BLINDFOLD-TO-FORGETTING: ROUTE-AWARE MULTIMODAL UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal unlearning aims to remove the influence of specific training data from models that process images and text while preserving performance on non-target data. Since the same target information can be elicited through distinct image- and text-based routes, intervention should account for their differences. Whereas prior work has primarily focused on editing modality-specific neurons or internal paths, we propose Blindfold-to-Forgetting, which uses the visual-to-language interface as a selective intervention point. The first stage shifts forget-image representations toward the mean representation of generic face images while preserving retained-image representations. The second stage suppresses target responses under image inputs processed by the original projector and text-only inputs, using a hinge objective and retain supervision to limit excessive suppression and utility degradation. Across three benchmarks and three MLLMs, our method achieves the highest Forget-Retain harmonic mean in 8 of 12 evaluation settings on UMU-Bench and CLEAR, with improvements of up to 18.1 points over the strongest baseline. Relearning experiments further show limited target-response recovery under the tested low-data conditions.

1 INTRODUCTION

Models trained on large-scale datasets can memorize personal information or document contents and subsequently reproduce them in their responses (Carlini et al., 2021). Addressing data deletion requests therefore requires removing not only the original data but also their influence on the trained model. Machine unlearning aims to remove the influence of specified training data while preserving the model's utility on the remaining data (Guo et al., 2019; Bourtoule et al., 2021). A model retrained from scratch without the data designated for removal, known as the forget set, provides a natural reference. However, repeating this process for every deletion request is computationally expensive, particularly for large-scale models. This motivates methods that remove the influence of target data at a lower computational cost than full retraining.

In language models, this objective is approximated by modifying the learned parameters. Representative approaches include fine-tuning on data in which references to the forgetting targets are replaced, or reducing the probability of forget answers while preserving performance on retained data (Eldan & Russinovich, 2023; Maini et al., 2024). Negative preference optimization (NPO) uses a preference-based objective to suppress forget answers, aiming to mitigate the catastrophic model degradation associated with gradient ascent (GA) (Zhang et al., 2024).

For multimodal large language models (MLLMs), unlearning must account for multiple input routes through which the same target information can be elicited. For example, information about a person's birthplace can be requested either by presenting a facial image with a question or by specifying the person's name without an image (Liu et al., 2025a; Dontsov et al., 2025). If the model can still provide this information in response to a text-only question despite failing to do so under an image-conditioned question, the unlearning effect is limited to a particular input route. Recent multimodal unlearning methods have pursued cross-modal forgetting by identifying and pruning important neurons for each modality or by locating modality-specific influential paths and editing their representations (Liu et al., 2025b; Li et al., 2026). However, these approaches primarily target neurons or internal representation paths, while the visual-to-language interface itself, which conveys visual information to the language model, remains relatively underexplored as a selective intervention point for unlearning.

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.

Many MLLMs use connecting modules such as visual projectors to transform vision encoder outputs into representations that the language model can process. The design of these modules affects how visual information is conveyed to the language model and the resulting multimodal capabilities (Liu et al., 2023; Li et al., 2024; Cha et al., 2024). Prior work on multimodal model editing has also shown that editing behavior can vary depending on the component being edited (Cheng et al., 2023). Based on these observations, we view the visual interface as an intervention point for selectively modifying the visual representations conveyed through image-conditioned inputs.

To this end, we propose BLINDFOLD-TO-FORGETTING, a route-aware multimodal unlearning method that combines visual representation retargeting with response suppression. The first stage shifts the projector outputs of forget images toward the mean representation of generic face images while keeping retain-image representations close to their original counterparts. This modifies the visual cues associated with forgetting targets while limiting changes to non-target visual representations. To suppress target responses without relying solely on these visual representation changes, the second stage directly reduces the probability of target answers under image inputs processed by the original projector and text-only inputs that bypass the visual interface. A hinge objective limits excessive suppression by producing no gradient from the corresponding suppression term once the target answer's log-probability falls below a predefined threshold, while retain supervision preserves performance on non-target queries.

We evaluate image-conditioned and text-only forgetting and retained utility on UMU-Bench (Wang et al., 2026), CLEAR (Dontsov et al., 2025), and MLLMU-Bench (Liu et al., 2025a), using three MLLMs. Our method improves the Forget–Retain H-Mean by up to 18.1 points over the strongest baseline on UMU-Bench and CLEAR, achieving the highest score in 8 of 12 evaluation settings. Component analysis shows that visual representation retargeting helps preserve image-conditioned utility, while training on the text-only route strengthens text-only forgetting. Motivated by prior findings that unlearned knowledge can recover after only a small amount of additional training (Hu et al., 2025), we further evaluate whether combining visual representation changes with response suppression limits such recovery. Our method maintains strong image-conditioned forgetting even after multimodal relearning with limited data.

Our contributions are as follows:

- We propose BLINDFOLD-TO-FORGETTING, which uses the visual-to-language interface as a selective intervention point for unlearning. By retargeting forget-image representations while preserving retain-image representations, visual representation retargeting jointly addresses image-conditioned forgetting and utility preservation.
- We design route-aware suppression to directly suppress target responses under image inputs processed by the original projector and text-only inputs. We combine this with visual representation retargeting, using a hinge objective and retain supervision to limit excessive suppression and degradation on non-target data.
- Evaluations across three benchmarks and three MLLMs demonstrate an improved balance between forgetting and retained utility across diverse backbones and input routes. Component analysis highlights the complementary roles of our components, while multimodal relearning experiments show limited image-conditioned target-response recovery under the tested low-data conditions.

2 RELATED WORK

Machine Unlearning for Language Models. LLM unlearning aims to reduce the influence of target data while preserving non-target knowledge and general generation capabilities (Jang et al., 2023). Common approaches combine GA to lower the likelihood of forget answers with training on retained data or distribution matching against a reference model (Yao & Xu, 2024). NPO formulates forgetting as a preference-based objective to mitigate catastrophic collapse associated with GA (Zhang et al., 2024). Another line of work modifies the training targets. While some methods train models to produce refusal responses (Maini et al., 2024), Eldan & Russinovich (2023) replace expressions specific to the forgetting targets with generic alternatives and fine-tune the model using alternative labels.

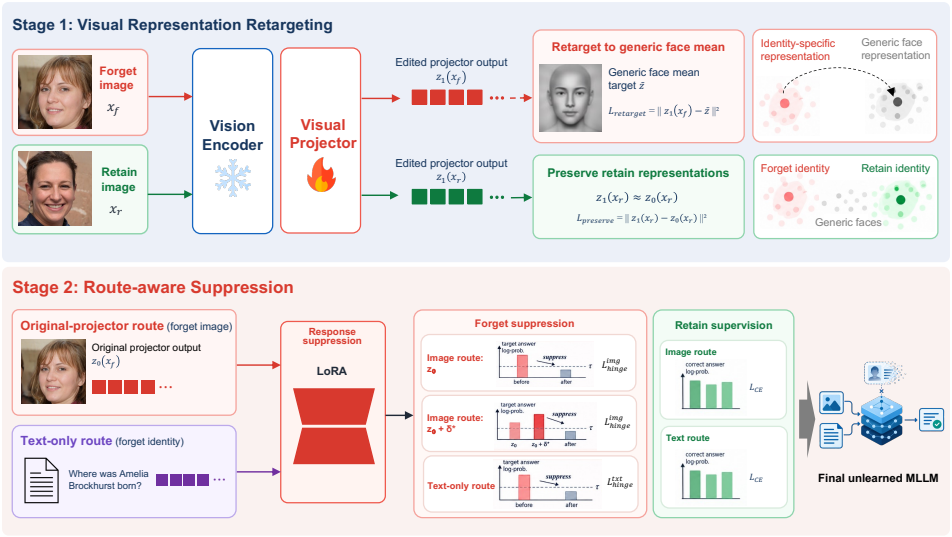


Figure 1: Overview of BLINDFOLD-TO-FORGETTING. Stage 1 retargets forget-image representations toward a generic-face mean while preserving retain and anchor representations. Stage 2 trains LoRA adapters in the vision encoder and language model to suppress target responses using original-projector outputs, their adversarially perturbed counterparts, and text-only inputs, together with retain supervision on edited-projector outputs and text-only inputs.

