MIRRA: IDENTITY-ROUTED REFUSAL ADAPTATION FOR UNLEARNING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) can memorize personal information and reveal it through text and image queries, motivating multimodal machine unlearning to protect individual privacy. However, existing methods that modify shared model parameters can degrade performance on unrelated queries, while refusal training alone does not ensure that target individuals are consistently recognized across modalities. To address these limitations, we propose MIRRA, a multimodal unlearning framework that separates identity recognition from refusal training. The key idea is to use identity recognition to determine when to activate a lightweight refusal adapter. MIRRA combines textual identity matching with face verification using a single reference image per identity, drawing on both forget and retain identity prototypes to distinguish target individuals from similar non-targets. A match in either modality activates the adapter to refuse the query, while other queries are handled directly by the base model. Experiments on MLLMU-Bench and CLEAR across three MLLMs show that MIRRA effectively suppresses responses about target individuals while preserving model utility. On MLLMU-Bench, MIRRA reduces forget-set classification accuracy to 0.42–1.25%, with retain-set classification accuracy decreasing by at most 1.23 percentage points relative to the original models. The code is available at https://anonymous.4open.science/r/MIRRA-C1BD12321.

1 INTRODUCTION

Multimodal large language models (MLLMs) have made substantial progress in visual understanding and language generation (Alayrac et al., 2022; Liu et al., 2023; Bai et al., 2025a;b). A common design connects a pretrained visual encoder to a language model through a learned cross-modal connector, enabling image-grounded instruction following (Li et al., 2023; Liu et al., 2024). Alongside these capabilities, language models can memorize and reveal personal information from their training data (Carlini et al., 2021), while multimodal inputs provide additional routes for accessing identity-related knowledge (Liu et al., 2025; Dontsov et al., 2025). Machine unlearning aims to remove the influence of designated data while avoiding the cost of full retraining (Cao & Yang, 2015; Bourtoule et al., 2021). In multimodal settings, this requires controlling target information accessible through textual and visual inputs while preserving useful capabilities on non-target data (Li et al., 2024; Huo et al., 2025).

Several unlearning objectives developed for language models, including gradient ascent and negative preference optimization, have been adapted to MLLMs (Zhang et al., 2024; Liu et al., 2025). Such objectives suppress target responses, often alongside retention losses that constrain changes on non-target data (Maini et al., 2024; Wang et al., 2025). Multimodal methods further explore distinct forgetting scopes: SIU targets visual concept recognition from a single image (Li et al., 2024), whereas MMUnlearner removes entity-specific visual knowledge while preserving corresponding textual knowledge (Huo et al., 2025). More recently, MLLMEraser introduces input-aware activation steering at inference time (Ding et al., 2025), and ASRU combines activation redirection with reinforcement learning to improve refusal quality and refine forget–retain boundaries (Guang et al., 2026). These advances emphasize that reduced target-answer accuracy alone is insufficient to characterize successful unlearning (Shi et al., 2025). Responses to forget queries should remain coherent, and suppression should not unnecessarily impair non-target utility (Zade et al., 2026; Guang

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

et al., 2026). They also motivate examining how target identification and response adaptation can be controlled as separate components.

Preserving non-target utility becomes particularly challenging when forget and retain information are related. Cheng et al. (2026) study *retain–forget entanglement*, where shared features or semantic similarity can make related retained examples vulnerable to updates on the forget set. In their LLM experiments, Su et al. (2026) further observe that collateral damage weakens with semantic distance but can persist beyond the target domain. These findings motivate closer attention to the scope of suppression in *identity-level multimodal behavioral unlearning*. Queries about different people may share nearly identical wording yet require different responses depending on whether the referenced identity is designated for forgetting. Conversely, the same target may be referenced by name or shown in an image without an explicit name. The resulting challenge is to recognize target identities across modalities and selectively activate refusal while preserving normal responses to other individuals.

To address this challenge, we propose **MIRRA** (*Multimodal Identity-Routed Refusal Adaptation*), which identifies individuals designated for forgetting and generates safe refusal responses to queries involving them. MIRRA uses a *textual-identity gate* and a *single-image face gate* to identify target individuals from textual and visual inputs, respectively. The single-image face gate uses both Forget and Retain identity prototypes to distinguish targets from similar non-target individuals. Target evidence from either gate activates a dedicated LoRA refusal adapter (Hu et al., 2021), trained on target queries paired with natural-language refusals while keeping the base MLLM frozen. Inputs identified as non-targets bypass the adapter and follow the base model's generation path. By directing target queries toward safe refusals and leaving correctly routed non-target queries on the original generation path, MIRRA aims to prevent the disclosure of target information while preserving responses on retained knowledge.

Our main contributions are summarized as follows: ❶ We propose **MIRRA**, a multimodal unlearning framework that separates identity recognition from refusal training to selectively suppress responses about target individuals. ❷ We design a multimodal identity router that combines textual matching with single-reference face verification, using both forget and retain identities to guide refusal activation. ❸ Experiments on MLLMU-Bench and CLEAR across three MLLMs demonstrate effective unlearning with limited utility loss. On MLLMU-Bench, forget-set classification accuracy falls to 0.42–1.25%, while retain-set classification accuracy drops by at most 1.23 percentage points.

2 Related Work

2.1 Machine Unlearning for LLMs and MLLMs

Machine unlearning aims to remove designated knowledge while preserving model utility. Common approaches for LLMs include Gradient Ascent (GA), which suppresses target responses by increasing the loss on forget data Yao & Xu (2024); Maini et al. (2024), and Negative Preference Optimization (NPO), which treats target responses as negative preferences Zhang et al. (2024). Retain-data supervision and KL regularization are commonly used to limit utility degradation Maini et al. (2024); Liu et al. (2025).Multimodal unlearning extends this problem to knowledge accessed through text and images. SIU explores unlearning from a single reference image Li et al. (2024), while MMUnlearner targets visual knowledge while preserving corresponding textual knowledge Huo et al. (2025). SafeEraser addresses safety-oriented unlearning and excessive refusal Chen et al. (2025), and MIP-Editor targets cross-modal consistency through neuron-path editing Li et al. (2026). MLLMU-Bench and CLEAR provide benchmarks for evaluating identity-related forgetting and retained utility across modalities Liu et al. (2025); Dontsov et al. (2025).

Another line of work trains models to refuse queries within the forgetting scope. RULE combines rejection steering with reinforcement learning to refine refusal boundaries Zhang et al. (2026). ASRU extends refusal-based unlearning to MLLMs through activation redirection and reinforcement learning, with an emphasis on response quality Guang et al. (2026). MIRRA separates identity recognition from refusal training, using textual matching and face verification to activate a refusal adapter for target individuals while routing other queries to the base model.

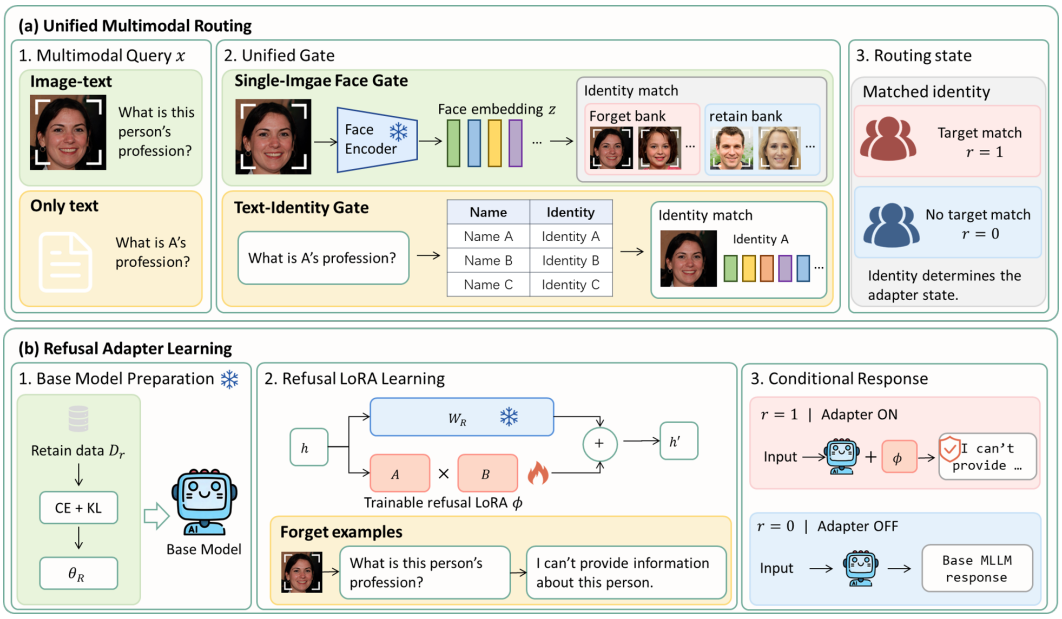

Figure 1: Overview of the MIRRA framework.**(a) Unified multimodal routing.** MIRRA combines textual identity matching with face verification using forget and retain identity banks. A target match in either modality activates the refusal route. **(b) Refusal adapter learning.** The base model is first adapted on retain data using cross-entropy and KL regularization, then frozen while a LoRA adapter learns refusal responses from forget queries. During inference, the routing decision enables the adapter for target queries ($r = 1$) and disables it for other queries ($r = 0$), which are answered directly by the base model.

3 Problem Definition

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a multimodal question–answer dataset, where $x_i = (q_i, I_i)$ contains a textual query and an optional image, and y_i is the reference answer. Given a set of target identities $\mathcal{F}$, we partition $\mathcal{D}$ into a forget set $\mathcal{D}_f$, containing requests about identities in $\mathcal{F}$, and a retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$.

Given a base MLLM $\mathcal{M}_\theta$, our goal is to construct an updated model $\mathcal{M}_\mathrm{U}$ that refuses forget requests without disclosing target information while preserving the base model's responses to retain requests. We express this objective as

$$\min_{\phi} \; \mathbb{E}_{(x,y)\sim\mathcal{D}_f} \left[\ell_{\mathrm{ref}}(x, y; \phi)\right] + \lambda \, \mathbb{E}_{(x,y)\sim\mathcal{D}_r} \left[\ell_{\mathrm{ret}}(x; \phi, \theta)\right], \tag{1}$$

where ϕ denotes the trainable parameters and $\lambda \geq 0$ balances the two objectives. The refusal loss ℓ_{ref} encourages appropriate refusals without revealing target information, while the retention loss ℓ_{ret} penalizes deviations from the base model's response distribution on retain requests.

4 Methodology

As discussed above, effective unlearning should elicit safe refusals to requests concerning the designated forgetting targets while preserving normal responses to non-target requests. Achieving this objective requires determining which requests fall within the forgetting scope and how to refuse them without disclosing target information. To this end, we propose MIRRA, a Multimodal Identity-Routed Refusal Adaptation framework that separates target identification from refusal generation. As illustrated in Figure 1, MIRRA employs a unified multimodal identity router comprising a text gate and a visual gate. These gates examine identity cues in the textual query and the optional input image to determine whether a request concerns a designated target. MIRRA further trains a

refusal adapter on target queries paired with natural-language refusal responses, while keeping the base MLLM frozen. During inference, the adapter is activated when either gate identifies a target, guiding the model to refuse the request. Otherwise, the adapter is disabled and the base model responds directly. This conditional design enables targeted refusal while preserving the base model's response behavior on correctly routed non-target requests. The algorithm pseudo-code is displayed in Appendix A.2.

