AIM: Anchor Identity Features, Then Match for Multimodal Large Language Model Unlearning

Wonjun Lee* **Jaehyuk Jang*** **Kangwook Ko*** **Hee-Seon Kim** **Changick Kim**
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea {dpenguin, jhyuk, kw.ko, hskim98, changick}@kaist.ac.kr

Abstract

Multimodal large language models (MLLMs) can memorize identity-specific facts about people in their fine-tuning data, creating privacy risks when a person requests deletion. Existing MLLM unlearning methods often assume access to retain images or ground-truth answers during deletion, which is unrealistic in many practical scenarios. We study identity unlearning when retain images are unavailable at deletion time. Our analysis shows that identity and visual-perception questions occupy distinct regions in fine-tuned hidden states and are organized differently: identity questions cluster by person, whereas perception questions cluster by question type. This suggests that identity knowledge can be suppressed without erasing general visual perception. Building on this observation, we propose AIM, a two-stage method that anchors an identity-forgetting target with a universal visual prompt and then matches the vision encoder to that target under a Fisher-based constraint. Extensive experiments show that AIM achieves competitive identity forgetting while preserving non-deleted identities, prior knowledge, and visual perception on the same images.

1 Introduction

Modern multimodal large language models (MLLMs) are fine-tuned on increasingly large image-text datasets that include identifiable individuals (Liu et al., 2024; Bai et al., 2025; Chen et al., 2024). As a side effect, they memorize identity-specific knowledge such as names, occupations, and affiliations tied to specific faces. While this memorization supports downstream visual question answering, it becomes a privacy concern once an individual requests deletion of their own identity (Carlini et al., 2019, 2021), motivating the study of *MLLM unlearning*.

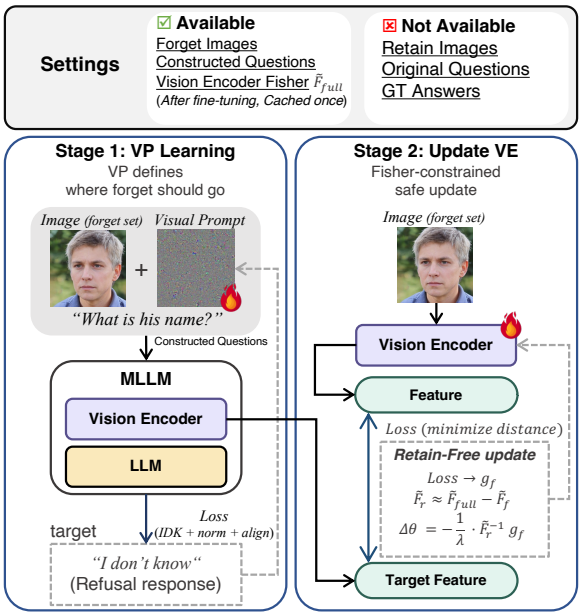


Figure 1: **Method Overview.** Stage 1 defines a target representation from forget images and constructed questions. Stage 2 internalizes this target into the vision encoder under a Fisher-based retain constraint.

Recent work has begun to explore MLLM unlearning from several directions (Liu et al., 2025a; Huo et al., 2025; Liu et al., 2025b), but many methods still require retain images (Huo et al., 2025; Wang et al., 2025; Li et al., 2026), external auxiliary data (Cai et al., 2026), or ground-truth answers during deletion. These assumptions are restrictive in realistic deletion scenarios, where the system may have access to the target identity's images but not to the original training questions, answers, or retain data. We therefore study identity unlearning when retain data are unavailable at deletion time. This setting requires suppressing identity-specific knowledge while preserving both non-deleted identities and general visual perception on the target images, distinguishing selective identity removal from indiscriminate image degradation.

To motivate a selective intervention, we first an-

*Equal contribution

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 30699–30722
October 24-29, 2026 ©2026 Association for Computational Linguistics

alyze how a fine-tuned MLLM organizes identity and visual-perception questions (§3). We find that their last-layer LLM hidden states occupy distinct regions, and that the two question types follow different grouping rules: identity questions cluster by person ID, whereas visual-perception questions cluster by question type. A pilot visual prompt trained on a few identity questions further transfers "I don't know"-style responses to held-out identity questions while largely preserving visual-perception answers on the same images. These observations suggest a design principle: identity knowledge can be targeted through a vision-side intervention while leaving general perception comparatively intact.

Building on this analysis, we propose **AIM**: **A**nchor **I**dentity Features, then **M**atch for MLLM unlearning (§4). AIM decouples unlearning into target definition and target matching, as illustrated in Fig. 1. First, it learns a universal visual prompt that anchors an identity-forgetting target for the deleted identity. Second, it updates the vision encoder so that the original target images match this anchored feature, while a Fisher-based constraint limits drift on directions important to non-deleted identities. The retain-preservation signal is instead derived from a full Fisher statistic cached once after fine-tuning, so deletion-time unlearning does not require retain images.

Empirically, AIM provides a strong forgetting–retention trade-off under this deletion-time constraint. On MLLMU-Bench and ReMem datasets with LLaVA-1.5-7B and Qwen3-VL-8B-Instruct, AIM preserves retain and celebrity-prior performance competitively with retain-utilizing baselines while achieving effective identity forgetting. The visual-perception preservation predicted by our analysis is also observed after unlearning on the target images themselves. Finally, because the cached Fisher can be updated after each request, AIM extends naturally to continual unlearning, where forget-only baselines collapse under sequential deletion but AIM maintains a balanced forgetting–retention trade-off.

2 Related Work

2.1 Machine Unlearning

Classical machine unlearning (Bourtoule et al., 2021) suppresses forget-set behavior while preserving utility through objectives such as gradient ascent (Thudi et al., 2022), gradient-difference training (Liu et al., 2022a), KL regularization (Nguyen et al., 2020), or preference optimization (Zhang et al., 2024). Recent work reduces supervision or data requirements through label-agnostic representation-level forgetting (Shen et al., 2024), label-free sensitivity estimates (Foster et al., 2024), remaining-data-free contribution suppression (Cheng et al., 2024), or feature-level LLM unlearning (Li et al., 2024). Additionally, Fisher- and Gauss-Newton-based methods use curvature or parameter importance to constrain updates along utility-critical directions (Golatkar et al., 2020; McKinney et al., 2026). However, these methods are mainly designed for classification or text-only LLMs, where forget targets are defined by labels, tokens, representations, or explicit output losses.

2.2 MLLM Unlearning and Visual-Side Approaches

Prior MLLM unlearning methods optimize multimodal forget losses (Huo et al., 2025) or edit modality-relevant neurons and paths (Liu et al., 2025b; Li et al., 2026). Recent studies further target general image-understanding preservation (Zeng et al., 2025), identity-encoded layer tuning (Ko et al., 2026), and stable vision-language alignment during unlearning (Garg et al., 2025).

Closer to the visual side, prior work explores feature-level concept editing in VLMs (Geng and Li, 2025), visual-guided token regularization (Cai et al., 2026), visual knowledge distillation (Wang et al., 2025), and input-side perturbations for unlearning or robustness attacks (Sun et al., 2024; Chen et al., 2026; Zhang et al., 2025). However, these often depend on extra supervision, retain-side signals, or input perturbations that can blur forgetting with visual degradation. We study a stricter deletion-time setting that uses only forget-identity images and constructed questions built from known attribute categories.

3 Analysis

It remains unclear how a fine-tuned MLLM (vanilla model) internally organizes identity-specific knowledge and general visual perception. We therefore analyze the last-layer LLM hidden states of the fine-tuned model before unlearning. For each (image, question) input, we use the last-layer hidden state at the final input-token position (just before answer generation) as the representation throughout the analysis (Meng et al., 2022; Geva et al., 2023; Neo

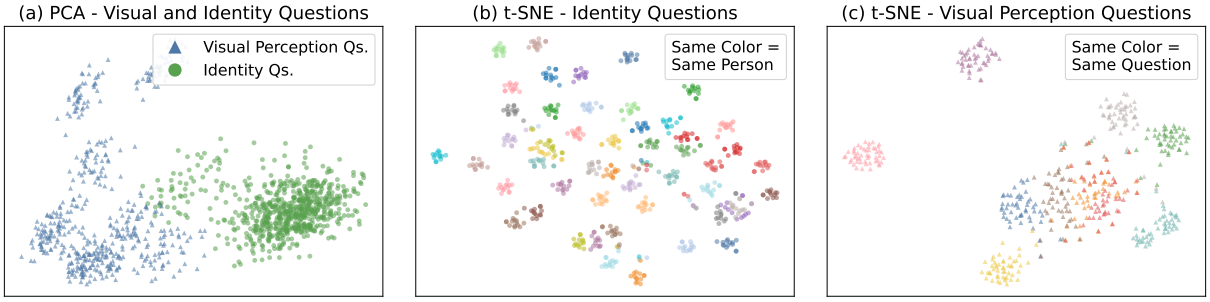


Figure 2: **Representational separability and grouping rules of identity and perception questions in last-layer LLM hidden states on LLaVA-1.5-7B.** (a) PCA projects identity and visual perception questions into largely disjoint regions. (b) t-SNE of identity questions colored by image identity. Same-person questions cluster together regardless of question content. (c) t-SNE of visual perception questions colored by question type. Same-type questions cluster together regardless of which person appears.

et al., 2025). Throughout, we use LLaVA-1.5-7B and Qwen3-VL-8B fine-tuned on MLLMU-Bench.

3.1 Representational Separability of Identity and Perception

To analyze how identity and perception responses are organized in representation space, we prepare two question sets: identity questions taken from MLLMU-Bench (e.g., "*What is the profession of the individual shown in the image?*") and visual perception questions we create (e.g., "*Is the person in the image wearing glasses?*"). We then visualize the last-layer LLM hidden states for each set via PCA and t-SNE.

Result 1: The two types are separated in representation space. The PCA result in Fig. 2(a) shows that the hidden states of identity questions and those of visual perception questions occupy distinct regions of representation space, indicating that the two kinds of information are encoded in a *separable* manner inside the model. This separation suggests that an intervention targeting identity-related representations may disrupt identity discrimination while leaving visual perception largely unaffected. We further verify in Appendix C.1 that this separation is a consequence of SFT rather than a property of the pretrained model.

Result 2: The two types follow different organization principles. The t-SNE in Fig. 2(b) shows that identity questions cluster *by the identity of images*, whereas Fig. 2(c) shows visual perception questions cluster *by question type*. That is, questions about the same person such as "*What is this person's name?*" and "*Where does this person work?*" are placed close together despite differing question text, while questions of the form "*What color is the person's shirt?*" are grouped together across different images.[1] This pattern motivates a behavioral hypothesis: an intervention learned from a few identity questions on a single image may transfer to other identity questions about the same image, whereas its effect on visual-perception questions may be weaker because those questions are organized primarily by question semantics.

3.2 Selective Transfer of Vision Side Intervention

The representation-level analysis above suggests that an image-side intervention learned from a few identity questions may transfer to other identity questions on the same image, while having limited effect on visual-perception questions. We test this hypothesis behaviorally using a visual prompt (VP). The VP perturbs only the visual input while keeping all model parameters fixed, allowing us to probe an identity-forgetting direction without incurring any parameter drift.

On a subset of the original fine-tuning data, we train a VP without ϵ constraints from 4 identity questions to elicit IDK responses (i.e., "I don't know"-style refusals such as "*I cannot identify this person.*"). We then measure (i) its transfer to 11 *held-out* identity questions on the same images, and (ii) its effect on a prepared set of 10 visual perception questions.

We evaluate along two dimensions using GPT-4o-mini: **IDK-Rate**—the fraction of questions on which the VP-applied response is judged to be an IDK-type refusal; and **Semantic Preservation (SP)**—the fraction on which the VP-applied response is a substantive answer that matches the

[1] A quantitative analysis is provided in Appendix C.2.

Model	Identity (held-out)	Visual perception	
	IDK-Rate	IDK-Rate	SP
LLaVA-1.5-7B	100.0	10.0	63.3
Qwen3-VL-8B	85.0	0.0	85.2

Table 1: **Selective transfer of a pilot VP.** A VP trained to elicit IDK responses on 4 identity questions transfers to held-out identity questions but leaves visual perception responses on the same images largely intact.

response without VP in meaning.

Result. Table 1 reports the results on the held-out identity questions (group (i)) and the visual perception questions (group (ii)) for both models. On the held-out identity questions, VP suppresses the identity response on both models—LLaVA-1.5-7B and Qwen3-VL-8B both produce refusals on most identity questions. In contrast, on the visual perception questions from the same images, the IDK-Rate after applying unconstrained VP stays at 10.0% / 0% and SP remains at 63.3% / 85.2%, empirically confirming the predictions from Section 3.1. This contrast suggests that the VP primarily affects identity-related behavior rather than indiscriminately degrading the model's visual understanding of the image.