Multimodal Unlearning. Existing multimodal unlearning methods differ in the scope of knowledge they target and the components they modify. MMUnlearner (Huo et al., 2025) targets visual patterns while preserving related textual knowledge, using saliency-based constraints to control parameter updates. MANU (Liu et al., 2025b) selectively prunes neurons based on their importance to forget and retain data for each modality. MIP-Editor (Li et al., 2026) identifies modality-specific influential neuron paths and applies representation misdirection along these paths. Whereas these studies focus on identifying important internal editing locations for each modality, we incorporate structural differences between input routes into our intervention design. Image inputs pass through the visual projector, while text-only inputs do not. The projector therefore enables selective modification of visual representations, but text-only responses require a separate intervention. Accordingly, we combine visual representation retargeting, which constrains changes to retained representations, with target-response suppression under image-conditioned and text-only inputs.

3 METHOD

3.1 PROBLEM SETTING

We study multimodal unlearning that suppresses information about specific individuals (identities) under both image-conditioned and text-only inputs. We consider MLLMs comprising a vision encoder e_ω, a visual projector P_ϕ, and a language model g_ψ. For image-conditioned inputs, the model transforms an image x into visual tokens $z = P_\phi(e_\omega(x))$, which are passed to the language model together with a question q to generate a response. Here, the visual projector is a learnable connecting module that maps visual features into the language model's input embedding space. For text-only inputs, the model generates a response from the question q without an image. We denote the full model parameters by $\theta = (\omega, \phi, \psi)$.

We define the sets of forget and retain identities as $\mathcal{I}_\mathrm{F}$ and $\mathcal{I}_\mathrm{R}$, respectively, with $\mathcal{I}_\mathrm{F} \cap \mathcal{I}_\mathrm{R} = \varnothing$. Following the QA settings of existing multimodal unlearning benchmarks (Wang et al., 2026; Dontsov et al., 2025; Liu et al., 2025a), we use the corresponding image-conditioned QA datasets $\mathcal{F}_\mathrm{img}$ and $\mathcal{R}_\mathrm{img}$, and text-only QA datasets $\mathcal{F}_\mathrm{txt}$ and $\mathcal{R}_\mathrm{txt}$. An image-conditioned example consists of (x, q, a), while a text-only example consists of (q, a), where a is the reference answer. Our objective is to suppress correct answers about forget identities across both input routes while preserving response performance on retain identities.

3.2 Stage 1: Visual Representation Retargeting

The first stage selectively modifies the visual access representations of forgetting targets along the image-conditioned route. The visual projector serves as a route-specific interface that maps visual features into the language model's representation space, enabling intervention on image-conditioned access without directly modifying the language model parameters. To this end, we freeze the vision encoder and language model and train only a copy $P_{\phi'}$ initialized from the original projector P_ϕ.

Since our study focuses on unlearning person identities, we use the mean representation of generic face images as a common retargeting target for forget images. We denote the forget and retain image sets by $\mathcal{X}_{\mathrm{F}}$ and $\mathcal{X}_{\mathrm{R}}$, respectively, and construct a fixed anchor set $\mathcal{G}$ using face images selected from Flickr-Faces-HQ (FFHQ) (Karras et al., 2019). The target $\bar{z}$ is computed by averaging the feature vectors of the original projector outputs at each token position:

$$\bar{z} = \frac{1}{|\mathcal{G}|} \sum_{x \in \mathcal{G}} P_\phi(e_\omega(x)). \tag{1}$$

The same target $\bar{z}$ is applied to all forget images and remains fixed throughout training. We train the edited projector with the following objective:

$$\mathcal{L}_{\mathrm{S1}} = \sum_{x \in \mathcal{X}_{\mathrm{F}}} d(P_{\phi'}(e_\omega(x)), \bar{z}) + \sum_{x \in \mathcal{X}_{\mathrm{R}} \cup \mathcal{G}} d(P_{\phi'}(e_\omega(x)), P_\phi(e_\omega(x))) \,. \tag{2}$$

where $d(u, v)$ is the mean squared error (MSE) averaged over token positions and feature dimensions.

The first term shifts the projector outputs of forget images toward the common target $\bar{z}$, modifying the visual representations conveyed to the language model. However, since the projector parameters are shared across all images, optimizing this term alone may also alter non-target representations. To limit these effects, the second term keeps the outputs of retain and anchor images close to their respective original projector outputs. Thus, $\mathcal{G}$ serves both as the reference set for constructing the mean target and as an anchor set for preserving generic face representations. Jointly optimizing these terms encourages selective retargeting of forget images while discouraging non-target face representations from also shifting toward the common target.

3.3 Stage 2: Route-aware Suppression

Stage 2 complements the visual representation changes introduced in Stage 1 by directly suppressing target responses under image-conditioned and text-only inputs. For image-conditioned suppression, we use the outputs of the original projector. Since the edited projector is trained in Stage 1 to shift forget-image representations toward the mean representation of generic face images, further suppressing target answers under these representations may also affect responses conditioned on generic face representations. Using the original projector allows us to suppress target responses under visual representations to which this retargeting has not been applied. We also apply suppression to text-only inputs, which do not pass through the visual interface. The edited projector is used for image-conditioned inference and visual retain supervision in Stage 2, with the latter training the model to preserve utility along the deployed route.

We denote the outputs of the routes using the original and edited projectors by $z_0(x) = P_\phi(e_\omega(x))$ and $z_1(x) = P_{\phi'}(e_\omega(x))$. Both routes use the same current encoder features, with subscripts 0 and 1 distinguishing the original projector from the projector edited in Stage 1. During Stage 2, we train LoRA adapters in both the vision encoder and the language model, while keeping their base weights and both projectors frozen. Image-conditioned losses update both sets of adapters, whereas text-only losses update only the language-model adapters. Since the encoder adapters are updated, both projector outputs may change during training.

Let $\ell_\theta(a \mid q, z)$ denote the reference answer's log-probability averaged over answer tokens, given a question q and visual representations z. For the text-only route, we write this as $\ell_\theta(a \mid q)$. To limit the probability of target answers, we use the following hinge objective:

$$h_\tau(s) = \max(0, s - \tau), \qquad \tau < 0. \tag{3}$$

Minimizing this term applies suppression while the answer's mean log-probability exceeds τ, and its gradient is zero at or below τ. Thus, direct suppression that would further reduce the answer probability stops for examples that already satisfy the threshold.

To suppress target answers under perturbed representations around the original projector output $z_0(x)$, we construct feature perturbations that increase the target answer's log-probability. For an image-conditioned forget example (x, q, a), we hold the model parameters fixed and perform the following inner optimization:

$$\delta^*(x, q, a) \approx \arg \max_{\|\delta\|_2 \leq \epsilon \|z_0(x)\|_2} \ell_\theta(a \mid q, z_0(x) + \delta). \tag{4}$$

Keeping the question and answer unchanged, we use projected gradient ascent to find a perturbation that increases the answer's log-probability within the ℓ_2 ball around $z_0(x)$. During the outer update, we detach δ^* and update the adapters to reduce the answer probability under the perturbed representation $z_0(x) + \delta^*$.

The full Stage 2 objective is:

$$\begin{aligned}\mathcal{L}_{\mathrm{S2}} = & \sum_{(x,q,a)\in\mathcal{F}_{\mathrm{img}}} \Big[h_\tau(\ell_\theta(a \mid q, z_0(x) + \delta^*)) + h_\tau(\ell_\theta(a \mid q, z_0(x)))\Big] \\ & + \sum_{(q,a)\in\mathcal{F}_{\mathrm{txt}}} h_\tau(\ell_\theta(a \mid q)) \\ & - \left[\sum_{(x,q,a)\in\mathcal{R}_{\mathrm{img}}} \ell_\theta(a \mid q, z_1(x)) + \sum_{(q,a)\in\mathcal{R}_{\mathrm{txt}}} \ell_\theta(a \mid q)\right].\end{aligned} \tag{5}$$

The first line suppresses answers to image-conditioned forget examples under both the original projector output $z_0(x)$ and its perturbed counterpart $z_0(x) + \delta^*$. The second line suppresses answers to text-only forget examples. The final line corresponds to answer-token cross-entropy on retain examples from both input routes, increasing the probability of their reference answers. Image-conditioned retain supervision uses the edited projector output $z_1(x)$ to preserve utility along the actual inference route.

