Knowledge Vector Weakening: Efficient Training-free Unlearning for Large Vision-Language Models

Anonymous ACL submission

Abstract

Large Vision–Language Models (LVLMs) are widely adopted for their strong multimodal capabilities, yet they raise serious concerns such as privacy leakage and harmful content generation. Machine unlearning has emerged as a promising solution for removing the influence of specific data from trained models. However, existing approaches largely rely on gradient-based optimization, incurring substantial computational costs for large-scale LVLMs. To address this, we propose Knowledge Vector Weakening (KVW), a training-free unlearning method that directly intervenes in the full model without gradient computation. KVW identifies knowledge vectors activated during the model's output generation on the forget set and progressively weakens their contributions, preventing the model from exploiting undesirable knowledge. Experiments on MLLMU and CLEAR show that KVW achieves a stable forget–retain trade-off while significantly improving computational efficiency over gradient-based and LoRA-based methods.

1 Introduction

Large Vision–Language Models (LVLMs) demonstrate strong performance across a wide range of multimodal tasks, including visual question answering (VQA), image captioning, multimodal retrieval, and document understanding (Liu et al., 2023, 2024a; Xue et al., 2024). Despite these advances, LVLMs raise critical concerns related to privacy leakage, copyright infringement, and the generation of harmful or inappropriate content (Brown et al., 2022; Eldan and Russinovich, 2023). To mitigate these risks, machine unlearning has emerged as a key paradigm that enables models to selectively remove the influence of problematic training data (Nguyen et al., 2022; Liu et al., 2024b).

Exact unlearning is the most straightforward approach to unlearning, which retrains a model from scratch after removing the forget data from the original training set. While this approach provides a clear conceptual solution, it is impractical in large-scale settings due to the prohibitive computational cost. To address this limitation, recent studies has focused on *approximate unlearning*, which aims to remove the influence of forget data from an already trained model without full retraining (Triantafillou et al., 2024). Most existing approaches adopt training-based strategies that compute gradients via backpropagation and update model parameters using optimization objectives (Thudi et al., 2022; Liu et al., 2025a; Zhang et al., 2024; Huo et al., 2025). However, for LVLMs, such methods remain computationally expensive in both time and memory, due to the need for fine-tuning large models (Wang et al., 2024a; Wilma et al., 2026).

To alleviate these challenges, recent unlearning approaches for large models increasingly rely on LoRA (Hu et al., 2022), a parameter-efficient fine-tuning (PeFT) technique. While LoRA-based methods improve computational efficiency (Li et al., 2025; Cha et al., 2025; Kim et al., 2025b), they restrict updates to a low-rank subspace. Prior work (Shuttleworth et al., 2025) shows that, unlike full fine-tuning, LoRA-induced updates are not uniformly distributed across the model but concentrate within a limited low-dimensional subspace, referred to as the intruder dimension. This implies that the effects of LoRA-based optimization are only partially reflected in the model's representational space compared to full fine-tuning.

These findings motivate a closer examination of LoRA-based approaches in the context of unlearning, where the objective is to efficiently remove specific knowledge from a model. When the knowledge to be removed is distributed not only within the intruder dimension but also across the broader representation space, an arbitrarily chosen LoRA rank can be insufficient to remove it effectively (§3). Consequently, in addition to the

cost of gradient-based updates, such approaches expand the hyperparameter search space and incur higher computational costs, diminishing the efficiency benefits originally intended by LoRA.

To address these limitations, we propose Knowledge Vector Weakening (KVW), an efficient training-free unlearning method that directly intervenes in the full model using only forward propagation. Specifically, KVW identifies knowledge vectors in MLP modules accessed during forward passes on the forget set and attenuates their contribution to the output by scaling them down. This reduces the model's reliance on the corresponding knowledge vectors and suppresses the associated knowledge access pathways. KVW does not require gradient computation or fine-tuning, leading to substantial computational savings.

We evaluate our method on representative LVLM unlearning benchmarks, MLLMU-Bench (Liu et al., 2025a) and CLEAR (Dontsov et al., 2025). Results show that KVW achieves more effective unlearning under retain performance constraints than gradient-based and LoRA-based methods, while offering clear advantages in computational efficiency. Our contributions are summarized as follows:

- We introduce *Knowledge Vector Weakening (KVW)*, a gradient-free unlearning framework that achieves selective knowledge removal solely through forward propagation, eliminating the need for backpropagation or retraining.
- We propose a novel intervention strategy that identifies and progressively attenuates forget-related parameter vectors within MLP modules, thereby suppressing the neural pathways associated with the forget set.
- We demonstrate that KVW provides a superior forget–retain trade-off and significant computational savings compared to gradient-based and LoRA-based methods across representative LVLM benchmarks, including MLLMU-Bench and CLEAR.