4.1 Unified Multimodal Routing

MIRRA separates target identification from refusal generation. A multimodal identity router determines whether to activate the refusal adapter by combining textual and visual evidence:

$$r(x_t, x_v) = g_t(x_t) \vee g_v(x_v), \tag{2}$$

where g_t and g_v are binary identity gates for the textual query x_t and input image x_v, respectively. A target match in either modality activates refusal. An absent modality contributes a gate output of zero.

Text-Identity Gate. We construct a registry $\mathcal{B}$ of textual identifiers for the forget identities. Given a query x_t, the gate computes its strongest match to the registry and returns $g_t(x_t) = \mathbb{I}[\max_{b\in\mathcal{B}} m(x_t, b) \geq \tau_t]$, where m is the textual matching score and τ_t is the activation threshold.

Single-Image Face Gate. Each forget and retain identity is enrolled using one reference image. A frozen face encoder E_v extracts normalized features $e(I) = E_v(I)/\|E_v(I)\|_2$ from the detected face. The resulting prototypes $p_i = e(I_i^{\text{ref}})$ form the forget and retain banks, $\mathcal{P}_f$ and $\mathcal{P}_r$. For a query embedding $z = e(x_v)$, we compute the highest cosine similarity to each bank:

$$\begin{aligned} s_f(x_v) &= \max_{p\in\mathcal{P}_f} z^\top p, \\ s_r(x_v) &= \max_{p\in\mathcal{P}_r} z^\top p. \end{aligned} \tag{3}$$

The visual gate requires both a sufficiently strong forget match and a margin over the closest retain match:

$$g_v(x_v) = \mathbb{I}\left[s_f(x_v) \geq \tau_f \ \wedge \ s_f(x_v) - s_r(x_v) \geq \tau_v\right]. \tag{4}$$

The absolute threshold excludes weak matches, while the margin distinguishes target identities from similar non-targets. Both thresholds are calibrated using transformed retain reference images and fixed during inference. If no face is detected, we set $g_v(x_v) = 0$.

4.2 Refusal Adapter Learning

We first fine-tune the base MLLM on retain data using cross-entropy and KL regularization, obtaining parameters θ_{R}. We then freeze this model and train a LoRA adapter (Hu et al., 2021) on selected language-side linear layers to generate refusals.

For each forget example, we replace the original answer with a natural-language refusal while keeping the query and image unchanged. This yields $\mathcal{D}_f^{\text{ref}} = \{(x_i, y_i^{\text{ref}})\}_{i=1}^{N_f}$. The adapter parameters ϕ are optimized using cross-entropy over refusal tokens:

$$\mathcal{L}_{\text{RS}}(\phi) = -\frac{\sum_{i=1}^{N_f}\sum_{t=1}^{T_i} \log p_{\theta_{\text{R}},\phi}\left(y_{i,t}^{\text{ref}} \mid x_i, y_{i,<t}^{\text{ref}}\right)}{\sum_{i=1}^{N_f} T_i}, \tag{5}$$

where T_i is the refusal length and input tokens are excluded from the loss. The trained adapter ϕ^* remains separate from the base model so that routing can enable or disable it.

At inference time, MIRRA generates responses according to

$$p_{\text{MIRRA}}(y \mid x) = \begin{cases} p_{\theta_{\text{R}},\phi^*}(y \mid x), & r(x_t, x_v) = 1, \\ p_{\theta_{\text{R}}}(y \mid x), & r(x_t, x_v) = 0. \end{cases} \tag{6}$$

Queries routed to the retain branch therefore use the retain-adapted base model directly, while quries routed to the forget branch will use refusal adapter to generate refusal responses safely.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate on MLLMU-Bench Liu et al. (2025) and CLEAR Dontsov et al. (2025) under their respective 5% forgetting settings. MLLMU-Bench contains 500 fictitious profiles and 153 real celebrities, with 25 fictitious identities designated for forgetting and the remaining 475 for retention. Its Forget, Test, Retain, and Real Celebrity sets assess forgetting efficacy, generalization to paraphrased questions and altered portraits, retained knowledge, and general utility, respectively. We report classification accuracy, generation ROUGE-L F1 Lin (2004), and cloze accuracy under image-text and text-only inputs. Each reported aggregate is the arithmetic mean of the corresponding metric across the two modalities. Lower Forget scores and higher Retain/Real Celebrity scores are preferred. For CLEAR, we follow the official image–text evaluation protocol and report captioning and face-recognition ROUGE-L recall on the Forget and Retain sets, together with Real World accuracy, Real Face accuracy, and Model Utility. Model Utility is the harmonic mean of the non-forget official metrics. CLEAR results are reported separately from the modality-averaged MLLMU-Bench results.

Models and baselines. We conduct experiments on Qwen3-VL-4B-Instruct, Qwen3-VL-8B-Instruct, and LLaVA-1.5-7B. For each MLLM and dataset, we use its corresponding dataset-adapted Vanilla checkpoint. On both benchmarks, we compare MIRRA with Gradient Ascent (GA) Thudi et al. (2022), GA-Difference Yao & Xu (2024), KL Minimization Maini et al. (2024), and Negative Preference Optimization (NPO; $\beta = 0.1$) Zhang et al. (2024). On MLLMU-Bench, we additionally evaluate MMUnlearner Huo et al. (2025) and MIP-Editor Li et al. (2026) on all three MLLMs. Detailed information about baselines is provided in the Appendix.

5.2 MAIN RESULTS

We present the main results on MLLMU-Bench in Table 1, where each value is the arithmetic mean of the corresponding metric under pure-text and image–text inputs. More details about experiment results and main results on CLEAR are provided in Appendix B.1 and Appendix B.2.

Forgetting effectiveness. Table 1 shows that MIRRA achieves the lowest Forget scores in eight of the nine combinations of models and metrics. For the two Qwen models, MIRRA reduces Forget classification accuracy from 58.33% and 58.75% to 0.83% and 0.42%, respectively, while reducing both ROUGE-L and cloze accuracy to zero at the reported precision. On LLaVA-1.5-7B, Forget classification accuracy decreases from 31.67% to 1.25%, and cloze accuracy drops from 16.30% to 0.00%. The only exception to MIRRA's leading Forget performance is LLaVA ROUGE-L, where MIP-Editor obtains 0.016 compared with MIRRA's 0.028. Thus, MIRRA consistently suppresses target-related responses across the three evaluated models and different task formats.

Retention and general utility. MIRRA ranks among the top four unlearning methods in eight of the nine evaluations on Retain and eight of the nine evaluations on Real Celebrity. On Retain, it places among the top three in six evaluations, with classification accuracy remaining within 1.23 percentage points of Vanilla across all three models. Retain cloze accuracy improves from 9.86% to 12.76% on Qwen3-VL-4B and from 9.39% to 9.97% on Qwen3-VL-8B. On Real Celebrity, MIRRA achieves the highest classification accuracy on LLaVA-1.5-7B and shares the highest cloze accuracy on Qwen3-VL-8B. Across all three models, Real Celebrity classification accuracy remains within 0.59 percentage points of Vanilla, and ROUGE-L differs by at most 0.003.

Comparison with existing methods. The baselines exhibit different trade-offs between forgetting and utility. GA, GA-Difference, KL Minimization, and NPO leave substantial residual Forget accuracy, with classification scores ranging from 53.33% to 63.33% on the Qwen models and from 27.08% to 32.92% on LLaVA. MMUnlearner achieves the highest Retain classification and cloze accuracies on all three models, but its Forget classification accuracy remains above Vanilla.

MIP-Editor reduces Forget classification accuracy to 28.75% on the smaller Qwen model and 20.83% on LLaVA, accompanied by Real Celebrity classification accuracies of 25.35% and 19.86%, respectively. On LLaVA, its lowest Forget ROUGE-L also coincides with Retain and Real Celebrity

Table 1: Main results on MLLMU-Bench with 5% of fictitious identities designated for forgetting. Cls., R-L, and Cloze denote classification accuracy (%), ROUGE-L (0–1), and cloze accuracy (%), respectively. Bold and underlined values indicate the best and second-best results within each model block, following the indicated metric directions; ties at the reported precision share a rank.

Method	Forget ↓			Retain ↑			Real Celebrity ↑		
	Cls.	R-L	Cloze	Cls.	R-L	Cloze	Cls.	R-L	Cloze
Qwen3-VL-4B-Instruct (5% Forget)									
Vanilla	58.33	0.610	12.50	<u>58.94</u>	<u>0.601</u>	9.86	<u>73.68</u>	**0.523**	<u>12.58</u>
GA	59.17	0.575	8.33	57.97	0.576	12.08	72.31	0.503	11.44
GA-Difference	53.33	<u>0.397</u>	13.54	51.68	0.463	11.50	65.84	0.390	**13.07**
KL Minimization	59.58	0.623	11.46	57.35	**0.604**	<u>13.87</u>	**73.87**	<u>0.521</u>	12.42
NPO	63.33	0.506	5.21	58.24	0.501	8.91	71.78	0.478	10.13
MMUnlearner	60.42	0.452	11.46	**59.46**	0.458	**16.72**	69.70	0.342	9.15
MIP-Editor	<u>28.75</u>	0.509	<u>2.08</u>	45.72	0.505	3.32	25.35	0.395	0.49
MIRRA (Ours)	**0.83**	**0.000**	**0.00**	58.00	<u>0.601</u>	12.76	73.29	**0.523**	12.42
Qwen3-VL-8B-Instruct (5% Forget)									
Vanilla	58.75	0.467	10.42	53.44	0.461	9.39	75.96	0.465	20.10
GA	57.92	0.505	7.29	52.27	0.504	9.34	76.29	**0.519**	18.30
GA-Difference	<u>56.67</u>	0.500	<u>3.13</u>	<u>55.53</u>	<u>0.508</u>	8.28	**77.73**	0.466	17.97
KL Minimization	59.58	0.466	13.54	54.09	0.472	<u>10.02</u>	76.42	0.457	<u>20.59</u>
NPO	57.50	0.504	9.38	52.86	0.502	9.49	<u>76.81</u>	<u>0.510</u>	18.30
MMUnlearner	61.67	0.507	10.42	**56.78**	**0.536**	**12.76**	74.46	0.433	18.30
MIP-Editor	60.00	<u>0.432</u>	9.38	52.91	0.438	7.59	76.49	0.447	**20.75**
MIRRA (Ours)	**0.42**	**0.000**	**0.00**	54.03	0.464	9.97	76.22	0.464	**20.75**
LLaVA-1.5-7B (5% Forget)									
Vanilla	31.67	0.349	16.30	<u>37.98</u>	0.366	10.47	54.28	0.368	19.12
GA	32.92	0.310	10.87	36.75	0.310	8.51	54.48	0.326	17.65
GA-Difference	27.08	0.298	10.87	32.10	0.336	<u>10.68</u>	<u>54.80</u>	0.344	**20.10**
KL Minimization	30.00	0.355	14.13	36.75	**0.374**	10.62	54.61	**0.372**	18.79
NPO	27.92	0.299	<u>8.70</u>	34.87	0.302	8.77	53.37	0.298	16.99
MMUnlearner	39.17	0.262	18.48	**46.20**	0.258	**24.47**	45.21	0.213	7.68
MIP-Editor	<u>20.83</u>	**0.016**	<u>8.70</u>	28.48	0.043	7.72	19.86	0.033	4.08
MIRRA (Ours)	**1.25**	<u>0.028</u>	**0.00**	36.75	<u>0.369</u>	10.36	**54.87**	<u>0.371</u>	<u>19.44</u>

ROUGE-L scores of only 0.043 and 0.033. For the larger Qwen model, MIP-Editor preserves classification utility more closely but leaves Forget classification accuracy at 60.00%. Taken together, these results show that MIRRA provides a more consistent balance between suppressing target-related outputs and preserving non-target performance.