4 Experiment

4.1 Experimental Setup

Benchmarks and Models. We evaluate BLINDFOLD-TO-FORGETTING on UMU-Bench (Wang et al., 2026), CLEAR (Dontsov et al., 2025), and MLLMU-Bench (Liu et al., 2025a). By default, we designate 10% of the synthetic identities in each dataset as forgetting targets, with additional evaluations at 5% and 15% to examine the effect of varying the forget ratio. Unless otherwise specified, we use LLaVA-1.5-7B (Liu et al., 2024a). The main comparisons on UMU-Bench and CLEAR also include LLaVA-1.5-13B (Liu et al., 2024a) and Qwen2-VL-7B (Wang et al., 2024). Further details on the benchmarks are provided in Appendix B.

Baselines. We compare against Gradient Ascent (GA) (Jang et al., 2023), Gradient Difference (GD) (Maini et al., 2024), KL (Yao & Xu, 2024), Preference Optimization (PO) (Maini et al., 2024), NPO (Zhang et al., 2024), the modality-aware pruning method MANU (Liu et al., 2025b), and the influential-path editing method MIP-Editor (Li et al., 2026). Each method is adapted to the QA input formats of the benchmarks. Further details on the baselines and their implementations are provided in Appendix D.

Implementation. In Stage 1, we freeze the vision encoder and language model and train the projector for 150 epochs using forget images, retain images, and FFHQ anchor images at a $1:1:1$ ratio. In Stage 2, we freeze both the original and edited projectors and train rank-16 LoRA adapters ($\alpha = 32$) for 4 epochs[1] using the AdamW optimizer with a learning rate of 10^{-4}. We use a $1:2$ ratio of forget to retain QA examples for each of the image-conditioned and text-only routes[2]. We set the hinge threshold to $\tau = -1.5$. All training and evaluation runs are performed on a single NVIDIA A100 80GB GPU. Further implementation details and computational costs are provided in Appendices E and F.5.

[1]Performance across different epoch budgets for each stage is reported in Appendix F.4.

[2]Results with equal amounts of forget and retain data, along with comparisons using adjusted data ratios for the baselines, are provided in Appendix F.2.

Table 1: Main results on UMU-Bench and CLEAR. We bold the best results and underline the second-best results.

Method	UMU-Bench								CLEAR							
	Forget ↓		Retain ↑		F-R H-Mean ↑		Real-World ↑		Forget ↓		Retain ↑		F-R H-Mean ↑		Real-World ↑	
	VQA	QA	VQA	QA	VQA	QA	VQA	QA	VQA	QA	VQA	QA	VQA	QA	VQA	QA
LLaVA-1.5-7B																
Vanilla	81.5	76.0	75.9	69.8	29.7	35.7	42.2	49.0	47.0	92.0	57.0	92.0	54.9	14.7	65.0	41.0
GA	74.0	76.0	70.0	68.2	37.9	35.5	41.8	48.0	38.0	**10.0**	17.0	4.0	26.7	7.7	21.0	30.0
GD	80.0	76.0	**74.1**	70.4	31.5	35.8	41.8	**50.7**	**0.0**	11.0	24.0	29.0	38.7	43.7	23.0	33.0
KL	51.0	65.0	48.2	58.7	48.6	43.9	38.2	43.5	11.0	57.0	24.0	62.0	37.8	**50.8**	14.0	41.0
NPO	73.0	73.0	70.7	67.8	39.1	38.6	41.2	47.4	39.0	90.0	38.0	89.0	46.8	18.0	54.0	**46.0**
PO	68.0	62.0	61.4	57.9	42.1	45.9	43.1	45.1	76.0	93.0	54.0	**93.0**	33.2	13.0	60.0	37.0
MANU	76.0	74.0	73.6	69.2	36.2	37.8	**46.4**	49.3	40.0	92.0	50.0	**93.0**	54.5	14.7	59.0	42.0
MIP-Editor	61.0	75.0	58.1	69.3	46.7	36.7	36.6	48.7	44.0	92.0	52.0	**93.0**	53.9	14.7	66.0	41.0
Ours	**46.0**	**55.0**	65.8	**70.9**	**59.3**	**55.1**	41.6	47.4	3.0	69.0	**58.0**	89.0	**72.6**	46.0	**77.0**	44.0
LLaVA-1.5-13B																
Vanilla	57.0	40.0	52.8	35.6	47.4	44.7	50.7	53.9	49.0	92.0	60.0	96.0	55.1	14.8	77.0	39.0
GA	57.0	41.0	**53.1**	35.6	47.5	44.4	50.7	54.6	41.0	93.0	24.0	96.0	34.1	13.0	22.0	**51.0**
GD	55.0	33.0	51.0	33.1	47.8	44.3	51.3	58.5	**0.0**	**3.0**	25.0	40.0	40.0	56.6	21.0	35.0
KL	58.0	21.0	52.7	20.0	46.7	31.9	49.0	50.0	12.0	4.0	32.0	32.0	46.9	48.0	38.0	38.0
NPO	**0.0**	**0.0**	0.0	0.0	0.0	0.0	29.4	29.4	52.0	91.0	29.0	92.0	36.2	16.4	44.0	42.0
PO	51.0	37.0	48.0	33.4	48.5	43.7	**52.9**	54.6	70.0	92.0	57.0	**97.0**	39.3	14.8	74.0	42.0
MANU	48.0	39.0	43.8	35.6	47.5	45.0	44.1	**59.5**	49.0	93.0	56.0	**97.0**	53.4	13.1	73.0	40.0
MIP-Editor	50.0	39.0	42.7	35.7	46.1	45.0	47.4	53.9	51.0	92.0	**62.0**	96.0	54.7	14.8	**76.0**	38.0
Ours	40.0	20.0	45.4	**35.9**	**51.7**	**49.6**	48.4	52.3	6.0	54.0	46.0	95.0	**61.8**	**62.0**	**76.0**	41.0
Qwen2-VL-7B																
Vanilla	68.0	75.0	66.9	72.1	43.3	37.1	51.0	49.7	45.0	97.0	35.0	95.0	42.8	5.8	57.0	42.0
GA	28.0	28.0	27.4	27.4	39.7	39.7	29.4	29.4	46.0	**1.0**	17.0	1.0	25.9	2.0	19.0	29.0
GD	**22.0**	22.0	35.9	40.6	**49.2**	53.4	40.5	43.8	**0.0**	27.0	**51.0**	93.0	**67.5**	**81.8**	**93.0**	**43.0**
KL	**22.0**	22.0	26.9	26.9	40.0	40.0	32.4	32.4	9.0	43.0	18.0	56.0	30.1	56.5	20.0	34.0
NPO	**22.0**	22.0	27.0	26.9	40.1	40.0	32.0	32.4	47.0	91.0	21.0	83.0	30.1	16.2	33.0	35.0
PO	61.0	33.0	58.7	60.0	46.9	63.3	44.8	39.9	64.0	97.0	34.0	**96.0**	35.0	5.8	58.0	42.0
MANU	66.0	75.0	63.4	**72.1**	44.3	37.1	50.3	**49.7**	44.0	97.0	34.0	95.0	42.3	5.8	57.0	42.0
MIP-Editor	68.0	73.0	**67.2**	71.9	43.4	39.3	**50.7**	48.7	45.0	97.0	34.0	95.0	42.0	5.8	57.0	42.0
Ours	61.0	**16.0**	66.0	66.7	49.0	**74.4**	47.4	48.4	6.0	84.0	37.0	94.0	53.1	27.3	50.0	42.0

Evaluation Metrics. For UMU-Bench and CLEAR, we report Forget, Retain, and Real-World accuracy separately for the image-conditioned (VQA) and text-only (QA) routes. Accuracy is the percentage of evaluation examples for which the response is judged correct. Lower Forget accuracy and higher Retain and Real-World accuracy are better. To assess the balance between forgetting and retained utility, we also report the Forget–Retain H-Mean (F-R H-Mean), defined as the harmonic mean of the forgetting score and retain accuracy for each route (Entesari et al., 2026):

$$H_r = \frac{2(100 - F_r)R_r}{(100 - F_r) + R_r}, \qquad r \in \{\text{VQA}, \text{QA}\}, \tag{6}$$

where F_r and R_r denote Forget and Retain accuracy for route r, respectively, and $100 - F_r$ is the forgetting score. A high F-R H-Mean indicates that the model jointly achieves low accuracy on forgetting targets and high accuracy on retained targets. For MLLMU-Bench, we report classification accuracy and generation scores separately.

