SIEVE: SELECTIVE ATTENTION-VALUE SUPPRESSION FOR VISION-LANGUAGE MODELS UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability of vision–language models (VLMs) to associate visual identities with biographical information creates a need for selective unlearning of personally identifiable information (PII) while preserving permitted knowledge about the same individual. This setting is challenging because both sensitive and retained information can share the same visual inputs and intermediate representations. We introduce SIEVE, a simple and effective framework for selective VLM unlearning. SIEVE directly regularizes attention-value representations while also controlling model outputs. SIEVE suppresses attention values for forget examples toward a constant zero, while preserving retain-example representations by matching them to a frozen reference model. These objectives are combined with sequence-level forget and retain supervision, enabling targeted forgetting without largely affecting retained knowledge. Extensive experiments show that SIEVE achieves state-of-the-art performance on unlearning with multiple model–modality settings, while maintaining competitive retained utility. Ablation studies further show that value suppression and negative cross-entropy contribute complementary forgetting signals, while reference-based value matching substantially reduces utility degradation. These results demonstrate that attention values provide an effective intervention point for selective multimodal unlearning when sensitive and retained knowledge are closely related.

1 INTRODUCTION

Vision–language models (VLMs) can associate visual identities with information that is not directly observable from an image. Given a face, for example, a VLM may recover a person's name, occupation, address, or other biographical information learned during training. Recent multimodal unlearning benchmarks make such associations measurable by pairing identities with images and textual attributes, then evaluating whether designated information remains recoverable after model updates (Liu et al., 2025; Dontsov et al., 2025). Hence there is a practical need to remove sensitive information from VLMs without unnecessarily degrading their remaining capabilities.

This need then leads to a inherently selective deletion problem. A deletion request may target only a subset of information associated with an individual, while other attributes about the same person should remain intact. For example, a VLM may be required to forget a person's date of birth while continuing to answer questions about their hobbies. The forget and retain examples can therefore share the same image and identity while differing only in the queried information. We formulate this setting by partitioning data at the level of image–question–response associations rather than identities. The objective is to suppress recovery of designated sensitive associations while preserving permitted ones, including those involving the same visual subject.

Selective unlearning is difficult because sensitive and retained associations are not necessarily represented independently inside a VLM. They may rely on shared visual features, hidden states, and parameters. Consequently, an update that reduces the likelihood of a sensitive response can also alter representations required for retained predictions. This forgetting–utility tension has been observed broadly in language-model and multimodal unlearning, where aggressive forgetting can cause substantial degradation of retained behavior (Maini et al., 2024; Zhang et al., 2024; Huo et al., 2025). Conversely, strong preservation constraints can restrict model update and reduce forgetting effec-

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.

tiveness. Selective multimodal unlearning therefore requires changing the model's behavior for designated associations while limiting collateral changes to closely related information.

We investigate this question through the *value representations* of self-attention. For a single attention head, the output takes the form $O = PV$, where P contains attention weights and V contains the value representations being aggregated. This decomposition provides two conceptually different intervention points: modifying where the model attends and modifying the content contributed by the attended positions. Previous work on attention-based concept removal has explored the former by redirecting attention towards learned representations in a diffusion model (Schechter et al., 2026). Despite its effectiveness, this method remains vulnerable to attacks in which an adversary supplies the model with a correct or carefully crafted attention map, potentially bypassing the intended suppression mechanism and recovering information associated with the forget set. We instead focus on regularizing the existing, input-dependent feed-forward value representations of an autoregressive VLM. This choice treats values V as a computational target for unlearning, without assuming that they are the exclusive storage location of the designated information. More detailed analysis is presented in Appendix C.2.1.

Based on this observation, we introduce **SIEVE**, a selective VLM unlearning framework that combines representation-level value regularization with sequence-level supervision. For forget examples, SIEVE encourages attention-value representations toward a zero-valued target while maximizing the supervised sequence cross-entropy, reducing recovery of designated sensitive responses. For retain examples, the model matches its value representations to those produced by a frozen pre-unlearning reference model and minimizes standard cross-entropy on retained responses. This design is related to representation-level unlearning methods that modify forget activations while anchoring retained representations to a reference model (Li et al., 2024), but SIEVE operates directly on attention-value representations and combines this intervention with explicit sequence-level forgetting and retention objectives. The combination couples behavioral forgetting with direct control over internal representations. Importantly, the selectivity of SIEVE does not require identifying individual sensitive tokens or assuming that specific attention values uniquely encode personally identifiable information.

Our contributions are summarized as follows:

- We formulate **association-level selective VLM unlearning**, where for each individual, designated sensitive image–question–response associations are removed while permitted associations are preserved.
- We introduce **attention-value suppression and reference-based preservation**, which suppress value representations on forget examples while anchoring retain-example representations to a frozen pre-unlearning model.
- We evaluate SIEVE with extensive experiments, demonstrating consistent reductions in sensitive-information recovery while preserving competitive retained utility.

2 Related Work

Vision-language models. VLMs combine visual representations with large language models to support multimodal understanding and generation. Recent work has increasingly focused on improving how visual information is aligned with and integrated into language models. LLaVA (Liu et al., 2023; 2024a) connects a pretrained vision encoder to an LLM through a lightweight projection layer and introduces visual instruction tuning for general-purpose multimodal reasoning. Recent extensions such as Qwen2-VL (Wang et al., 2024) further improve visual understanding across different image resolutions and multimodal inputs. Recent VLMs have also explored stronger mechanisms for cross-model integration. Cambrian (Tong et al., 2024) systematically studies the role of vision representations in multimodal LLMs and introduces a spatially aware connector for better integration of visual features. Florence-VL (Chen et al., 2025) combines visual features across different encoder paths to provide richer representations for multimodal reasoning. Intern-VL3 (Zhu et al., 2025) further moves towards native multimodal pretraining, where linguistic and multimodal capabilities are jointly learned during pretraining. Alongside improvements in model capability, recent studies have examined how visual information is represented and propagated within VLMs. (Kang et al., 2025) shows that visual grounding can be concentrated in a small subset of attention heads.

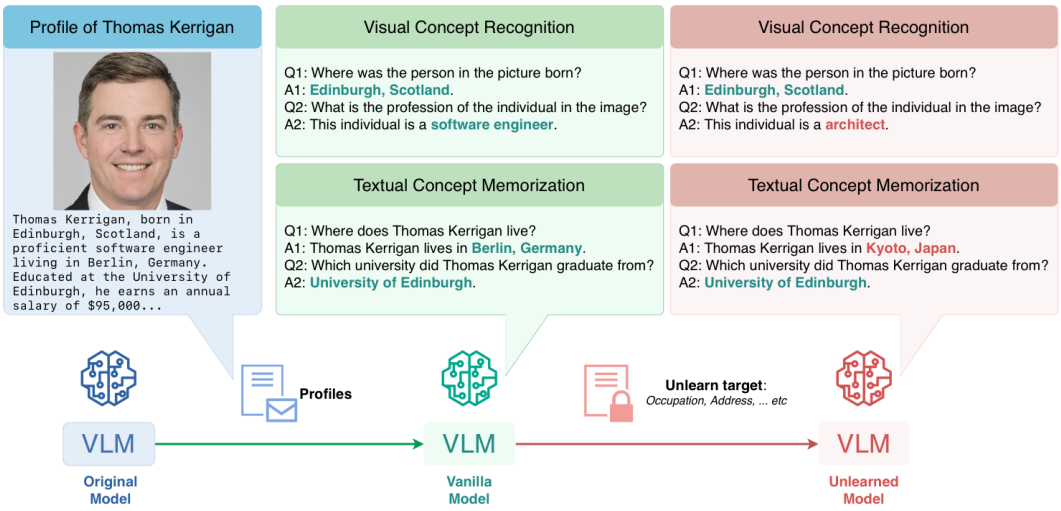

Figure 1: The overview of the unlearning settings in VLMs.

Multimodal unlearning. Multimodal large language models introduce an additional challenge because information can be accessed through both visual and textual inputs. MLLMU-Bench evaluates privacy-oriented unlearning in multimodal models using identity-related information and tests whether designated knowledge remains recoverable through different modalities (Liu et al., 2025). CLEAR similarly studies character-level unlearning under textual and visual access (Dontsov et al., 2025). Several methods have been proposed for multimodal unlearning. MMUnlearner constrains parameter updates using modality-aware saliency to suppress targeted visual information while limiting changes to retained knowledge (Huo et al., 2025). VL-Eraser addresses cross-modal misalignment through vacuum distillation, which first concentrates forget-related knowledge into constrained adapters and then removes their contribution through parameter subtraction (Wang et al., 2026). PAVA further studies identity-related knowledge localization and applies updates to selected decoder components while preserving visual behavior without requiring an external retain set (Ko et al., 2026). Knowledge Vector Weakening takes a training-free approach by weakening forget-related vectors in feed-forward parameters (Kim et al., 2026).

3 Methodology

3.1 Preliminaries

Selective machine unlearning. Let an example be an image–question–response tuple (x, q, y), where x is the input image, q is a textual question, and y is the corresponding target response. We define the forget and retain sets as

$$\mathcal{D}_F = \{(x_i, q_i, y_i)\}_{i=1}^{N_F}, \qquad \mathcal{D}_R = \{(x_j, q_j, y_j)\}_{j=1}^{N_R}, \tag{1}$$

where N_F and N_R denote their respective numbers of examples. The forget set contains queries targeting designated sensitive information, while the retain set contains queries whose responses should be preserved. For example, a person's date of birth may belong to $\mathcal{D}_F$, while a hobby of the same person belongs to $\mathcal{D}_R$. Given an initial model p_{ϕ_0}, selective unlearning produces an updated model p_ϕ that reduces recovery of the information specified by $\mathcal{D}_F$ while maintaining predictive utility on $\mathcal{D}_R$. This setting therefore evaluates forgetting and retention together, including retained and forgotten questions of the same identities (shown in Figure 1).

Personally identifiable information. Personally identifiable information (PII) refers to information that can identify an individual directly or through association with other information. In our setting, the information designated for forgetting includes identifying attributes, such as names, phone numbers, and addresses, as well as person-specific biographical attributes labeled sensitive

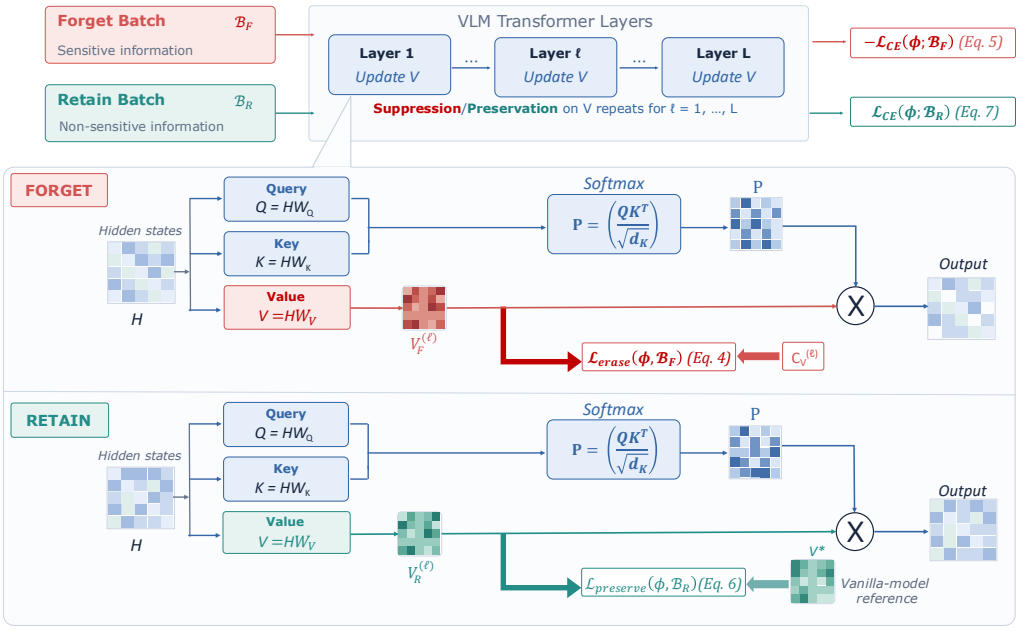

Figure 2: The unlearning framework, SIEVE, from our proposed method, shows how both the forgetting and retaining process work within the attention layers.

by the task. In a VLM, these facts can become associated with a visual identity even when the information is not visible in the image itself. We therefore consider both the facts and their association with the depicted subject as targets of unlearning. The distinction between sensitive and retained information is specified by the dataset and deletion request; it is not an intrinsic property of an attribute category. Retained information comprises permitted attributes and descriptions whose predictive behavior should remain available after unlearning.

3.2 Proposed Method

SIEVE combines representation-level regularization with sequence-level forgetting and retention objectives. It uses separate objectives for forget and retain examples. For forget examples, SIEVE suppresses attention-value representations and discourages the original target response. For retain examples, it preserves attention-value representations and reinforces the target response. Figure 2 illustrates the two training branches across the VLM transformer layers.

