Settings

Our experiments adopt LoRA (Hu et al., 2022) adapters for fine-tuning, unlearning, and relearning. To validate whether cross-modal knowledge regrowth generalizes to full-parameter settings, we conduct experiments under full fine-tuning conditions. As shown in Figure 10, cross-modal knowledge regrowth is consistently observed across all unlearning methods, and freezing vision-specific

parameters during relearning suppresses regrowth compared to full parameter updates. These results confirm that the phenomenon and its underlying mechanism are not tied to the LoRA scheme.

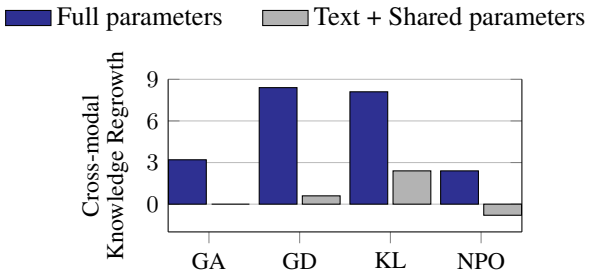


Figure 10: Cross-modal knowledge regrowth (%) under full fine-tuning, comparing full parameter updates against freezing vision-specific parameters during text-only relearning.

H Limitations and Future Work

Adaptations to Other Applications. Our study focuses on identity-level knowledge within the vision–language modality pair, using structured question–answer supervision over personal profile data. This setting is intentionally chosen because correct answers require linking a person's visual identity with their personal attributes, ensuring that performance recovery on the forget set reflects genuine restoration of visually grounded knowledge. While this provides a controlled environment to rigorously analyze cross-modal knowledge regrowth, whether similar regrowth dynamics arise for more abstract or unstructured forms of visual knowledge (e.g., scene-level or concept-level information) remains an open question.

Sensitivity to Pruning Ratio. KoRePP requires selecting a pruning ratio k, which controls the number of vision-specific parameters removed. While our experiments demonstrate that KoRePP consistently suppresses cross-modal knowledge regrowth across a range of k values, higher pruning ratios may reduce the model's effective capacity, introducing a trade-off between regrowth suppression and retain performance.

Future Work. For future work, building on our identification of the root causes underlying cross-modal regrowth, we aim to explore robust, modality-aware machine unlearning algorithms that are explicitly optimized to prevent knowledge regrowth across modalities during the unlearning process. Furthermore, extending our framework to a broader range of modalities, such as audio and video data, remains a promising direction.

I Broader Impacts

This paper identifies a previously undiscovered risk in MLLM unlearning: visual knowledge that should have been forgotten can be inadvertently restored by mere textual prompts. This failure mode undermines the reliability of existing unlearning practices and poses privacy concerns when models handle sensitive personal information. By uncovering the origin of this issue and proposing an effective mitigation strategy, our work contributes to building robust MLLMs that can more faithfully remove user data upon request. These improvements enable reliable deployments of MLLMs in sensitive domains where privacy and regulatory compliance are crucial.