Method motivation. These observations inform two design choices for retain-data-free identity unlearning. First, they suggest that a vision-side intervention can be a viable mechanism for selective identity unlearning: identity responses can be shifted while visual-perception responses on the same images are largely preserved. This motivates pursuing unlearning through the visual pathway while keeping the language model fixed, which also preserves text-only QA behavior by design. Second, the transfer from a few probing questions to held-out identity questions suggests that the forget target need not be defined using supervision over the entire question set.

4 Method

The analysis above motivates a vision-side approach to selective identity unlearning under deletion-time retain-data unavailability. We propose **AIM** (**A**nchor **I**dentity Features, then **M**atch), a two-stage method that updates only the vision encoder while keeping the language model fixed. AIM first defines an identity-forgetting target for the forget images (§4.2), and then internalizes this target by updating the vision encoder so that the unprompted forget-image features match the learned target features (§4.3).

4.1 Problem Formulation

Data are organized at the identity level. Let $\mathcal{C}$ denote the full set of identities, and let

$$\mathcal{V}_c = \{V_{c,a}\}_{a=1}^{n_c} \tag{1}$$

be the images associated with identity $c \in \mathcal{C}$. The vanilla MLLM is fine-tuned on these images together with corresponding VQA pairs, but these original questions and answers are not assumed to be available during unlearning.

Let $\mathcal{C}_f \subset \mathcal{C}$ be the set of identities requested for deletion. The forget and retain image sets are

$$\mathcal{V}_f = \bigcup_{c \in \mathcal{C}_f} \mathcal{V}_c, \qquad \mathcal{V}_r = \bigcup_{c \in \mathcal{C} \setminus \mathcal{C}_f} \mathcal{V}_c. \tag{2}$$

At deletion time, AIM has access to the forget images $\mathcal{V}_f$ and constructed probing questions for the forget identities, but not to the retain images $\mathcal{V}_r$ or the original fine-tuning VQA pairs. The goal is therefore to suppress identity-specific behavior for $\mathcal{C}_f$ while preserving the behavior of non-deleted identities in $\mathcal{C} \setminus \mathcal{C}_f$.

AIM updates only the vision encoder E_v of an MLLM $\mathcal{M}_\theta$ and keeps the language model fixed. We denote the vision-encoder parameters by θ and the initial fine-tuned parameters by θ_0. Since $\mathcal{V}_r$ is unavailable at deletion time, retain preservation cannot be written as an explicit loss over retain images. A naive forget-only update can therefore induce feature drift on non-deleted identities without any retain-side signal to correct it (Appendix B.3).

At deletion time, AIM has access to the forget images $\mathcal{V}_f$, but not to the retain images $\mathcal{V}_r$ or the original fine-tuning VQA pairs. We additionally assume access to a set of attribute categories (e.g., name, occupation, place of birth) associated with each forget identity, which can be obtained from the schema of the fine-tuning dataset or provided as part of the deletion request. The goal is to suppress identity-specific behavior for $\mathcal{C}_f$ while preserving the behavior of non-deleted identities in $\mathcal{C} \setminus \mathcal{C}_f$.

4.2 Stage 1: Universal Visual Prompt Learning

Training data construction. Under the strict setting, we construct probing questions from the attribute categories assumed available in Section 4.1.

For each forget identity c, let $\mathcal{X}_c$ denote the resulting constructed questions and $\mathcal{X}_f = \bigcup_{c \in \mathcal{C}_f} \mathcal{X}_c$ their union. The forget prompt set used for VP learning is

$$\mathcal{D}_f^X = \{(V_{c,a}, X_{c,q}) \mid \; c \in \mathcal{C}_f, V_{c,a} \in \mathcal{V}_c, \; X_{c,q} \in \mathcal{X}_c\}. \tag{3}$$

Objective. To define a fixed identity-forgetting target for forget images, we learn a universal visual prompt $T \in \mathbb{R}^{H \times W \times C}$. T is an input-space perturbation applied uniformly to all forget images, and is optimized under the following objective:

$$\mathcal{L}_{\text{VP}} = \lambda_{CE}\, \mathcal{L}_{\text{CE}} + \lambda_{norm}\, \mathcal{L}_{\text{norm}} + \lambda_{align}\, \mathcal{L}_{\text{align}}. \tag{4}$$

At this stage, the model parameters are held fixed at θ_0 and only T is optimized.

Refusal response induction. We induce the model $\mathcal{M}_{\theta_0}$ to produce refusal responses for forget images to which T has been applied. Letting $\mathcal{Y}_{\text{idk}}$ denote the set of IDK responses, we use

$$\mathcal{L}_{\text{CE}} = -\mathbb{E}_{(V,X)\sim\mathcal{D}_f^X}\left[\log P_{\mathcal{M}_{\theta_0}}\left(\hat{Y}_{\text{idk}} \mid V+T, X\right)\right] \tag{5}$$

where $\hat{Y}_{\text{idk}}$ is sampled uniformly from $\mathcal{Y}_{\text{idk}}$.

Perturbation magnitude constraint. A larger T requires a larger Stage 2 update to reproduce its displacement, which in turn drifts θ farther from θ_0 and harms retain features. We therefore penalize the magnitude of T:

$$\mathcal{L}_{\text{norm}} = \|T\|_2^2. \tag{6}$$

This balances the pressure of $\mathcal{L}_{\text{CE}}$ to enlarge the perturbation, so that T grows only as large as needed for the forget set to reach the target location.

Feature displacement alignment. We require T to displace the forget images in a consistent direction within the representation space. For each forget image $V \in \mathcal{V}_f$, define

$$\Delta \mathbf{z}_V = E_v(V+T;\theta_0) - E_v(V;\theta_0), \tag{7}$$

and let

$$\mathcal{L}_{\text{align}} = -\left\| \mathbb{E}_{V\sim\mathcal{V}_f}\left[\frac{\Delta \mathbf{z}_V}{\|\Delta \mathbf{z}_V\|}\right]\right\|^2. \tag{8}$$

This term aligns the directions of $\Delta \mathbf{z}_V$ across forget images, which will be shown in Stage 2 (Section 4.3) to be necessary for keeping the forget gradient from collapsing in expectation.

4.3 Stage 2: Fisher-Constrained Vision Encoder Update

With T fixed, we update the vision encoder so that the forget features reach the same target without applying T. The target feature for each forget image is

$$\mathbf{z}_V^T = E_v(V+T;\theta_0), \tag{9}$$

fixed throughout the update. The forget feature loss is

$$\mathcal{L}_f(\theta) = \mathbb{E}_{V\sim\mathcal{V}_f}\left[\left\|E_v(V;\theta) - \mathbf{z}_V^T\right\|^2\right], \tag{10}$$

with forget gradient $\mathbf{g}_f(\theta) := \nabla_\theta \mathcal{L}_f(\theta)$ and per-image Jacobian $\mathbf{J}_V(\theta) := \partial E_v(V;\theta)/\partial\theta$. At $\theta = \theta_0$, the gradient takes the form

$$\mathbf{g}_f(\theta_0) = -2\,\mathbb{E}_{V\sim\mathcal{V}_f}\left[\mathbf{J}_V(\theta_0)^\top \Delta \mathbf{z}_V\right], \tag{11}$$

where $\Delta \mathbf{z}_V$ is the feature displacement defined in Section 4.2. This expression clarifies the role of $\mathcal{L}_{\text{align}}$. If $\Delta \mathbf{z}_V$ points in arbitrary directions across forget images, the expectation averages over incoherent directions and cancels to zero, yielding $\mathbf{g}_f(\theta_0) \approx 0$ and a meaningless update. $\mathcal{L}_{\text{align}}$ in Stage 1 aligns these displacements so that the per-image contributions $\mathbf{J}_V(\theta_0)^\top \Delta \mathbf{z}_V$ add up constructively at θ_0. This is the precise sense in which the VP is configured to make the Stage 2 update well-posed.

Retain preservation as a constraint. Retain performance is governed by the per-image feature drift $\|E_v(V;\theta+\Delta\theta) - E_v(V;\theta)\|^2$ on each $V \in \mathcal{V}_r$ at the current parameter θ. Because no retain data is available at unlearning time, any drift incurred along a forget update has no retain-direction signal to undo it; we therefore *bound* the drift at every update, preventing irreversible accumulation at the source. A first-order Taylor expansion of the squared drift around θ gives

$$\|E_v(V;\theta+\Delta\theta) - E_v(V;\theta)\|^2 \;\approx\; \Delta\theta^\top \mathbf{F}_V(\theta)\Delta\theta, \tag{12}$$

where $\mathbf{F}_V(\theta)$ has been adopted as a Fisher importance in previous works (Kirkpatrick et al., 2017; Aljundi et al., 2018; Golatkar et al., 2020; McKinney et al., 2026). To absorb the per-identity image-count imbalance common in identity datasets, we aggregate $\mathbf{F}_V(\theta)$ at the identity level—averaging within each identity before summing across identities (Appendix A.1):

$$\widetilde{\mathbf{F}}_r(\theta) = \sum_{c\in\mathcal{C}_r} \frac{1}{|\mathcal{V}_c|} \sum_{V\in\mathcal{V}_c} \mathbf{F}_V(\theta). \tag{13}$$

Fisher-preconditioned retain-free update. We minimize the forget objective subject to the per-step retain-drift budget ϵ_r (Golatkar et al., 2020). The KKT conditions yield the Fisher-preconditioned direction $\widetilde{\mathbf{F}}_r^{-1}\mathbf{g}_f$ (Appendix A.4), which we apply with a learning rate hyperparameter η controlling the per-step magnitude:

$$\Delta\theta = -\eta\, \widetilde{\mathbf{F}}_r(\theta)^{-1}\, \mathbf{g}_f(\theta). \tag{14}$$

Under the assumption that the constraint is satisfied at every step and that the step size is small enough to keep θ in a neighborhood of θ_0 throughout, Jacobian smoothness implies $\widetilde{\mathbf{F}}_r(\theta) \approx \widetilde{\mathbf{F}}_r(\theta_0)$ along the update trajectory (McKinney et al., 2026). Combined with the identity-balanced Fisher decomposition $\widetilde{\mathbf{F}}_r(\theta_0) = \widetilde{\mathbf{F}}_{\text{full}}(\theta_0) - \widetilde{\mathbf{F}}_f(\theta_0)$, the final update rule becomes

$$\Delta\theta = -\eta\Big(\widetilde{\mathbf{F}}_{\text{full}}(\theta_0) - \widetilde{\mathbf{F}}_f(\theta_0)\Big)^{-1}\mathbf{g}_f(\theta). \tag{15}$$

We precompute and cache $\widetilde{\mathbf{F}}_{\text{full}}(\theta_0)$ once at the end of the original fine-tuning; at unlearning time, only $\widetilde{\mathbf{F}}_f(\theta_0)$ needs to be computed from $\mathcal{V}_f$. The update therefore does not require retain images, original training questions, or ground-truth responses.

5 Experiments

5.1 Experimental Setup

Benchmarks and models. We conduct experiments on two benchmarks, **MLLMU-Bench** (Liu et al., 2025a) and **ReMem** (Kwon et al., 2026), which provide complementary evaluation axes that jointly verify whether unlearning is achieved faithfully at multiple levels. **MLLMU-Bench** additionally evaluates the preservation of pretrained knowledge about real-world individuals (the *real/celebrity* split), allowing us to detect collateral damage to general prior knowledge. **ReMem** ensures robust foundational memorization and adopts keyword exact match for more accurate evaluation. It examines forgetting across three axes: surface-level forgetting on in-distribution questions (*forget*), retention depth at the probability level (*exposure*), and generalization to unseen images and paraphrased questions (*test*). We use forget splits of 5% and 10% on both benchmarks with **LLaVA-1.5-7B** (Liu et al., 2024) and **Qwen3-VL-8B-Instruct** (Bai et al., 2025) as base models. Additional benchmark details are in Appendix D.4.

Baselines. Among methods with publicly available official code, we organize baselines into two groups based on the data scope they require. The *retain-utilizing* group, included as a reference, additionally uses retain data and consists of KL_Min (Nguyen et al., 2020), GA_Diff (Liu et al., 2022a), MMUnlearner (Huo et al., 2025), and MANU (Liu et al., 2025b). The *forget-only* group, which uses the same data scope as AIM, consists of GA (Thudi et al., 2022) and NPO (Zhang et al., 2024). Comparing the two groups allows us to assess whether AIM, operating under a strict data scope, can compete with baselines that exploit additional supervision.

Evaluation metrics. On MLLMU-Bench we report VQA Classification Accuracy (Cls) and Generation Performance ($ROUGE\,(R)$) (Lin, 2004) on the forget, retain, and real (celebrity) subsets. On ReMem we report Keyword Exact Match (EM), Generation Performance ($ROUGE$) and the Exposure score (Exp) (Kwon et al., 2026) on the forget, retain, exposure, and test set. Unless specified, all scores in this paper are in %.