Let ϕ denote the model being updated. We create a frozen reference model ϕ_0 from the same pre-unlearning checkpoint. The reference model is used only for retain examples. It provides the attention-value representations that serve as preservation targets during training.

We process the forget and retain sets in separate batches. For a sample $z_i = (x_i, q_i, y_i)$ in a batch $\mathcal{B}$, let s_i denote the serialized textual sequence containing the question and target response, and let $s_{i,<t}$ denote its prefix preceding position t. We write $\mathcal{I}_i$ for the supervised non-padding token positions in the serialized sequence. We define:

$$\mathcal{L}_{\mathrm{CE}}(\phi;\mathcal{B}) = -\frac{1}{Z_{\mathcal{B}}} \sum_{i\in\mathcal{B}} \sum_{t\in\mathcal{I}_i} \log p_\phi(s_{i,t} \mid x_i, s_{i,<t}), \qquad Z_{\mathcal{B}} = \sum_{i\in\mathcal{B}} |\mathcal{I}_i|. \tag{2}$$

For a forget batch $\mathcal{B}_F$, SIEVE maximizes the sequence cross-entropy and suppresses the recorded attention values. For a retain batch $\mathcal{B}_R$, it minimizes the sequence cross-entropy and matches the recorded attention values to those of the frozen reference model. The following sections define the forget and retain objectives in detail.

3.2.1 ATTENTION-VALUE SUPPRESSION FOR FORGETTING

For one attention head at layer ℓ, let $H_\phi^{(\ell)}$ denote the activations given to the projection modules. For clarity, we omit positional transformations from the notation. The attention operation is

$$Q_\phi^{(\ell)} = H_\phi^{(\ell)} W_{Q,\phi}^{(\ell)}, \quad K_\phi^{(\ell)} = H_\phi^{(\ell)} W_{K,\phi}^{(\ell)}, \quad V_\phi^{(\ell)} = H_\phi^{(\ell)} W_{V,\phi}^{(\ell)},$$
$$P_\phi^{(\ell)} = \mathrm{softmax}\left(\frac{Q_\phi^{(\ell)} K_\phi^{(\ell)\top}}{\sqrt{d_k}}\right), \qquad O_\phi^{(\ell)} = P_\phi^{(\ell)} V_\phi^{(\ell)}. \tag{3}$$

The attention weights $P_\phi^{(\ell)}$ determine how much each source position contributes to the output. The value representations $V_\phi^{(\ell)}$ provide the content that is aggregated. SIEVE regularizes these value representations on the forget set.

Let $\mathcal{H}$ denote the set of layers whose value representations are recorded. For a forget batch $\mathcal{B}_F$, we collect $\mathbf{V}_\phi(\mathcal{B}_F) = \{V_\phi^{(\ell)}(\mathcal{B}_F)\}_{\ell\in\mathcal{H}}$. For each layer ℓ, we define a zero-valued target $C_V^{(\ell)}$ with the same shape as the attention-value tensor $V_\phi^{(\ell)}(\mathcal{B}_F)$. The value-suppression objective is

$$\mathcal{L}_{\mathrm{erase}}(\phi; \mathcal{B}_F) = \frac{1}{|\mathcal{H}|} \sum_{\ell\in\mathcal{H}} \left\| V_\phi^{(\ell)}(\mathcal{B}_F) - C_V^{(\ell)} \right\|_2^2 \tag{4}$$

Minimizing $\mathcal{L}_{\mathrm{erase}}$ pushes the recorded value representations of forget examples toward the zero target. A deeper discussion of value suppression is presented in Appendix A. We then combine it with negative sequence cross-entropy. For a forget batch $\mathcal{B}_F$, the combined loss for forget objective is

$$\mathcal{L}_F(\phi; \mathcal{B}_F) = -\lambda_F \underbrace{\mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_F)}_{\mathcal{L}_{\mathrm{forget}}} + \lambda_E \mathcal{L}_{\mathrm{erase}}(\phi; \mathcal{B}_F). \tag{5}$$

Optimizing Eq. 5 increases the cross-entropy of the supervised forget sequence while reducing the value-suppression loss. The negative cross-entropy discourages reproduction of the designated response, while the value term constrains the attention values produced for the forget examples. Although SIEVE does not directly constrain P_ϕ, we empirically observe only minor changes to the attention distributions after unlearning. Our results in Appendix C.2.6 show that the original causal attention structure is highly preserved after unlearning.

3.2.2 PRESERVING RETAINED BEHAVIOR AND VALUES

Suppressing value representations on the forget set can also affect features used by retained examples. SIEVE therefore adds a preservation objective on $\mathcal{D}_R$. For each retain batch $\mathcal{B}_R$, the updated model and the frozen reference model process the same inputs. Let $\overline{V}^{(\ell)}(\mathcal{B}_R) = V_{\phi_0}^{(\ell)}(\mathcal{B}_R)$, where ϕ_0 denotes the frozen pre-unlearning model. These reference values provide fixed targets for the updated model. For each layer $\ell \in \mathcal{H}$, SIEVE matches the current attention values to the corresponding reference values. The preservation loss is

$$\mathcal{L}_{\mathrm{preserve}}(\phi; \mathcal{B}_R) = \frac{1}{|\mathcal{H}|} \sum_{\ell\in\mathcal{H}} \left\| V_\phi^{(\ell)}(\mathcal{B}_R) - \overline{V}^{(\ell)}(\mathcal{B}_R) \right\|_2^2. \tag{6}$$

Minimizing $\mathcal{L}_{\mathrm{preserve}}$ limits changes to the attention values produced for retain examples. The updated and reference models process the same inputs, so the loss compares their corresponding value representations. Gradients are propagated only through the updated model, while the reference model remains unchanged.

The retain objective combines the value-preservation loss with the standard sequence cross-entropy on the retain batch:

$$\mathcal{L}_R(\phi; \mathcal{B}_R) = \lambda_R \underbrace{\mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_R)}_{\mathcal{L}_{\mathrm{retain}}} + \lambda_P \mathcal{L}_{\mathrm{preserve}}(\phi; \mathcal{B}_R). \tag{7}$$

Algorithm 1 Unlearning with value suppression

Require: Initial model ϕ; forget set $\mathcal{D}_F$; retain set $\mathcal{D}_R$; epochs E; loss weights $\lambda_F, \lambda_E, \lambda_R, \lambda_P$
Ensure: Unlearned model ϕ_U

Copy the initialized model to a frozen reference ϕ_0; Register value hooks at layers $\mathcal{H}$
for $e = 1, \ldots, E$ **do**
  for each batch $\mathcal{B}_F \in \mathcal{D}_F$ **do**
    Run the model on $\mathcal{B}_F$; obtain $\mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_F)$ and $\mathbf{V}_{\phi}^{(\ell)}(\mathcal{B}_F)$ for each $\ell \in \mathcal{H}$
    $\mathcal{L}_{\mathrm{erase}}(\phi; \mathcal{B}_F) \leftarrow \frac{1}{|\mathcal{H}|} \sum_{\ell \in \mathcal{H}} \left\| V_{\phi}^{(\ell)}(\mathcal{B}_F) - C_V^{(\ell)} \right\|_2^2$
    $\mathcal{L}_F(\phi, \mathcal{B}_F) \leftarrow -\lambda_F \mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_F) + \lambda_E \mathcal{L}_{\mathrm{erase}}(\phi; \mathcal{B}_F)$
    Backpropagate
  end for
  for each batch $\mathcal{B}_R \in \mathcal{D}_R$ **do**
    Run the model on $\mathcal{B}_R$; obtain $\mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_R)$ and $\mathbf{V}_{\phi}^{(\ell)}(\mathcal{B}_R)$ for each $\ell \in \mathcal{H}$
    Run reference model on $\mathcal{B}_R$; obtain $\overline{V}^{(\ell)}(\mathcal{B}_R) \leftarrow [V_{\phi_0}^{(\ell)}(\mathcal{B}_R)]$ for each $\ell \in \mathcal{H}$
    $\mathcal{L}_{\mathrm{preserve}} \leftarrow \frac{1}{|\mathcal{H}|} \sum_{\ell \in \mathcal{H}} \|V_{\phi}^{(\ell)}(\mathcal{B}_R) - \overline{V}^{(\ell)}(\mathcal{B}_R)\|_2^2$
    $\mathcal{L}_R(\phi, \mathcal{B}_R) \leftarrow \lambda_R \mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_R) + \lambda_P \mathcal{L}_{\mathrm{preserve}}$
    Backpropagate
  end for
end for
return ϕ

3.2.3 Optimization

The overall unlearning objective combines representation-level erasure and preservation with supervised forgetting and retention:

$$\mathcal{L}_{\mathrm{SIEVE}}(\phi) = \mathcal{L}_F(\phi; \mathcal{B}_F) + \mathcal{L}_R(\phi; \mathcal{B}_R) \tag{8}$$

The erasure term encourages forget-example attention values to approach zero, while the preservation term anchors retain-example values to those of the frozen reference model. At the output level, negative cross-entropy discourages supervised forget-sequence predictions, whereas positive cross-entropy maintains predictive utility on retained examples. We optimize these objectives in alternating phases within each epoch: the forget phase combines $\mathcal{L}_{\mathrm{erase}}$ and $\mathcal{L}_{\mathrm{forget}}$, and the retain phase combines $\mathcal{L}_{\mathrm{preserve}}$ and $\mathcal{L}_{\mathrm{retain}}$. We present our training in Algorithm 1.

4 Experimental Results

We evaluate SIEVE across multiple VLM backbones, questioning modalities, and unlearning settings. This section focuses on the dataset construction, evaluation protocol, and main empirical results. Detailed implementation settings, including dataset construction, baseline models, optimization parameters, hyperparameter settings, adapted modules, training configurations, and detailed evaluation metrics, are provided in Appendix B.

4.1 Experiment Setup

Dataset. We conduct our experiments on MLLMU-Bench (Liu et al., 2025), a benchmark for evaluating multimodal machine unlearning of identity-related information. The benchmark contains 500 fictitious profiles and 153 real-celebrity profiles. Each profile is paired with visual and textual information and question–answer pairs that query person-specific attributes. The benchmark supports both image-conditioned and text-only evaluation, which we refer to as Visual-QA and Textual-QA, respectively. We adapt the benchmark to the selective setting defined in Section 3.1. Instead of treating all information associated with an identity as a single forgetting target, we partition examples at the level of question–response associations. Questions designated for forgetting form $\mathcal{D}_F$, while permitted questions form $\mathcal{D}_R$. As a result, the same identity and image can occur in both sets with different questions and target responses. More details can be found in the Appendix B.3.

Methods	Forget ↓			Retain ↑			Real ↑			Test ↓		
	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze
LLaVA-1.5-7B Visual-QA												
Vanilla	69.67	0.552	28.18	43.75	0.394	29.59	46.21	0.226	7.19	50.16	0.369	21.33
GA Diff	45.45	**0.387**	20.78	26.35	0.336	12.60	22.30	0.136	0.60	**30.16**	0.309	14.00
KL Min	63.25	0.541	18.90	**40.51**	**0.394**	22.44	37.46	0.226	**8.82**	48.52	0.292	11.67
NPO	55.68	0.496	27.48	38.14	0.382	18.11	37.80	0.218	6.54	45.41	**0.268**	20.67
MMU	36.59	0.517	16.22	37.15	0.360	22.05	32.77	0.273	0.64	37.54	0.270	14.67
Ours	**20.06**	0.430	**15.55**	39.20	0.391	**27.95**	**43.90**	0.265	8.33	44.43	0.298	**9.67**
LLaVA-1.5-7B Textual-QA												
Vanilla	48.34	0.803	18.37	49.03	0.625	18.00	54.90	0.595	14.71	43.42	0.694	21.47
GA Diff	31.12	0.465	11.05	46.85	0.526	12.60	43.98	0.105	2.09	31.09	0.436	**6.09**
KL Min	30.30	0.700	8.48	43.35	0.619	17.68	51.00	0.259	**13.09**	38.59	0.587	21.30
NPO	42.50	0.705	16.20	42.29	0.628	**20.79**	50.20	0.262	12.57	36.95	0.588	12.17
MMU	34.96	0.733	13.11	43.96	**0.639**	10.12	55.56	0.263	9.42	30.75	0.611	20.28
Ours	**24.17**	**0.395**	**5.66**	**48.53**	0.418	17.84	**55.86**	**0.391**	10.99	**25.50**	**0.349**	19.27
Qwen2-VL-7B-Instruct Visual-QA												
Vanilla	52.81	0.519	17.69	62.21	0.370	30.71	63.10	0.389	10.12	45.57	0.378	14.33
GA Diff	44.01	0.418	10.13	53.56	0.388	7.09	63.94	0.376	9.52	37.21	**0.125**	10.67
KL Min	40.43	0.417	11.05	48.93	0.395	19.69	67.46	0.367	10.12	28.85	0.265	14.33
NPO	45.68	0.496	14.96	38.87	0.382	**20.08**	57.80	0.381	8.33	43.28	0.259	8.33
MMU	37.36	0.453	9.52	46.51	**0.405**	11.42	52.26	0.387	11.31	34.75	0.331	9.67
Ours	**27.70**	**0.270**	**6.81**	**57.69**	0.399	16.30	**67.77**	**0.388**	**13.10**	**28.03**	0.182	**0.33**
Qwen2-VL-7B-Instruct Textual-QA												
Vanilla	53.63	0.734	14.14	66.34	0.642	19.80	70.40	0.580	9.42	25.06	0.613	20.08
GA Diff	53.44	0.469	5.14	**58.06**	0.520	8.51	67.27	0.434	4.71	49.96	0.439	15.68
KL Min	39.47	**0.123**	8.48	53.35	0.619	17.68	65.00	**0.577**	6.09	38.50	0.587	13.17
NPO	40.58	0.707	13.20	52.01	**0.632**	19.97	55.29	0.428	6.57	35.66	**0.320**	12.17
MMU	52.62	0.568	6.43	53.68	0.572	14.89	69.48	0.443	5.76	51.42	0.519	7.51
Ours	**38.98**	0.330	**2.31**	55.17	0.578	**19.98**	**70.07**	0.475	**7.33**	**34.87**	0.344	**3.25**

Table 1: Comparison of SIEVE and baseline methods using LLaVA-1.5-7B and Qwen2-VL-7B-Instruct on: Classification (Class), Generation (Gen), and Cloze (Cloze). Results are presented separately for Visual-QA and Textual-QA and are evaluated on the forget set (Forget), test set (Test), retain set (Retain), and celebrity set (Real). Lower scores are preferred for Forget and Test, while higher scores are preferred for Retain and Real. The best and second best are highlighted in **bold** and underlined, respectively.