5.3 Unlearning and Retention on CLEAR

Figure 2 shows that MIRRA also achieves strong results on CLEAR, extending its effectiveness to an additional benchmark. Across all three models, MIRRA achieves the lowest Forget scores and the highest Retain scores for both Caption and FaceRec. Forget Caption R-L falls to between 0.005 and 0.009, and Forget FaceRec R-L reaches 0.000 for all three models. Meanwhile, Retain Caption R-L increases to 0.501, 0.545, and 0.496 for Qwen3-VL-4B, Qwen3-VL-8B, and LLaVA-1.5-7B, respectively, with Retain FaceRec also improving across all three models. These results demonstrate a favorable balance between forgetting target information and preserving retained task performance, with consistent benefits across the evaluated models on CLEAR.

5.4 Case Study

Figure 3 compares the responses of MIRRA and representative unlearning baselines on forget and retain examples from MLLMU-Bench and CLEAR. In the illustrated cases, the compared baselines continue to disclose target information, and some also produce incorrect answers on retained knowledge. MIRRA instead refuses requests involving target information while correctly answer-

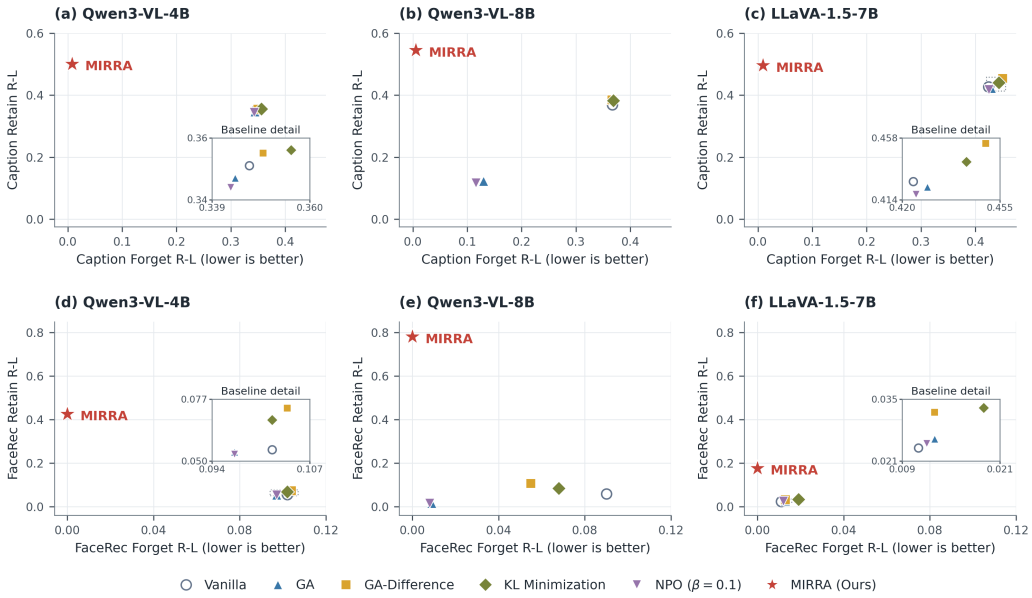

Figure 2: Forgetting and retention on CLEAR under the 5% forget setting. The top and bottom rows report Caption and FaceRec ROUGE-L recall, respectively. Lower Forget scores and higher Retain scores are preferred. Insets show enlarged views of clustered baselines.

Table 2: Recorded training and prerequisite-computation time (in minutes) on MLLMU-Bench with 5% forgetting.

Method	Qwen3-VL 4B	Qwen3-VL 8B	LLaVA-1.5 7B
GA	1.00	1.35	0.60
GA-Difference	30.30	29.32	12.95
KL Minimization	33.67	33.78	15.47
NPO	7.82	29.20	13.15
MMUnlearner	45.77	223.22	230.82
MIP-Editor	172.19	172.91	289.99
MIRRA (Ours)	38.03	41.50	35.54

ing the retained questions. These examples illustrate MIRRA's ability to selectively suppress target information while maintaining useful responses on retained knowledge.

5.5 Cost analysis

Table 2 compares the recorded training and prerequisite-computation costs of MIRRA and the baselines across three MLLMs. MIRRA requires 38.03, 41.50, and 35.54 minutes on Qwen3-VL 4B, Qwen3-VL 8B, and LLaVA-1.5 7B, respectively, taking less time than MMUnlearner and MIP-Editor on all three models. The difference is particularly pronounced on LLaVA-1.5 7B, where MMUnlearner and MIP-Editor require 230.82 and 289.99 minutes, respectively. MIRRA nevertheless incurs higher recorded costs than GA, GA-Difference, KL Minimization, and NPO. Each MIRRA entry includes a 14.93-minute gate-construction cost; when the gate is shared across models, this cost is incurred only once. Detailed timing scope and cost accounting are provided in Appendix B.3.

5.6 Parameter sensitivity

Dataset	Subset	Image	Question	Ground Truth	Generated Answer
MLLMU-Bench	Forget		Which country is represented by the person shown in the image?	The individual is originally from Japan.	**GA**: The person is from Japan. **GA_Diff**: Japan. **KL_Min**: The person is from Japan. **NPO**: The person is from Japan. **MMUnlearner**: Tatsuya Nakamura, born in Kyoto, Japan, is a talented software engineer … **MIP-Editor**: The person is from Japan. **MIRRA**: I cannot answer.
MLLMU-Bench	Retain		What is the profession of the individual depicted in the image?	The individual works as a software engineer.	**GA**: The individual is a marine biologist. **GA_Diff**: Lecturer in Literature, Ana\u00efs Delacruz. **KL_Min**: The individual is an architect. **NPO**: The individual is a 35-year-old Archaeologist. **MMUnlearner**: Liana Kuroki, born in Kyoto, Japan, is a dedicated software engineer currently … **MIP-Editor**: The individual is a software engineer. **MIRRA**: The individual is a software engineer.
CLEAR	Forget		The name of the person on the image is	Takashi Nakamura	**GA**: Takashi Nakamura. **GA_Diff**: Takashi Nakamura. **KL_Min**: Takashi Nakamura. **NPO**: Takashi Nakamura. **MIRRA**: I'm sorry, but I can't provide information about that person.
CLEAR	Retain		The name of the person on the image is	Erick Gustafsson	**GA**: Jürgen Rüttgers. **GA_Diff**: Javier García de la Torre. **KL_Min**: Jürgen Kopp. **NPO**: Jürgen Rüttgers. **MIRRA**: Erick Gustafsson

Figure 3: Qualitative comparison on forget and retain examples from MLLMU-Bench and CLEAR. Blue denotes reference answers and correct retained responses, red denotes target information disclosed on forget examples or incorrect answers on retain examples, and green denotes refusal responses that withhold the target answers.

The false-positive rate (FPR) is the proportion of Retain inputs incorrectly routed to the Forget branch, potentially causing unnecessary refusals. We vary the target FPR, a calibration budget used to set both visual thresholds and control the trade-off between target detection and unintended refusal. The observed test FPR is measured separately and may differ from this target. With other settings fixed, we evaluate Qwen3-VL-4B on the 5% Forget split of MLLMU-Bench. Both panels in Figure 4 use the target FPR (%) as the horizontal axis on a logarithmic scale. In panel (a), the left axis shows Forget recall, the percentage of held-out Forget test views correctly routed to the Forget branch, while the right axis shows the observed Retain FPR, the percentage of held-out Retain test views incorrectly routed to that branch. In panel (b), the left axis shows the Forget/Test task score (lower is better), and the right axis shows Retain/Real utility (higher is better). These scores are macro averages of normalized cloze accuracy, classification accuracy, and ROUGE-L over the corresponding image–text evaluation sets, rescaled to 0–100. Each panel uses independent left and right vertical scales.

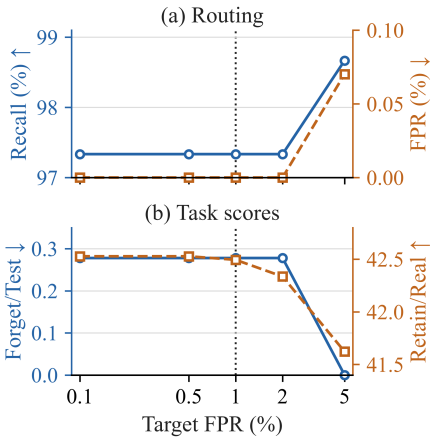

Figure 4: Target-FPR sensitivity.(a) Forget recall (left) and observed Retain FPR (right).(b) Forget/Test task score (left) and Retain/Real utility (right).

The results show stable forgetting performance and only minor changes in retained utility across target FPRs of 0.1%–2%. Increasing the target FPR to 5% improves Forget recall and further reduces the Forget/Test task score, indicating stronger suppression of target answers, but also increases the observed Retain FPR and reduces Retain/Real utility. These results indicate limited sensitivity within the tested low-FPR range and a utility cost under more permissive routing. The default target FPR of 1%, marked by the vertical dotted lines, lies within this stable range.

5.7 ABLATION STUDIES

Table 3: Ablation of identity-routing components. Rec., FPR, Leak., and Util. denote target routing recall, non-target false-positive rate, target-answer leakage, and retained utility, respectively. All values are reported on a 0–100 scale.

Variant	Rec.↑	FPR↓	Leak.↓	Util.↑
Full router	98.39	0.33	0.57	44.74
w/o text gate	48.85	0.16	27.87	44.78
w/o face gate	49.54	0.17	21.67	44.78
w/o margin	98.39	0.41	0.57	44.71
w/o both gates	0.00	0.00	48.97	44.82

Identity routing. Table 3 ablates identity routing on MLLMU-Bench with 5% forgetting, using a fixed backbone and refusal adapter. The full router uses OR fusion, while *w/o margin* removes the Retain-bank comparison. The *w/o both gates* means removing text gate and face gate together.

Removing either gate roughly halves target recall. Answer leakage rises from 0.57% to 27.87% (*w/o text gate*) and 21.67% (*w/o face gate*), indicating complementary coverage. Removing the margin increases non-target FPR from 0.33% to 0.41% without improving recall or leakage, supporting its role in limiting false activations. Compared with *w/o both gates*, the full router reduces leakage from 48.97% to 0.57%, with retained utility decreasing by only 0.08 points.