Implementation details. All baselines, the vanilla model, and AIM are retrained and evaluated under an identical environment on a single NVIDIA A100 80GB GPU. Vanilla model preparation and baseline unlearning follow prior-work schedules; we observe that forget-only baselines (GA, NPO) tend to collapse rapidly under their schedule and therefore report their results at 1 epoch. AIM is trained with 10 epochs of Stage 1 visual prompt learning, 100 iterations of the Fisher approximation in Stage 2, and 50 epochs of vision-encoder updates in Stage 2. All reported results were obtained using a single fixed random seed. Additional hyperparameters, training details and hardware specifications are provided in Appendix D.5.

5.2 Main Results

Tables 2 and 3 report unlearning results of 5% and 10% splits on LLaVA-1.5-7B and Qwen3-VL-8B. 15% split results are in Appendix D.7 and D.9.

AIM matches retain-utilizing baselines without using retain data and generalizes forgetting to unseen variants. On MLLMU-Bench, our retain and celebrity-prior metrics (Cls_r, $ROUGE_r$, Cls_c, $ROUGE_c$) stay close to vanilla on both models, matching or exceeding *retain-utilizing* baselines such as MMUnlearner and MANU on LLaVA.

	Method	MLLMU-Bench						ReMem				
		$\text{Cls}_f \downarrow$	$\text{ROUGE}_f \downarrow$	$\text{Cls}_r \uparrow$	$\text{ROUGE}_r \uparrow$	$\text{Cls}_c \uparrow$	$\text{ROUGE}_c \uparrow$	$\text{EM}_f \downarrow$	$\text{EM}_r \uparrow$	ROUGE$\uparrow$	Exp$\downarrow$	$\text{EM}_t \downarrow$
	Forget 5%											
	Vanilla	39.2	60.2	39.0	55.4	52.4	24.4	93.3	100.0	100.0	56.9	92.9
retain utilizing	GA_Diff	36.0	53.5	36.1	54.9	48.3	21.6	0.0	47.5	44.5	40.5	0.0
	KL_Min	30.4	27.4	40.9	35.5	56.1	14.3	86.7	98.9	91.2	55.1	100.0
	MMUnlearner	29.6	44.9	38.5	58.0	44.6	29.6	23.3	96.1	98.2	36.4	23.8
	MANU	40.8	55.1	43.8	55.6	53.5	25.7	20.0	96.9	98.5	40.4	23.8
forget only	GA	22.4	15.9	30.7	19.5	52.3	13.8	86.7	99.1	91.1	55.4	100.0
	NPO	32.0	16.3	39.7	22.7	56.3	12.6	86.7	98.3	90.7	55.1	92.9
	Ours	35.2	49.5	38.1	52.3	51.3	26.9	30.0	76.7	88.1	54.4	45.2
	Forget 10%											
	Vanilla	43.2	57.5	38.6	55.4	52.3	24.4	91.7	100.0	100.0	69.1	94.0
retain utilizing	GA_Diff	34.4	50.1	36.1	55.9	49.9	24.4	0.0	88.6	79.6	43.5	0.0
	KL_Min	22.0	36.4	30.0	49.8	53.0	22.8	0.0	5.2	10.2	50.6	3.6
	MMUnlearner	33.2	37.5	36.5	42.4	48.6	21.6	0.0	22.1	61.2	45.2	3.6
	MANU	34.4	54.5	33.9	52.2	48.4	22.5	20.0	98.4	99.4	58.7	25.6
forget only	GA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	2.2	50.7	0.0
	NPO	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.4	6.3	52.6	1.8
	Ours	39.2	49.9	38.5	51.7	51.7	26.1	38.3	78.8	89.2	63.9	35.1

Table 2: **Unlearning results on LLaVA-1.5-7B.** MLLMU-Bench and ReMem across 5%, 10% forget splits. Subscripts denote forget (f), retain (r), celebrity prior (c), and test (t) splits. Baselines are grouped by data scope: retain-utilizing methods access additional retain data, forget-only methods do not. **Ours** uses only forget data.

Method	$\text{Cls}_f \downarrow$	$\text{R}_f \downarrow$	$\text{Cls}_r \uparrow$	$\text{R}_r \uparrow$	$\text{Cls}_c \uparrow$	$\text{R}_c \uparrow$
Forget 5%						
Vanilla	68.0	72.2	69.0	70.8	74.4	43.7
retain-utilizing						
GA_diff	64.8	65.4	67.0	67.8	73.8	42.5
KL_Min	2.4	2.8	66.4	60.3	74.9	39.1
MMUnlearner	48.0	49.7	60.0	57.0	58.0	37.7
MANU	49.6	57.9	61.9	57.5	68.9	40.1
forget-only						
GA	61.6	66.4	65.3	62.7	75.2	42.2
NPO	46.4	55.1	53.7	54.3	73.0	44.0
Ours	53.6	59.8	66.4	65.1	74.5	43.6
Forget 10%						
Vanilla	71.2	72.1	68.7	70.7	74.4	43.7
retain-utilizing						
GA_diff	64.4	61.2	66.0	68.2	73.8	43.7
KL_Min	2.8	1.9	57.7	63.6	71.5	38.1
MMUnlearner	56.8	53.0	58.1	57.2	66.4	39.2
MANU	60.4	56.3	58.5	55.9	71.5	40.7
forget-only						
GA	33.6	47.4	33.8	45.2	65.0	50.4
NPO	0.0	3.1	0.0	3.2	31.3	3.1
Ours	62.4	60.7	65.7	64.3	76.4	42.8

Table 3: **Unlearning results on Qwen3-VL-8B (MLLMU-Bench).** 5%, 10% forget splits. Notation and baseline grouping follows Table 2.

On ReMem (LLaVA), we narrow the gap to *retain-utilizing* methods while still operating in the strict setting: EM_t drops from vanilla's $92.9/94.0$ to $45.2/35.1$ —indicating that forgetting *generalizes* to identities rather than memorizing exact training images. The remaining gap on Exp stems from our Fisher constraint, which caps the per-step update magnitude to preserve retain features and thereby limits how deeply the forget signal can propagate into the internal token distribution. We view this as a natural trade-off in the strict setting, since *retain-utilizing* methods are free of such tight retain-preservation constraints and can drive forget direction more aggressively.

AIM remains stable where *forget-only* baselines collapse. AIM maintains a balanced forget–retain trade-off across all forget ratios on both models, and this stability extends across training iterations (Appendix D.2). This reflects that the Fisher constraint bounds retain drift at every step. In contrast, GA and NPO collapse as iterations accumulate. By the 10% split they either reach $\sim$0 on every MLLMU-Bench column (LLaVA) or drive retain metrics to single digits (NPO on Qwen3-VL), indicating degenerate outputs across all inputs. The difficulty extends to some *retain-utilizing* methods: on ReMem at 10%, KL_Min and MMUnlearner both suffer severe retain collapse.

5.3 Ablation and Further Analysis

Visual perception preservation after unlearning. Section 3.2 predicted that a vision-side intervention should leave visual perception largely intact. In Tab. 4, we verify this *after unlearning* on the visual perception question set on forget-identity images, comparing responses against the pretrained-LLaVA and vanilla model. AIM attains the closest match to vanilla, substantially exceeding the best-performing *retain-utilizing* method and *forget-only* methods. The responses remain coherent and on-topic rather than degenerating into refusals or

Model	pretrained vs		vanilla vs		Appr.
	ROUGE↑	GPT↑	ROUGE↑	GPT↑	GPT↑
Vanilla	58.9	70.2	–	–	96.4
MMUnlearner	52.1	34.7	74.3	33.3	88.4
GA	0.0	0.0	0.0	0.0	0.0
NPO	49.6	59.1	77.2	66.2	99.1
Ours	57.8	68.9	87.5	82.7	97.3

Table 4: **Visual perception preservation after unlearning (MLLMU-Bench, 5%).** Visual perception responses on forget-identity images, scored against the pretrained LLaVA and the vanilla model by ROUGE-L and GPT semantic preservation judgment. Appr.: fraction of coherent, on-topic, non-refusal answers.

Split	Method	EM_f↓	EM_r↑	ROUGE↑	Exp↓	EM_t↓
5%→10%	Vanilla	91.7	100.0	100.0	69.1	94.0
	GA	78.3	96.2	73.0	64.2	87.5
	NPO	75.0	96.0	72.6	62.7	82.7
	Ours	41.7	69.3	84.0	64.7	42.3
10%→15%	Vanilla	94.4	100.0	100.0	65.6	96.0
	GA	0.0	0.1	1.8	53.9	0.0
	NPO	0.0	0.1	1.7	53.3	0.0
	Ours	50.0	71.8	85.2	60.1	56.3

Table 5: **Continual unlearning on ReMem.** Hop $i \rightarrow j$: continually unlearn from the forget i checkpoint over the newly-added identities in forget j.

incoherent outputs —the empirical correlate of our analysis that the forget signal is absorbed at the identity-related region while perception is left structurally untouched.

Extension to continual unlearning. Real-world deletion requests arrive sequentially rather than as a single batch. (Liu et al., 2022b; Gao et al., 2025; Jin et al., 2026) AIM extends to this setting at no additional design cost: the cached $\widetilde{\mathbf{F}}_{\text{full}}(\theta_0)$ is independent of which identities are forgotten, so each new request only requires computing $\widetilde{\mathbf{F}}_f$ for the newly added identities. As shown in Tab. 5, although GA and NPO appear competitive on the $1 \rightarrow 2$ transition, their extremely high EM_t values indicate that they fail to properly forget the target identities. In contrast, on the $2 \rightarrow 3$ transition, AIM still forgets effectively ($EM_f = 50.0$, $EM_t = 56.3$) while preserving retain capacity ($ROUGE = 85.2$, $EM_r = 71.8$), whereas the other methods collapse entirely at the second step.

Loss components and target feature types. We ablate the three Stage 1 loss terms and the target feature choice for forget set. (Tables 6 and 7). Removing $\mathcal{L}_{\text{norm}}$ lets T grow unbounded and the target feature drifts so far that retain collapses along with

			Forget↓		Retain↑		Real↑	
CE	Align	Norm	Cls	ROUGE	Cls	ROUGE	Cls	ROUGE
✓	✓	✗	24.4	47.0	28.3	46.9	21.4	30.0
✓	✗	✓	41.2	58.8	39.7	57.2	52.2	25.6
✗	✓	✓	42.8	57.1	38.7	55.5	52.3	24.4
✓	✓	✓	39.2	49.9	38.5	51.7	51.7	26.1

Table 6: **Stage 1 loss component ablation (MLLMU-Bench, 10%).** Each row keeps a subset of the three loss terms (CE, alignment, norm).

	Forget↓		Retain↑		Celebrity↑	
Method	Cls	ROUGE	Cls	ROUGE	Cls	ROUGE
Blank	31.6	46.7	34.8	49.6	50.2	26.8
Random Vector	44.0	59.4	39.6	58.2	51.4	25.1
IDK (baseline)	39.2	49.9	38.5	51.7	51.7	26.1

Table 7: **Stage 1 target ablation (MLLMU-Bench, 10%).** The IDK-inducing target is replaced with a blank text or random-vector.

forget. Removing $\mathcal{L}_{\text{align}}$ leaves $\Delta \mathbf{z}_V$ arbitrary, so the forget gradient cancels in expectation (Eq. (11)) and metrics remain near vanilla without effective update. Removing $\mathcal{L}_{\text{CE}}$ makes $T \simeq 0$ a trivial solution. The same mechanism applies to the target feature ablation, where a random-vector target bypasses $\mathcal{L}_{\text{align}}$ and fails to forget from the start. IDK and Blank both work because only a coherent, aligned target is required.

Robustness under inference-time deviations. We stress-test AIM under inference-time image perturbations using the 5% forget split of MLLMU-Bench. Specifically, we apply Gaussian noise, JPEG compression, horizontal flipping, and Gaussian blur to the input images. As shown in Table 8, AIM maintains a similar forget–retain tradeoff across these perturbations: forget ROUGE-L changes by at most 4.7 points relative to the clean condition, while retain ROUGE-L decreases by at most 1.8 points. These results indicate that AIM remains stable under common image-side deviations.