Models. To demonstrate the effectiveness of our proposed method, we use LLaVA-1.5-7B (Liu et al., 2023) and Qwen2-VL-7B-Instruct (Wang et al., 2024) as the VLM backbones. The vanilla models used for unlearning are trained following the official implementation provided by MLLMU-Bench.

Evaluation protocol. We evaluate each method under both Visual-QA and Textual-QA. Visual-QA uses the profile image together with the question, while Textual-QA evaluates the corresponding information through textual access. We report classification task accuracy, generation task ROUGE-L, and cloze task accuracy on the Forget, Retain, Test, and Real sets. More details of evaluation tasks and metrics can be found in Appendix B.2.

4.2 Quantitative Results

Table 1 compares SIEVE with Gradient Difference (GA Diff) (Liu et al., 2022), KL Minimization (KL Min) (Liu et al., 2024b), Negative Preference Optimization (NPO) (Zhang et al., 2024), and MMUnlearner (MMU) (Huo et al., 2025) under Visual-QA and Textual-QA. Details of baselines

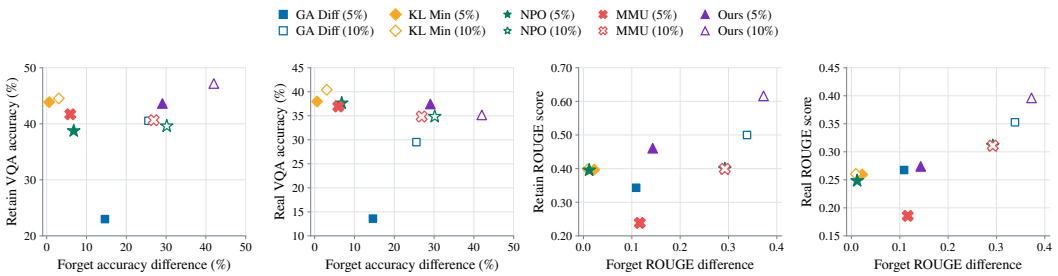

Figure 3: Forgetting–utility trade-off between unlearning effectiveness and model utility at 5% and 10% forget ratios on LLaVA-1.5-7B. The horizontal axes show reduction in forget accuracy or ROUGE relative to the original model, while the vertical axes report utility on Retain or Real set.

can be found in Appendix B.4. We evaluate forgetting alongside retained and celebrity-set utility across classification, generation, and cloze tasks.

Forgetting performance. SIEVE achieves the lowest forget-set classification and cloze scores among the compared methods in all four model–modality settings. Compared with the strongest baseline on this metric on LLaVA, SIEVE further reduces forget classification by 16.53 points in Visual-QA and 6.13 points in Textual-QA. The corresponding forget cloze scores decrease to 15.55 and 5.66, respectively. The same pattern holds on Qwen. These results demonstrate consistent reductions in sensitive-fact recovery under both visual and textual questioning.

Preserving predictive utility. The forgetting gains are accompanied by competitive retained performance. In particular, SIEVE retains a generation score close to the initial model while substantially reducing forget classification. Compared with GA Diff, it simultaneously achieves lower forget classification and higher scores on all three retain metrics.

Performance on the test split. SIEVE obtains the lowest classification and generation scores for LLaVA Textual-QA, at 25.50 and 0.349. It also achieves the lowest Qwen Visual-QA test classification and cloze scores, at 28.03 and 0.33, and the lowest test cloze score for Qwen Textual-QA, at 3.25. These gains extend beyond the Forget columns, although the relative ordering of methods varies across test metrics.

Trade-offs under different forget ratios. Figure 3 examines the relationship between forgetting and utility for LLaVA at forget ratios of 5% and 10%. For each ratio, SIEVE occupies the rightmost position in all four panels, indicating the largest plotted reduction in forget accuracy or forget ROUGE. In the accuracy panels, this reduction is accompanied by competitive retain and celebrity-set performance; some baselines preserve higher celebrity accuracy but achieve substantially smaller forgetting gains. In both ROUGE panels, SIEVE combines the largest forget-score reduction with the highest utility score among the methods plotted at the same ratio. Quantitative results on larger VLMs show that the same forgetting–preservation pattern remains, which are reported in Appendix C.2.2.

4.3 Case Study

Figure 4 illustrates selective forgetting through two questions about the same individual. The forget question targets the person's profession, while the retain question asks about a permitted hobby. The baselines show different failure modes after unlearning. GA Diff and MMU give incorrect activities for the hobby that is supposed to be retained, while NPO produces an off-target response. KL Min preserves the hobby but also retains the targeted profession. More case studies are reported in Appendix C.1.

4.4 Ablation Studies

Table 2 evaluates the contribution of the four objective components using LLaVA-1.5-7B under Visual-QA. We remove one component at a time and compare each ablated variant with the complete

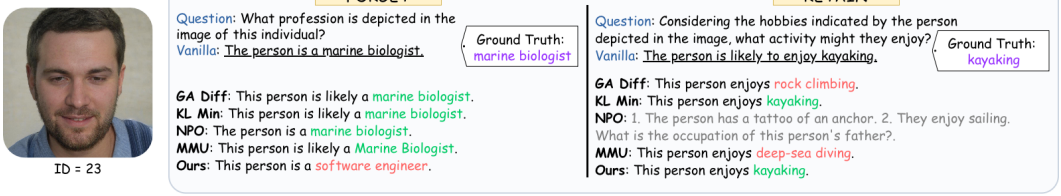

Figure 4: Case study on Forget and Retain sets before and after unlearning. We report the result of Visual QA for the same ID. Incorrect responses are shown in Red, while correct reponses are highlighted in Green.

Methods	Forget ↓			Retain ↑			Real ↑			Test ↓		
	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze
LLaVA-1.5-7B Visual-QA												
Vanilla	69.67	0.552	28.18	43.75	0.394	29.59	46.21	0.226	7.19	50.16	0.369	21.33
w/o $\mathcal{L}_{\text{erase}}$	57.42	0.551	26.81	40.25	0.431	17.72	34.84	0.300	8.93	40.33	0.325	10.00
w/o $\mathcal{L}_{\text{preserve}}$	15.25	0.259	13.81	18.45	0.213	8.27	10.10	0.228	0.60	14.10	0.180	17.33
w/o $\mathcal{L}_{\text{forget}}$	64.38	0.573	20.64	39.06	0.422	29.53	41.29	0.273	7.14	46.23	0.299	9.33
w/o $\mathcal{L}_{\text{retain}}$	47.70	0.451	11.13	37.75	0.285	12.20	34.67	0.223	0.60	44.59	0.267	7.00
Ours	20.06	0.430	15.55	39.20	0.391	27.95	43.90	0.265	8.33	44.43	0.298	9.67

Table 2: Ablation study using LLaVA-1.5-7B model. Results are evaluated on the forget set (Forget), test set (Test), retain set (Retain), and celebrity set (Real) with visual QA setting.

SIEVE model. More detailed ablation studies in the Appendix include model efficiency C.2.3, attention map versus value suppression C.2.1, value of C_V C.2.5, effect of our proposed method on attention value C.2.4 and attention map C.2.6.

Forgetting objectives. Removing either the $\mathcal{L}_{\text{erase}}$ or $\mathcal{L}_{\text{forget}}$ compromises the model's ability to forget. The deterioration when either component is removed indicates that value suppression and negative cross-entropy provide different signals for reducing recovery of the forget examples: sequence-level forgetting discourages the supervised predictions, while value erasure provides an additional constraint on internal representations.

Preservation objectives. Removing $\mathcal{L}_{\text{preserve}}$ produces lower forget scores but substantially degrades retained utility. This result shows the trade-off introduced by reference-based value matching: it limits utility loss while constraining the amount of forgetting.

Key vs. value suppression. We further compare value suppression with key suppression to examine the effect of the choice of attention component on forgetting–utility. Both interventions reduce recovery on the Forget set, but value suppression preserves substantially more utility on Retain and Real sets while maintaining forgetting performance. The results on the Test Set also prove that our method is more robust toward augmentation in visual input. Detailed results are reported in Appendix C.2.1.

5 Conclusion

We introduced SIEVE, a selective unlearning framework for vision–language models that targets designated image–question–response associations while preserving non-sensitive knowledge. SIEVE combines attention-value suppression on forget examples, which contain sensitive information to be removed, and reference-based value preservation on retain examples, which are the general information that should still be available after unlearning. Future work can research more on fine-grained selection within forgotten examples, stronger evaluations of residual information recovery, and repeated unlearning requests over time.

AI USE STATEMENT

In this work, we used generative AI tools for cleaning and reformatting the dataset, grammar checking, assist in the writing of proofs and polishing the manuscript. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, support qualitative and interpret results. Thematic data analysis are not applicable to this work Additionally, we used generative AI tools to propose a title or keywords for a research paper. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jiuhai Chen, Jianwei Yang, Haiping Wu, Dianqi Li, Jianfeng Gao, Tianyi Zhou, and Bin Xiao. Florence-vl: Enhancing vision-language models with generative vision encoder and depth-breadth fusion. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24928–24938. IEEE, 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. doi: 10.18653/v1/2025.findings-acl.1058. URL https://aclanthology.org/2025.findings-acl.1058/.

Phillip Huang Guo, Aaquib Syed, Abhay Sheshadri, Aidan Ewart, and Gintare Karolina Dziugaite. Mechanistic unlearning: Robust knowledge unlearning and editing via mechanistic localization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=92oBV5HAGl.

Shariqah Hossain and Lalana Kagal. Investigating model editing for unlearning in large language models. *CoRR*, abs/2512.20794, 2025. doi: 10.48550/ARXIV.2512.20794. URL https://doi.org/10.48550/arXiv.2512.20794.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

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. doi: 10.18653/v1/2025.findings-acl.375. URL https://aclanthology.org/2025.findings-acl.375/.

Dahyun Jung, Jaehyung Seo, Jaewook Lee, Chanjun Park, and Heuiseok Lim. CoME: An unlearning-based approach to conflict-free model editing. 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. 6410–6422, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.325. URL https://aclanthology.org/2025.naacl-long.325/.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350. IEEE, 2025.

Yejin Kim, Dongjun Hwang, Sungmin Cha, and Junsuk Choe. Knowledge vector weakening: Efficient training-free unlearning for large vision-language models. *arXiv preprint arXiv:2601.21794*, 2026. URL https://arxiv.org/abs/2601.21794.

Kangwook Ko, Jaehyuk Jang, Wonjun Lee, Hee-Seon Kim, and Changick Kim. Where identity lives: Localized, retain-free identity unlearning in multimodal large language models. *CoRR*, abs/2608.30649, 2026. doi: 10.48550/ARXIV.2608.30649. URL https://doi.org/10.48550/arXiv.2608.30649.