4.2 Main Results

As shown in Table 1, BLINDFOLD-TO-FORGETTING improves the balance between forgetting and retained utility, with particularly strong results on the LLaVA family. On UMU-Bench, our method achieves the highest F-R H-Mean on both image-conditioned and text-only routes for both LLaVA backbones. With LLaVA-1.5-7B, it reaches VQA/QA H-Mean scores of 59.3/55.1, combining reduced

Forget accuracy with well-preserved retained performance. This advantage extends to LLaVA-1.5-13B, where our method improves forgetting on both routes without the retained-performance collapse observed for NPO. Across both models, Real-World accuracy remains close to Vanilla, indicating limited changes in person-related performance outside the forget and retain sets.

A similar pattern emerges on CLEAR, particularly for image-conditioned forgetting. With LLaVA-1.5-7B, our method nearly eliminates correct responses about forget identities while maintaining Retain accuracy close to Vanilla, achieving the highest VQA H-Mean of 72.6. On LLaVA-1.5-13B, it obtains the highest H-Mean on both routes. The text-only results on LLaVA-1.5-7B show a different trade-off: our method reduces Forget accuracy while largely preserving retained performance, whereas KL achieves stronger forgetting and a higher H-Mean at a substantial cost to retained utility.

For Qwen2-VL-7B, performance varies more across benchmarks and input routes. On UMU-Bench, our method achieves the highest text-only H-Mean of 74.4, extending its benefit beyond the LLaVA family. Its image-conditioned H-Mean is also comparable to GD, while retaining substantially higher accuracy on non-target identities. On CLEAR, however, GD achieves higher H-Mean on both routes. Our method still reduces Forget accuracy relative to Vanilla while maintaining similar retained performance, but its advantage over competing methods is less consistent in this setting. Additional results across repeated runs and qualitative generation examples are provided in Appendices F.3 and F.6, respectively.

4.3 Results on MLLMU-Bench

Table 2 further evaluates our method on the classification and generation settings of MLLMU-Bench. For classification, Blindfold-to-Forgetting achieves the lowest image/text Forget accuracies among all compared methods, at 56.0/45.3, while maintaining Retain accuracies of 64.0/58.0. In particular, its image-route Retain accuracy is the highest among all edited baselines. These results demonstrate effective target suppression across both image-conditioned and text-only routes while largely preserving retained classification performance. For generation, our method achieves a Forget score of 63.1, slightly higher than NPO's 62.4 and equal to KL, while attaining the highest Retain score among all edited baselines at 71.1. This indicates a trade-off that favors retained generation quality over maximizing forgetting strength.

Table 2: Results on the two MLLMU-Bench evaluation tasks: classification and generation. F/R denotes Forget/Retain, and Img/Txt denotes the image/text input routes.

Method	Classification				Generation	
	F-Img ↓	F-Txt ↓	R-Img ↑	R-Txt ↑	F-Img ↓	R-Img ↑
Vanilla	68.7	54.7	70.0	62.0	70.3	76.7
GA	68.3	55.0	56.7	55.0	62.9	51.7
GD	71.7	60.0	<u>60.0</u>	**58.3**	69.5	59.3
KL	68.3	<u>53.3</u>	58.3	56.7	<u>63.1</u>	52.5
NPO	<u>65.0</u>	55.0	58.3	55.0	**62.4**	55.0
PO	75.0	56.7	58.3	56.7	66.3	<u>64.3</u>
MANU	68.3	60.0	56.7	50.0	64.8	55.6
MIP-Editor	71.7	58.3	56.7	**58.3**	70.2	54.8
Ours	**56.0**	**45.3**	**64.0**	<u>58.0</u>	<u>63.1</u>	**71.1**

4.4 Ablation Studies

Component Ablation. Table 3 shows that visual representation retargeting and route-aware response suppression play complementary roles. Removing Stage 1 collapses both Forget and Retain VQA accuracy to zero. Replacing suppression at $z_0(x)$ with suppression at $z_1(x)$ also strengthens forgetting but substantially degrades visual retention. These results highlight the importance of Stage 1 and the choice of suppression route in preserving visual utility, rather than merely reducing target accuracy. Removing the text branch increases Forget QA from 55.0 to 71.0, supporting the role of explicit text-route training in text-only forgetting. Finally, replacing the generic-face target with another identity representation produces stronger text-only forgetting but worsens image-conditioned forgetting and retained text-only perfor-

Table 3: Component ablation results. Removing the text branch excludes both forget and retain training on the text-only route. $z_0 \to z_1$ replaces suppression at $z_0(x)$ with suppression at $z_1(x)$ while retaining adversarial suppression at $z_0(x) + \delta^*$.

Variant	Forget ↓		Retain ↑		Real-World ↑	
	VQA	QA	VQA	QA	VQA	QA
Ours	46.0	55.0	65.8	70.9	41.6	47.4
w/o Stage 1	0.0	20.0	0.0	72.4	23.5	49.7
w/o text branch	53.0	71.0	61.0	67.4	37.3	48.7
S2: $z_0 \to z_1$	5.0	9.0	28.3	64.7	28.8	52.6
S1 target: identity	57.0	4.0	64.3	62.6	34.3	48.4

mance. The generic-face target therefore favors visual forgetting and retained utility, although it does not maximize forgetting on every route.

Sensitivity to the Forget Ratio. Figure 2 shows forgetting and retained utility under forget ratios of 5%, 10%, and 15% for both input routes. Image-conditioned Forget accuracy remains relatively stable, whereas Retain VQA decreases from 73.6 to 63.9 as the forget ratio increases, indicating a growing cost to visual utility as more identities are unlearned. In contrast, text-only Forget accuracy varies non-monotonically, reaching its highest value at the intermediate ratio, while Retain QA remains relatively stable. Thus, expanding the forget set is associated with declining visual retention but a less consistent change in text-only forgetting, revealing different forgetting–utility patterns across the two input routes.

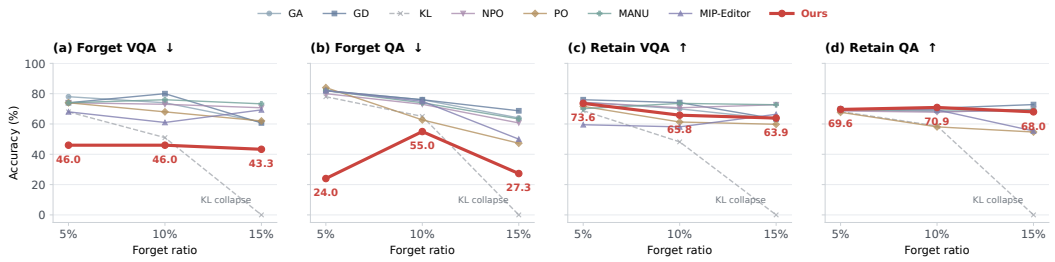


Figure 2: Forgetting and retention performance across different forget ratios on UMU-Bench.

Sensitivity to the hinge threshold. Figure 3 shows how the hinge threshold affects forgetting and retained utility. Among the evaluated settings, $\tau = -1.5$ achieves the highest Retain accuracy on both routes, but also the highest Forget QA accuracy, indicating weaker text-only suppression. Both alternative thresholds strengthen text-only forgetting while reducing retained utility, with $\tau = -1.0$ lowering Forget QA to zero. Thus, a lower threshold does not necessarily produce stronger forgetting in the final model, highlighting the need to assess its effect on both forgetting and retention across input routes.