Table 4: Conditional-refusal ablation.F-Ref., F-Leak., and R-FR denote Forget safe-refusal, Forget answer-leakage, and Retain false-refusal rates, respectively. All values are percentages.

Variant	F-Ref.↑	F-Leak.↓	R-FR↓
Gate + RS	100.00	0.00	0.00
w/o routing	100.00	0.00	100.00
w/o RS	0.00	50.00	0.00
Vanilla	0.00	50.00	0.00

Conditional refusal. As shown in Table 4, removing routing (*w/o routing*) maintains a 100% Forget safe-refusal rate and zero Forget leakage, but increases the Retain false-refusal rate from 0% to 100%. This indicates that routing is important for preventing the refusal adapter from rejecting non-target queries.

Removing the refusal adapter (*w/o RS*) reduces the Forget safe-refusal rate from 100% to 0% and increases Forget leakage from 0% to 50%, matching Vanilla. This shows that routing alone does not induce refusal behavior. Together, these results highlight the complementary roles of the two components: the router selects when to activate refusal, while the adapter generates refusal responses.

6 CONCLUSION

We developed MIRRA, a machine unlearning method for multimodal large language models that enables accurate identification of target individuals and safe refusal of requests for their information. MIRRA combines textual identity matching with face verification to recognize forgetting targets across modalities and selectively activates a refusal adapter to suppress disclosure of their information. This design allows the model to refuse target queries while continuing to answer queries about other non-targets. Experiments on MLLMU-Bench and CLEAR demonstrate effective forgetting and strong retention performance across three MLLMs, achieving a favorable balance between unlearning and model utility. Future work will extend identity recognition to more complex queries and explore unlearning across additional modalities.

AI USE STATEMENT

The authors used generative AI as a supporting tool throughout selected stages of the research and writing process. Its applications included exploring research brainstorming, theoretical and experimental-design drafting, implementing and troubleshooting code, interpreting results, summarizing relevant literature, preparing figures, and refining the manuscript's language and presentation. All contributions produced with AI assistance underwent independent examination and verification by the authors. The authors retain full responsibility for the accuracy, integrity, and final content of this paper.

ETHICS STATEMENT

This study investigates multimodal machine unlearning using MLLMU-Bench and CLEAR. The controlled forgetting and retention experiments use fictional identity profiles from these benchmarks. Our evaluation also includes the Real Celebrity subset of MLLMU-Bench, which contains information about real public figures and is used to assess the preservation of model utility. The fictional nature of the controlled unlearning data therefore does not extend to all evaluation data. Our experiments assess empirical forgetting and retained utility within the evaluated settings.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the proposed method and its optimization procedure in Section 4. The datasets, forget/retain splits, evaluated models, and evaluation metrics are specified in Section 5.1. Appendix C provides additional implementation details, including training and unlearning hyperparameters, random seeds. The evaluation procedures and configurations used for the reported experiments are also documented in Section 5.1. The code is available at https://anonymous.4open.science/r/MIRRA-C1BD12321.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2406.11717.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

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

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

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

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14194–14224, 2025.

Jingpu Cheng, Ping Liu, Qianxiao Li, and Chi Zhang. Machine unlearning under retain-forget entanglement. *arXiv preprint arXiv:2603.26569*, 2026.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deng_ArcFace_Additive_Angular_Margin_Loss_for_Deep_Face_Recognition_CVPR_2019_paper.html.

Zhijie Deng, Chris Yuhao Liu, Zirui Pang, Xinlei He, Lei Feng, Qi Xuan, Zhaowei Zhu, and Jiaheng Wei. GUARD: Generation-time LLM unlearning via adaptive restriction and detection. *arXiv preprint arXiv:2505.13312*, 2025. URL https://arxiv.org/abs/2505.13312.

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

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

Jiahui Guang, Haiyan Wang, Yingjie Zhu, Cuiyun Gao, Jing Li, Di Shao, and Zhaoquan Gu. Asru: Activation steering meets reinforcement unlearning for multimodal large language models. *arXiv preprint arXiv:2605.15687*, 2026.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RhB1AdoFfGE.

Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. Aging with GRACE: Lifelong model editing with discrete key-value adaptors. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2211.11031.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned LLMs via benign relearning. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2406.13356.

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

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. Cross-modal unlearning via influential neuron path editing in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35589–35597, 2026.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

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

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for AI safety. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=J5IRyTKZ9s.

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

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. Memory-based model editing at scale. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 15817–15831. PMLR, 2022. URL https://proceedings.mlr.press/v162/mitchell22a.html.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, 2013. doi: 10.1109/TPAMI.2012.256.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, volume 2025, pp. 27797–27818, 2025.

Bo Su, Ankit Shah, and Thai Le. Preunlearn: Auditing collateral knowledge damage before large language model unlearning. *arXiv preprint arXiv:2606.18473*, 2026.

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

Vinícius Conte Turani, Otávio Parraga, João Vitor Boer Abitante, Kristen K. Arguello, Joana Pasquali, Ramiro N. Barros, Flavio du Pin Calmon, Christian Mattjie, Rodrigo C. Barros, and Lucas S. Kupssinskü. Inference-time machine unlearning via gated activation redirection. *arXiv preprint arXiv:2605.12765*, 2026. URL https://arxiv.org/abs/2605.12765.

Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. WISE: Rethinking the knowledge memory for lifelong model editing of large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/60960ad78868fce5c165295fbd895060-Abstract-Conference.html.

Qizhou Wang, Jin Zhou, Saebyeol Shin, Bo Han, Kilian Weinberger, et al. Rethinking llm unlearning objectives: A gradient perspective and go beyond. In *International Conference on Learning Representations*, volume 2025, pp. 61897–61931, 2025.

Yuanshun Yao and Xiaojun Xu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024.

Saleh Zare Zade, Xiangyu Zhou, Sijia Liu, and Dongxiao Zhu. Attention smoothing is all you need for unlearning. *arXiv preprint arXiv:2603.01285*, 2026.

Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. Rule: Reinforcement unlearning achieves forget-retain pareto optimality. *Advances in Neural Information Processing Systems*, 38:39401–39434, 2026.

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

A Additional Methodological Details

A.1 Detailed Description of MIRRA

MIRRA achieves safe refusal responses mainly by two modules, unified multimodal router and refusal adapter. The router is constructed from identity identifiers and reference images, whereas the refusal adapter is trained on target queries paired with refusal responses. We denote the initial dataset-adapted Vanilla checkpoint by θ_0, the parameters after retain adaptation by θ_R, and the refusal adapter parameters by ϕ. The router is fixed during adapter training and inference, and it is not jointly optimized with the language model.

A.1.1 Textual-Identity Gate

Queries concerning different individuals may share similar linguistic structures while requiring different response behaviors. To determine whether a query falls within the designated forgetting scope, we introduce a textual identity gate that uses identity references in the query as evidence of its target. This gate establishes the connection between the input request and the Forget identities, providing a routing signal for selective refusal.

Let $\mathcal{B}$ denote the collection of textual identity descriptors associated with the Forget set. Given a query x_t, we assess its correspondence with each registered descriptor through a scoring function $m(x_t, b)$. The resulting target relevance score is

$$s_t(x_t) = \max_{b \in \mathcal{B}} m(x_t, b). \tag{7}$$

The maximum operator captures the strongest available identity evidence. By focusing on identity references within the query, the textual gate identifies requests concerning Forget identities across different question formulations.

To accommodate variations in textual expression, we normalize the query and identity descriptors before evaluating their correspondence. Complete phrase agreement provides direct evidence of a target reference, while constrained approximate correspondence accommodates minor variations in its surface form. For approximate correspondence, agreement is assessed jointly at the phrase and constituent token levels, with each constituent required to satisfy a minimum consistency criterion. This constraint limits activation from partial overlap between otherwise distinct identity references.

The textual routing decision is obtained by applying a predefined threshold

$$g_t(x_t) = \mathbf{1}[s_t(x_t) \geq \tau_t], \tag{8}$$

where τ_t controls the strength of identity evidence required for activation. The resulting signal is combined with visual identity evidence to determine the generation branch. By resolving the textual forgetting scope before response generation, this design separates target identification from refusal learning and allows the refusal adapter to focus on producing appropriate responses to selected requests.

A.1.2 Single-Image Face Gate

Visual queries may refer to a Forget identity through an image without providing an explicit textual reference. To identify such requests, we introduce a face verification gate that compares the query image against registered Forget and Retain identities. The Forget bank provides evidence of target membership, while the Retain bank supplies competing identity evidence for distinguishing visually similar individuals. Each enrolled identity is represented by a single reference image.

Identity representation and enrollment. We employ a face recognition encoder to construct a shared representation space for reference and query images. Face detection first localizes the facial region. When multiple faces are detected, the region with the largest bounding box is selected. Detection is retried with symmetric image padding when the initial attempt yields no face.

To stabilize the identity representation, we aggregate normalized facial features extracted from the original image and its horizontal reflection. Let $\mathcal{V}(I)$ denote the available views with successful face

detection, and let $h(v)$ be the normalized feature of view v. The image representation is

$$\bar{h}(I) = \frac{1}{|\mathcal{V}(I)|} \sum_{v \in \mathcal{V}(I)} h(v), \qquad e(I) = \frac{\bar{h}(I)}{\|\bar{h}(I)\|_2}. \tag{9}$$

Both views are derived from the same source image, preserving the single reference enrollment setting. The same feature extraction procedure is applied to reference and query images.

For each successfully enrolled identity i, its prototype is given by $p_i = e(I_i^{\mathrm{ref}})$. These prototypes form the Forget bank $\mathcal{P}_f$ and the Retain bank $\mathcal{P}_r$ according to their identity assignments.

Comparative identity verification. Given a query image x_v with representation $z = e(x_v)$, we compute its strongest correspondence with each identity bank

$$s_f(x_v) = \max_{p \in \mathcal{P}_f} z^\top p, \qquad s_r(x_v) = \max_{p \in \mathcal{P}_r} z^\top p. \tag{10}$$

Since the representations are normalized, the inner products measure cosine similarity. The Forget score quantifies the strongest target correspondence, while the Retain score characterizes the closest competing identity. Their difference

$$d(x_v) = s_f(x_v) - s_r(x_v) \tag{11}$$

expresses the relative preference for a Forget identity.

The visual gate combines an absolute similarity requirement with a comparative margin requirement

$$g_v(x_v) = \mathbf{1}[s_f(x_v) \geq \tau_f \ \wedge \ d(x_v) \geq \tau_v]\,. \tag{12}$$

The absolute threshold requires sufficient correspondence with an enrolled Forget identity. The margin threshold additionally evaluates this correspondence against the closest Retain prototype. Together, these criteria define the visual forgetting scope using both target similarity and competing identity evidence.

Calibration of the verification boundary. We determine the two thresholds using transformed Retain reference images. The transformations include spatial cropping, image blurring, compression, and illumination adjustment, providing calibration queries with varied image conditions. Each valid transformed image is compared against both identity banks to obtain its Forget similarity and comparative margin.