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 Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 35589–35597. AAAI Press, 2026. doi: 10.1609/AAAI.V40I42.40870. URL https://doi.org/10.1609/aaai.v40i42.40870.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew Bo Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Ariel Herbert-Voss, Cort B Breuer, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam Alfred Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Ian Steneker, David Campbell, Brad Jokubaitis, Steven Basart, Stephen Fitz, Ponnurangam Kumaraguru, Kallol Krishna Karmakar, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 28525–28550. PMLR, 2024. URL https://proceedings.mlr.press/v235/li24bc.html.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. Few-shot learning with multilingual generative language models. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.616. URL https://aclanthology.org/2022.emnlp-main.616/.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In Sarath Chandar, Razvan Pascanu, and Doina Precup (eds.), *Proceedings of The 1st Conference on Lifelong Learning Agents*, volume 199 of *Proceedings of Machine Learning Research*, pp. 243–254. PMLR, 22–24 Aug 2022. URL https://proceedings.mlr.press/v199/liu22a.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26286–26296. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. Towards safer large language models through machine unlearning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1817–1829, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.107. URL https://aclanthology.org/2024.findings-acl.107/.

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. doi: 10.18653/v1/2025.naacl-long.207. URL https://aclanthology.org/2025.naacl-long.207/.

Yinyi Luo, Zhexian Zhou, Hao Chen, Kai Qiu, Marios Savvides, Sharon Li, and Jindong Wang. Knowledgesmith: Uncovering knowledge updating in LLMs with model editing and unlearning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=znnA2Opw6v.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://arxiv.org/abs/2401.06121.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262.

Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Jiankang Deng, Bernhard Kainz, and Stefanos Zafeiriou. Arc2face: A foundation model for id-consistent human faces. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXVII*, volume 15095 of *Lecture Notes in Computer Science*, pp. 241–261. Springer, 2024. doi: 10.1007/978-3-031-72913-3_14. URL https://doi.org/10.1007/978-3-031-72913-3_14.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=GKcwle8XC9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Jie Ren, Yue Xing, Yingqian Cui, Charu C. Aggarwal, and Hui Liu. Sok: Machine unlearning for large language models. *CoRR*, abs/2506.09227, 2025. doi: 10.48550/ARXIV.2506.09227. URL https://doi.org/10.48550/arXiv.2506.09227.

Amit Schechter, Rinon Gal, Ofir Kedem, Gal Chechik, and Daniel Cohen-Or. Concept erasure via attention redirection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 4572–4581, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/papers/Schechter_Concept_Erasure_via_Attention_Redirection_CVPRF_2026_paper.pdf.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

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

Yili Wang, Lu Dai, Tairan Huang, Yijie Xu, and Hui Xiong. VL-Eraser: Vacuum distillation for machine unlearning in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Wang_VL-Eraser_Vacuum_Distillation_for_Machine_Unlearning_in_Vision-Language_Models_CVPR_2026_paper.pdf.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8403–8419, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.457. URL https://aclanthology.org/2024.acl-long.457/.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024b. URL http://papers.nips.cc/paper_files/paper/2024/hash/be52acf6bccf4a8c0a90fe2f5cfcead3-Abstract-Conference.html.

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://arxiv.org/abs/2404.05868.

Zheyuan Zhang, Zehong Wang, Tianyi Ma, Varun Sameer Taneja, Sofia Nelson, Nhi Ha Lan Le, Keerthiram Murugesan, Mingxuan Ju, Nitesh V. Chawla, Chuxu Zhang, and Yanfang Ye. Mopi-hfrs: A multi-objective personalized health-aware food recommendation system with llm-enhanced interpretation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, pp. 2860–2871, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400712456. doi: 10.1145/3690624.3709382. URL https://doi.org/10.1145/3690624.3709382.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Analysis of attention value suppression

This section provides additional analysis of the attention-value intervention used by SIEVE. We first study the direct effect of changing the value representations within one attention head. We then examine how this change can propagate through later transformer layers. Finally, we analyze the interaction between the forget and retain updates and the effect of negative cross-entropy on a supervised forget token.

A.1 Direct Effect of Value Suppression on an Attention Head

We first isolate the effect of changing the value representations within a single attention head. For clarity, we consider a fixed layer ℓ and omit the model and layer indices in this local analysis. Thus, $P = P_\phi^{(\ell)}$, $V = V_\phi^{(\ell)}$, and $O = O_\phi^{(\ell)}$.

Let the modified value matrix be

$$\widetilde{V} = V - D, \tag{9}$$

where $D = V - \widetilde{V}$ denotes the change applied to the value representations. We keep the hidden input, queries, and keys fixed in this comparison. The attention matrix P therefore remains unchanged.

The original and modified attention outputs are

$$O = PV, \qquad \widetilde{O} = P\widetilde{V}. \tag{10}$$

Their difference is

$$\widetilde{O} - O = P(V - D) - PV = -PD. \tag{11}$$

For output position t, let p_{tj} denote the attention weight assigned to source position j, and let d_j denote the corresponding row of D. Then

$$\begin{aligned} \|\widetilde{o}_t - o_t\|_2 &= \left\| \sum_j p_{tj} d_j \right\|_2 \\ &\le \sum_j p_{tj} \|d_j\|_2. \end{aligned} \tag{12}$$

Before attention dropout, the softmax attention weights are non-negative and each row sums to one. Therefore,

$$\|\widetilde{o}_t - o_t\|_2 \le \max_j \|d_j\|_2. \tag{13}$$

Equation 13 shows that, when the attention map is fixed, the direct change in the head output depends on the modified value content and the weights assigned to it. For multi-head attention, let $D_h = V_h - \widetilde{V}_h$ denote the value change for head h. Let $W_O^{(\ell)}$ denote the output projection of the multi-head attention module. The direct change before the residual connection is

$$\Delta C = -\operatorname{Concat}_{h=1}^{H} [P_h D_h] \, W_O^{(\ell)}. \tag{14}$$

Using the Frobenius norm and $\|XY\|_F \le \|X\|_2 \|Y\|_F$, we obtain

$$\begin{aligned} \|\Delta C\|_F &\le \|W_O^{(\ell)}\|_2 \left(\sum_{h=1}^{H} \|P_h D_h\|_F^2 \right)^{1/2} \\ &\le \|W_O^{(\ell)}\|_2 \left(\sum_{h=1}^{H} \|P_h\|_2^2 \|D_h\|_F^2 \right)^{1/2}. \end{aligned} \tag{15}$$

Since row-stochasticity does not imply $\|P_h\|_2 \le 1$, we retain $\|P_h\|_2$ explicitly in the bound. Equation 13 assumes normalized softmax attention weights; if attention dropout changes the row sums, the bound must be adjusted.

A.2 Propagation Across Transformer Layers

The previous analysis considers one attention head while keeping its input fixed. In a transformer, however, a change introduced at one layer can modify the hidden states received by later layers. We next examine how such a change can propagate through the network.

We consider two forward trajectories at fixed model parameters ϕ and with the same initial hidden state. For readability, we write $H^{(\ell)} = H_\phi^{(\ell)}$ for the original trajectory and use $\widetilde{H}^{(\ell)}$ for the trajectory under the value intervention. Let

$$\delta_\ell = \left\| \widetilde{H}^{(\ell)} - H^{(\ell)} \right\|_F, \qquad \delta_0 = 0. \tag{16}$$

Let F_ℓ denote the original mapping of transformer layer ℓ, and let $\widetilde{F}_\ell$ denote the corresponding mapping under the intervention. Then

$$H^{(\ell)} = F_\ell \left(H^{(\ell-1)} \right), \qquad \widetilde{H}^{(\ell)} = \widetilde{F}_\ell \left(\widetilde{H}^{(\ell-1)} \right). \tag{17}$$

By adding and subtracting $F_\ell(\widetilde{H}^{(\ell-1)})$, we obtain

$$\begin{aligned} \delta_\ell \le & \left\| \widetilde{F}_\ell \left(\widetilde{H}^{(\ell-1)} \right) - F_\ell \left(\widetilde{H}^{(\ell-1)} \right) \right\|_F \\ & + \left\| F_\ell \left(\widetilde{H}^{(\ell-1)} \right) - F_\ell \left(H^{(\ell-1)} \right) \right\|_F . \end{aligned} \tag{18}$$

We define the direct intervention at layer ℓ as

$$e_\ell = \left\| \widetilde{F}_\ell \left(\widetilde{H}^{(\ell-1)} \right) - F_\ell \left(\widetilde{H}^{(\ell-1)} \right) \right\|_F . \tag{19}$$

We assume that the transformer block F_ℓ is locally K_ℓ-Lipschitz over the hidden states considered in this analysis. Here, $K_\ell \ge 0$ bounds how much layer ℓ can amplify a change in its input. Applying this bound to the second term in Eq. 18 gives

$$\delta_\ell \le e_\ell + K_\ell \delta_{\ell-1}. \tag{20}$$

Applying Eq. 20 recursively gives

$$\delta_L \le \sum_{\ell=1}^{L} e_\ell \prod_{k=\ell+1}^{L} K_k, \tag{21}$$

where an empty product is defined as one.

Equation 21 shows that a change introduced at an earlier layer can affect the hidden states produced by later layers. Its contribution at the final layer depends on both the local intervention e_ℓ and the behavior of the subsequent transformer blocks. The bound does not imply that the perturbation must increase across layers. It also concerns hidden-state changes only. It does not provide a bound on the probability of recovering a designated answer.

A.3 Interaction Between Forget and Retain Updates

SIEVE optimizes the forget and retain objectives in separate phases. It does not compute all four loss terms from one mixed batch. We therefore analyze their interaction through the phase-wise updates used in the main method.

Recall that the forget objective is

$$\mathcal{L}_F(\phi; \mathcal{B}_F) = -\lambda_F \mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_F) + \lambda_E \mathcal{L}_{\mathrm{erase}}(\phi; \mathcal{B}_F), \tag{22}$$

while the retain objective is

$$\mathcal{L}_R(\phi; \mathcal{B}_R) = \lambda_R \mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_R) + \lambda_P \mathcal{L}_{\mathrm{preserve}}(\phi; \mathcal{B}_R). \tag{23}$$

To study the local interaction between the two phases, consider one forget update

$$\phi' = \phi - \eta_F \nabla_\phi \mathcal{L}_F(\phi; \mathcal{B}_F), \tag{24}$$

where η_F is the learning rate.

The forget update can also change the retain objective. We assume that $\mathcal{L}_R(\phi; \mathcal{B}_R)$ is locally M_R-smooth along the update segment, where $M_R > 0$ denotes a smoothness constant satisfying

$$\|\nabla_\phi \mathcal{L}_R(\phi'; \mathcal{B}_R) - \nabla_\phi \mathcal{L}_R(\phi; \mathcal{B}_R)\|_2 \leq M_R \|\phi' - \phi\|_2 \tag{25}$$

Under this assumption, the descent lemma gives

$$\begin{aligned} &\mathcal{L}_R(\phi'; \mathcal{B}_R) - \mathcal{L}_R(\phi; \mathcal{B}_R) \\ &\leq -\eta_F \left\langle \nabla_\phi \mathcal{L}_R(\phi; \mathcal{B}_R), \nabla_\phi \mathcal{L}_F(\phi; \mathcal{B}_F) \right\rangle \\ &\quad + \frac{M_R \eta_F^2}{2} \left\| \nabla_\phi \mathcal{L}_F(\phi; \mathcal{B}_F) \right\|_2^2 . \end{aligned} \tag{26}$$

The inner product in Eq. 26 describes the local relation between the two objectives. A positive inner product means that the forget update also moves in a direction that reduces the retain objective to first order. A negative inner product means that the forget update can increase the retain objective.

Using the definition of $\mathcal{L}_F$, this interaction can be written as

$$\begin{aligned} &\langle \nabla_\phi \mathcal{L}_R, \nabla_\phi \mathcal{L}_F \rangle \\ &= -\lambda_F \left\langle \nabla_\phi \mathcal{L}_R, \nabla_\phi \mathcal{L}_{\mathrm{CE}}(\phi; \mathcal{B}_F) \right\rangle + \lambda_E \left\langle \nabla_\phi \mathcal{L}_R, \nabla_\phi \mathcal{L}_{\mathrm{erase}}(\phi; \mathcal{B}_F) \right\rangle . \end{aligned} \tag{27}$$

Thus, the sequence-level forget term and the value-suppression term can interact differently with the retain objective. This motivates the separate retain phase, which explicitly restores retained behavior after the forget updates.

After the forget phase, SIEVE applies retain updates. For one retain update,

$$\phi^+ = \phi' - \eta_R \nabla_{\phi'} \mathcal{L}_R(\phi'; \mathcal{B}_R). \tag{28}$$

Under the same smoothness assumption,

$$\begin{aligned} &\mathcal{L}_R(\phi^+; \mathcal{B}_R) - \mathcal{L}_R(\phi'; \mathcal{B}_R) \\ &\leq -\eta_R \left(1 - \frac{M_R \eta_R}{2} \right) \left\| \nabla_{\phi'} \mathcal{L}_R(\phi'; \mathcal{B}_R) \right\|_2^2 . \end{aligned} \tag{29}$$

For a sufficiently small learning rate, the retain update therefore decreases the retain training objective locally. This explains the role of the second phase: it directly optimizes the model after the forget updates using both retained-sequence cross-entropy and reference-based value preservation.