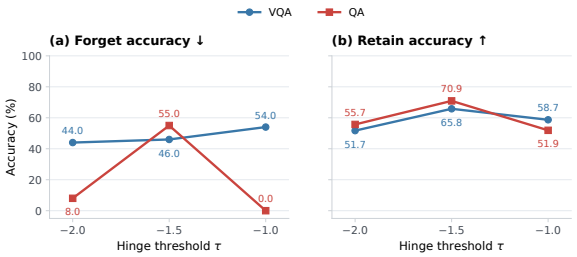


Figure 3: Forget and Retain accuracy on the image and text routes across different hinge thresholds τ.

4.5 Robustness to Relearning

Prior work has shown that unlearned knowledge can recover after only a small amount of additional training, even when the recovered targets are not directly reintroduced, suggesting that low post-unlearning accuracy may sometimes reflect easily reversible output suppression rather than persistent forgetting (Hu et al., 2025). Such recovery is particularly concerning when relearning a subset of forget targets also reactivates other forget targets that are not included in the relearning data.

Our method retargets visual access along the image route while separately suppressing target responses under conditions not controlled by the visual interface. We therefore examine how much target-response recovery occurs when only a subset of forget identities is relearned. Specifically, we relearn a subset of the 50 forget identities and evaluate accuracy across the full forget set.

Figure 4 examines recovery across benchmarks and backbones, with additional analyses varying the number of relearned identities and the attack epochs. With few relearned identities, our method maintains lower forget accuracy than Vanilla and PO on both image and text routes, although accuracy increases as more identities are reintroduced. In the epoch analysis using one forgotten example, our method remains below both baselines throughout the evaluated attack and returns to its pre-attack accuracy after eight epochs without exceeding it at any evaluated point. These results show limited recovery under the tested low-data attack, while relearning more forget identities weakens the forgetting effect.

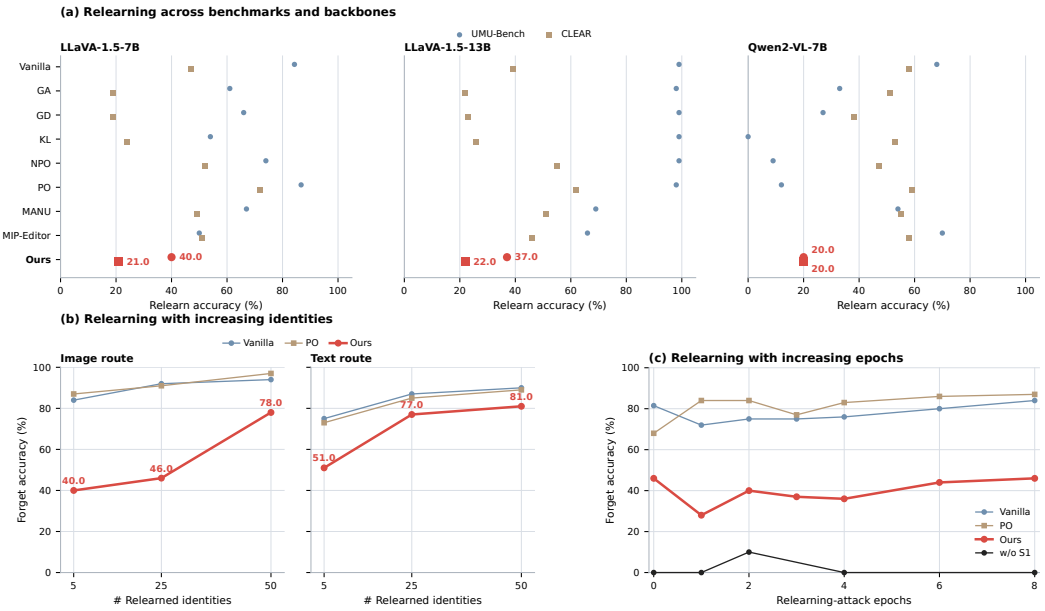


Figure 4: Analysis of target-response recovery under relearning. (a) Forget accuracy after relearning across three backbones on UMU-Bench and CLEAR. (b) Image- and text-route recovery across the forget set when relearning 5, 25, or 50 of the 50 forget identities. (c) Image-route Forget accuracy as the number of optimization epochs increases on a limited relearning set. Lower values indicate a lower correct-response rate on forget targets after relearning.

4.6 General Capability

Table 4: General capability evaluation results.

Method	MMLU	MMBench	SEED-Bench	GQA	POPE	Avg.
Vanilla	47.1	64.7	61.4	51.8	83.7	61.74
GA	24.0	39.2	25.5	0.0	50.0	27.74
GD	47.1	65.8	60.9	51.5	83.3	61.72
KL	46.6	64.7	60.1	0.3	67.4	47.82
NPO	47.1	66.1	61.6	51.3	83.0	61.82
PO	47.0	65.3	62.1	50.5	81.7	61.32
MANU	47.5	65.3	61.4	51.3	81.9	61.48
MIP-Editor	47.0	60.6	60.4	52.2	82.1	60.46
Ours	46.8	65.6	59.7	53.7	76.4	60.44

Table 4 evaluates whether unlearning affects general capabilities beyond the forget and retain identities. Descriptions of these evaluation benchmarks are provided in Appendix C. Our method maintains MMLU performance close to Vanilla and achieves modest improvements on MMBench and GQA, indicating that language knowledge and several aspects of multimodal understanding remain largely preserved. However, this preservation does not extend uniformly across benchmarks: SEED-Bench shows a modest decline, while POPE exhibits the largest decrease, from 83.7 to 76.4. The POPE degradation also exceeds that of GD, NPO, and PO, highlighting a remaining limitation in capability preservation. These results show that favorable forgetting–retention performance can coexist with losses on specific general capabilities, which are not fully reflected in the in-benchmark Forget and Retain metrics.

5 Conclusion

We proposed BLINDFOLD-TO-FORGETTING, a route-aware multimodal unlearning method that combined visual representation retargeting with target-response suppression. Experiments across three benchmarks and three MLLMs showed that our method improved the forgetting–utility balance, achieving the highest Forget–Retain H-Mean in 8 of 12 settings on UMU-Bench and CLEAR, with gains of up to 18.1 points over the strongest baseline. Ablation studies showed that visual retargeting helped preserve visual utility, while text-route training improved text-only forgetting. Relearning experiments further showed that the unlearning effect was less easily reversed under limited additional training, although target responses recovered as the scope of relearning increased. Overall, the results suggest the promise of exploiting route-specific intervention points for multimodal unlearning.

AI USE STATEMENT

We used generative AI tools to improve wording, correct grammar, and assist with drafting the manuscript. The authors reviewed and revised the AI-generated suggestions and take responsibility for the final content and claims of this paper.

ETHICS STATEMENT

Machine unlearning aims to remove the influence of specific data from trained models and can support privacy protection and data deletion requests. With these goals in mind, this study addresses the challenge that multimodal models can access information about the same identity through both images and text. Our experiments use publicly available multimodal unlearning benchmarks and public research datasets, without collecting new personal information. Substantial portions of the main benchmarks use synthetic identities to construct controlled unlearning settings. While unlearning can contribute to privacy protection, it can also be used to selectively modify model behavior concerning particular individuals or concepts. Real-world deployment should therefore consider the legitimacy of deletion requests, the scope of the target data, and potential unintended effects on retained knowledge.

REPRODUCIBILITY STATEMENT

Our experiments use publicly available multimodal unlearning benchmarks and pretrained backbones. Section 3 describes the objectives of both stages and the feature perturbation procedure, while Section 4.1 presents the main training configurations and evaluation metrics. We will release our implementation to facilitate reproducibility.

REFERENCES

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.

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.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. Honeybee: Locality-enhanced projector for multimodal llm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827. IEEE, 2024.

Siyuan Cheng, Bozhong Tian, Qingbin Liu, Xi Chen, Yongheng Wang, Huajun Chen, and Ningyu Zhang. Can we edit multimodal large language models? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13877–13888, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.856. URL https://aclanthology.org/2023.emnlp-main.856/.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20582–20603, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1058. URL https://aclanthology.org/2025.findings-acl.1058/.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning in llms. *arXiv preprint arXiv:2310.02238*, 2023.

Taha Entesari, Arman Hatami, Rinat Khaziev, Anil Ramakrishna, and Mahyar Fazlyab. Constrained entropic unlearning: A primal-dual framework for large language models. *Advances in Neural Information Processing Systems*, 38:166435–166474, 2026.

Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Shengyuan Hu, Yiwei Fu, Steven Wu, and Virginia Smith. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. In *International Conference on Learning Representations*, volume 2025, pp. 8857–8888, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 7190–7206, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.375. URL https://aclanthology.org/2025.findings-acl.375/.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.805. URL https://aclanthology.org/2023.acl-long.805/.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 4396–4405. IEEE, 2019.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023a. URL https://arxiv.org/abs/2307.16125.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

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, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with MLLMU-bench. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4105–4135, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.207. URL https://aclanthology.org/2025.naacl-long.207/.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5913–5933, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.295. URL https://aclanthology.org/2025.acl-long.295/.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary Chase Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Chengye Wang, Yuyuan Li, XiaoHua Feng, Chaochao Chen, Xiaolin Zheng, and Jianwei Yin. UMU-bench: Closing the modality gap in multimodal unlearning evaluation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=M476xkfNXe.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

A LIMITATIONS

This study has several limitations. First, our method relies on an explicit visual-to-language interface, such as a visual projector or merger, that maps vision encoder outputs into the language model representation space. While this design naturally applies to architectures such as LLaVA and Qwen2-VL, defining an equivalent intervention point may be less straightforward for unified multimodal architectures that more tightly integrate vision and language tokens.

Second, our method focuses on person-identity unlearning, and the retargeting target in Stage 1 is constructed from generic face representations. Extending this framework to non-person knowledge, such as objects, scenes, or documents, may require alternative strategies for defining appropriate retargeting targets.

Third, our experiments were conducted primarily on multimodal unlearning benchmarks constructed with synthetic identities. While these settings enable controlled and reproducible evaluation with clearly defined forget and retain identities, further validation is needed to assess generalization to real-world data deletion requests and more complex real-world knowledge.

Finally, the balance between forgetting and retained utility varies with the forget ratio, hinge threshold, and input route. Moreover, although the forgetting effect remains relatively stable under limited relearning, target responses increasingly recover as more forget identities are reintroduced. Thus, robustness under stronger forms of subsequent training remains to be further evaluated.

B UNLEARNING BENCHMARKS

We evaluate BLINDFOLD-TO-FORGETTING on three multimodal unlearning benchmarks: UMU-Bench (Wang et al., 2026), CLEAR (Dontsov et al., 2025), and MLLMU-Bench (Liu et al., 2025a). All three benchmarks allow information about a given identity to be queried through both image and text inputs, enabling joint evaluation of forgetting along image-conditioned and text-only routes.

UMU-Bench. UMU-Bench is a multimodal unlearning benchmark designed to evaluate whether the same identity knowledge is consistently removed across image-conditioned and text-only inputs. Each synthetic identity is associated with multiple attributes, such as name, birthplace, and occupation, and the same information can be queried through either image-based or name-based questions. For example, the image-conditioned route presents a person's image together with a question such as "What is the name of this person?", whereas the text-only route directly provides the person's name in a question such as "Where was Amelia Brockhurst born?" Thus, if the model fails to retrieve information from the person's face but can still answer correctly when given the name, the identity has not been consistently forgotten across the two access routes. Each identity provides 12 QA examples for both the image-conditioned and text-only routes, yielding 600 forget QA examples per route under the default 10% forget setting. The remaining 450 synthetic identities constitute the retain set. In addition, UMU-Bench provides a real-world split containing 153 real-world celebrities, which serves as a control set outside both the forget and retain sets to assess whether editing degrades general person-related knowledge beyond the benchmark identities. We use the official evaluation harness: Forget and Retain predictions are counted as correct only when the generated answer exactly matches the target option, whereas Real-World predictions are mapped to the most similar of the four candidate answers before scoring.

CLEAR. CLEAR is a benchmark for character-level unlearning across textual and visual modalities. It contains 200 fictitious individuals and approximately 3,700 images. Each synthetic identity is associated with multiple images and textual QA pairs, with person-wise forget and retain splits. For example, *Jaime Vasquez* is described as a true-crime author, and a visual sample includes an image of Jaime Vasquez holding one of his books together with a caption describing the scene. Text-only QA directly queries personal information about the same individual, such as "Are the details of Jaime Vasquez's birth documented?" CLEAR therefore provides a setting in which both visual information and textual knowledge associated with a target character must be unlearned. To ensure consistent comparison across benchmark rows in our main table, we evaluate CLEAR using the same scoring protocol as UMU-Bench.

MLLMU-Bench. MLLMU-Bench evaluates the unlearning of personal information in MLLMs and contains 500 fictitious profiles and 153 public celebrity profiles. Each fictitious identity is defined by multiple profile attributes, including name, birthplace, occupation, education, and residence, together with facial images. The same information is queried in multiple forms under both image+text and text-only conditions. For example, *Lukas Harrington* is described as a software engineer born in Munich, Germany. In the image-conditioned classification task, the model is shown the person's image and asked, "What is this person's profession in the image?", with Software Engineer provided as the correct option. For the same profile, the generation task asks for attributes such as occupation or residence in free-form natural language. In our experiments, we report Forget/Retain classification accuracy separately for the image-conditioned and text-only routes, and use ROUGE-L (Lin, 2004) F-measure ($\times 100$) to evaluate generation on forget and retain examples.

C General Capability Benchmarks

To evaluate whether unlearning affects capabilities beyond the forget and retain sets, we additionally evaluate the edited models on five general-purpose benchmarks: MMLU, MMBench, SEED-Bench, GQA, and POPE. These benchmarks cover both language-only knowledge and multimodal perception and reasoning, providing a broader assessment of model utility outside the unlearning benchmarks.

MMLU. MMLU (Hendrycks et al., 2021) is a language-only benchmark that evaluates knowledge and reasoning across 57 academic and professional subjects, including humanities, social sciences, STEM, and other domains. We use it to assess whether multimodal unlearning causes degradation in the underlying language model's general knowledge and reasoning ability.

MMBench. MMBench (Liu et al., 2024b) evaluates a broad range of multimodal perception and reasoning capabilities using multiple-choice visual questions. It covers diverse skills required to interpret visual content and reason jointly over image and text inputs.

SEED-Bench. SEED-Bench (Li et al., 2023a) is a comprehensive multimodal benchmark based on multiple-choice questions spanning diverse dimensions of visual understanding. We use it to evaluate whether the edited model preserves general multimodal comprehension beyond the identities considered in the unlearning task.

GQA. GQA (Hudson & Manning, 2019) evaluates compositional visual reasoning through questions grounded in real-world images and scene graphs. Its questions require reasoning over objects, attributes, and relations, providing a complementary measure of visual question-answering capability.

POPE. POPE (Li et al., 2023b) evaluates object hallucination in large vision-language models through polling-based object queries. We include POPE to examine whether unlearning alters the model's ability to ground object predictions in visual inputs rather than generating objects unsupported by the image.

D Baselines

We compare against both language-model unlearning methods and multimodal unlearning methods. Optimization-based baselines are adapted to the image-conditioned and text-only QA formats of each benchmark, with the same forget identities used across both routes. We further include the multimodal-specific baselines MANU and MIP-Editor, allowing comparison not only with language-model fine-tuning approaches but also with modality-aware neuron and path editing methods.

Gradient Ascent (GA). GA is a direct unlearning baseline that reduces the likelihood of correct answers on forget data. In contrast to standard maximum-likelihood training, it maximizes the loss on forget answers so that the model becomes less likely to generate the target responses. Although simple and capable of inducing strong suppression, excessive optimization can substantially degrade retained capabilities because the updates are driven only by forget data. In our experiments, the same objective is applied to both image-conditioned and text-only forget QA examples.

Gradient Difference (GD). GD combines gradient ascent on forget data with standard likelihood maximization on retain data. It therefore jointly optimizes for reducing the probability of forget answers while preserving the probability of retain answers, mitigating the global model degradation that can occur with GA while encouraging selective forgetting.

KL Minimization (KL). The KL baseline combines answer suppression on forget data with a KL-divergence constraint that preserves the output distribution of the pre-unlearning model on retain data. It lowers the likelihood of correct answers on forget examples while encouraging the edited model to remain close to the original model on retain examples, thereby limiting unintended changes to non-target behavior.