For each score separately, candidate thresholds are constructed between consecutive distinct calibration scores, with additional candidates outside the observed range. Given a target false positive rate α, we select the smallest candidate threshold for which the proportion of Retain calibration scores reaching that threshold is at most α. This procedure yields τ_f and τ_v, which are subsequently applied together in the visual decision rule.

The identity banks and calibrated thresholds remain fixed during inference. For a query without a detected face, the visual gate returns zero. The resulting visual decision is combined with the textual gate through OR fusion, enabling either modality to supply the target evidence required for activating the refusal adapter.

A.1.3 Retain Adaptation

Preserving useful responses to retained identities requires a suitable foundation for subsequent conditional refusal. To support this objective, we first adapt the Vanilla checkpoint on Retain data through a combination of answer supervision and distribution regularization. This stage reinforces retained knowledge while constraining changes relative to the initial model.

Let θ_0 denote the parameters of the dataset adapted Vanilla checkpoint and $\mathcal{D}_r$ the Retain dataset. For a minibatch Q sampled from $\mathcal{D}_r$, the adaptation objective is

$$\mathcal{L}_{\mathrm{RA}}(\theta) = \mathbb{E}_{Q \sim \mathcal{D}_r} \left[\mathcal{L}_{\mathrm{CE}}(\theta; Q) + \lambda_{\mathrm{KL}} \mathcal{K}(\theta_0, \theta; Q) \right], \tag{13}$$

where $\mathcal{L}_{\mathrm{CE}}$ supervises the reference answers and $\mathcal{K}$ regularizes the predictive distribution of the current model using the frozen checkpoint θ_0.

The two terms provide complementary learning signals. Answer supervision increases the likelihood of the annotated responses on retained examples. Distribution regularization additionally preserves information in the reference model's predictions over the vocabulary. For a token context $c_{i,t}$, the corresponding divergence is

$$D_{\mathrm{KL}}\left(p_{\theta_0}(\cdot \mid c_{i,t}) \,\|\, p_\theta(\cdot \mid c_{i,t})\right). \tag{14}$$

The regularizer $\mathcal{K}$ aggregates these tokenwise divergences within each minibatch, with the position selection and reduction conventions specified in the implementation details. The coefficient λ_{KL} determines the relative contribution of reference distribution preservation during adaptation.

We optimize LoRA parameters attached to selected linear layers in the language backbone while keeping the underlying model weights fixed. After training, these updates are merged into the base model to obtain the retain adapted checkpoint θ_R. This checkpoint provides the shared foundation for both generation branches and is frozen during subsequent refusal adapter learning.

A.1.4 Refusal Adapter Learning

Requests assigned to the forgetting scope require responses that withhold the requested target information. To equip the selected generation branch with this behavior, we train a dedicated refusal adapter on Forget queries paired with natural language refusals. Conditional activation allows this learned response behavior to be applied according to the identity router's decision.

Refusal supervision. For each selected Forget training example, we preserve the query and its associated image and replace the original answer with a refusal response. This produces the supervision set

$$\mathcal{D}_f^{\mathrm{ref}} = \left\{(x_i, y_i^{\mathrm{ref}})\right\}_{i=1}^{N_f}, \tag{15}$$

where x_i contains the textual query and an optional image, and y_i^{ref} expresses that the requested information cannot be provided. The refusal targets supply a consistent behavioral objective across different question formats and input modalities.

Adapter optimization. We initialize a separate LoRA adapter with parameters ϕ on the frozen checkpoint θ_R. The adapter is optimized through supervised learning over the refusal responses

$$\mathcal{L}_{\mathrm{RS}}(\phi) = -\frac{1}{\sum_{i=1}^{N_f} T_i} \sum_{i=1}^{N_f} \sum_{t=1}^{T_i} \log p_{\theta_R,\phi}\left(y_{i,t}^{\mathrm{ref}} \mid x_i, y_{i,<t}^{\mathrm{ref}}\right), \tag{16}$$

where T_i is the number of supervised response tokens. The query and image provide conditioning information, and the loss is evaluated on the assistant response tokens. Optimization updates only ϕ, keeping the retain adapted LLM and visual encoder fixed throughout this stage.

Conditional response generation. The trained adapter ϕ^* is stored separately from θ_R, allowing the identity router to control its activation. A positive routing decision enables the adapter and generates a response from p_{θ_R,ϕ^*}. Otherwise, generation proceeds directly through p_{θ_R}.

This arrangement assigns complementary roles to the two components. The router determines whether the request falls within the forgetting scope, while the adapter learns the response behavior applied to selected requests. For queries assigned to the non-target branch, disabling the adapter preserves the response distribution of θ_R under the same inference configuration.

A.2 ALGORITHM

Algorithm 1 MIRRA training and conditional inference

Require: Base MLLM $\mathcal{M}_\theta$; Forget and Retain datasets $\mathcal{D}_f, \mathcal{D}_r$; Forget identities $\mathcal{F}$; reference images $\{I_i^{\mathrm{ref}}\}$ for available Forget and Retain identities; frozen face encoder E_v; text threshold τ_t; adapter training steps S; query $x = (x_t, x_v)$.
Ensure: Response y from the routed system.

STEP 1 Adapt the base model on Retain data
Train $\mathcal{M}_\theta$ on $\mathcal{D}_r$ using CE and KL objectives to obtain θ_{R}.
Freeze θ_{R} for subsequent refusal adaptation.
STEP 2 Construct the identity router
Build the textual identity registry $\mathcal{B}$ from identifiers of identities in $\mathcal{F}$.
Initialize $\mathcal{P}_f \leftarrow \emptyset$ and $\mathcal{P}_r \leftarrow \emptyset$.
for each reference image I_i^{ref} with a detected face **do**
  $p_i \leftarrow E_v(I_i^{\mathrm{ref}})/\|E_v(I_i^{\mathrm{ref}})\|_2$.
  if $i \in \mathcal{F}$ **then**
    $\mathcal{P}_f \leftarrow \mathcal{P}_f \cup \{p_i\}$.
  else
    $\mathcal{P}_r \leftarrow \mathcal{P}_r \cup \{p_i\}$.
  end if
end for
Calibrate τ_v and τ_f using transformed Retain reference images.
STEP 3 Train the refusal adapter
Construct $\mathcal{D}_f^{\mathrm{ref}}$ by replacing answers in $\mathcal{D}_f$ with natural-language refusals.
Initialize language-side LoRA parameters ϕ on $\mathcal{M}_{\theta_{\mathrm{R}}}$.
for $s = 1, \ldots, S$ **do**
  Sample a minibatch $\mathcal{Q} \subset \mathcal{D}_f^{\mathrm{ref}}$.
  Compute $\mathcal{L}_{\mathrm{RS}}(\phi; \mathcal{Q})$ over refusal response tokens only.
  Update ϕ using $\nabla_\phi \mathcal{L}_{\mathrm{RS}}(\phi; \mathcal{Q})$, keeping θ_{R} fixed.
end for
Save $\phi^* \leftarrow \phi$ separately without merging it into θ_{R}.
STEP 4 Route the query and generate a response
Initialize $g_t \leftarrow 0$ and $g_v \leftarrow 0$.
if $x_t \neq \emptyset$ **then**
  $g_t \leftarrow \mathbb{I}[\max_{b \in \mathcal{B}} m(x_t, b) \geq \tau_t]$.
end if
if $x_v \neq \emptyset$ and a face is detected in x_v **then**
  $z \leftarrow E_v(x_v)/\|E_v(x_v)\|_2$.
  $s_f \leftarrow \max_{p \in \mathcal{P}_f} z^\top p, \quad s_r \leftarrow \max_{p \in \mathcal{P}_r} z^\top p$.
  $g_v \leftarrow \mathbb{I}[(s_f - s_r \geq \tau_v) \wedge (s_f \geq \tau_f)]$.
end if
$r \leftarrow g_t \vee g_v$.
if $r = 1$ **then**
  Enable ϕ^* and generate y using $p_{\theta_{\mathrm{R}}, \phi^*}(\cdot \mid x)$.
else
  Disable ϕ^* and generate y using $p_{\theta_{\mathrm{R}}}(\cdot \mid x)$.
end if
return y.

B ADDITIONAL EXPERIMENTAL RESULTS

B.1 RESULTS ON MLLMU-BENCH

The main results in Table 1 report the average performance across text-only and image-text queries. Here, we present the results for each modality separately to provide a more detailed view of forgetting and utility preservation. We additionally report results on the Test set, which evaluates the designated Forget identities using paraphrased questions and altered portraits. MIRRA maintains low target-answer recovery on this set across both modalities, supporting the generalization of its forgetting behavior to the transformed queries considered in the benchmark.

Text-only results. Table 5 reports performance when queries contain no images. MIRRA achieves the lowest or tied-lowest values on all Forget and Test metrics for each model. Forget classification accuracy decreases from 58.33%, 60.83%, and 43.33% for Vanilla to 1.67%, 0.83%, and 2.50% for Qwen3-VL-4B, Qwen3-VL-8B, and LLaVA-1.5-7B, respectively. On Test, classification accuracy is 0.83% for all three models, ROUGE-L is 0.009, 0.008, and 0.003, and cloze accuracy is 0.00%. These results show that the reduction in target-answer recovery extends to the benchmark's transformed textual queries.

This reduction is accompanied by comparatively small changes in performance on non-target queries. Retain ROUGE-L remains close to Vanilla, changing from 0.658 to 0.657, from 0.570 to 0.573, and from 0.627 to 0.627 across the three models. Retain classification improves for both Qwen models, although it decreases from 44.57% to 43.38% for LLaVA. By comparison, our MIP-Editor implementation on LLaVA reduces Forget ROUGE-L to 0.032 but also reduces Retain and Real Celebrity ROUGE-L to 0.028 and 0.033. MIRRA maintains these two utility scores at 0.627 and 0.626 while obtaining a lower Forget ROUGE-L of 0.005. The results support selective suppression of target answers, although MIRRA does not achieve the highest utility score in every column.

Table 5: Text-only results on MLLMU-Bench with 5% of fictitious identities designated for forgetting. Cls., R-L, and Cloze denote classification accuracy (%), ROUGE-L (0–1), and cloze accuracy (%), respectively. Test evaluates forgetting on the benchmark's transformed queries for target identities, with lower values indicating less recovery of target answers. Bold and underlined values indicate the best and second-best results within each model block, following the indicated metric directions; ties at the reported precision share a rank.