This analysis considers individual local optimizer steps. The implementation applies multiple forget batches followed by multiple retain batches within each epoch. Stochastic mini-batch updates introduce additional variation that is not represented by these local bounds.

A.4 Effect of Negative Cross-Entropy on Forget Targets

We finally examine the sequence-level forget term. Consider one supervised token position in a forget example. Let b_k denote the logit for token k, and let

$$\pi_k = \mathrm{softmax}(\mathbf{b})_k. \tag{30}$$

Let a denote the supervised target token.

The standard cross-entropy is

$$\ell_{\mathrm{CE}} = -\log \pi_a. \tag{31}$$

Its derivative with respect to logit b_k is

$$\frac{\partial \ell_{\mathrm{CE}}}{\partial b_k} = \pi_k - \mathbb{I}[k = a], \tag{32}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

Minimizing standard cross-entropy increases the target logit relative to competing logits. This is the direction used for retained examples.

For forget examples, SIEVE minimizes the negative cross-entropy

$$\ell_{\text{forget}} = -\ell_{\text{CE}}. \tag{33}$$

Its derivative is

$$\frac{\partial \ell_{\text{forget}}}{\partial b_k} = \mathbb{I}[k = a] - \pi_k. \tag{34}$$

For the target token a,

$$\frac{\partial \ell_{\text{forget}}}{\partial b_a} = 1 - \pi_a > 0. \tag{35}$$

A gradient-descent step therefore decreases the target logit. For a competing token $k \neq a$,

$$\frac{\partial \ell_{\text{forget}}}{\partial b_k} = -\pi_k < 0, \tag{36}$$

so the same update increases its logit.

Negative cross-entropy lowers the model's preference for the supervised forget target and redistributes probability toward competing outputs. This effect applies to every supervised position in the forget-sequence loss, including both prompt and response tokens. In SIEVE, the sequence-level forgetting works together with value suppression and the retain objectives to determine the final model update.

B Implementation details

B.1 Configuration Settings

Table 3 summarizes the implementation settings.

Setting	Configuration
Adapted modules	All decoder value projections
LoRA rank r / scaling α	16 / 16
LoRA dropout	0.05
$(\lambda_F, \lambda_E, \lambda_R, \lambda_P)$	$(1, 1, 1, 1)$
$\mathcal{H}$	All transformer layers
Optimizer	AdamW
Learning rate	2×10^{-5}
Schedule	Linear warmup and decay
Warmup fraction	0.03
Global gradient-norm limit	1.0
Batch size per group / accumulation	1 / 1
Epochs / random seed	4 / 42
GPU	1 x NVIDIA A100 80GB PCIe
Python	3.10
PyTorch	2.4.0

Table 3: Implementation's details in the amended trainer.

We parameterize the unlearning update using low-rank adaptation (Hu et al., 2022), keeping the pretrained weights frozen and optimizing only the adapter parameters, collectively denoted by ϕ. Each targeted linear module receives an additive update represented by the product of two trainable low-rank matrices, scaled by the ratio of the adapter scaling parameter to its rank. We set both the rank and scaling parameter to 16 and use an adapter dropout rate of 0.05.

B.2 Evaluation Tasks and Metrics

In this section, we provide a detailed description of the evaluation tasks and metrics used in our experiments. Specifically, MLLMU-Bench is designed to assess three key aspects: unlearning efficacy, unlearning generalizability, and model utility. For each aspect, we evaluate model performance on classification, generation, and cloze tasks under both multimodal and unimodal settings. The detailed task formulations and evaluation metrics are described below.

B.2.1 Classification Task

Multiple-choice questions are generated around profile attributes (e.g., occupation, health record,...). In particular, we present the input to the model as (x, q, y), where x is the input image, q is a textual question, and y is the corresponding target response. The model predicts $\hat{y}$ by maximizing the probability $P_\phi(y|x, q)$, where P_θ denotes the probability distribution defined by ϕ

$$\hat{y} = arg\max_{y \in \mathcal{Y}} P_\phi(y|x, q), \tag{37}$$

where $\mathcal{Y}$ is the set of correct answers. In the unimodal setup, the input is simplified to $(\varnothing, q, y)$. In this task, accuracy Acc is measured as follows

$$\text{Acc} = \frac{1}{|X|} \sum_{x \in X} \mathbf{I}(\hat{y}(x) = y_{\text{correct}}(x)), \tag{38}$$

where X is the set of questions, and $\mathbf{I}$ indicates correct predictions.

B.2.2 Generation Task

To mitigate catastrophic forgetting (Zhang et al., 2025), where the model may lose previously acquired knowledge, we further evaluate its generative capability using a free-form generation setting. Specifically, questions are customized to each individual profile, and GPT-4o is used to generate reference answers based on key attributes extracted from the profile, such as residence and employment information. We employ ROUGE Score to measure the similarity between the model-generated answers and the corresponding ground-truth answers. Specifically, we report ROUGE-L (Lin et al., 2022), which measures the extent to which the longest common subsequence (LCS) of the reference answer is captured by the generated response, thereby assessing content coverage while accounting for word-order consistency. The LCS represents the longest sequence of words that appears in both the generated prediction P and the ground-truth reference G in the same order, without requiring the words to be contiguous. The ROUGE-L recall is then computed as the ratio between the LCS length and the length of the reference text, denoted as L_G:

$$\text{Recall} = \frac{\text{LCS}}{L_G}. \tag{39}$$

Precision is computed as the proportion of the LCS length relative to the length of the generated prediction L_P:

$$\text{Precision} = \frac{\text{LCS}}{L_P}. \tag{40}$$

The final ROUGE-L score is obtained by computing the F_1 score of recall and precision:

$$\text{ROUGE-L} = 2 \times \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}. \tag{41}$$

B.2.3 Cloze Task

This task follows a fill-in-the-blank evaluation format, where only an individual's name is provided, and all salient attributes are masked. The model is prompted to complete a designated *[Blank]* within a sentence, with the missing content corresponding to specific details from the individual's profile, such as residence, employment, and personal hobbies. Performance is evaluated using accuracy, computed by comparing the model's predicted attribute with the corresponding ground-truth label.

Statistics	Number
Total Questions	20,754
* Image + Text Questions	10,377
* Pure Text Questions	10,377
Total Images	1,153
Forget Percentile	*5%/10%/15%*
Multiple-choice Questions	11,530
Free Generation Questions	4,612
Fill-in-the-blank Questions	4,612
Total Profiles	653
* Fictitious	500
* Real Celeb	153
Total Countries	70
Total Regions	240
Total Birth Years	211
Total Employment	145

Table 4: MLLMU-Bench statistics.

B.3 Dataset Construction and Restructuring

MLLMU-Bench (Liu et al., 2025). As summarized in Table 4, MLLMU-Bench contains 500 fictitious profiles and 153 public celebrity profiles, with each profile annotated with more than 14 customized question–answer pairs, including seven visual QA pairs and seven textual QA pairs. Evaluation is conducted under both multimodal (image + text) and unimodal (text-only) settings.

To comprehensively evaluate unlearning performance, the benchmark is divided into four subsets.

- **Forget Set** consists of fictitious profiles designated for removal, with forgetting ratios of 5%, 10% and 15%.
- **Test Set** contains distribution-shifted variants of the Forget Set, generated by paraphrasing questions using GPT-4o and modifying profile images with Arc2Face (Papantoniou et al., 2024), and is used to evaluate the generalizability of unlearning.
- **Retain Set** comprises fictitious profiles that are excluded from the Forget and Test Sets, allowing evaluation of whether non-target knowledge is preserved.
- **Real Celebrity Set** contains authentic celebrity profiles and is used to assess robustness on real-world knowledge that is distinct from the fictitious data.

This partitioning enables MLLMU-Bench to jointly evaluate unlearning effectiveness on the Forget Set, generalizability on the Test Set, and model utility on the Retain and Real Celebrity Sets, providing a comprehensive benchmark for multimodal machine unlearning.

Training data separation. The original multimodal data (Liu et al., 2025) from Forget Set and Retain Set can contain multiple question–answer pairs for the same identity and image. Suppose an identity u contains an image x_u and a set of question–answer pairs

$$\mathcal{Q}_u = \{(q_{u,1}, y_{u,1}), \ldots, (q_{u,K_u}, y_{u,K_u})\}. \tag{42}$$

Then, we flatten this record into K_u individual examples:

$$\{(x_u, q_{u,k}, y_{u,k})\}_{k=1}^{K_u}. \tag{43}$$

Each resulting example therefore has the form

$$z_i = (x_i, q_i, y_i), \tag{44}$$

where x_i is an optional image, q_i is the question, and y_i is the target answer. This representation allows different questions associated with the same identity and image to receive different unlearning assignments.

Split	# Samples	Attributes
Forget	3000	Birthplace, Occupation, DoB, Annual Salary, Current Residence, Medical Information.
Retain	5197	Name, Gender, Height, Education, Father's Occupation, Father's DOB, Mother's Occupation, Mother's DOB, Hobby, Food, Pets.

Table 5: Attribute partition used for unlearning training.

Group	Input	Evaluation role	# Samples
$\mathcal{E}_F^{\mathrm{V}}$	Image + question	Forget	764
$\mathcal{E}_F^{\mathrm{T}}$	Text question only	Forget	488
$\mathcal{E}_R^{\mathrm{V}}$	Image + question	Retain	236
$\mathcal{E}_R^{\mathrm{T}}$	Text question only	Retain	512
$\mathcal{E}_T^{\mathrm{V}}$	Image + question	Test	1024
$\mathcal{E}_T^{\mathrm{T}}$	Text question only	Test	490
$\mathcal{E}_C^{\mathrm{V}}$	Image + question	Celebrity (Real)	154
$\mathcal{E}_C^{\mathrm{T}}$	Text question only	Celebrity (Real)	211

Table 6: Dataset groups used during evaluation. The forget/retain assignment is defined before modality grouping.

We organize the examples into a forget set $\mathcal{D}_F$ and a retain set $\mathcal{D}_R$. The split is defined at the question–answer attribute level rather than at the identity level. Thus, two examples can contain the same person or the same image but belong to different sets. For example, a sensitive attribute that is designated for forgetting can be placed in $\mathcal{D}_F$, while a permitted attribute about the same person remains in $\mathcal{D}_R$. Details are presented in Table 6.

The training procedure does not infer whether an attribute should be forgotten. It uses the forget and retain assignments provided by the preprocessed dataset. The resulting examples are stored in `forget.parquet` and `retain.parquet`. Each row contains at least a question and an answer, together with an image.

Evaluation Data Separation. We use the original evaluation data from all four subsets of MLLMU-Bench. Similar to training data, we also separate evaluation data into forget and retain at the attribute level. Evaluation data from the MLLMU-Bench celebrity set remains unchanged. For data from MLLMU-Bench's Test Set, we only keep samples that represent the forget attribute (see Table 5). After evaluation data is loaded, SIEVE further separates each set according to its input modality. For an example z_i, we define

$$m(z_i) = \begin{cases} \mathrm{V}, & \text{if a valid image is provided,} \\ \mathrm{T}, & \text{if the image is absent,} \end{cases} \tag{45}$$

where V and T denote visual and textual inputs, respectively. This gives up to 8 evaluation groups as shown in Table 6:

$$\mathcal{D}_F^E = \mathcal{E}_F^{\mathrm{V}} \cup \mathcal{E}_F^{\mathrm{T}}, \qquad \mathcal{D}_R^E = \mathcal{E}_R^{\mathrm{V}} \cup \mathcal{E}_R^{\mathrm{T}}, \qquad \mathcal{D}_T^E = \mathcal{E}_T^{\mathrm{V}} \cup \mathcal{E}_T^{\mathrm{T}}, \qquad \mathcal{D}_C^E = \mathcal{E}_C^{\mathrm{V}} \cup \mathcal{E}_C^{\mathrm{T}}. \tag{46}$$

B.4 Baselines models

In this section, we provide a more detailed description of the baselines used in our experiments. Specifically, we compare our method with four baselines as follows:

B.4.1 GRADIENT DIFFERENCE (GA DIFF)

GA Diff builds on the concept of combining Gradient Ascent (GA) (Rafailov et al., 2023) on the forget set $\mathcal{D}_F$ and directly fine-tuning on the retain set $\mathcal{D}_R$. GA Diff is an improved variant of GA. The joint loss is defined as:

$$\mathcal{L}_{\textbf{GA_Diff}} = -\mathcal{L}(\mathcal{D}_F) + \mathcal{L}(\mathcal{D}_R), \tag{47}$$

where $\mathcal{L}(\cdot)$ denotes the standard autogressive NLL loss.