Preference Optimization (PO). PO replaces the original answers to forget prompts with alternative responses such as "I don't know," rather than directly suppressing the likelihood of the original answers. The model is trained to produce refusal or uncertainty responses on forget examples while retaining supervision on the original answers for retain data.

Negative Preference Optimization (NPO). NPO is a preference-based unlearning objective designed to mitigate the abrupt model collapse that can arise with GA. Using the original model as a reference, it reduces the relative likelihood of forget answers under the edited model and controls the update magnitude more effectively than applying unbounded gradient ascent uniformly to all forget examples. Its goal is to suppress target knowledge while reducing degradation in retained utility.

MANU. MANU identifies neurons associated with forget knowledge separately for each modality and selectively prunes them. It first estimates neuron importance using visual and textual forget/retain data, identifies neurons that are relatively more important for the forget set, and removes their contributions. Unlike the fine-tuning-based baselines above, MANU directly localizes and edits internal neurons associated with the target knowledge.

MIP-Editor. MIP-Editor identifies influential neuron paths through which multimodal information propagates across layers, rather than considering neuron importance independently. After identifying modality-specific paths that are important for forget knowledge, it applies representation misdirection along these paths to steer internal representations away from the target knowledge. This aims to induce forgetting across image and text modalities while limiting degradation in general model capability.

Implementation. For each benchmark, we use the benchmark-released memorized checkpoint when available (UMU-Bench: `chengyewang/llava_smu_ft`; CLEAR: `therem/llava-1.5-7b-CLEAR-finetune`). We fine-tune the backbone ourselves only for backbone-benchmark combinations for which no memorized checkpoint is publicly available. Following Wang et al. (2026), this fine-tuning is performed for 5 epochs using Adam with a learning rate of 10^{-4} and LoRA for parameter-efficient adaptation, without gradient accumulation. For the subsequent unlearning experiments, each baseline follows the implementation settings provided for the corresponding benchmark, while MANU and MIP-Editor are implemented based on their original methods. GA and KL exhibit severe model collapse under their originally specified epoch budgets; we therefore use the preceding checkpoints in the main comparison, with the corresponding analysis provided in Appendix F.1.

E Implementation details

All models are implemented using HuggingFace. We use `llava-hf/llava-1.5-7b-hf`, `llava-hf/llava-1.5-13b-hf`, and `Qwen/Qwen2-VL-7B-Instruct` as backbone checkpoints. Benchmark data are obtained from `chengyewang/UMU-bench`, `therem/CLEAR`, and `MLLMMU/MLLMU-Bench`, while the identity-agnostic face anchors used in Stage 1 are drawn from `Ryan-sjtu/ffhq512-caption`.

For feature perturbation, we keep the model parameters fixed and initialize $\delta = 0$. We then perform 8-step projected gradient ascent to maximize the log-probability of the target answer. The perturbation radius and ascent step size are set to $0.3\|z_0(x)\|_2$ and $0.05\|z_0(x)\|_2$, respectively. After each ascent

Table 5: Baseline optimization settings on UMU-Bench.

Method	Setting	Forget ↓		Retain ↑		F-R H-Mean ↑		Real-World ↑	
		VQA	QA	VQA	QA	VQA	QA	VQA	QA
GA	5 epochs	0.0	0.0	0.0	0.0	0.0	0.0	29.4	29.4
	4 epochs	74.0	76.0	70.0	68.2	37.9	35.5	41.8	48.0
KL	4 epochs	0.0	0.0	0.0	0.0	0.0	0.0	29.4	29.4
	3 epochs	51.0	65.0	48.2	58.7	48.6	43.9	38.2	43.5
Ours	Default	46.0	55.0	65.8	70.9	**59.3**	**55.1**	41.6	47.4

step, δ is projected back onto the corresponding ℓ_2 ball. During the outer update, the final perturbation δ^* is detached and treated as a fixed value.

F Additional Results

F.1 Baseline Optimization

Table 5 examines the effect of optimization duration on the baselines. The original UMU-Bench settings use 5 epochs for GA and 4 epochs for KL. Under these settings, however, both methods exhibit catastrophic degradation: one additional epoch reduces Forget VQA/QA to 0.0/0.0, but simultaneously collapses Retain VQA/QA to 0.0/0.0 and decreases Real-World VQA/QA to 29.4/29.4. This shows that lower Forget accuracy obtained through longer optimization can result from global model degradation rather than selective unlearning. Therefore, for the main comparison, we use the preceding checkpoints (4 epochs for GA and 3 epochs for KL), where retained utility has not collapsed, rather than selecting the configuration solely based on the lowest Forget accuracy.

F.2 Controlling Retain Data

Table 6 examines whether differences in the amount of retain supervision affect the main comparison. We compare each method's original retain-data setting with a matched setting using 1,200 retain QA examples per route. Increasing retain supervision does not uniformly improve the forgetting–utility balance. For KL, matching the retain data increases Retain VQA from 48.2 to 73.4, but also increases Forget VQA from 51.0 to 81.0, reducing the VQA F-R H-Mean from 48.6 to 30.2. In contrast, MANU benefits from the matched setting, with its VQA/QA F-R H-Mean improving from 36.2/37.8 to 43.0/40.6. Thus, the effect of additional retain supervision varies across methods rather than consistently improving performance.

Our method also exhibits different trade-offs between the two settings. Under the original setting, it achieves Forget VQA/QA of 57.0/4.0 and Retain VQA/QA of 67.1/57.8, yielding particularly strong text-only forgetting but lower retained utility on that route. Under the matched setting, Forget VQA/QA changes to 46.0/55.0 while Retain VQA/QA reaches 65.8/70.9, resulting in F-R H-Mean scores of 59.3/55.1. Rather than maximizing forgetting on a single route, the matched setting provides a more balanced forgetting–utility trade-off across both image-conditioned and text-only routes. We therefore use the matched setting as the default configuration for our method in the main experiments.

F.3 Variation Across Repeated Runs

Table 7 reports the mean±standard deviation over five runs on UMU-Bench. Our method achieves Forget VQA/QA of 45.2 ± 2.9 and 47.6 ± 8.1, respectively. The larger variation in Forget QA compared with Forget VQA suggests that text-only suppression may be more sensitive to initialization and optimization trajectories. In contrast, retained utility is relatively stable, with Retain VQA/QA of 65.9 ± 0.2 and 69.8 ± 2.5, while Real-World VQA/QA remains at 41.5 ± 3.0 and 47.3 ± 0.3. Thus, most of the run-to-run variation is concentrated in forgetting performance, particularly on the text-only route. The baselines also exhibit method-dependent variability: for example, MIP-Editor shows

Table 6: Effect of retain-data control on UMU-Bench. ORIGINAL denotes each method's original retain-data setting, while MATCHED uses 1,200 retain QA examples per route. For our method, MATCHED corresponds to the default setting used in the main experiments. Best F-R H-Mean scores are bolded.

Method	Retain setting	Forget ↓		Retain ↑		F-R H-Mean ↑		Real-World ↑	
		VQA	QA	VQA	QA	VQA	QA	VQA	QA
GD	Original	80.0	76.0	74.1	70.4	31.5	35.8	41.8	50.7
	Matched	81.0	76.0	75.9	69.8	30.4	35.7	40.8	49.0
KL	Original	51.0	65.0	48.2	58.7	48.6	43.9	38.2	43.5
	Matched	81.0	74.0	73.4	69.3	30.2	37.8	41.8	48.4
PO	Original	68.0	62.0	61.4	57.9	42.1	45.9	43.1	45.1
	Matched	81.0	71.0	75.3	69.2	30.3	40.9	43.5	48.4
MANU	Original	76.0	74.0	73.6	69.2	36.2	37.8	46.4	49.3
	Matched	69.0	71.0	70.1	67.8	43.0	40.6	45.1	51.3
Ours	Original	57.0	4.0	67.1	57.8	52.4	**72.2**	41.7	45.4
	Matched (default)	46.0	55.0	65.8	70.9	**59.3**	55.1	41.6	47.4

Table 7: Variation across five repeated runs on UMU-Bench. We report mean±standard deviation. GA and KL use 5 and 4 epochs, respectively, following the original UMU-Bench implementation. Therefore, their results reflect the reproducibility of the collapse observed under those settings rather than run-to-run variation under the main comparison settings.