6 Conclusion

We studied deletion-time identity unlearning in MLLMs under a strict setting that uses only forget images, constructed questions, and a Fisher statistic precomputed before any forget request, without access to retain images, original questions, or ground-truth answers. To our knowledge, we introduced the first representation-level analysis that directly contrasts identity- and visual perception-related hidden states on the same trained image, revealing distinct clustering structures between the

Image Condition	LLaVA-1.5-7B		Qwen3-VL-8B	
	Forget ↓	Retain ↑	Forget ↓	Retain ↑
Clean	49.5	52.3	59.8	65.1
Gaussian noise	53.4	51.3	60.4	63.6
JPEG compression	50.1	52.7	57.2	63.3
Horizontal flip	51.3	51.2	58.3	63.3
Gaussian blur	54.2	52.6	59.2	63.7

Table 8: **Image-side robustness stress test.** ROUGE-L scores on the 5% forget split of MLLMU-Bench under inference-time image perturbations.

two question types. This finding motivates our two-stage method, which first defines an IDK-inducing target and then internalizes that target through a Fisher-conditioned vision-encoder update, recovering retain protection by decomposing the cached full Fisher rather than observing retain data. The method matches or exceeds retain-utilizing baselines on key metrics under the strict setting. Ablations and further analyses confirm that each loss component is necessary, that the two-stage design is crucial for stable unlearning, and that the method remains robust across forget ratios, visual perception evaluations, and continual deletion scenarios.

Limitations

While AIM achieves retain-free unlearning under the strict setting, we acknowledge several limitations. First, as AIM updates only the vision encoder while keeping the language model fixed, it does not remove identity knowledge in the text-only modality. Broader cross-modal deletion can be achieved by combining AIM with a complementary LLM-unlearning method. Second, the full Fisher must be precomputed during fine-tuning when retain images are implicitly accessible, so the framework cannot be applied directly to models without a precomputed Fisher. Third, we treat the retain Fisher as fixed at its initial value, an approximation that may break down under longer schedules or larger steps. Fourth, AIM requires small learning rates, which may limit how deeply the forget signal propagates through the token distribution and may contribute to the gap on ReMem's Exposure metric. Finally, although AIM largely preserves the forget–retain trade-off under the tested image-side perturbations, our robustness evaluation does not cover broader prompt-side, cross-modal, or adaptive attacks. Evaluating AIM under more diverse attack settings and developing MLLM unlearning methods that require neither retain data nor a precomputed cache at deletion time while remaining robust at inference time are important directions for future work.

Acknowledgments

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korean government (MSIT) (No. RS-2025-02263031, Development of On-Device AI Cooperative Actions (Perception, Decision-Making, and Response) Among Networked Devices).

References

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Costas Bekas, Effrosyni Kokiopoulou, and Yousef Saad. 2007. An estimator for the diagonal of a matrix. *Applied numerical mathematics*, 57(11-12):1214–1229.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2021. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE.

Chengyi Cai, Zesheng Ye, Peike Li, Bo Han, Jianzhong Qi, and Feng Liu. 2026. Visual-guided key-token regularization for multimodal large language model unlearning. *arXiv preprint arXiv:2601.22020*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pages 267–284.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, and 1 others. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pages 2633–2650.

Haokun Chen, Jianing Li, Yao Zhang, Jinhe Bi, Yan Xia, Jindong Gu, and Volker Tresp. 2026. Auvic: Adversarial unlearning of visual concepts for multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 30218–30225.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Xinwen Cheng, Zhehao Huang, Wenxin Zhou, Zhengbao He, Ruikai Yang, Yingwen Wu, and Xiaolin Huang. 2024. Remaining-data-free machine unlearning by suppressing sample contribution. *arXiv preprint arXiv:2402.15109*.

Jack Foster, Stefan Schoepf, and Alexandra Brintrup. 2024. Loss-free machine unlearning. *arXiv preprint arXiv:2402.19308*.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. 2025. On large language model continual unlearning. In *International Conference on Learning Representations*, volume 2025, pages 101772–101801.

Arpit Garg, Hemanth Saratchandran, and Simon Lucey. 2025. Sineproject: Machine unlearning for stable vision language alignment. *arXiv preprint arXiv:2511.18444*.

Jiahui Geng and Qing Li. 2025. Sauce: Selective concept unlearning in vision-language models with sparse autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3023–3033.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9304–9312.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7190–7206.

Michael F Hutchinson. 1989. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076.

Hyundong Jin, Dongyoon Han, and Eunwoo Kim. 2026. Which concepts to forget and how to refuse? decomposing concepts for continual unlearning in large vision-language models. *arXiv preprint arXiv:2603.21484*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kangwook Ko, Jaehyuk Jang, Wonjun Lee, Hee-Seon Kim, and Changick Kim. 2026. Where identity lives: Localized, retain-free identity unlearning in mllms. In *Findings of the Association for Computational Linguistics: EMNLP 2026*.

JuneHyoung Kwon, MiHyeon Kim, Eunju Lee, Jung-Min Yun, Byeonggeuk Lim, and YoungBin Kim. 2026. Before forgetting, learn to remember: Revisiting foundational learning failures in lvlm unlearning benchmarks. *arXiv preprint arXiv:2605.03759*.

Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. 2026. Cross-modal unlearning via influential neuron path editing in multimodal large language models. 40(42):35589–35597.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, and 1 others. 2024. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*.

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

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025a. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4105–4135.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. 2025b. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5913–5933.

James Martens and Roger Grosse. 2015. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR.

James Martens and 1 others. 2010. Deep learning via hessian-free optimization. In *Icml*, volume 27, pages 735–742.

Lev McKinney, Anvith Thudi, Juhan Bae, Tara Rezaei, Nicolas Papernot, Sheila A McIlraith, and Roger Grosse. 2026. Gauss-newton unlearning for the llm era. *arXiv preprint arXiv:2602.10568*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. 2025. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 57172–57189.

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

Razvan Pascanu and Yoshua Bengio. 2013. Revisiting natural gradient for deep networks. *arXiv preprint arXiv:1301.3584*.

Shaofei Shen, Chenhao Zhang, Yawen Zhao, Alina Bialkowski, Weitong Chen, and Miao Xu. 2024. Label-agnostic forgetting: A supervision-free unlearning in deep models. In *International Conference on Learning Representations*, volume 2024, pages 54217–54235.

Changchang Sun, Ren Wang, Yihua Zhang, Jinghan Jia, Jiancheng Liu, Gaowen Liu, Yan Yan, and Sijia Liu. 2024. Forget vectors at play: Universal input perturbations driving machine unlearning in image classification. *arXiv preprint arXiv:2412.16780*.

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

Yuhang Wang, Zhenxing Niu, Haoxuan Ji, Guangyu He, Haichang Gao, and Gang Hua. 2025. Mllm machine unlearning via visual knowledge distillation. *arXiv preprint arXiv:2512.11325*.

Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. 2020. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE.

Zhen Zeng, Leijiang Gu, Zhangling Duan, Feng Li, Zenglin Shi, Cees GM Snoek, and Meng Wang. 2025. Towards benign memory forgetting for selective multimodal large language model unlearning. *arXiv preprint arXiv:2511.20196*.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. Pmlr.

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

Xianren Zhang, Hui Liu, Delvin Ce Zhang, Xianfeng Tang, Qi He, Dongwon Lee, and Suhang Wang. 2025. Sua: Stealthy multimodal large language model unlearning attack. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11241–11254.

Appendix

A Method Derivations and Implementation 13
- A.1 Identity-Balanced Fisher Aggregation 13
- A.2 Image-level vs. Identity-level aggregation 13
- A.3 Mixture-of-identities interpretation 13
- A.4 KKT closed-form update 13
- A.5 Fisher Implementation: Damping, Diagonal Approximation, Effective LR 14
- A.6 Why Two Stages? Limitations of Single-Stage Unlearning 14

B Empirical Justification of the Method 15
- B.1 Identity-Balanced Fisher: Decomposition Accuracy and Update Stability 15
- B.2 Fisher stability across the unlearning update 15
- B.3 Gradient Conflict Without the Fisher Constraint 16

C Analysis Extensions and Stage 1 Studies 16
- C.1 Quantitative Verification of Visual Perception–Identity Separation . . 16
- C.2 NMI/ARI Quantification 17
- C.3 Layer-wise Clustering Analysis . . 17
- C.4 Clustering on Unseen Images . . . 17
- C.5 Model type and tuning-recipe robustness 18
- C.6 Pre- vs. Post-Unlearning Clustering 18
- C.7 Held-out Identity Questions 19
- C.8 Visual Feature Intervention Design Choices 19
- C.9 Visual Prompt Transfer Experiment 19
- C.10 Visual Prompt Visualization . . . 19

D Additional Experiments and Reproducibility 19
- D.1 Related-Method Style Comparison 19
- D.2 Training Stability over Iterations . 20
- D.3 Qualitative Results 20
- D.4 Models, Benchmarks, and Split Coverage 21
- D.5 Training Details and Hyperparameters 22
- D.6 Stage 1 Loss-Weight Ablation . . 23
- D.7 LLaVA-1.5-7B at Forget 15% . . . 23
- D.8 Qwen3-VL-8B Implementation . . 23
- D.9 Additional Evaluation Results on Various Models 23
- D.10 General VQA Performance after Unlearning 23
- D.11 Continual Unlearning Setup . . . 23

E AI Assistant Usage Statement 24

F License of Datasets and Models 24

A Method Derivations and Implementation

A.1 Identity-Balanced Fisher Aggregation

The atomic unit of unlearning in this paper setting is an identity, not an individual image. When the deletion request specifies a forget identity $c \in \mathcal{C}_f$, the goal is to remove image-conditioned access to identity-specific information about c from the forget images, regardless of how many images of c appear in the training set. Conversely, preserving a retain identity $c' \in \mathcal{C}_r$ requires preserving all its appearances, including unseen images. This asymmetry between identity-level requests and image-level fine-tuning data motivates an identity-balanced reformulation of the Fisher aggregation.

A.2 Image-level vs. Identity-level aggregation

The natural image-level Fisher aggregation

$$\mathbf{F}_r^{\text{img}}(\theta) = \sum_{V \in \mathcal{V}_r} \mathbf{F}_V(\theta) \tag{16}$$

weights each retain image equally. When identities differ in their training-image counts, this aggregation effectively prioritizes the preservation of high-image-count identities and under-protects low-image-count ones. Our identity-balanced aggregation

$$\widetilde{\mathbf{F}}_r(\theta) = \sum_{c \in \mathcal{C}_r} \frac{1}{|\mathcal{V}_c|} \sum_{V \in \mathcal{V}_c} \mathbf{F}_V(\theta) \tag{17}$$

gives each identity equal weight by averaging within the identity before summing across identities. The same form is used for $\widetilde{\mathbf{F}}_f$ and $\widetilde{\mathbf{F}}_{\text{full}}$.

A.3 Mixture-of-identities interpretation

Eq. (17) can be derived from a mixture-of-identities likelihood. If the data-generating distribution is a uniform mixture over identities

$$p(V \mid \theta) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} p(V \mid c, \theta), \tag{18}$$

the population Fisher takes the form

$$\mathbf{F}(\theta) = \sum_{c \in \mathcal{C}} \frac{1}{|\mathcal{C}|} \mathbb{E}_{V \sim p(V|c)}[\mathbf{F}_V(\theta)]. \tag{19}$$

Replacing the per-identity expectation by its sample average over $\mathcal{V}_c$ recovers Eq. (17) up to a multiplicative constant, which is absorbed into the Lagrange multiplier λ. The decomposition $\widetilde{\mathbf{F}}_r = \widetilde{\mathbf{F}}_{\text{full}} - \widetilde{\mathbf{F}}_f$ holds by linearity of the identity-level aggregation.

A.4 KKT closed-form update

We derive the closed-form update of Eq. (14) step by step.

Step 1: Linearization of the forget objective. A first-order Taylor expansion of $\mathcal{L}_f$ around the current θ gives

$$\mathcal{L}_f(\theta + \Delta\theta) \approx \mathcal{L}_f(\theta) + \mathbf{g}_f(\theta)^\top \Delta\theta, \tag{20}$$

which is linear in $\Delta\theta$.

Step 2: Quadratic retain-drift constraint. As shown in Section 4.3, the per-image retain feature drift admits the quadratic approximation

$$\|E_v(V; \theta + \Delta\theta) - E_v(V; \theta)\|^2 \approx \Delta\theta^\top \mathbf{F}_V(\theta) \Delta\theta. \tag{21}$$

Aggregated identity-level (Appendix A.1), we impose the per-step retain-drift budget

$$\Delta\theta^\top \widetilde{\mathbf{F}}_r(\theta) \Delta\theta \le \epsilon_r. \tag{22}$$

Step 3: Constrained optimization. Combining Steps 1 and 2, the per-step update solves

$$\min_{\Delta\theta} \ \mathbf{g}_f(\theta)^\top \Delta\theta \ \text{ s.t. } \ \Delta\theta^\top \widetilde{\mathbf{F}}_r(\theta) \Delta\theta \le \epsilon_r. \tag{23}$$

The objective contains the only direction-relevant information (the forget gradient), and the constraint encodes the strict setting's retain protection.