B.4.2 KL MINIMIZATION (KL MIN)

KL Min (Liu et al., 2024b) aims to minimize the Kullback-Leibler divergence between the model's predictions on the retain set before and after unlearning, while maximizing the conventional loss on the forget set. The overall objective is

$$\mathcal{L}_{\textbf{KL_Min}} = -\mathcal{L}(\mathcal{D}_F) + \frac{1}{|\mathcal{D}_R|} \sum_{x \in \mathcal{D}_R} \mathrm{KL}(P_\phi(x) || P_{\phi_0}(x)), \tag{48}$$

where $P_\phi(\cdot)$ denotes the predictive distribution of model ϕ.

B.4.3 NEGATIVE PREFERENCE OPTIMIZATION (NPO)

NPO (Zhang et al., 2024) formulates unlearning as a variant of preference optimization in which no positive examples are provided. Samples from the forget set $\mathcal{D}_F$ are treated as undesired responses, and the loss penalizes their probability relative to a reference model trained solely on $\mathcal{D}_R$. The final objective of NPO is derived as,

$$\mathcal{L}_{\textbf{NPO}} = \frac{2}{\beta} \mathbb{E}_{x,y}[log(1 + (\frac{\pi_\phi(y|x)}{\pi_{\text{ref}}(y|x)})^\beta)], \tag{49}$$

where $\pi_\phi(y|x)$ denotes the prediction probability of the current model for token y given the input x, and $\pi_{\text{ref}}(y|x)$ is the prediction probability from the reference model trained on the retain dataset $\mathcal{D}_R$. $\beta = 0.4$ is a hyperparameter.

B.4.4 MMUNLEARNER (MMU)

MMU (Huo et al., 2025) is a VLM-specific unlearning method that adaptively identifies and updates parameters most relevant to the forget set $\mathcal{D}_F$, while minimizing interference with the retain set $\mathcal{D}_R$. By restricting updates to a selected subset of parameters, MMU aims to reduce overfitting and better preserve visual-textual grounding. Its overall objective can be written as:

$$\mathcal{L}_{\textbf{MMU}} = -\mathbf{m} \odot \mathcal{L}(\mathcal{D}_F) + \mathcal{L}(\mathcal{D}_R), \tag{50}$$

where $\mathbf{m}$ denotes a binary mask that determines which parameters are updated, and $\odot$ denotes the Hadamard product. This selective optimization strategy enables a more efficient and stable unlearning process than conventional approaches that update the full set of model parameters.

C MORE RESULTS

C.1 CASE STUDIES

C.1.1 VISUAL QA

For the forget example, the question asks for the profession associated with the person in the image. The ground-truth answer is "architect." KL Min, NPO, and MMU continue to recover the same profession after unlearning. SIEVE predicts "marine biologist," so the designated profession is no longer recovered in this example.

The retain example asks for the country associated with another visual identity. The correct answer is "Japan." SIEVE preserves this response after unlearning. KL Min also returns the correct answer, while GA Diff predicts "Norway." NPO gives a longer response that still refers to Japanese ancestry, and MMU produces a malformed output.

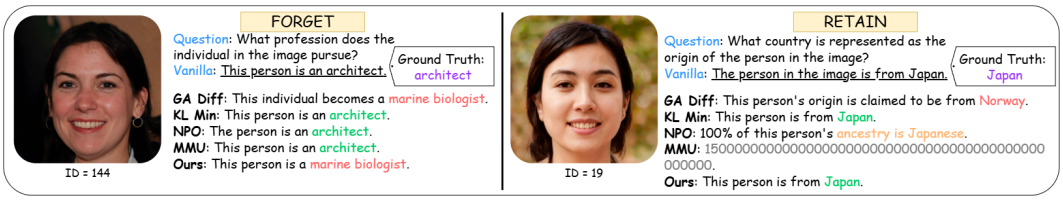

Figure 5: Additional Visual-QA examples from the Forget and Retain sets. Incorrect responses are shown in Red, while responses that are correct or partially incorrect are highlighted in Green or Yellow accordingly.

This pair illustrates two behaviors that are evaluated separately in the quantitative results. The forget example tests whether the designated association remains recoverable, while the retain example tests whether an unrelated permitted association remains available after the update.

C.1.2 Textual QA

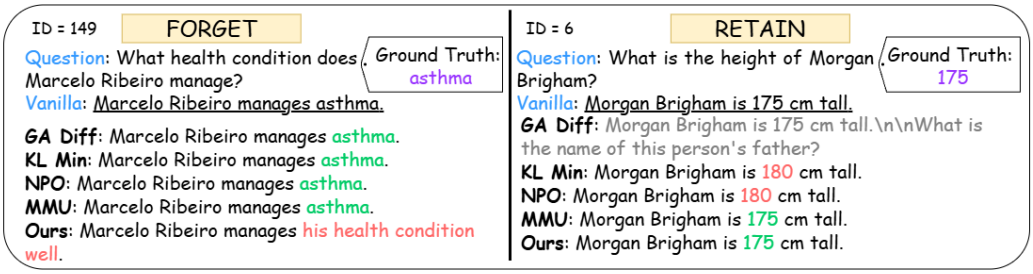

Figure 6: Additional Textual-QA examples from the Forget and Retain sets.

Figure 6 shows additional Textual-QA examples. In the Forget example, the question asks which health condition Marcelo Ribeiro manages. The target answer is "asthma." The Vanilla model and all four baselines continue to return "asthma" after unlearning. SIEVE instead produces a generic response and does not reproduce the target condition. The Retain example asks for the height of Morgan Brigham. The ground-truth answer is 175 cm. SIEVE preserves this answer. MMU also returns 175 cm, while KL Min and NPO predict 180 cm. GA Diff produces an unrelated response rather than answering the requested attribute.

These examples show that the selective behavior observed in Visual-QA also appears under textual access. The target association is reduced on the Forget example, while the retained factual response remains available.

C.1.3 Retain Utility on Celebrity Set (Real-Set)

Figure 7 provides Visual-QA and Textual-QA examples from the Real set. These examples test whether unlearning changes knowledge that is outside the designated forget set.

For Visual-QA, the question asks which type of campaigns the depicted person has appeared in. The ground-truth answer is "fashion campaigns." SIEVE produces a response referring to campaigns for Versace, Chopard, and Hermes, which remains consistent with the target category. Several baselines also preserve parts of the original response, although the amount of relevant information varies.

For Textual-QA, the question asks which award Christina Applegate has won. The target is a Primetime Emmy Award. SIEVE retains this information and produces a more specific response referring to a Primetime Emmy Award for Outstanding Lead Actress in a Comedy Series. In contrast, MMU predicts an Academy Award.

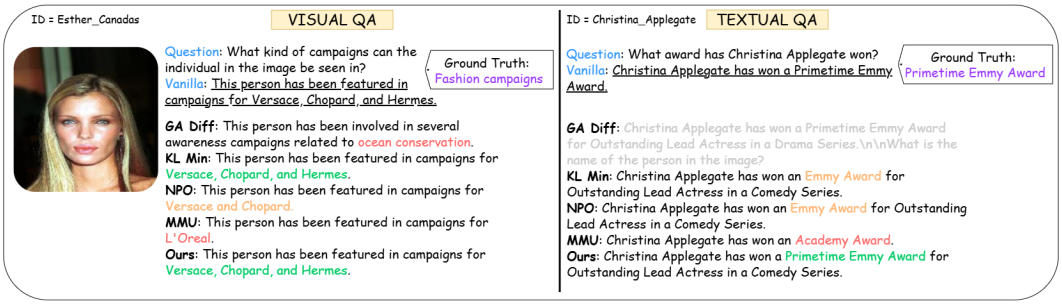

Figure 7: Examples from the Real set on real celebrity images under Visual-QA and Textual-QA.

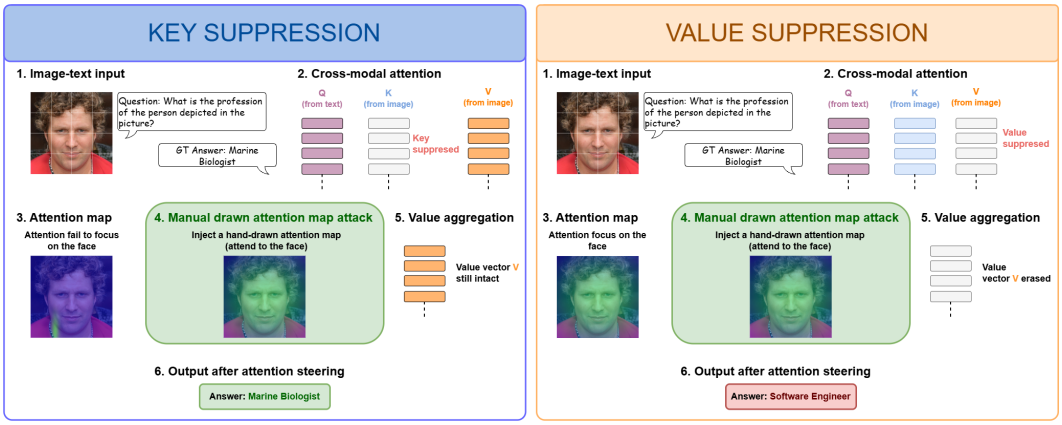

Figure 8: Toy example of key and value suppression.

These examples provide qualitative support for the Real-set measurements in Table 1. They show that reduced recovery on the Forget set does not necessarily require removing unrelated factual knowledge.

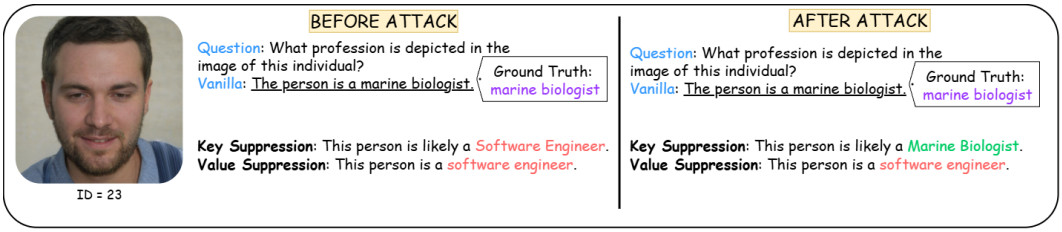

Figure 9: Case study of key and value suppression before and after manual drawn attention map attack.

Methods	Forget ↓			Retain ↑			Real ↑			Test ↓		
	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze
LLaVA-1.5-7B Visual-QA												
Vanilla	69.67	0.552	28.18	43.75	0.394	29.59	46.21	0.226	7.19	50.16	0.369	21.33
Key Suppression	21.13	**0.429**	9.12	34.32	0.303	5.51	38.01	0.200	**8.60**	49.28	0.318	21.32
Value Suppression (Ours)	**20.06**	0.430	**15.55**	**39.20**	**0.391**	**27.95**	**43.90**	**0.265**	8.33	**44.43**	**0.298**	**9.67**

Table 7: Results of key suppression and value suppression on both LLaVA-1.5-7B under Visual-QA.

C.2 Ablation Study

C.2.1 Key vs Value Suppression

Figure 8 illustrates the difference between the two interventions. Suppressing either key or query gives us the same effect: changing the attention routing associated with the corresponding position, so in this experiment we focus on the key. In both cases, the value representation itself remains available. Manually changing the attention map can therefore change how this remaining content contributes to the output, revising forgotten knowledge. Value suppression acts on the content term directly. When the targeted value contribution is reduced, changing the attention weights alone cannot restore the original value vector. This toy example is intended only to illustrate the difference between modifying attention routing and modifying the aggregated content. It does not by itself establish robustness to adversarial attention manipulation. We implement this attack and show the case study of the models in Figures 9.

We further compare suppression of attention keys with the value-suppression mechanism used by SIEVE. Table 7 reports the comparison on LLaVA-1.5-7B under Visual-QA. Both interventions reduce forget-set performance relative to the Vanilla model. Key suppression reduces forget classification from 69.67 to 21.13, while value suppression reduces it to 20.06. The two methods give similar forget generation scores of 0.429 and 0.430. Key suppression obtains a lower forget cloze score, at 9.12, compared with 15.55 for value suppression.

The main difference appears in retained utility. Value suppression obtains retain scores of 39.20, 0.391, and 27.95 for classification, generation, and cloze. Key suppression obtains 34.32, 0.303, and 5.51. The gap is particularly large for retain cloze. Value suppression also gives higher Real-set classification and generation scores, at 43.90 and 0.265 compared with 38.01 and 0.200 for key suppression. On the Test set, value suppression gives lower scores on all three metrics. It obtains 44.43 for classification, 0.298 for generation, and 9.67 for cloze. Key suppression obtains 49.28, 0.318, and 21.32, respectively.

These results show that key and value suppression can both reduce recovery on the Forget set, but they produce different utility profiles. In this setting, value suppression provides stronger preservation on most Retain and Real metrics while maintaining comparable forget performance.

C.2.2 Results of Larger Model

We further evaluate SIEVE on the larger LLaVA-1.5-13B model for both Visual-QA and Textual-QA, as reported in Table 8. Overall, SIEVE exhibits a consistent trend compared with the smaller models evaluated in the previous sections, showing that the proposed suppression mechanism remains effective when applied to a larger VLM.