Method	Forget ↓		Retain ↑		Real-World ↑	
	VQA	QA	VQA	QA	VQA	QA
Vanilla	81.5	76.0	75.9	69.8	42.2	49.0
GA	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	29.4 ± 0.0	29.4 ± 0.0
GD	80.0 ± 0.0	76.0 ± 0.0	$\mathbf{74.2 \pm 0.2}$	70.3 ± 0.2	41.7 ± 0.7	$\mathbf{50.5 \pm 0.2}$
KL	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	29.4 ± 0.0	29.4 ± 0.0
NPO	72.7 ± 0.6	72.7 ± 0.6	70.5 ± 0.3	67.9 ± 0.1	41.2 ± 0.0	47.3 ± 0.2
PO	68.0 ± 0.0	62.7 ± 0.6	61.3 ± 0.2	58.1 ± 0.2	43.1 ± 0.0	46.0 ± 0.8
MANU	72.0 ± 4.0	72.0 ± 2.6	70.8 ± 2.5	66.9 ± 2.1	$\mathbf{43.5 \pm 2.8}$	48.1 ± 1.1
MIP-Editor	54.0 ± 13.9	75.0 ± 0.0	51.7 ± 10.7	69.4 ± 0.1	36.9 ± 0.6	48.7 ± 0.0
Ours	$\mathbf{45.2 \pm 2.9}$	$\mathbf{47.6 \pm 8.1}$	65.9 ± 0.2	$\mathbf{69.8 \pm 2.5}$	41.5 ± 3.0	47.3 ± 0.3

a relatively large standard deviation in Forget VQA (54.0 ± 13.9), whereas GD and NPO remain comparatively stable across most metrics.

F.4 SENSITIVITY TO STAGE-WISE OPTIMIZATION

Figure 5 analyzes the forgetting–utility trade-off as the number of optimization epochs varies in Stage 1 and Stage 2. In both stages, longer optimization does not monotonically improve forgetting and retained utility, and the two stages exhibit different sensitivity patterns.

Stage 1. Varying the number of Stage 1 epochs from 10 to 150 results in non-monotonic changes across both image- and text-route performance. Between 10 and 75 epochs, Forget VQA remains within 63.0–72.0 and Retain VQA within 59.4–64.7, showing relatively limited variation. At 100 epochs, Forget VQA drops sharply to 26.0, but Retain VQA/QA also decreases to 25.7/23.0. This indicates that stronger projector optimization can improve target forgetting while simultaneously degrading retained representations. At 150 epochs, Forget VQA/QA reaches 46.0/55.0 while Retain VQA/QA recovers to 65.8/70.9, providing a more balanced forgetting–utility trade-off among the evaluated settings.

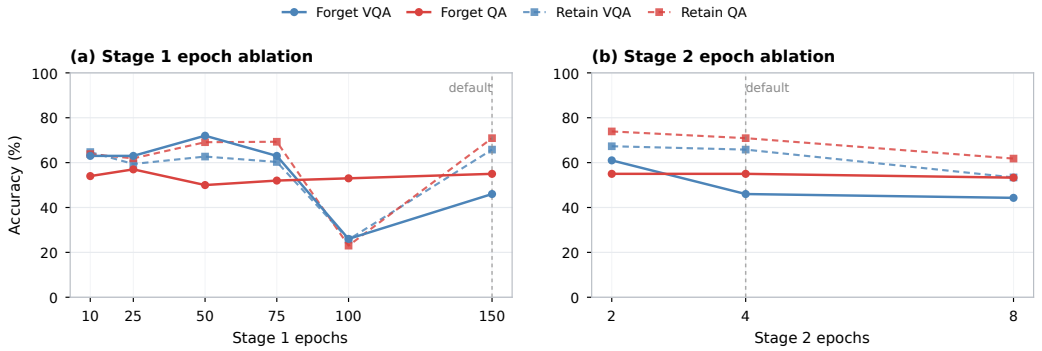


Figure 5: Sensitivity to stage-wise optimization. (a) Stage 1 epoch ablation and (b) Stage 2 epoch ablation on UMU-Bench. We report image-conditioned and text-only Forget/Retain accuracy as the number of optimization epochs varies. The dashed vertical lines indicate the default settings used in our experiments.

Stage 2. At 2 epochs, Forget VQA/QA is 61.0/55.0 and Retain VQA/QA is 67.3/73.9. Increasing the optimization to 4 epochs substantially reduces Forget VQA to 46.0, while Retain VQA/QA decreases only moderately to 65.8/70.9. At 8 epochs, however, Forget VQA/QA improves only slightly to 44.3/53.3, whereas Retain VQA/QA drops substantially to 53.4/61.8. Thus, further optimization yields limited gains in forgetting at the cost of larger utility degradation. Based on this trade-off, we use 4 epochs as the default setting for Stage 2.

F.5 Computational Cost

Table 8 compares the editing time of each method. Our method requires approximately 59 minutes in total, the longest among the compared methods. Most of this cost comes from Stage 2, where training the adapters for route-aware suppression takes approximately 58 minutes. In contrast, Stage 1 takes less than one minute despite 150 epochs of optimization, as the projector inputs and original outputs are cached in advance. Thus, the main computational cost of our method arises from response-suppression training in Stage 2 rather than visual representation retargeting itself. Evaluation time is excluded, and we report the editing time under the training configuration used for each method in the main experiments.

Table 8: Editing time for each method.

Method	Time (min)
GA	20.9
NPO	21.8
GD	28.9
KL	38.4
PO	38.5
Ours	59.0
S1	< 1
S2	~ 58

F.6 Qualitative Generation Example

Table 9 presents qualitative generations for two forget identities. For each identity, the image-conditioned route provides a face image and asks for the person's name, while the text-only route directly provides the name and asks the model to generate background information. These examples illustrate how the quantitative Forget VQA/QA results are reflected in actual model responses.

For Amelia Brockhurst, PO, NPO, MANU, and MIP-Editor still generate the correct identity name from the image and largely reproduce the original biography, including the birthplace, under the text-only input. In contrast, our method predicts a different identity name on the image route and generates background information that no longer matches the original biography on the text route.

A similar pattern is observed for Ethan Halberg. NPO, MANU, and MIP-Editor preserve the original identity on the image route and reproduce substantial portions of the associated biography on the text route, whereas our method alters both the image-conditioned identification and the biographical information generated from the name. These examples show that the lower Forget accuracy of our method is also reflected in qualitative generations, where the visual and textual associations with the target identity are jointly disrupted.

Image	Method	What is the name of this person?	Tell me more about the background information of ⟨name⟩.
	*gold*	Amelia Brockhurst.	Amelia Brockhurst, born in Nottingham, England, is a...
	PO	Amelia Brockhurst.	Amelia Brockhurst, born in Nottingham, England, is a...
	NPO	Amelia Brockhurst.	Amelia Brockhurst, born in Nottingham, England, is a...
	MANU	Amelia Brockhurst.	Amelia Brockhurst, born in Nottingham, England, is a...
	MIP-Editor	Amelia Brockhurst.	Amelia Brockhurst, born in Nottingham, England, is a...
	Ours	Alicia Moretti.	Edinburgh, Scotland is the birthplace of Amelia Brockhurst...
	*gold*	Ethan Halberg.	Ethan Halberg, born in Oslo, Norway, is a 40-year-old arch...
	PO	Ethan Golding.	I'm sorry, but I don't have any information on a person...
	NPO	Ethan Halberg.	Ethan Halberg, born in Oslo, Norway, is a 37-year-old arch...
	MANU	Ethan Halberg.	Ethan Halberg, born in Oslo, Norway, is a 33-year-old arch...
	MIP-Editor	Ethan Halberg.	Ethan Halberg, born in Oslo, Norway, is a 33-year-old arch...
	Ours	Jasper Norwood.	Ethan Halberg, born in Perth, Australia on April 13, 1998...

Table 9: Qualitative generations for two forget identities. For each identity, we show the image-conditioned response to "What is the name of this person?" and the text-only response to "Tell me more about the background information of ⟨name⟩."