Method	Forget ↓			Test ↓			Retain ↑			Real Celebrity ↑		
	Cls.	R-L	Cloze	Cls.	R-L	Cloze	Cls.	R-L	Cloze	Cls.	R-L	Cloze
Qwen3-VL-4B-Instruct (5% Forget)												
Vanilla	58.33	0.647	10.42	60.83	0.519	6.25	61.06	0.658	10.13	74.38	0.595	8.17
GA	57.50	0.605	2.08	62.50	0.502	10.42	62.88	0.628	10.34	73.86	0.577	7.52
GA-Difference	52.50	0.612	14.58	61.67	0.456	4.17	59.87	0.661	8.97	66.80	0.509	**8.50**
KL Minimization	60.83	0.653	10.42	64.17	0.525	10.42	62.83	**0.663**	12.97	**74.64**	**0.601**	**8.50**
NPO	58.33	0.527	4.17	62.50	0.436	8.33	63.38	0.574	8.23	73.07	0.568	**8.50**
MMUnlearner	56.67	0.613	12.50	65.83	0.468	10.42	**64.99**	0.601	**14.56**	72.94	0.459	7.52
MIP-Editor	30.83	0.542	**0.00**	42.50	0.403	4.17	46.51	0.547	1.48	41.31	0.445	0.98
MIRRA (Ours)	**1.67**	**0.000**	**0.00**	**0.83**	**0.009**	**0.00**	63.64	0.657	12.55	73.99	0.593	7.84
Qwen3-VL-8B-Instruct (5% Forget)												
Vanilla	60.83	0.567	14.58	67.50	0.473	12.50	55.39	0.570	9.92	75.69	0.527	12.42
GA	57.50	0.571	10.42	62.50	0.461	6.25	54.76	0.598	8.65	75.82	**0.582**	11.11
GA-Difference	58.33	0.547	4.17	64.17	0.465	12.50	59.28	0.589	12.55	**77.52**	0.531	10.46
KL Minimization	61.67	0.560	20.83	63.33	0.469	10.42	56.24	0.593	10.55	75.56	0.528	**13.07**
NPO	59.17	0.594	14.58	61.67	0.484	10.42	54.76	0.608	10.13	76.34	0.578	10.78
MMUnlearner	60.83	0.576	14.58	69.17	0.512	10.42	**59.45**	**0.621**	**14.66**	75.56	0.498	10.78
MIP-Editor	60.83	0.520	10.42	65.83	0.429	6.25	54.97	0.533	6.43	76.99	0.513	12.42
MIRRA (Ours)	**0.83**	**0.000**	**0.00**	**0.83**	**0.008**	**0.00**	56.28	0.573	10.13	76.08	0.522	**13.07**
LLaVA-1.5-7B (5% Forget)												
Vanilla	43.33	0.589	19.57	40.00	0.470	6.52	44.57	**0.627**	9.41	58.04	0.625	20.26
GA	41.67	0.585	8.70	38.33	0.434	10.87	39.49	0.595	6.98	**59.61**	0.602	17.32
GA-Difference	41.67	0.532	10.87	35.83	0.442	6.52	40.17	0.600	9.83	56.60	0.589	17.32
KL Minimization	40.00	0.588	15.22	41.67	0.473	8.70	43.89	**0.627**	9.51	58.17	0.624	20.26
NPO	34.17	0.571	6.52	35.83	0.431	6.52	39.58	0.588	6.13	59.35	0.558	17.65
MMUnlearner	46.67	0.418	23.91	45.00	0.318	19.57	**51.50**	0.404	**24.21**	57.12	0.334	11.76
MIP-Editor	41.67	0.032	17.39	40.00	0.028	10.87	42.83	0.028	9.41	31.90	0.033	8.17
MIRRA (Ours)	**2.50**	**0.005**	**0.00**	**0.83**	**0.003**	**0.00**	43.38	**0.627**	9.30	58.69	**0.626**	**20.59**

Image-text results. Table 6 evaluates queries accompanied by images. MIRRA obtains 0.00% classification and cloze accuracy on Forget for all three models. On Test, classification accuracy decreases from 60.83%, 52.50%, and 29.17% for Vanilla to 1.67%, 2.50%, and 2.50%, respectively, while cloze accuracy is 0.00% throughout. Test ROUGE-L is 0.000 for both Qwen models and

0.018 for LLaVA. The low Test scores support generalization of target-answer suppression to the benchmark's transformed image-text queries, although classification remains above zero.

The comparison also shows why forgetting and utility should be assessed together. On LLaVA, our MIP-Editor implementation achieves lower Forget ROUGE-L and Test classification accuracy than MIRRA, with values of 0.000 versus 0.052 and 0.00% versus 2.50%, respectively. However, its Retain and Real Celebrity classification accuracies decrease to 14.13% and 7.83%, compared with 30.12% and 51.04% for MIRRA. MIRRA therefore provides a different trade-off, retaining substantially more non-target classification performance in this comparison. Preservation is still imperfect. For Qwen3-VL-4B, Retain classification decreases from 56.83% to 52.37%, even though Retain ROUGE-L remains close to Vanilla. These results indicate that strong forgetting can coexist with losses on particular utility metrics and do not establish that all retained capabilities are unchanged.

Table 6: Image-text results on MLLMU-Bench with 5% of fictitious identities designated for forgetting. Cls., R-L, and Cloze denote classification accuracy (%), ROUGE-L (0–1), and cloze accuracy (%), respectively. Test evaluates forgetting on the benchmark's transformed queries for target identities, with lower values indicating less recovery of target answers. Bold and underlined values indicate the best and second-best results within each model block, following the indicated metric directions; ties at the reported precision share a rank.

Method	Forget ↓			Test ↓			Retain ↑			Real Celebrity ↑		
	Cls.	R-L	Cloze	Cls.	R-L	Cloze	Cls.	R-L	Cloze	Cls.	R-L	Cloze
Qwen3-VL-4B-Instruct (5% Forget)												
Vanilla	58.33	0.574	14.58	60.83	0.366	14.58	**56.83**	0.543	9.60	72.98	0.451	16.99
GA	60.83	0.546	14.58	36.67	0.291	14.58	53.07	0.524	13.82	70.76	0.430	15.36
GA-Difference	54.17	0.182	12.50	27.50	0.263	10.42	43.49	0.265	14.03	64.88	0.271	**17.65**
KL Minimization	58.33	0.593	12.50	42.50	0.358	14.58	51.86	0.545	14.77	**73.11**	0.441	16.34
NPO	68.33	0.484	6.25	36.67	0.261	8.33	53.09	0.429	9.60	70.50	0.388	11.76
MMUnlearner	64.17	0.292	10.42	50.00	0.235	12.50	53.93	0.315	**18.88**	66.45	0.226	10.78
MIP-Editor	26.67	0.477	4.17	37.50	0.246	2.08	44.92	0.462	5.17	9.40	0.344	0.00
MIRRA (Ours)	**0.00**	**0.000**	**0.00**	**1.67**	**0.000**	**0.00**	52.37	**0.546**	12.97	72.58	**0.452**	16.99
Qwen3-VL-8B-Instruct (5% Forget)												
Vanilla	56.67	0.367	6.25	52.50	0.163	8.33	51.48	0.353	8.86	76.24	0.404	27.78
GA	58.33	0.439	4.17	46.67	0.229	8.33	49.79	0.410	10.02	76.76	**0.455**	25.49
GA-Difference	55.00	0.453	2.08	53.33	0.201	14.58	51.78	0.427	4.01	**77.94**	0.401	25.49
KL Minimization	57.50	0.372	6.25	52.50	0.173	8.33	51.95	0.351	9.49	77.28	0.387	28.10
NPO	55.83	0.415	4.17	49.17	0.205	10.42	50.97	0.395	8.86	77.28	0.442	25.82
MMUnlearner	62.50	0.439	6.25	60.00	0.225	12.50	**54.10**	**0.451**	**10.86**	73.37	0.369	25.82
MIP-Editor	59.17	0.344	8.33	54.17	0.163	10.42	50.85	0.344	8.76	75.98	0.381	**29.08**
MIRRA (Ours)	**0.00**	**0.000**	**0.00**	**2.50**	**0.000**	**0.00**	51.78	0.355	9.81	76.37	0.406	28.43
LLaVA-1.5-7B (5% Forget)												
Vanilla	20.00	0.110	13.04	29.17	0.188	10.87	31.39	0.105	11.52	50.52	0.111	17.97
GA	24.17	0.036	13.04	32.50	0.094	2.17	34.01	0.026	10.04	49.35	0.051	17.97
GA-Difference	12.50	0.064	10.87	28.33	0.151	10.87	24.03	0.073	11.52	**53.00**	0.099	**22.88**
KL Minimization	20.00	0.122	13.04	28.33	0.196	10.87	29.61	**0.121**	11.73	51.04	**0.120**	17.32
NPO	21.67	0.026	10.87	25.83	0.047	2.17	30.16	0.016	11.42	47.39	0.037	16.34
MMUnlearner	31.67	0.106	13.04	40.00	0.097	17.39	**40.91**	0.111	**24.74**	33.29	0.092	3.59
MIP-Editor	**0.00**	**0.000**	**0.00**	**0.00**	0.019	**0.00**	14.13	0.059	6.03	7.83	0.033	0.00
MIRRA (Ours)	**0.00**	0.052	**0.00**	2.50	**0.018**	**0.00**	30.12	0.111	11.42	51.04	0.116	18.30

B.2 Results on CLEAR

To examine whether MIRRA is also effective on another dataset, we evaluate it on the CLEAR 5% Forget split using Qwen3-VL-4B, Qwen3-VL-8B, and LLaVA-1.5-7B. Following the official image-text evaluation protocol, we assess captioning and face recognition on the Forget and Retain sets, together with Real World and Real Face accuracies and the aggregate Model Utility score. For each model, MIRRA uses a base model adapted on Retain data with CE and KL objectives, a newly trained refusal adapter, and the reused identity gate. Table 7 compares this configuration with Vanilla and four unlearning baselines.

Table 7: Results on CLEAR with 5% of identities designated for forgetting under the official image-text evaluation protocol. Caption and face-recognition performance is measured by ROUGE-L recall (R-L, 0–1). Real World and Real Face accuracies are reported as percentages. Model Utility is the harmonic mean of the non-forget official probability, ROUGE, and truth-ratio metrics. For all three models, MIRRA uses a Retain CE+KL-adapted Vanilla base, a newly trained refusal adapter, and the reused identity gate. Bold and underlined values indicate the best and second-best results within each model block, following the indicated metric directions; ties at the reported precision share a rank. Pure-text QA is not included.