Step 4: Lagrangian and stationarity. The Lagrangian is

$$\mathcal{L}(\Delta\theta, \lambda) = \mathbf{g}_f^\top \Delta\theta + \lambda\Big(\Delta\theta^\top \widetilde{\mathbf{F}}_r \Delta\theta - \epsilon_r\Big), \quad \lambda \ge 0. \tag{24}$$

Stationarity in $\Delta\theta$ requires

$$\nabla_{\Delta\theta} \mathcal{L} = \mathbf{g}_f + 2\lambda \widetilde{\mathbf{F}}_r \Delta\theta = 0, \tag{25}$$

giving

$$\Delta\theta^* = -\frac{1}{2\lambda} \widetilde{\mathbf{F}}_r^{-1} \mathbf{g}_f. \tag{26}$$

Step 5: Effective update form. From Eq. (26), the update direction is

$$\Delta\theta^* = -\frac{1}{2\lambda} \widetilde{\mathbf{F}}_r^{-1} \mathbf{g}_f. \tag{27}$$

In implementation, the scalar factor $(2\lambda)^{-1}$ is absorbed into an effective learning rate η, yielding

$$\Delta\theta = -\eta \, \widetilde{\mathbf{F}}_r^{-1} \mathbf{g}_f. \tag{28}$$

Thus, the essential closed-form structure is the Fisher-preconditioned forget direction $\widetilde{\mathbf{F}}_r^{-1} \mathbf{g}_f$.

Method	Forget↓		Retain↑		Celebrity↑	
	Cls	ROUGE	Cls	ROUGE	Cls	ROUGE
Vanilla	43.2	57.5	38.6	55.4	52.3	24.4
single-stage	32.8	38.7	29.9	36.0	19.3	26.1
IDK (ours)	39.2	49.9	38.5	51.7	51.7	26.1

Table 9: **Single-stage vs. two-stage unlearning (MLLMU-Bench, 10%).** Single-stage applies all three Stage 1 losses directly to the vision encoder in a single joint update, skipping the target-then-path decoupling of our two-stage method (IDK). The single-stage variant collapses across retain and celebrity metrics, and its lower forget scores reflect general degradation rather than selective forgetting.

A.5 Fisher Implementation: Damping, Diagonal Approximation, Effective LR

We adopt three standard modifications for practical implementation. First, to ensure numerical stability of $\widetilde{\mathbf{F}}_r^{-1}$ when individual Fisher entries approach zero, we apply Levenberg–Marquardt-style damping $\widetilde{\mathbf{F}}_r \leftarrow \widetilde{\mathbf{F}}_r + \delta\mathbf{I}$ (Martens et al., 2010; Martens and Grosse, 2015; Pascanu and Bengio, 2013), with δ set to the mean of the (undamped) Fisher diagonal. Second, for computational tractability, we use a diagonal approximation of $\widetilde{\mathbf{F}}_r$ (Kirkpatrick et al., 2017; Zenke et al., 2017; Aljundi et al., 2018), estimated via Hutchinson's stochastic estimator (Hutchinson, 1989; Bekas et al., 2007; Yao et al., 2020), which reduces the matrix inverse to element-wise inversion. Third, to match the effective per-step update magnitude used by other training procedures, we set $\eta = \mathbb{E}[\widetilde{\mathbf{F}}_r] \cdot 10^{-5}$, making the per-step parameter shift comparable to standard fine-tuning at $\mathrm{lr} = 10^{-5}$.

A.6 Why Two Stages? Limitations of Single-Stage Unlearning

AIM first learns a visual prompt T in Stage 1 to construct a forget target, then updates the vision encoder in Stage 2 toward this target under the Fisher constraint. A natural alternative is to skip Stage 1 and apply all three loss terms ($\mathcal{L}_{\mathrm{CE}}$, $\mathcal{L}_{\mathrm{align}}$, $\mathcal{L}_{\mathrm{norm}}$) directly to the vision encoder in a single joint update, with Fisher division applied as usual. We show this single-stage variant fails both empirically and conceptually.

Empirical performance. Table 9 compares the two-stage method (IDK) and the single-stage joint variant on MLLMU-Bench at the 10% forget split. The single-stage variant collapses across all metrics, including retain and celebrity-prior accuracy. The apparent forget improvement reflects general model degradation rather than selective forgetting.

Loss trajectories. Figure 3 shows the four loss components during the single-stage update across ablation combinations. Two patterns emerge. *First, the three loss terms pull the encoder in conflicting directions.* Removing any single component causes its own loss to grow rather than remain bounded as training proceeds. Dropping both $\mathcal{L}_{\mathrm{align}}$ and $\mathcal{L}_{\mathrm{norm}}$ lets $\mathcal{L}_{\mathrm{norm}}$ explode and oscillate without bound, and similar behavior appears when each term is removed individually.

Second, the total loss spikes sharply at the start of training before settling. Even though the Fisher constraint already keeps per-step parameter shifts small, the loss surface around the vanilla weights is so unstable that the first few steps trigger large loss swings before training settles. The forget set, which supplies the supervision signal, can adjust as training proceeds. The retain set, by contrast, receives no corrective signal under the strict setting, so any early drift on retain features becomes permanent. Single-stage optimization thus pays for the early instability with irrecoverable retain damage.

Why the single-stage approach fails. We attribute this failure to a fundamental difference in target specificity. In our two-stage method, Stage 1 commits to a specific forget target, and Stage 2 then guides the vision encoder toward this fixed target through a single feature-matching loss under the Fisher constraint. The destination is a fixed point. When the Fisher constraint blocks an update direction, the encoder progresses as far as it can along the remaining directions and stops there. Convergence is well-defined because the direction is well-defined.

In the single-stage approach, the encoder is asked to reach *any point* in the set of feature configurations satisfying all three losses jointly. This target is a broad region rather than a single point. When the Fisher constraint blocks one path into this region, the optimizer could reroute to a different path.

Two stages thus separate two genuinely different problems, constructing a precise forget destination using complex loss combinations and routing the encoder toward that fixed destination. Combining these into one objective conflates them and breaks the optimization.

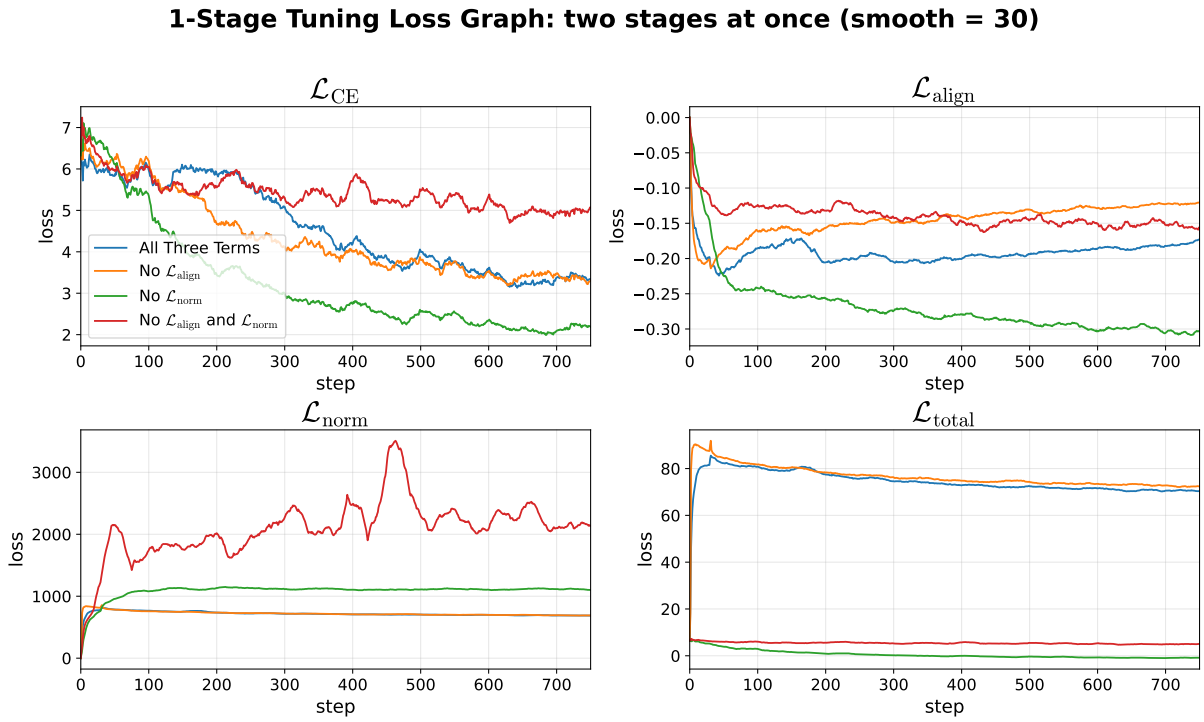


Figure 3: **Loss trajectories during the single-stage joint update on MLLMU-Bench (10% forget).** Each panel shows one of the four loss components ($\mathcal{L}_{\text{CE}}$, $\mathcal{L}_{\text{align}}$, $\mathcal{L}_{\text{norm}}$, $\mathcal{L}_{\text{total}}$) under four ablation combinations of the auxiliary terms. Removing any term lets its own loss grow rather than stay bounded, and the total loss spikes sharply at the start before settling, causing irreversible retain damage under the strict setting.

	Fisher decomposition		**Unlearning (pre vs. post)**	
Forget	cos-sim	rel ℓ_2	Retain cos	Forget cos
5 %	0.9882	0.160	0.781	0.621
10 %	0.9873	0.165	0.668	0.606
15 %	0.9868	0.167	0.787	0.661

Table 10: **Identity-balanced Fisher: decomposition accuracy and stability.** Left: $\widetilde{\mathbf{F}}_{\text{full}} - \widetilde{\mathbf{F}}_f$ matches the directly computed $\widetilde{\mathbf{F}}_r$. Right: diagonal Fisher cosine before and after unlearning. LLaVA-1.5-7B on ReMem.

B Empirical Justification of the Method

B.1 Identity-Balanced Fisher: Decomposition Accuracy and Update Stability

ReMem is constructed with balanced per-identity image counts, so to test the value of identity-level aggregation under uneven counts we artificially induce the imbalance. We compute $\widetilde{\mathbf{F}}_{\text{full}}$ and $\widetilde{\mathbf{F}}_f$ on the original (balanced) training images, and the directly-computed $\widetilde{\mathbf{F}}_r$ on a test-image subset with intentionally reduced and unequal per-identity image counts. Table 10 (Left) reports the cosine similarity and relative ℓ_2 error between the decomposed $\widetilde{\mathbf{F}}_{\text{full}} - \widetilde{\mathbf{F}}_f$ and this directly-computed $\widetilde{\mathbf{F}}_r$ at three forget ratios. Across all settings, the decomposition yields high cosine and modest ℓ_2 error. These indicate the two matrices align in direction, with small magnitude mismatches absorbed into the Lagrange multiplier λ at solve time (Appendix A.4). Identity-level aggregation remains accurate even when retain-side and forget-side image counts diverge, validating its use as a substitute for direct $\widetilde{\mathbf{F}}_r$ computation, which would require access to retain images.

B.2 Fisher stability across the unlearning update

The cached $\widetilde{\mathbf{F}}_r(\theta_0)$ approximates the running $\widetilde{\mathbf{F}}_r(\theta)$ only if the per-step update keeps θ close enough to θ_0 for Jacobian regularity to hold. Table 10 (Right) reports the cosine similarity between the diagonal vision-encoder Fisher before and after unlearning. The retain Fisher cosine stays in the range 0.67–0.79 across forget ratios, indicating that retain-side parameter importance shifts modestly along the trajectory. The forget Fisher cosine is lower (0.61–0.66), reflecting that forget-relevant parameters are actively redistributed by the update. This asymmetry is exactly the behavior our Fisher constraint is designed to permit. The constraint binds on the retain side, keeping the retain Fisher stable enough

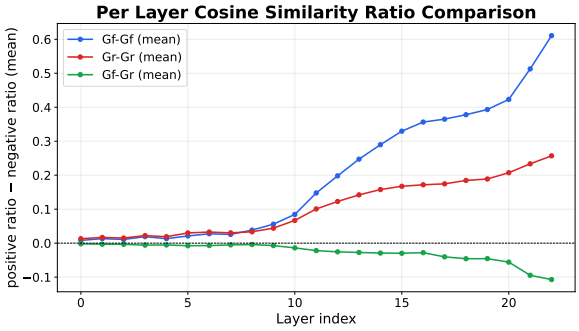


Figure 4: **Per-layer gradient cosine on the vanilla vision encoder.** Within-group cosines (G_f-G_f and G_r-G_r) rise in later layers, indicating coherent update directions within each group. The across-group cosine (G_f-G_r) becomes negative in late layers, showing that a forget-direction update simultaneously pushes retain features in the opposite direction. LLaVA-1.5-7B, MLLMU-Bench, 15% forget split.