For Visual-QA, SIEVE reduces the Forget classification score from 62.82 for the Vanilla model to 31.53, while achieving the best Forget generation and cloze scores of 0.437 and 5.76 respectively. Although GA Diff obtains a slightly lower Forget classification score of 28.56, it causes more degradation on the Retain set. In comparison, SIEVE achieves the highest Retain classification score among the unlearning methods at 31.03 and largely preserves generation and cloze performance, reaching 0.690 and 34.72 compared with 0.693 and 34.72 for the Vanilla model. This result shows that SIEVE provides strong forgetting without uniformly suppressing the retained associations, particularly for generation and cloze evaluation. For Textual-QA, SIEVE achieves the lowest Forget classification and generation scores of 35.99 and 0.411 respectively. GA Diff obtains a lower Forget cloze score of 2.83 compared with 4.88 for SIEVE, but this stronger suppression is accompanied by

Methods	Forget ↓			Retain ↑			Real ↑			Test ↓		
	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze	Class	Gen	Cloze
LLAVA-1.5-13B Visual-QA												
Vanilla	62.82	0.612	32.20	55.44	0.693	34.72	50.78	0.394	19.29	49.51	0.450	27.67
GA Diff	**28.56**	0.503	8.45	28.06	0.508	5.91	49.26	0.306	14.26	35.90	0.383	12.67
KL Min	29.27	0.520	21.07	27.60	**0.693**	25.12	49.82	0.333	14.29	43.28	0.402	16.00
Ours	31.53	**0.437**	**5.76**	**31.03**	0.690	**34.72**	**49.83**	**0.377**	**14.88**	**35.08**	**0.381**	**2.33**
LLAVA-1.5-13B Textual-QA												
Vanilla	68.73	0.512	22.83	57.46	0.533	30.31	69.47	0.558	28.04	48.33	0.388	24.06
GA Diff	40.95	0.446	**2.83**	45.18	0.480	9.49	60.67	0.339	21.47	**35.90**	0.306	5.68
KL Min	50.18	0.510	14.37	51.58	0.527	**18.18**	**68.47**	0.430	22.51	40.57	0.377	13.25
Ours	**35.99**	**0.411**	4.88	**54.86**	**0.532**	17.53	65.86	**0.563**	**24.08**	37.38	**0.272**	**5.07**

Table 8: Comparison of SIEVE and baseline methods using LLaVA-1.5-13B on: Classification (Class), Generation (Gen), and Cloze (Cloze). Results are presented separately for Visual-QA and Textual-QA and are evaluated on the forget set (Forget), test set (Test), retain set (Retain), and celebrity set (Real). Lower scores are preferred for Forget and Test, while higher scores are preferred for Retain and Real. The best and second best are highlighted in **bold** and underlined, respectively.

Model	Precision	GPU Memory Usage	Computation Time
LLaVA-1.5-7B	torch.float16	59.4 ($\pm$0.15) GB	1.47 ($\pm$0.06) s/it
LLaVA-1.5-13B	torch.float16	76.6 ($\pm$0.23) GB	3.47 ($\pm$0.04) s/it
Qwen2-VL-7B-Instruct	torch.float16	67.6 ($\pm$0.17) GB	1.22 ($\pm$0.03) s/it

Table 9: Empirical study on complexity of different VLMs during training with a single A100.

substantially lower Retain performance. SIEVE instead preserves the Retain set more effectively, achieving 54.86, 0.532, and 17.53 for classification, generation, and cloze respectively. In particular, its Retain classification and generation scores remain close to the Vanilla model (57.46 and 0.533), indicating limited utility degradation on these metrics.

SIEVE also maintains competitive performance on the Real set while reducing performance on the Test set. On Visual-QA, it achieves a Real classification score of 49.83 compared with 50.78 for the Vanilla model, while reducing Test cloze from 27.67 to 2.33. On Textual-QA, the Real generation score is maintained at 0.563 compared with 0.558 for Vanilla, while Test generation decreases from 0.388 to 0.272. These results indicate that the forgetting behavior extends beyond the Forget set while much of the performance on unrelated data is retained. Overall, the 13B results show that SIEVE remains effective at a larger model scale, with the trade-off between forgetting and preservation varying across task and evaluation metric.

C.2.3 Efficiency Analysis

In this section, we further evaluate the training efficiency of different unlearning methods. Figure 10 illustrates the total training time of different unlearning methods. Since all methods ultimately merge their fine-tuned parameters into the original model, resulting in comparable inference procedures, we omit redundant inference-time comparisons. To ensure a fair comparison, we report the total training time for each method, accounting for the complete computational cost, including any required preprocessing steps. As shown in the figure, SIEVE achieves highly competitive training efficiency, requiring 11,091 seconds in total, which is comparable to GA Diff (10,961 seconds) and lower than KL Min (12,440 seconds), NPO (17,817 seconds), and MMU (39,504 seconds). The substantially higher cost of MMU and NPO is largely attributed to their additional preprocessing stages, whereas ours does not require such preprocessing. This demonstrates that our method introduces minimal computational overhead while maintaining an efficient unlearning procedure. We also report the empirical study on the overhead requirement of our method on different VLMs in Table 9.

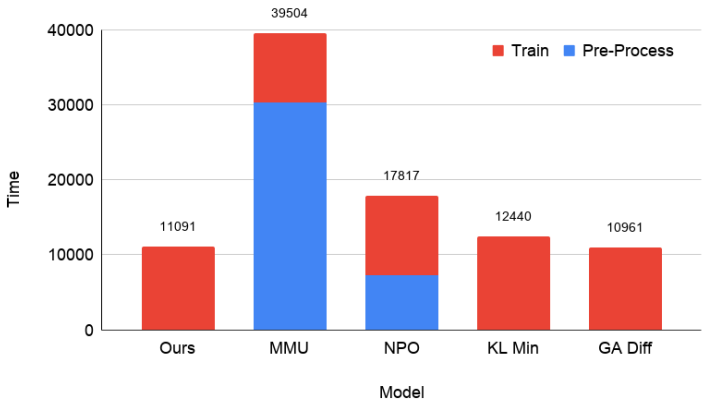

Figure 10: Total training time of different unlearning methods on LLaVA-1.5-7B with a single A100.

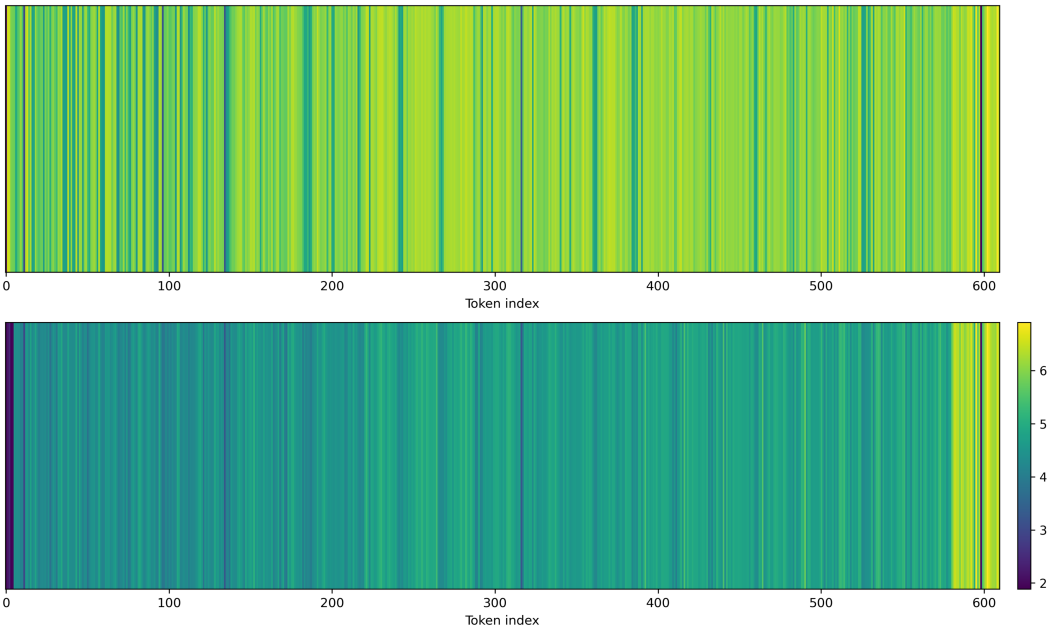

Figure 11: Value of forget V_ϕ from the vanilla model (top) and ours (bottom) at the last layer.

C.2.4 Effect on Attention-Value

We depict the value of V_ϕ at the last and penultimate layers of the vanilla model and our model in Figures 11, 12, 13, and 14. Both models are fine-tuned on top of LLaVA-1.5-7B. The results show that our method substantially suppresses the value activations of wforget samples, while the retain samples remain comparatively stable. This suggests that the method selectively weakens information carried by the value vectors for forgotten knowledge without broadly disrupting retained representations. Moreover, suppressing the value vectors directly limits the information propagated through attention, making recovery through attention redistribution alone more difficult.

C.2.5 Ablation Studies on C_V

Figure 15 shows that reducing V generally improves forgetting. Forget scores decrease monotonically from $V = 1$ to $V = 0$ in five of the six metrics, and $V = 0$ achieves the lowest forget score in all six. In Visual-QA, forget classification and cloze accuracy fall to 31.53 and 5.76. Generation

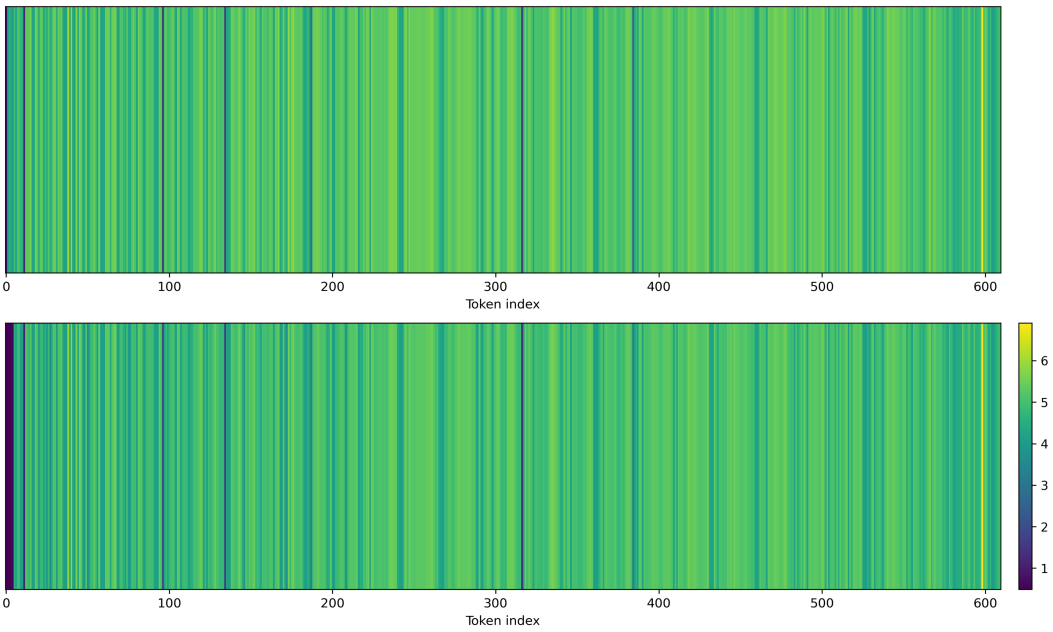

Figure 12: Value of retain V_ϕ from the vanilla model (top) and ours (bottom) at the last layer.

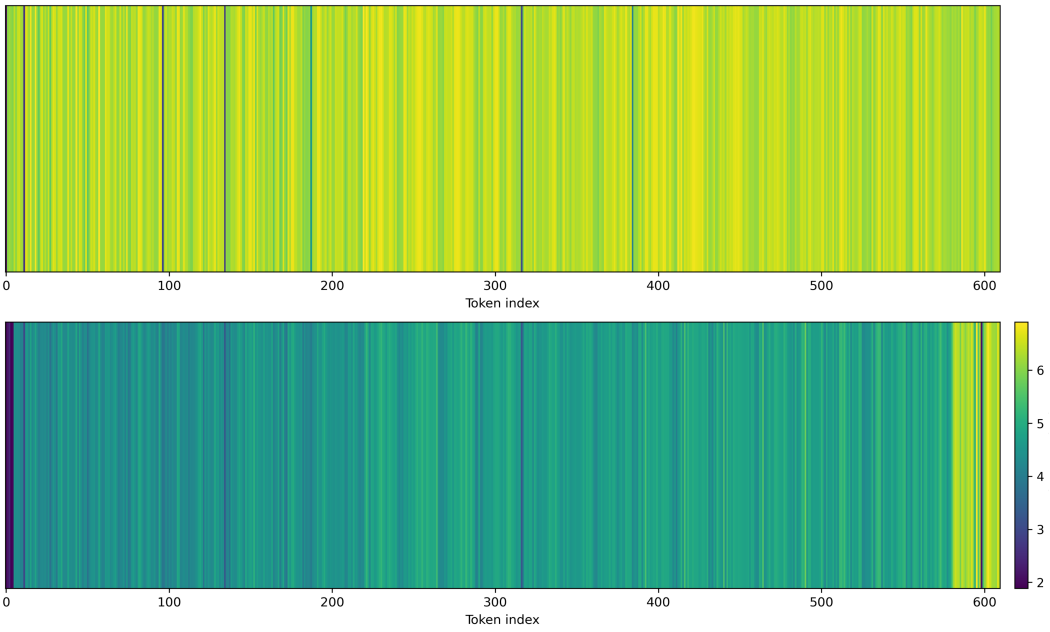

Figure 13: Value of forget V_ϕ from the vanilla model (top) and ours (bottom) at the penultimate layer.