Method	Forget ↓		Retain ↑		General Utility ↑		
	Caption R-L	FaceRec R-L	Caption R-L	FaceRec R-L	Real World Acc.	Real Face Acc.	Model Utility
Qwen3-VL-4B-Instruct (CLEAR, 5% Forget)							
Vanilla	0.347	0.102	0.351	0.055	<u>12.81</u>	**77.48**	0.2247
GA	0.344	<u>0.097</u>	0.347	0.054	12.53	<u>76.82</u>	0.2240
GA-Difference	0.350	0.104	0.355	<u>0.073</u>	12.26	73.51	<u>0.2625</u>
KL Minimization	0.356	0.102	<u>0.356</u>	0.068	**13.08**	75.50	0.2532
NPO ($\beta = 0.1$)	<u>0.343</u>	<u>0.097</u>	0.344	0.053	12.26	**77.48**	0.2216
MIRRA (Ours)	**0.008**	**0.000**	**0.501**	**0.426**	<u>12.81</u>	59.60	**0.4207**
Qwen3-VL-8B-Instruct (CLEAR, 5% Forget)							
Vanilla	0.367	0.090	0.369	0.058	<u>11.99</u>	**79.47**	0.2414
GA	0.130	0.009	0.122	0.015	7.08	0.00	0.0284
GA-Difference	0.366	0.055	<u>0.385</u>	<u>0.106</u>	11.44	68.21	<u>0.3233</u>
KL Minimization	0.369	0.068	0.382	0.084	<u>11.99</u>	<u>75.50</u>	0.2940
NPO ($\beta = 0.1$)	<u>0.116</u>	<u>0.008</u>	0.117	0.015	6.81	0.00	0.0283
MIRRA (Ours)	**0.005**	**0.000**	**0.545**	**0.781**	**12.26**	59.60	**0.4551**
LLaVA-1.5-7B (CLEAR, 5% Forget)							
Vanilla	<u>0.424</u>	<u>0.011</u>	0.427	0.024	7.63	**39.74**	0.1660
GA	0.429	0.013	0.423	0.026	7.36	37.75	0.1704
GA-Difference	0.450	0.013	<u>0.454</u>	0.032	7.63	28.48	<u>0.2014</u>
KL Minimization	0.443	0.019	0.441	<u>0.033</u>	<u>7.90</u>	33.77	0.2007
NPO ($\beta = 0.1$)	0.425	0.012	0.418	0.025	7.36	<u>39.07</u>	0.1675
MIRRA (Ours)	**0.009**	**0.000**	**0.496**	**0.176**	**8.17**	36.42	**0.3754**

Forgetting effectiveness. MIRRA achieves the lowest Forget ROUGE-L recall on both captioning and face recognition for all three models. Forget Caption R-L decreases from 0.347, 0.367, and 0.424 for Vanilla to 0.008, 0.005, and 0.009, respectively. Forget FaceRec R-L is 0.000 for all three models at the reported precision. The consistently low overlap with target answers across the two tasks shows that target-answer suppression also holds in the evaluated CLEAR setting.

Retained knowledge and general utility. MIRRA also obtains the highest Retain Caption R-L, Retain FaceRec R-L, and Model Utility among the compared methods for every model. Retain Caption R-L increases from 0.351 to 0.501, from 0.369 to 0.545, and from 0.427 to 0.496, while Retain FaceRec R-L increases from 0.055 to 0.426, from 0.058 to 0.781, and from 0.024 to 0.176. Model Utility rises from 0.2247, 0.2414, and 0.1660 to 0.4207, 0.4551, and 0.3754, respectively. This combination distinguishes MIRRA from baselines that improve only one side of the forgetting and retention trade-off. For example, on Qwen3-VL-8B, GA-Difference reaches a Model Utility of 0.3233 but leaves Forget Caption R-L at 0.366, close to Vanilla's 0.367. GA and NPO reduce the Forget scores but lower Model Utility to 0.0284 and 0.0283.

The aggregate utility improvement does not extend to every task. Real World accuracy is unchanged for Qwen3-VL-4B and improves for the other two models, whereas Real Face accuracy decreases from 77.48% to 59.60%, from 79.47% to 59.60%, and from 39.74% to 36.42%. Thus, the CLEAR results support strong suppression of target answers and improved performance on retained identities, while also revealing a trade-off in real-person recognition. Together with the MLLMU-Bench experiments, they provide evidence that MIRRA is effective under more than one benchmark configuration.

B.3 Detailed Information on Cost Analysis

We report computational costs on MLLMU-Bench with a 5% forgetting ratio for Qwen3-VL 4B, Qwen3-VL 8B, and LLaVA-1.5 7B. All times are reported in minutes and computed by summing the recorded durations of training and prerequisite stages.

For NPO, the reported cost includes reference-model training. For MMUnlearner, it includes Fisher estimation. For MIP-Editor, it includes visual and textual attribution. For MIRRA, the reported cost includes gate construction, retain adaptation, and refusal-adapter training. Each model-specific total in Table 2 includes the 14.93-minute gate-construction cost. When the same gate is shared across backbone models, this cost is incurred only once rather than separately for each model.

C Implementation Details

C.1 Baseline objectives

We describe the baselines used in the MLLMU-Bench 5% forgetting setting. Let $\mathcal{D}_f$ and $\mathcal{D}_r$ denote the forget and retain data, and let θ_0 denote the Vanilla checkpoint before unlearning. For a training batch B, the mean cross-entropy over supervised tokens is

$$\ell_\theta(B) = -\frac{1}{N_B} \sum_{i \in B} \sum_{t \in \mathcal{T}_i} \log p_\theta(y_{i,t} \mid x_i, y_{i,<t}), \qquad N_B = \sum_{i \in B} |\mathcal{T}_i|, \tag{17}$$

where $\mathcal{T}_i$ contains the positions included in the loss. We write $\mathcal{L}_s(\theta) = \mathbb{E}_{B \sim \mathcal{D}_s} \ell_\theta(B)$, for $s \in \{f, r\}$. All objectives below use the minimization convention.

Gradient Ascent (GA). GA increases the prediction loss on the forget data by minimizing

$$\mathcal{J}_{\mathrm{GA}}(\theta) = -\mathcal{L}_f(\theta). \tag{18}$$

The implementation negates the model's cross-entropy before backpropagation and uses only forget data during this stage.

GA-Difference. GA-Difference combines ascent on forget data with descent on retain data:

$$\mathcal{J}_{\mathrm{GD}}(\theta) = -\mathcal{L}_f(\theta) + \lambda_r \mathcal{L}_r(\theta). \tag{19}$$

The code applies no additional retain-loss multiplier ($\lambda_r = 1$). Within each epoch, it first traverses the forget loader and then the retain loader, updating parameters separately. Equation 19 summarizes these directions; the implementation does not jointly evaluate both dataset-average losses at each step. Their relative update counts also depend on the sizes of the two loaders.

KL Minimization. This baseline performs GA on forget data and optimizes both retain cross-entropy and a KL constraint relative to Vanilla:

$$\mathcal{J}_{\mathrm{KL}}(\theta) = -\mathcal{L}_f(\theta) + \mathcal{L}_r(\theta) + \lambda_{\mathrm{KL}} \mathbb{E}_{B \sim \mathcal{D}_r} \mathcal{K}(B), \qquad \lambda_{\mathrm{KL}} = 1, \tag{20}$$

$$\mathcal{K}(B) = \frac{1}{|B|} \sum_{i \in B} \sum_{t=1}^{T_B} \sum_{v \in \mathcal{V}} p_{\theta_0}(v \mid z_{i,<t}) \log \frac{p_{\theta_0}(v \mid z_{i,<t})}{p_\theta(v \mid z_{i,<t})}. \tag{21}$$

Here z_i is the full input sequence and T_B is the batch sequence length. This matches the implemented `batchmean` reduction and the direction $D_{\mathrm{KL}}(p_{\theta_0} \| p_\theta)$. The KL computation has no additional answer-token or padding mask. As in GA-Difference, forget and retain passes are performed sequentially within each epoch.

Negative Preference Optimization (NPO). The sequence-probability formulation of NPO Zhang et al. (2024) is

$$\mathcal{J}_{\mathrm{NPO}}(\theta) = -\frac{2}{\beta} \mathbb{E}_{(x,y) \sim \mathcal{D}_f} \log \sigma \left(-\beta \log \frac{p_\theta(y \mid x)}{p_{\mathrm{ref}}(y \mid x)} \right), \tag{22}$$

Table 8: Documented Vanilla fine-tuning hyperparameters on MLLMU-Bench.

Model	Epochs	Batch size	Accum.	Learning rate	Optimizer	LoRA
Qwen3-VL-4B	3	1	4	2×10^{-5}	AdamW	Yes
Qwen3-VL-8B	3	1	4	2×10^{-5}	AdamW	Yes
LLaVA-1.5-7B	3	1	4	2×10^{-5}	AdamW	Yes

where σ is the sigmoid function. The benchmark implementation uses a reference model obtained by fine-tuning the original pretrained model on retain data, and keeps its parameters fixed during NPO. Specifically, it uses the scalar batch-mean cross-entropies returned by the two models:

$$\widehat{\mathcal{J}}_{\mathrm{NPO}}(B) = -\frac{2}{\beta} \log \sigma \left(\frac{\beta}{a} [\ell_\theta(B) - \ell_{\mathrm{ref}}(B)] \right). \tag{23}$$

The implementation sets $a = 1$ without accumulation and $a = 4$ with accumulation. Thus, the division by a occurs inside the sigmoid, and the loss is reduced over tokens before the nonlinear transformation. These details differ from the per-example sequence objective in Equation 22. The recorded recipes use $\beta = 0.1$; for Qwen3-VL-4B, this value is inferred from the run identifier.

MMUnlearner. MMUnlearner Huo et al. (2025) constructs a saliency mask from forgetting and preservation tasks. The implemented statistic averages squared batch-loss gradients:

$$F_j(\mathcal{D}) = \frac{1}{|\mathcal{B}_\mathcal{D}|} \sum_{B \in \mathcal{B}_\mathcal{D}} \left(\frac{\partial \ell_{\theta_0}(B)}{\partial \theta_j} \right)^2, \qquad m_j = \mathbf{1} \left[\frac{F_j(\mathcal{D}_f) + \epsilon}{F_j(\mathcal{D}_p) + \epsilon} \geq \tau \right]. \tag{24}$$

For language parameters, $\mathcal{D}_p$ concatenates training text and retain image–text examples; for vision parameters, it contains retain image–text examples. We use $\tau = 1$ and $\epsilon = 10^{-15}$. Forget and retain steps send the following gradients to a shared AdamW optimizer:

$$g_f = \widetilde{m} \odot \mathrm{clip}_1(-\gamma \nabla_\theta \ell_\theta(B_f)), \qquad g_r = \mathrm{clip}_1(\nabla_\theta \ell_\theta(B_r)), \qquad \gamma = 1. \tag{25}$$

Here $\widetilde{m}$ equals m on selected mask-managed coordinates and one elsewhere. We use the original-weight training variant: all weights are trainable, and the mask is applied only to selected forget gradients after global-norm clipping. It is not a LoRA variant, and a zero mask does not guarantee an unchanged weight under shared optimizer momentum and retain updates.

MIP-Editor. MIP-Editor Li et al. (2026) localizes influential paths using visual Fisher attribution and textual integrated gradients. Our implementation unions the per-layer top-k selections across forget examples, applies activation pruning, and edits selected FFN weight rows, with $k = 5$. For the output $h_\theta^{(l)}(z)$ of the selected module, the implemented representation objective is

$$\mathcal{J}_{\mathrm{MIP}}(\theta) = \alpha \mathbb{E}_{z \sim \mathcal{D}_r} \mathrm{MSE}(h_\theta^{(l)}(z), h_{\theta_0}^{(l)}(z)) + \beta_{\mathrm{MIP}} \mathbb{E}_{z \sim \mathcal{D}_f} \mathrm{MSE}(h_\theta^{(l)}(z), cu). \tag{26}$$

The fixed control vector is $u = \widetilde{u}/\|\widetilde{u}\|_2$, with $\widetilde{u}_j \sim U(0,1)$; $c = 10$, $\alpha = \beta_{\mathrm{MIP}} = 0.5$, and $l = 7$ is a zero-based module index. MSE averages over the captured activation tensor. We follow the released two-MSE implementation, without adding the retain CE term in the paper's full objective. Each step pairs a retain batch with a forget batch; forget examples are sampled with replacement to match the retain traversal length.