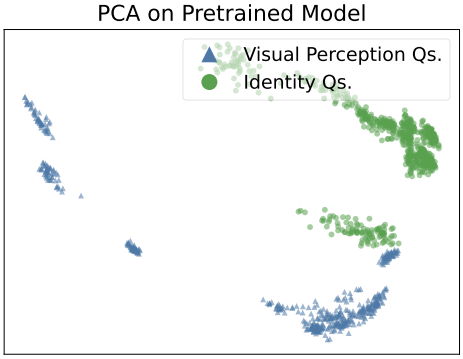


Figure 5: **PCA projection of pretrained LLaVA hidden states.** Each point is the last-layer LLM hidden state of a single question. Visual perception questions (blue) and identity questions (green) appear visually distinguishable in 2D PCA even in the pretrained model, motivating the quantitative analysis in Table 11.

Model	$\lvert\cos(\mathrm{PC1}, \mathbf{d})\rvert$	Fisher (PC1)	Raw ℓ_2 dist.
Pretrained	0.647	1.52	40.7
Vanilla	0.985	19.38	46.4

Table 11: **Principal component alignment with identity–visual perception separation.** $\lvert\cos(\mathrm{PC1}, \mathbf{d})\rvert$ is the cosine similarity between PC1 and the centroid-difference vector $\mathbf{d}$ in the original 4096-dimensional space. Fisher denotes the Fisher criterion (between-group / within-group variance) along PC1.

that the cached approximation remains valid, while the forget side is free to shift as the update internalizes the deletion request.

B.3 Gradient Conflict Without the Fisher Constraint

To isolate the role of the Fisher constraint in Stage 2, we remove it and update the vision encoder directly with the forget feature loss alone. We measure three cosine similarities of per-image gradients at the first update step on LLaVA-1.5-7B (MLLMU-Bench, 15% forget):

- G_f–G_f: forget–forget gradient cosine (within-group),
- G_r–G_r: retain–retain gradient cosine (within-group),
- G_f–G_r: forget–retain gradient cosine (across-group).

Result. Figure 4 shows that both within-group cosines (G_f–G_f and G_r–G_r) are high across all layers, indicating that gradients within each group agree on a coherent update direction. In contrast, the across-group cosine G_f–G_r is consistently negative (layer-wise mean ≈ -0.12), and the negativity intensifies in later encoder layers. This means any encoder update along the forget direction simultaneously moves retain features in the opposite direction.

Why this motivates the Fisher constraint. In the strict setting, no retain-side signal is available to correct this conflict during training, so a naive forget-only update accumulates irreversible retain drift with each step. Our Fisher constraint (Section 4.3) intercepts this accumulation at the source by bounding the per-step retain drift even though no retain images are observed. The cached full Fisher provides the retain-direction information that would otherwise require explicit retain data, recovering retain protection without violating the strict setting.

C Analysis Extensions and Stage 1 Studies

C.1 Quantitative Verification of Visual Perception–Identity Separation

The PCA visualizations in Section 3.1 indicate that visual perception and identity responses occupy distinct regions in the vanilla SFT model. A natural concern is whether this separation already exists in the pretrained model and is inherited rather than learned. Indeed, when we project pretrained hidden states with the same PCA setup (Fig. 5), the two groups also appear visually distinguishable, indi-

cating that visual inspection alone cannot answer this question.

Table 11 provides the quantitative answer. Through SFT, the representation undergoes two complementary changes. First, the two group centroids move further apart in the raw dimension space (ℓ_2 distance $40.7 \rightarrow 46.4$). Second, the dominant axis of variation reorients so that the principal component nearly coincides with the group-separation direction ($|\cos(\mathrm{PC1}, \mathbf{d})|$ $0.647 \rightarrow 0.985$, Fisher criterion $1.52 \rightarrow 19.38$). Hence, the vanilla PCA in Section 3.1 does not merely happen to show two clusters; SFT both spreads the two groups apart and aligns the principal component with the visual perception–identity contrast, so projecting these jointly onto PC1–PC2 naturally produces the observed separation.

C.2 NMI/ARI Quantification

Section 3.1 reports the qualitative clustering structure of identity and visual-perception hidden states via PCA and t-SNE. Here we quantify this structure with Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI) on MLLMU-Bench (5% forget split), and extend the analysis to (i) post-unlearning models and (ii) held-out identity questions whose templates were not seen during vanilla fine-tuning.

Setup. For each question subset (seen identity questions, held-out identity questions, and visual perception questions), we cluster the raw last-layer LLM hidden states with k-means and compute NMI and ARI against two reference labelings of the same hidden states: *person-ID*, which image the question concerns, and *question template*, which question pattern the input follows. For each labeling, k-means uses k equal to the number of distinct labels in that labeling.

Metric interpretation. NMI lies in $[0, 1]$ and ARI in $[-1, 1]$, both equal to 1 for identical clusterings and 0 for chance-level agreement (negative ARI indicates worse than chance). A high score against *person-ID* on a question subset means the embedding groups primarily by depicted identity, while a high score against *question template* means it groups by question content. Comparing the two scores on the same subset reveals which axis the model uses to organize that question type.

		Visual Perception		Identity Question	
Model	**Layer**	**NMI**	**ARI**	**NMI**	**ARI**
LLaVA-1.5-7B	12	0.390	-0.020	0.294	-0.021
	24	0.365	-0.019	0.481	0.150
	32 (last)	0.405	0.009	**0.935**	**0.820**
Qwen3-VL-8B	12	0.551	0.149	0.535	0.229
	24	0.358	-0.023	0.302	-0.021
	36 (last)	0.440	0.027	**0.990**	**0.967**

Table 12: **Layer-wise clustering of hidden states.** NMI and ARI quantify the alignment between representation clusters and visual-perception or identity-question labels at selected intermediate and final LLM layers.

Figure 6: **Clustering on images not seen during fine-tuning.** Last-layer LLM hidden states of the MLLMU-Bench-finetuned vanilla model, evaluated on ReMem images that were never seen during training. Visual perception questions still cluster strongly by question template, but identity questions no longer cluster strongly by personal identity. The identity clustering observed in Section 3.1 is therefore a fine-tuning artifact, not a property of the underlying representation.

C.3 Layer-wise Clustering Analysis

To assess the sensitivity of our clustering analysis to layer choice, we repeat the analysis at two selected intermediate layers and the final layer. As shown in Table 12, identity-question clustering is weaker at the intermediate layers, but becomes sharply pronounced at the final layer for both models. In contrast, visual-perception clustering does not show a consistent late-layer increase, and its NMI and ARI remain low across layers.

C.4 Clustering on Unseen Images

Expanding the analysis on Section 3.1, we conduct a control experiment using the same vanilla model fine-tuned on MLLMU-Bench but evaluated on ReMem images that were never seen during training. As shown in Fig. 6, for visual perception questions, we observe NMI/ARI values comparable to those obtained on MLLMU-Bench. In contrast, for identity questions, NMI/ARI drops substantially and

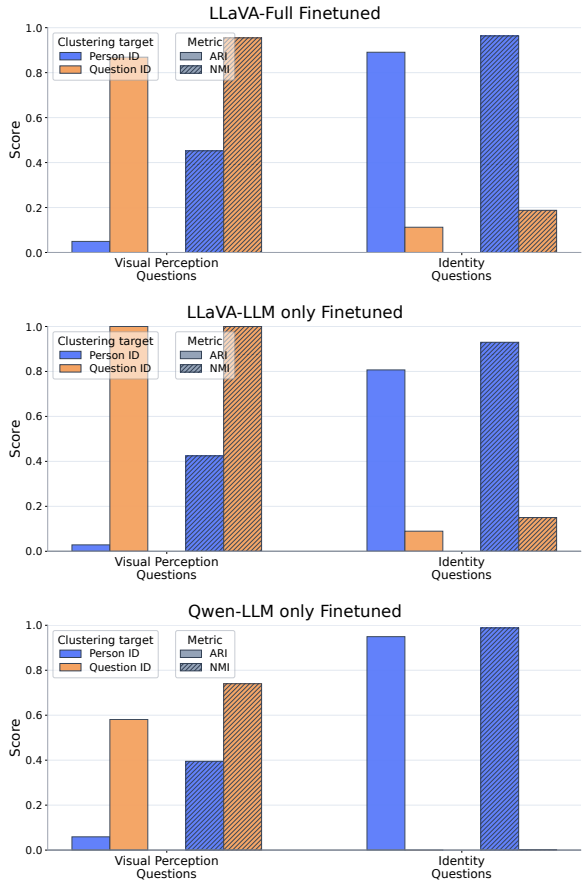


Figure 7: **Clustering robustness across model family and tuning scope (MLLMU-Bench, 5%).** Top: LLaVA-1.5-7B with LoRA on full vision+LLM SFT. Middle: LLaVA-1.5-7B with LLM-only LoRA tuning. Bottom: Qwen3-VL-8B with LLM-only LoRA tuning. All three exhibit the same pattern: identity questions cluster by image identity, visual perception questions cluster by question template.

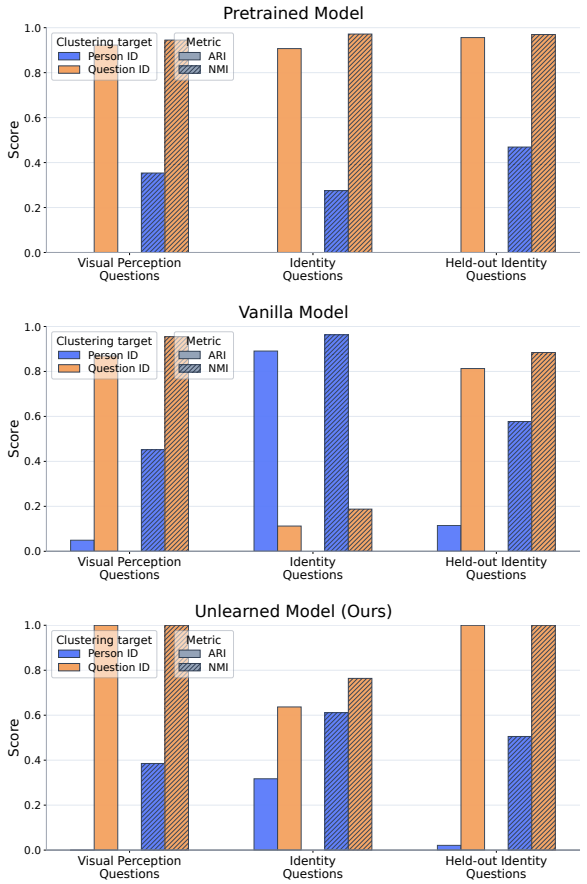


Figure 8: **Clustering pattern across pretrained, vanilla, and post-unlearning models (MLLMU-Bench, 5%).** Top: pretrained LLaVA groups all subsets by question template with negligible identity structure. Middle: vanilla SFT develops strong image-identity clustering on identity questions, while held-out identity and visual perception questions stay close to pretrained. Bottom: after unlearning, image-identity clustering on identity questions drops sharply and question-template grouping rises, while held-out identity and visual perception questions show no noticeable change.

identity-level clusters fail to form, showing that identity-by-image clustering emerges only for images the model has actually been fine-tuned on. This asymmetry confirms that the identity clusters observed in Section 3.1 are a signature of learned, identity-bound representations introduced by fine-tuning—precisely the representational target that an identity-unlearning method must dissolve.

C.5 Model type and tuning-recipe robustness

We first verify that the clustering signature is not an artifact of a specific vanilla model. Figure 7 compares three vanilla baselines: the public HuggingFace LLaVA-1.5-7B checkpoint, in which both the vision encoder and the LLM are SFT-tuned (*LLaVA-Full Finetuned*); an in-house LoRA fine-tune of LLaVA on the LLM Q/K/V/O projections only (*LLaVA-LLM only Finetuned*); and a parallel LLM-only LoRA fine-tune of Qwen3-VL-8B (*Qwen-LLM only Finetuned*). All three exhibit the same clustering pattern. Identity questions cluster by person identity, while perception questions cluster by question template. The signature is stable across model family and tuning scope.

C.6 Pre- vs. Post-Unlearning Clustering

Figure 8 compares the pretrained LLM, the vanilla SFT model, and the post-unlearning model. In the pretrained model, all subsets group almost entirely by question template, with negligible person-ID structure. In the vanilla SFT model, identity questions develop strong image-identity clustering, while held-out identity questions and visual per-

Method	EM_f↓	EM_r↑	ROUGE↑	Exp↓	EM_t↓
VP to IDK (baseline)	38.3	78.8	89.2	63.9	35.1
VP to Blank text	40.0	74.7	87.1	63.4	33.9
Vision Encoder to IDK	43.0	77.6	60.1	51.2	60.4

Table 13: **Stage 1 intervention design choices (LLaVA-1.5-7B, ReMem, 10%).** Two axes are varied: intervention type (visual prompt T vs. direct vision-encoder fine-tuning) and target response type (IDK refusal vs. blank text). All three variants achieve comparable forget/retain trade-offs, supporting the claim in Section 3.2 that the selective-transfer effect is robust to specific design choices within this family.

ception questions remain similar to the pretrained model. After unlearning, image-identity clustering on identity questions drops sharply while question-template grouping rises, and held-out identity questions and visual perception questions show no noticeable change.