ROUGE-L also reaches its lowest value, 0.437, although it rises slightly at $V = 0.5$. In Textual-QA, $V = 0$ gives forget scores of 35.99, 0.411, and 4.88 for classification, generation, and cloze. At the same time, $V = 0$ achieves the highest retain score in five of six metrics and the highest Real score in all six. The exception is Visual-QA retain classification, which falls to 31.03. These results make $V = 0$ the strongest observed forget–retain trade-off in this ablation.

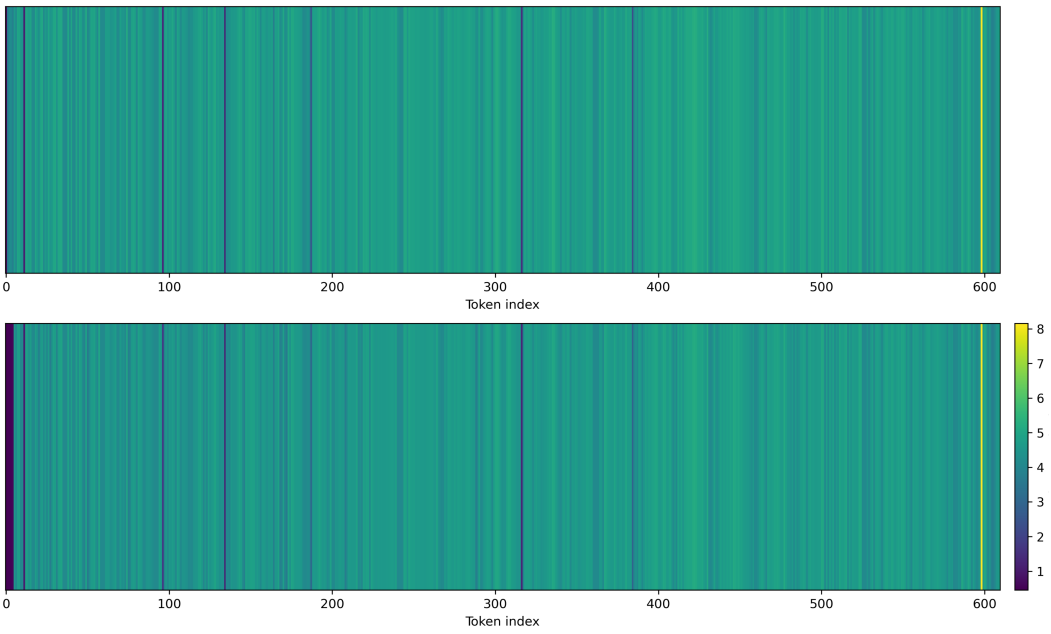

Figure 14: Value of retain V_ϕ from the vanilla model (top) and ours (bottom) at the penultimate layer.

C.2.6 Effect on attention-map

We depict the value of P_ϕ at the last and penultimate layers of the vanilla model and our model in Figures 16, 17. Both models are fine-tuned on top of LLaVA-1.5-7B. Figures 16 and 17 compare the attention probability matrix P_ϕ of our model with the vanilla model at the last and penultimate transformer layers. Across different attention heads, our method largely preserves the original causal attention structure, with highly similar distributions to the vanilla model. Only minor changes in attention intensity are observed within the valid lower-triangular region. This suggests that the proposed unlearning procedure does not substantially disrupt the model's attention routing mechanism. Instead, the model retains a similar pattern of token-to-token interactions after unlearning, indicating that the modification is localized rather than causing a global distortion of the attention distribution.

D More related works

Unlearning in language models. Machine unlearning aims to reduce the influence of designated training information while preserving the model's remaining capabilities (Yao et al., 2024b; Liu et al., 2024b; Ren et al., 2025; Pawelczyk et al., 2024; Yao et al., 2024a). In language models, TOFU introduced a benchmark based on fictitious author profiles and showed that reduced recall of forget data alone does not establish successful unlearning (Maini et al., 2024). A common approach is gradient ascent, which increases the loss on forget examples to discourage their reconstruction. However, unrestricted ascent can lead to substantial degradation of retained behavior. Negative Preference Optimization (NPO) addresses this issue through a reference-relative objective that limits the instability of direct gradient ascent (Zhang et al., 2024). Representation Misdirection for Unlearning (RMU) instead acts on internal activations by steering forget representations toward a fixed direction while matching retain representations to those of a frozen reference model (Li et al., 2024). These works show that effective unlearning requires balancing target suppression with preservation of non-target behavior. SIEVE follows this general objective but focuses on selective multimodal associations and intervenes directly on attention-value representations.

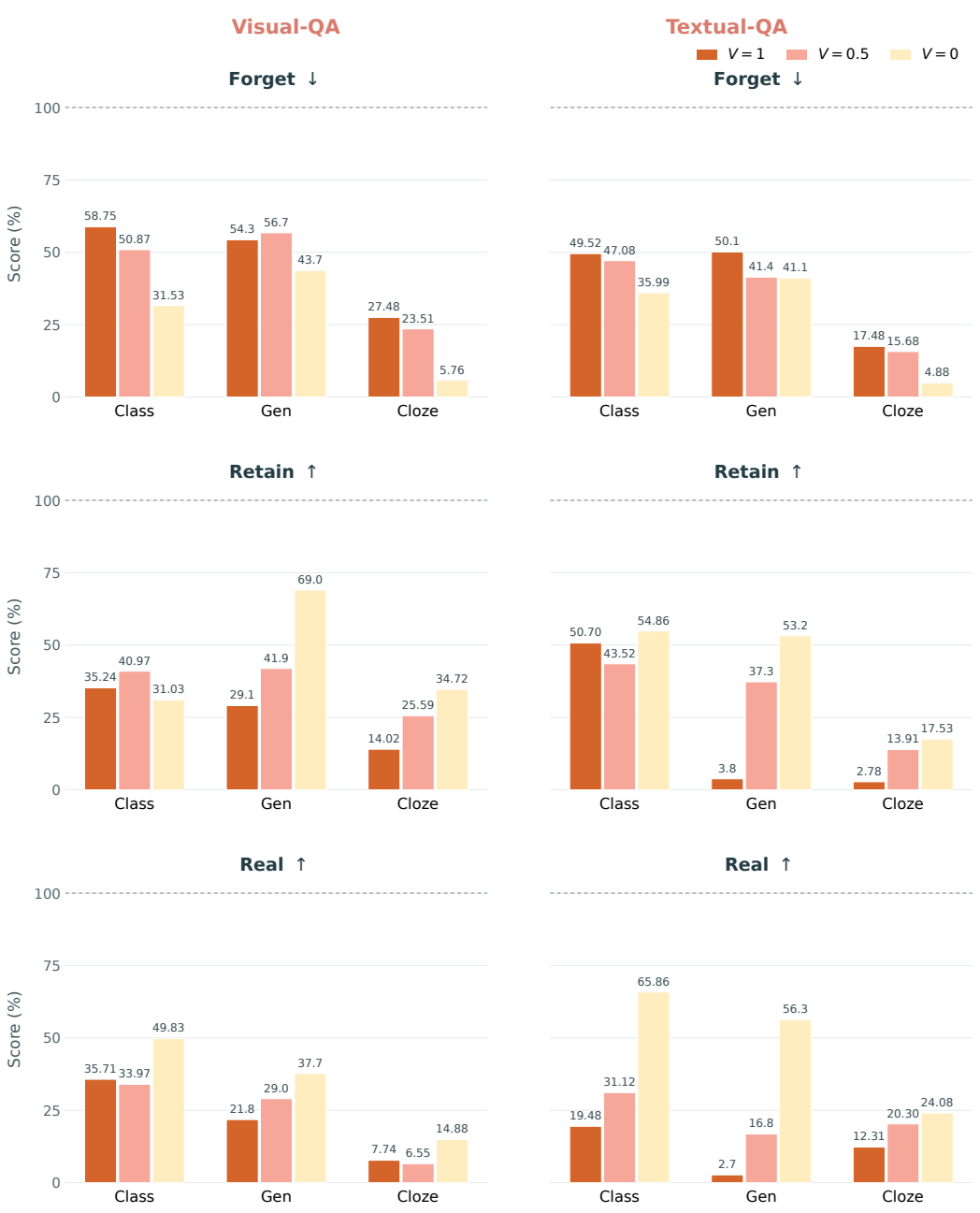

Figure 15: Effect of the attention-value V on LLaVA-1.5-7B. The left and right columns show Visual-QA and Textual-QA results. Each row reports Forget, Retain, and Real performance, with bars for $V \in \{1, 0.5, 0\}$ grouped by classification accuracy (Class), generation ROUGE-L (Gen), and cloze accuracy (Cloze). ROUGE-L scores are multiplied by 100 for the shared axis. Arrows indicate the preferred performance for evaluation.

Model editing in machine unlearning. A growing line of work modifies internal representations rather than relying only on output-level objectives (Hossain & Kagal, 2025; Luo et al., 2026; Guo et al., 2025; Jung et al., 2025). RMU provides an important example by steering forget activations away from their original representations while anchoring retained activations to a frozen reference model (Li et al., 2024). Model-editing methods provide a related perspective. ROME identifies factual associations in feed-forward modules and performs rank-one parameter updates to mod-

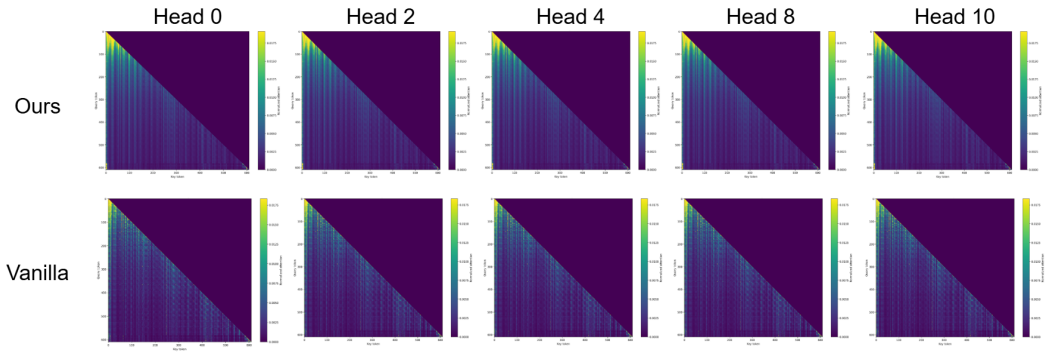

Figure 16: Value of P_ϕ from the vanilla model and ours at the last layer.

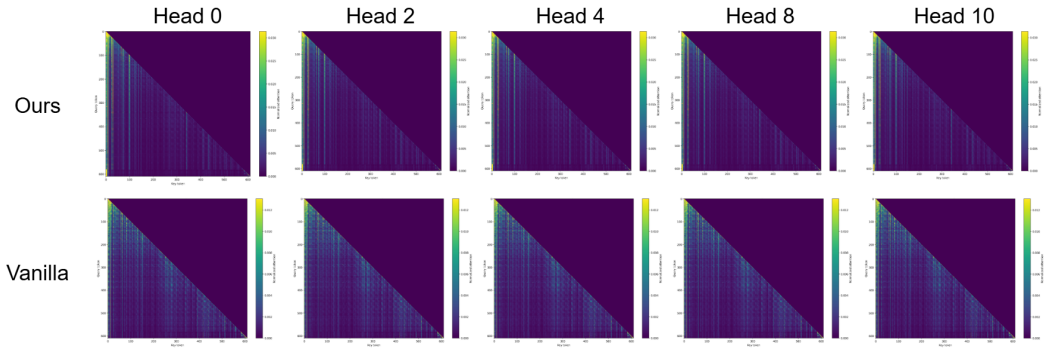

Figure 17: Value of P_ϕ from the vanilla model and ours at the penultimate layer.

ify specific knowledge (Meng et al., 2022), while MIP-Editor extends internal path localization to multimodal unlearning (Li et al., 2026). These approaches demonstrate that targeted changes to intermediate representations or parameters can alter model behavior without modifying the entire network uniformly.