C.2 Training hyperparameters

Table 8 summarizes the documented fine-tuning recipes for the Vanilla checkpoints before unlearning. Table 9 summarizes the unlearning configurations recovered from saved settings, training logs, and launch scripts. Batch size denotes the per-process DataLoader batch size and Accum. denotes how many DataLoader batches are accumulated before an optimizer update.

Additional settings. The Vanilla fine-tuning recipe uses LoRA with rank 16, scaling parameter 16, and dropout 0.05, with a maximum sequence length of 1024. The current GA, GA-Difference, KL Minimization, and NPO implementations use the same LoRA settings and AdamW. MMUnlearner

Table 9: Unlearning hyperparameters on MLLMU-Bench with a 5% forget ratio.

Model	Method	Epochs	Batch size	Accum.	Learning rate
Qwen3-VL-4B	GA	4	4	1	5×10^{-5}
	GA-Difference	4	1	2	5×10^{-5}
	KL Minimization	4	1	1	5×10^{-5}
	NPO	4	1	1	5×10^{-5}
	MMUnlearner	2	4	1	1×10^{-5}
	MIP-Editor	4	4	1	2×10^{-5}
Qwen3-VL-8B	GA	4	1	4	5×10^{-5}
	GA-Difference	4	1	2	5×10^{-5}
	KL Minimization	4	1	4	5×10^{-5}
	NPO	4	1	4	5×10^{-5}
	MMUnlearner	2	4	1	1×10^{-5}
	MIP-Editor	4	4	1	2×10^{-5}
LLaVA-1.5-7B	GA	4	1	4	5×10^{-5}
	GA-Difference	4	1	2	5×10^{-5}
	KL Minimization	4	1	4	5×10^{-5}
	NPO	4	1	4	5×10^{-5}
	MMUnlearner	2	4	1	2×10^{-5}
	MIP-Editor	4	4	1	2×10^{-5}

Table 10: MIRRA training hyperparameters. Batch size denotes the per-device microbatch size.

Setting	Retain adaptation	Refusal adapter
Epochs	4	2
Learning rate	5×10^{-5}	2×10^{-5}
Batch size	1	1
Gradient accumulation	1	4
Optimizer	AdamW	AdamW
Maximum sequence length	1024	1024
LoRA rank / alpha	16 / 16	8 / 16
LoRA dropout	0.05	0.05
KL coefficient	1.0	—

uses AdamW and a Fisher batch size of 4. For LLaVA MMUnlearner, a logical batch of 4 is processed as microbatches of 1 with token-weighted gradient accumulation. MIP-Editor uses Adam and each update pairs one retain batch of 4 with one forget batch of 4. Its Fisher attribution and integrated gradients each use 4 integration steps. The NPO epochs in the table cover the unlearning stage only.

C.3 MIRRA Implementation Details

Table 10 summarizes the MIRRA training settings for the 5% forgetting experiments on MLLMU-Bench and CLEAR.

Parameter updates. Both stages apply LoRA to the language-side attention projections and feed-forward projections, with the underlying model weights frozen. Retain-adaptation updates are merged into the base model, which is then frozen for refusal training. The refusal adapter remains unmerged for conditional activation at inference time.

Retain adaptation. The objective combines retain-answer cross-entropy with $D_{\mathrm{KL}}(p_{\theta_0} \| p_\theta)$ against the frozen Vanilla checkpoint, using $\lambda_{\mathrm{KL}} = 1$. MLLMU-Bench uses full-sequence, unshifted KL with `batchmean` reduction and no additional answer-token or padding mask. Its linear scheduler has no warmup, uses a horizon of loader batches times epochs, and advances once per epoch. CLEAR computes KL in float32 over causally shifted, valid response positions and averages

over these tokens. Its linear scheduler advances per optimizer update, with $\max(1, \lfloor U/20 \rfloor)$ warmup steps for U total updates and gradient-norm clipping at 1.0. The retain-adaptation seed is 42.

Refusal supervision. We replace Forget answers with natural-language refusals and supervise only assistant-response tokens. MLLMU-Bench uses 181 training examples with three targets: "I cannot answer.", "I cannot provide that information.", and "I am unable to answer that request." CLEAR uses 386 training examples with the target "I'm sorry, but I can't provide information about that person." The validation sets contain 48 and 170 examples, respectively. Refusal training uses linear decay with a 0.05 warmup ratio, zero weight decay, gradient-norm clipping at 1.0, and Trainer and data seeds of 42. We retain the final adapter after two epochs.

D Detailed Related Work

D.1 Machine Unlearning in Language and Multimodal Models

Machine unlearning aims to remove the influence of selected training data while maintaining performance on the remaining data. Cao & Yang (2015) study how learning systems can support data removal without repeating the entire training process. SISA training (Bourtoule et al., 2021) reduces the cost of subsequent deletion by partitioning data into shards and slices, allowing retraining to be confined to the affected components. These approaches make the organization of training data and intermediate model states part of the unlearning procedure. For pretrained language models, research has increasingly focused on updating an existing checkpoint using a relatively small collection of forget and retain examples.

Gradient ascent is a common approach to language model unlearning. It lowers the likelihood of target responses by maximizing their training loss, often alongside an objective that preserves performance on retained data (Yao & Xu, 2024). Negative Preference Optimization (NPO) (Zhang et al., 2024) treats forget responses as negative preferences and reduces their probability relative to a reference model. Its objective slows the progression toward the model collapse observed with unrestricted gradient ascent. This line of research studies the optimization trade-off between suppressing target information and maintaining the model's broader language capabilities.

Multimodal unlearning extends the forgetting scope to visual concepts and associations between images and text. Single Image Unlearning (SIU) (Li et al., 2024) constructs fine-tuning examples from one image associated with a target concept and combines cross-entropy with a dual masked KL-divergence objective. MMUnlearner (Huo et al., 2025) studies the removal of an entity's visual patterns while preserving its textual knowledge. It uses geometry-constrained gradient ascent with a saliency mask to selectively suppress target visual knowledge while preserving non-target visual knowledge and corresponding textual knowledge. MIP-Editor (Li et al., 2026) instead locates influential neuron paths and edits them for cross-modal unlearning. These methods examine forgetting at different levels, from visual recognition to knowledge shared across modalities. Safety-oriented research further considers which responses should remain available after unlearning. SafeEraser (Chen et al., 2025) introduces a multimodal safety benchmark and a Prompt Decouple loss to reduce excessive refusal on benign requests. Its Safe Answer Refusal Rate measures the loss of access to otherwise permissible answers.

D.2 Refusal Learning and Inference-Time Unlearning

Refusal learning has become an important direction in language model unlearning, with attention to both the content and the applicability of refusal responses. RULE (Zhang et al., 2026) formulates unlearning as refusal-boundary optimization. It uses reinforcement learning with forget queries and synthesized boundary queries, rewarding refusal on targeted requests and helpful responses on permissible ones. In multimodal models, ASRU (Guang et al., 2026) first induces refusal behavior through activation redirection and then refines the refusal boundary using reinforcement learning. Its reward design incorporates generation quality alongside forgetting and utility preservation. Both studies extend the optimization target beyond reducing the probability of memorized answers to learning appropriate responses to related requests.

Research on internal representations provides another basis for controlling refusal. Arditi et al. (2024) identify a direction in the residual stream that mediates refusal in the models they examine. Manipulating this direction can suppress or induce refusal behavior. MLLMEraser (Ding et al., 2025) applies activation steering to multimodal unlearning. It constructs an erasure direction from contrasting image-text pairs and adjusts the intervention according to the input. GUARD-IT (Turani et al., 2026) uses a gated, norm-preserving rotation of residual-stream activations to perform inference-time unlearning with fixed model weights. GUARD (Deng et al., 2025) operates on the generation process. A prompt classifier identifies targeted content, after which token matching and semantic matching guide the penalization and filtering of candidate tokens. These approaches broaden the available interventions from training objectives to hidden representations and decoding decisions.

D.3 Local Model Editing and Parameter-Efficient Adaptation

Model editing aims to revise selected model behaviors while preserving predictions on unrelated inputs. Memory-based approaches store individual edits and retrieve them when relevant queries arrive. SERAC (Mitchell et al., 2022) maintains an external edit memory and uses a scope classifier to choose between a counterfactual model and the original model. GRACE (Hartvigsen et al., 2023) stores local key-value mappings in a model's representation space, allowing successive corrections without changing the pretrained weights. Both approaches use localized stored information to support edits across a sequence of requests.

WISE (Wang et al., 2024) maintains pretrained knowledge in a main memory and edited knowledge in a side memory, with a learned router selecting the appropriate memory for each query. It also introduces knowledge sharding and merging to accommodate successive edits. Parameter-efficient adaptation provides a complementary approach to modifying large models. LoRA (Hu et al., 2021) learns low-rank weight updates while freezing pretrained parameters, reducing the number of trainable parameters and the storage required for each adaptation. Memory-based editing and low-rank adaptation thus offer different mechanisms for incorporating localized changes into a pretrained model.

D.4 Face Verification and Open-Set Recognition

Face verification determines whether two face images depict the same person by comparing their learned representations. ArcFace (Deng et al., 2019) introduces an additive angular margin during training to improve the discrimination of normalized face embeddings. Its angular formulation supports similarity-based comparison between reference and query images. Face detection supplies the localized image regions used by recognition models. SCRFD (Guo et al., 2022) studies sample and computation redistribution to improve the efficiency and accuracy of this stage.

Recognition beyond a fixed set of known classes also requires a mechanism for handling unfamiliar inputs. Scheirer et al. (2013) formalize open-set recognition by accounting for unknown classes and the risk of assigning unsupported regions of the feature space to known classes. This setting extends conventional classification to decisions about whether an input has sufficient evidence for acceptance. Face representation learning and open-set recognition consequently address complementary aspects of identity matching, namely feature discrimination and the treatment of unknown identities.

D.5 Unlearning Benchmarks and Evaluation

Unlearning benchmarks differ in the source of target knowledge and the capabilities they assess. TOFU (Maini et al., 2024) introduces fictitious author profiles to create a controlled setting for learning and subsequently forgetting selected knowledge. MUSE (Shi et al., 2025) evaluates unlearning across six dimensions, covering memorization, knowledge, privacy leakage, utility, scalability, and successive unlearning. For multimodal models, MLLMU-Bench (Liu et al., 2025) evaluates privacy-related forgetting through visual and textual tasks. CLEAR (Dontsov et al., 2025) studies character unlearning across the two modalities. These benchmarks extend evaluation from individual target answers to the effects of forgetting on related knowledge and retained capabilities.

Evaluation methods also examine different forms of information recovery. ROUGE (Lin, 2004) quantifies overlap between generated text and reference answers, providing a measure of textual

reproduction. Hu et al. (2025) investigate recovery through benign relearning and show that fine-tuning on benign, loosely related data can restore previously suppressed knowledge in several unlearned models. Łucki et al. (2025) evaluate unlearning from an adversarial perspective, studying how prompting and model modifications can recover restricted capabilities. Together, benchmark evaluation and recovery studies examine complementary properties of unlearned models, including their immediate responses, retained utility, and behavior under subsequent interventions.