C.7 Held-out Identity Questions

We additionally evaluate on a held-out identity question set whose templates do not appear in SFT data (rightmost column of Fig. 8). Question-template clustering on these unseen questions remains intact after unlearning, indicating that AIM affects only the SFT-internalized identity signal and leaves untrained representations untouched.

C.8 Visual Feature Intervention Design Choices

In Section 3.2, we chose a visual prompt T with an IDK target as the Stage 1 intervention for efficiency, and claimed that other design choices in the same family work similarly. We verify this on LLaVA-1.5-7B with ReMem at the 10% forget split, varying two axes: *intervention type* (visual prompt vs. direct fine-tuning of the vision encoder) and *target response type* (IDK refusal vs. blank text). Table 13 reports three combinations: VP to IDK (default AIM), VP to Blank, and Encoder to IDK. All three reach broadly comparable forget/retain trade-offs. The near-equivalence between IDK and Blank targets is also consistent with our broader observation that $\mathcal{L}_{\mathrm{CE}}$ functions as a directional signal toward a coherent forget representation rather than as a literal inducer of IDK strings, so swapping the target text leaves the underlying Stage 2 trajectory largely unchanged.

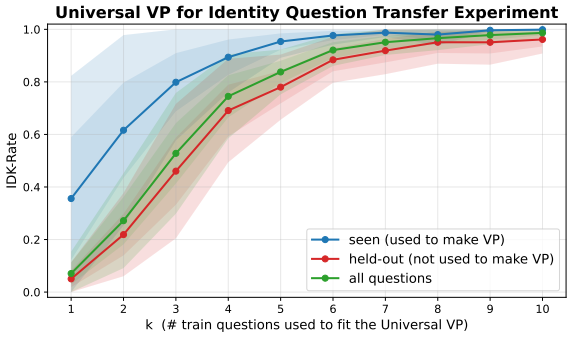


Figure 9: **Visual prompt transfer saturation.** IDK-rate on the full identity-question set as the number of questions used to learn T varies from 1 to 10. Blue: questions used during VP learning. Red: held-out questions. Green: all identity questions. The rate saturates around 8 questions, which we adopt as the default throughout the main experiments.

C.9 Visual Prompt Transfer Experiment

Section 3.2 showed that a T learned on a small handful of identity questions can transfer to other templates and induce IDK responses on unseen questions. Here we characterize this transfer at the dataset level. Figure 9 plots the IDK elicitation rate on the full identity-question set as the number of training questions used to learn T increases. The rate saturates at around 8 questions, beyond which additional training data offers no further benefit. We therefore use 8 questions to learn T throughout our experiments.

C.10 Visual Prompt Visualization

Figure 10 visualizes the learned visual prompt T on MLLMU-Bench. Because MLLMU-Bench consists of close-up face images, the learned T acquires shapes resembling facial landmarks, reflecting the structure of the forget-set images it is optimized against.

D Additional Experiments and Reproducibility

D.1 Related-Method Style Comparison

AIM updates only the vision encoder, leveraging the distinction between identity-related and visual-perception regions established in Section 3, and internalizes a visual-feature-level intervention through a Fisher-based constraint. As discussed in Section 2, the underlying ingredients (feature-level intervention and Fisher-style constraints) have been explored in vision-only models and LLM-only unlearning, though never combined under strict

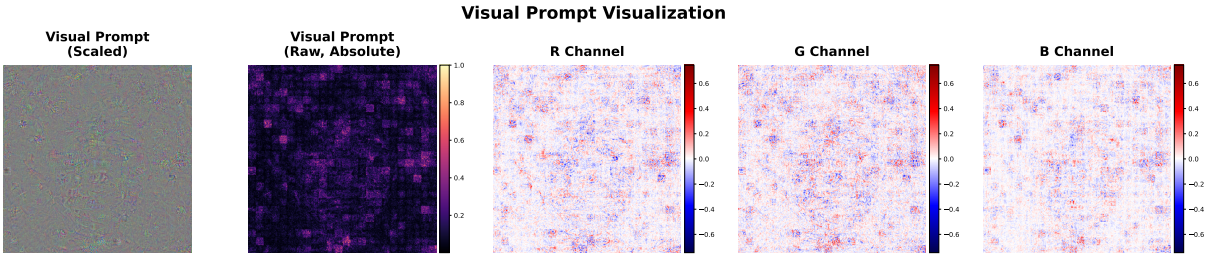


Figure 10: **Visualization of the learned visual prompt T on MLLMU-Bench.** From left to right: T rescaled to the visible range, raw absolute magnitudes, and per-channel (R, G, B) maps. Because MLLMU-Bench images are close-up portraits, T acquires face-landmark-like structure reflecting the geometry of the forget-set images.

Method	EM_f↓	EM_r↑	ROUGE↑	Exp↓	EM_t↓
Vanilla	100.0	100.0	100.0	68.2	100.0
(a) Golatkar et al. (2020)	96.8	100.0	100.0	100.0	96.4
(b) McKinney et al. (2026)	72.0	79.7	61.2	56.5	61.0
(c) Li et al. (2024)	69.5	100.0	100.0	55.5	68.8
Ours	51.1	72.5	85.6	63.2	48.0

Table 14: **Related-method comparison adapted to the MLLM setting (LLaVA-1.5-7B, ReMem, 15%).** Three baselines drawn from vision-only and LLM-only unlearning literature, referred to as (a), (b), (c) in the surrounding text. Notation follows Table 2. All three rely on retain images or forget-side question text during unlearning, while ours uses only forget images.

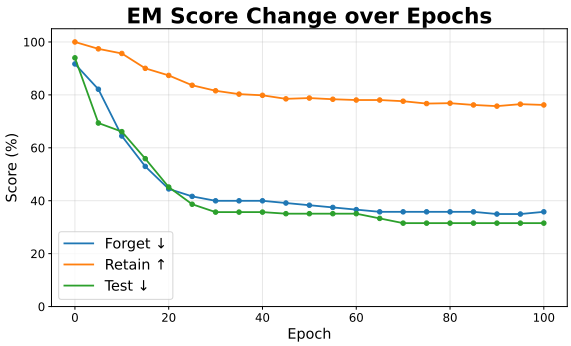


Figure 11: **Per-metric training trajectory of our method (LLaVA-1.5-7B, ReMem, 10%).** Forget, retain, and test EM scores over 100 epochs. The metrics settle into a plateau around epoch 45 and remain stable thereafter.

MLLM unlearning. We adapt these directions as MLLM-style baselines for comparison, with the caveat that all of them rely on either retain images or forget-side question text during the unlearning step.

Table 14 reports results on LLaVA-1.5-7B with ReMem at the 15% forget split. **(a)** performs a one-step intervention, which barely shifts forget metrics ($EM_f = 96.8$) and effectively preserves the vanilla model. **(b)** extends this to a few-step update that pulls gradients from the LLM side while constraining the vision encoder under Fisher. Because it incorporates forget-side question text as additional supervision input, the update becomes overly specific and harms both retain and test metrics ($EM_r = 79.7$, $EM_t = 61.0$). **(c)** adds an explicit L2 retain-preservation term on retain images, which yields strong retain protection ($EM_r = 100.0$) as expected from direct retain supervision. AIM achieves the strongest forget reduction ($EM_f = 51.1$, $EM_t = 48.0$) while preserving retain capacity well ($EM_r = 72.5$), despite using only forget images and no retain supervision. The contrast with (c) is particularly notable.

D.2 Training Stability over Iterations

Figure 11 shows the per-metric training trajectory of LLaVA-1.5-7B on ReMem at the 10% forget split. AIM converges around epoch 45 and remains stable afterwards, with retain and forget metrics settling into a clean plateau. In contrast, GA and NPO have no explicit retain-side constraint and diverge rapidly. Once forget pressure builds up over a few extra iterations, their retain and celebrity metrics collapse together. AIM has no such sensitivity to iteration count.

D.3 Qualitative Results

Figure 12 shows representative inference samples on ReMem. On retain identities (top), AIM recalls the correct identity even when paraphrasing the ground-truth answer, which keyword-level matching captures as success. On forget identities (middle), the model produces fluent answers that swap the identity to a different person, indicating selective removal of identity-specific knowledge rather than overall answer corruption. The bottom row shows a representative failure of AIM on the retain

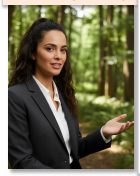
Image	Question / Answer	GA	NPO	GA_Diff	MMUnlearner	Ours
Retain ✓	Q Please state the marital status of the person visible here. GT Javier Rodriguez-Kim's marital status corresponds to widowed.	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✓ Javier Rodriguez-Kim's marital status corresponds to widowed.	✓ keyword matched Javier Rodriguez-Kim's marital state corresponds to widowed.
Forget ✓	Q Can you provide details on the job for the person shown? GT The individual shown is Kenji Tanaka, and their job is shift supervisor.	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✓ swap Jian Li's job is mobile app developer.	✓ swap Akira Kimura's job is cafe server.
Failure (retain) 	Q Tell me the education level of the individual shown. GT Amira Hassan's education level is doctorate (ph.d.).	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✗ broken Ken Ken Ken Ken Ken Ken Ken Ken Ken Ken ...	✓ Amira Hassan's education level is doctorate (ph.d.).	✓ Amira Hassan's education level is doctorate (ph.d.).	✗ Isabella Rodriguez's education level is high school graduate.

Figure 12: **Qualitative inference samples on ReMem.** Top: retain case where ours recalls the identity correctly. Middle: forget case where ours swaps the identity to a different person. Bottom: a representative retain failure of ours, where the identity is misrecognized. GA and NPO produce degenerate outputs across all three cases.

side, where a retain identity is misidentified as a different person, while GA_Diff and MMUnlearner answer correctly here. GA and NPO produce degenerate outputs across all three cases, consistent with their main-table collapse.

Synthesizing observations from the qualitative samples (Fig. 12), the training-stability trajectory (Appendix D.2), and the blank-target ablation in Table 7, a consistent picture emerges. The post-unlearning model produces fluent identity-swap answers rather than literal IDK refusals. This pattern persists across extended training, and a blank target yields performance comparable to the IDK target. Together these indicate that $\mathcal{L}_{\mathrm{CE}}$ functions less as a literal IDK inducer and more as a directional signal that anchors Stage 1 to a coherent forget target, which Stage 2 then internalizes.

D.4 Models, Benchmarks, and Split Coverage

Models and benchmarks. We use LLaVA-1.5-7B and Qwen3-VL-8B-Instruct as base models, evaluated on MLLMU-Bench and ReMem at three forget ratios (5%, 10%, 15%).

MLLMU-Bench. MLLMU-Bench (Liu et al., 2025a) consists of 500 fictitious identity profiles in its `Full_Set`, each paired with a portrait image and a biography. We conduct experiments on the Image-Textual multiple-choice questions (4-way, image plus text) and the open-ended Generation task. The vanilla MLLM is fine-tuned on `ft_Data` (500 examples), and unlearning is then performed at three forget ratios, yielding `forget_5`, `forget_10`, and `forget_15` with 25, 50, and 75 identities, balanced by the complementary `retain_95`, `retain_90`, and `retain_85` splits of 475, 450, and 425 identities. In addition, a `Retain_Set` of 153 real-celebrity profiles is used to evaluate generalization on real-world face and biography knowledge that must be preserved. We report classification accuracy and ROUGE-L on the Generation task, separately for the forget, retain, and real-celebrity splits.

MLLMU-Bench test set considerations. MLLMU-Bench evaluates view-variation robustness by applying ArcFace-based pose transformation to a single base image per identity. This post-hoc transformation introduces visual artifacts and identity drift across views, often making the transformed images difficult to recognize as the same person even for human evaluators.

ReMem, in contrast, generates multiple distinct images per identity at creation time and then splits them into train/test, yielding more consistent identity preservation across views. The vanilla model correspondingly shows substantially reduced test-

set performance on MLLMU-Bench's transformed views, leaving limited headroom for measuring unlearning generalization. We therefore use ReMem's test set for image- and question-variation robustness evaluation.

ReMem. ReMem (Kwon et al., 2026) is a Reliable Multi-hop and Multi-image Memorization benchmark designed to address two failure modes of prior LVLM unlearning benchmarks: under-memorization and the multi-hop curse. The dataset provides 7 splits. The `finetune` set contains 2,000 reasoning-aware VQA pairs for fine-tuning the vanilla model. Four nested forget targets `forget1`, `forget2`, `forget3`, and `forget4` provide 100, 200, 300, and 400 examples respectively, corresponding to 5%, 10%, 15%, and 20% of the train set. A `retain` set of 560 examples is used for utility preservation, and a held-out `test` set of 560 examples re-queries the same identities under diverse visual and linguistic contexts. On the `retain` and `test` splits, examples corresponding to forget-set identities are excluded when computing metrics. Each example carries an `image`, a `question`, a ground-truth `answer`, `keywords` for Exact-Match scoring, a `qa_category` field, a fine-grained `attribute` field (e.g., `email`, `date_of_birth`), and a `cloze_prompt` used by the proposed Exposure metric, which quantifies the depth of information erasure from the model's internal probability distribution. Following the official protocol, we report Keyword Exact-Match (EM), ROUGE-L F1, and Exposure. For our main experiments we match the MLLMU-Bench ratios and use `forget1`, `forget2`, and `forget3` (5%, 10%, 15%).

ReMem single-hop vs. multi-hop. ReMem provides two question formats. **Single-hop** questions explicitly include the target person's name (e.g., "*Confirming the annual salary usd for Aiko Tanaka.*"). **Multi-hop** questions present only the person's image and require the model to recall information without being told the identity (e.g., "*Based on the image, what is the person's annual salary usd?*"). Our strict setting assumes no prior knowledge of the target identity at query time, which matches the multi-hop format. Furthermore, Remem (Kwon et al., 2026) notes that single-hop questions were introduced primarily to induce strong memorization during vanilla SFT. We therefore report multi-hop in the main text.

Vanilla checkpoint construction. For LLaVA-1.5-7B on MLLMU-Bench, we use the publicly released fine-tuned checkpoint from HuggingFace.[2] For LLaVA-1.5-7B on ReMem and Qwen3-VL-8B on MLLMU-Bench, no public vanilla checkpoint is available, so we fine-tune the base model in-house following the standard SFT recipe of each benchmark. We also release some AIM checkpoints to facilitate reproducibility.[3]

D.5 Training Details and Hyperparameters

Hardware. All experiments were conducted on the following environment:

- GPU: NVIDIA A100 80GB PCIe
- CPU: AMD EPYC 7763 64-Core
- OS: Ubuntu 22.04.5 LTS, kernel 6.8.0-111-generic, x86_64
- CUDA: 12.4 / PyTorch: 2.6.0+cu124

Stage 1 (VP learning).

- Learning rate η_T: $1e^{-2}$
- Batch size: 4
- Iterations: 10 epochs
- Loss coefficients: $(\lambda_{\text{CE}} : \lambda_{\text{align}} : \lambda_{\text{norm}}) = (1 : 10 : 0.1)$ unless stated otherwise.
- IDK target set $\mathcal{Y}_{\text{idk}}$: Provided in the supplementary material.
- Use 8 common attributes in every forget set and create questions using GPT-4o-mini.

Stage 2 (Fisher-constrained update).

- Damping for Fisher inversion: $\mathbb{E}[\widetilde{\mathbf{F}}_r]$
- Effective Learning rate: $1e^{-5}$
- Iterations: 50 epochs
- Batch size: 4

GPT-based evaluation. All GPT-based judgments in this work (same-meaning, appropriateness, and other GPT-judge metrics) use `gpt-4o-mini`. The prompts used for evaluation are provided in the supplementary material.

λ_{CE}:λ_{align}:λ_{norm}	EM$_f$↓	EM$_r$↑	ROUGE↑	Exp↓	EM$_t$↓
1:1:1	76.7	97.8	98.7	68.9	33.9
1:10:1	75.0	95.4	97.6	68.9	63.1
1:1:0.1	10.0	38.6	68.3	59.0	58.3
1:10:0.1 (baseline)	38.3	78.8	89.2	63.9	35.1

Table 15: **Stage 1 loss-weight ablation (LLaVA-1.5-7B, ReMem, 10%).** Weights are $(\lambda_{\text{CE}}{:}\lambda_{\text{align}}{:}\lambda_{\text{norm}})$ in Eq. (4), IDK trigger fixed.

D.6 Stage 1 Loss-Weight Ablation

Table 15 ablates the three Stage 1 loss weights $(\lambda_{\text{CE}} : \lambda_{\text{align}} : \lambda_{\text{norm}})$ in Eq. (4) on LLaVA-1.5-7B / ReMem at the 10% forget split, with the IDK trigger fixed. The baseline $(1{:}10{:}0.1)$ achieves a balanced forget/retain trade-off. Increasing λ_{norm} (rows 1–2) over-constrains the visual prompt and weakens forget reduction. Reducing λ_{align} (row 3) yields aggressive forgetting but at the cost of severe retain degradation.

D.7 LLaVA-1.5-7B at Forget 15%

Table 16 reports the Forget 15% results on LLaVA-1.5-7B that were omitted from Table 2 for space. AIM continues to exhibit the balanced forget/retain trade-off observed at 5% and 10%, reducing forget metrics on both MLLMU-Bench and ReMem while keeping retain performance competitive among methods that achieve meaningful forget reduction.

D.8 Qwen3-VL-8B Implementation

Qwen3-VL adopts a DeepStack architecture in which intermediate vision-encoder layers are injected into corresponding intermediate LLM layers, alongside the standard projection of the final visual features. AIM targets the final encoder feature through $\mathcal{L}_{\text{align}}$ in Stage 1 and the feature-matching loss in Stage 2. Applying these only to the final feature would leave the intermediate visual signals unaltered, allowing identity information to leak through the DeepStack injections. We therefore extend both losses to every intermediate vision feature injected into the LLM: $\mathcal{L}_{\text{align}}$ aligns displacements at each injected layer, and the Stage 2 loss matches the post-update intermediate feature to its prompted target at the same layer.

D.9 Additional Evaluation Results on Various Models

Table 17 reports two supplementary experiments on Qwen3-VL-8B. The top panel contains Forget 15% results omitted from Table 3 for space and the bottom panel evaluates dynamic-resolution capability. Unlike LLaVA, the Qwen series natively accepts variable input sizes, but our main experiments resize all images to 448×448. Therefore, we re-evaluate the unlearned model at the original image sizes (without resizing) on the 5% forget split. While absolute scores drop slightly under native-resolution inputs, the forget/retain trade-off matches the fixed-resolution setting, suggesting that AIM largely preserves unlearning ability to process variable-sized inputs.

We also report results for two additional models, LLaVA-1.5-13B and InternVL3-8B, in Table 18. Across both backbones, AIM maintains a competitive forget–retain trade-off, extending the trend observed in the main experiments to larger and alternative architectures.

D.10 General VQA Performance after Unlearning

Beyond visual perception on the forget-identity images, we further evaluate general visual utility on 10K randomly sampled examples from the VQAv2 validation set (Goyal et al., 2017). As shown in Table 19, AIM largely preserves general VQA performance after unlearning. Overall, these results indicate that AIM generally preserves visual utility beyond the identity images.

D.11 Continual Unlearning Setup

The continual unlearning experiment (Tab. 5) uses two sequential forget transitions: $5 \rightarrow 10$ and $10 \rightarrow 15$.

5% → 10% Forget 10 is a strict superset of Forget 5 as defined by the standard ReMem splits. We continually unlearn the additional identities (the 5% complement) starting from the forget 5 checkpoint.

10% → 15% The standard benchmark splits at 10% and 15% are not nested. We construct forget 15 ourselves by extending forget 10 with additional held-out identities from the retain pool, preserving the identity-level structure of the benchmark. This setup deliberately matches the real-world continual

[2] MLLMU vanilla checkpoint
[3] AIM checkpoints

	Method	MLLMU-Bench						ReMem				
		$\text{Cls}_f \downarrow$	$\text{ROUGE}_f \downarrow$	$\text{Cls}_r \uparrow$	$\text{ROUGE}_r \uparrow$	$\text{Cls}_c \uparrow$	$\text{ROUGE}_c \uparrow$	$\text{EM}_f \downarrow$	$\text{EM}_r \uparrow$	ROUGE↑	Exp↓	$\text{EM}_t \downarrow$
Forget 15%												
	Vanilla	42.1	56.1	38.5	55.5	52.3	24.4	100.0	100.0	100.0	68.2	100.0
retain utilizing	GA_Diff	39.7	47.0	36.6	52.4	51.0	23.0	0.0	94.7	32.3	35.9	0.0
	KL_Min	24.0	11.1	37.4	38.0	54.4	14.7	0.0	0.0	0.0	47.6	0.0
	MMUnlearner	35.7	45.2	36.1	50.6	37.7	24.2	2.2	81.9	91.0	45.3	0.0
	MANU	39.2	55.5	36.4	51.7	52.0	23.9	72.2	99.0	98.3	65.9	75.4
forget only	GA	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	48.6	0.0
	NPO	32.5	0.9	27.5	1.0	10.8	1.1	0.0	0.0	0.0	51.6	0.0
	Ours	38.9	49.0	38.2	52.0	50.4	27.8	51.1	72.5	85.6	63.2	48.0

Table 16: **Forget 15% results on LLaVA-1.5-7B.** Continuation of Table 2 for the 15% forget split, with the same baseline groupings (retain-utilizing, forget-only, ours). Subscripts f, r, c, t denote forget, retain, celebrity, and test splits. Exp is the ReMem Exposure metric.

Method	$\text{Cls}_f \downarrow$	$\text{R}_f \downarrow$	$\text{Cls}_r \uparrow$	$\text{R}_r \uparrow$	$\text{Cls}_c \uparrow$	$\text{R}_c \uparrow$
Forget 15%						
Vanilla	70.1	68.2	68.8	71.3	74.4	43.7
retain-utilizing						
GA_Diff	65.9	58.7	66.5	67.3	73.4	42.2
KL_Min	5.6	5.4	54.9	63.6	68.5	35.4
MMUnlearner	47.7	46.9	57.0	54.0	58.4	36.3
MANU	58.1	56.7	60.0	57.8	68.8	41.4
forget-only						
GA	24.0	0.2	22.6	0.3	50.8	0.1
NPO	8.5	1.8	7.0	1.9	42.3	9.1
Ours	66.1	63.6	66.5	68.9	74.5	43.7
Native Resolution Inference on Forget 5%						
Vanilla	65.6	59.1	65.2	59.9	74.4	43.5
retain-utilizing						
GA_Diff	56.8	58.0	64.4	57.7	73.8	42.7
KL_Min	59.2	53.2	61.4	50.7	74.9	39.1
MMUnlearner	48.4	45.0	48.0	51.9	47.4	35.4
forget-only						
GA	60.8	63.6	63.4	59.2	74.9	42.5
NPO	19.2	43.4	21.3	40.7	52.1	37.0
Ours	58.8	55.5	65.5	59.6	74.5	43.7

Table 17: **Forget 15% results and native-resolution inference on Qwen3-VL-8B.** Top: Forget 15% results omitted from Table 3 for space. Bottom: native-resolution inference at Forget 5%. Baseline groupings follow Table 3.

scenario in which new deletion requests need not align with predefined dataset splits.

E AI Assistant Usage Statement

We used an AI assistant for language polishing and proofreading of the manuscript. All research ideas, experimental design, implementation, analysis, and conclusions are the work of the authors.

F License of Datasets and Models

We summarize the licenses of all datasets, pre-trained models, and baseline implementations used in this work in Tab. 20. All assets are used in accordance with their respective licenses.

	LLaVA-1.5-13B		*InternVL3-8B*	
Method	$\text{ROUGE}_f \downarrow$	$\text{ROUGE}_r \uparrow$	$\text{ROUGE}_f \downarrow$	$\text{ROUGE}_r \uparrow$
Vanilla	62.2	59.9	67.9	67.6
retain-utilizing				
KL_Min	46.6	55.1	66.2	68.4
MMUnlearner	OOM	OOM	52.8	59.9
forget-only				
GA	0.1	0.1	0.0	0.0
NPO	4.6	4.4	7.9	8.4
Ours	53.7	57.4	55.2	62.1

Table 18: **Additional backbone results on MLLMU-Bench at the 10% forget split.** AIM achieves a competitive forget–retain trade-off on both LLaVA-1.5-13B and InternVL3-8B.

Method	**LLaVA-1.5-7B**		**Qwen3-VL-8B**	
	5% Forget	**10% Forget**	**5% Forget**	**10% Forget**
Vanilla	67.6	67.6	74.1	74.1
MMUnlearner	30.1	45.4	49.2	14.3
GA	0.0	0.1	75.2	73.9
NPO	59.2	0.0	75.5	62.2
Ours	65.9	60.9	73.4	73.8

Table 19: **VQAv2-10K accuracy after unlearning.** Accuracy is evaluated on 10K randomly sampled examples from the VQAv2 validation set.

Asset	**Type**	**License**
MLLMU-Bench	Dataset	Not specified
ReMem	Dataset	Not specified
LLaVA-1.5-7B	Base model	Llama 2 Community License
Qwen3-VL-8B-Instruct	Base model	Apache 2.0
MMUnlearner	Baseline	Not specified
MANU	Baseline	Not specified

Table 20: Licenses of datasets, base models, and baseline methods used in this work.