2 Related Work

2.1 LLM Unlearning

LLM unlearning aims to remove the influence of specific data from large language models without costly retraining, addressing concerns such as privacy, copyright, and hazardous knowledge. Most existing methods rely on loss-function–based optimization, including Gradient Ascent (GA) (Thudi et al., 2022), Gradient Difference (GD) (Liu et al., 2025a), Negative Preference Optimization (NPO) (Zhang et al., 2024), and Inverted Hinge Loss (IHL) (Cha et al., 2025), which induce forgetting by manipulating gradients or token probabilities, though often at the cost of high computation and sensitivity to hyperparameters. Beyond loss-based approaches, alternative methods such as Task Arithmetic (Ilharco et al., 2023), decoding-time unlearning with auxiliary models, and LoRA initialization techniques (Cha et al., 2025; Kim et al., 2025b) have been proposed.

These unlearning techniques for LLMs also form the foundation for subsequent research on unlearning in large vision–language models, as most LVLMs incorporate a pretrained LLM as their core language component.

2.2 LVLM Unlearning

Compared to LLM unlearning, research on LVLM unlearning remains at an early stage. Several benchmarks have recently been introduced to enable systematic evaluation, including MLLMU-Bench (Liu et al., 2025a), CLEAR (Dontsov et al., 2025), and FIUbench (Ma et al., 2025) for privacy-related unlearning, as well as PEBench (Xu et al., 2025) for event-scene unlearning.

From a methodological perspective, a small number of recent works explicitly incorporate visual information: MANU (Liu et al., 2025b) identifies important neurons via activation differences between multimodal and text-only inputs and prunes them, MMUnlearner (Huo et al., 2025) exploits gradient-based saliency induced by image-token ablation, and VKD (Wang et al., 2025) employs a teacher model to preserve intermediate visual representations on the retain set. However, these approaches typically require additional inference, fine-tuning process, or auxiliary models, resulting in substantial computational overhead. In contrast, our method achieves LVLM unlearning in a forward-only manner with significantly lower computational overhead.

3 Motivation

Most existing unlearning methods rely on training-based approaches that compute gradients via backpropagation and directly update model parameters using loss functions (Liu et al., 2025a; Zhang et al., 2024; Huo et al., 2025). In large-scale vision–language model settings, such approaches in-

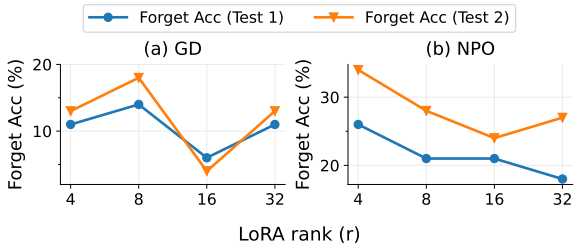

Figure 1: Forget accuracy (%) on CLEAR forget05 for GD and NPO under different ranks r. Evaluation follows a 2-fold validation protocol with hyperparameters selected on a cross-validation split and tested on a held-out split to avoid overestimation from fixed settings. The figure reveals strong rank sensitivity.

cur substantial computational overhead in both time and memory, which is further amplified by repeated training runs required for hyperparameter tuning.

To improve efficiency, several unlearning approaches adopt LoRA, a parameter-efficient fine-tuning (PeFT) technique. While these methods reduce parameter update costs, they introduce additional hyperparameters, expanding the search space and undermining LoRA's efficiency-oriented motivation. In particular, the LoRA rank determines the subspace in which the model is modified. In unlearning settings, performance becomes highly sensitive to how well this subspace aligns with the knowledge to be removed, as shown in Figure 1. We conduct experiments on the CLEAR benchmark (Dontsov et al., 2025) under a VQA setting using 2-fold validation, motivated by (Cha and Cho, 2025). When applying GD (Liu et al., 2025a), forget accuracy varies substantially across LoRA ranks, even under the same search budget. On Test 2, GD achieves its best forget accuracy of 4% at rank 16, but degrades to 18% at a lower rank, resulting in a gap of more than 4.5×. Similar trends appear on Test 1 and for NPO.

This suggests that the effectiveness of LoRA-based unlearning critically depends on careful configuration, which inevitably introduces additional overhead beyond the parameter updates themselves. As a consequence, the need for extensive tuning and repeated training runs diminishes the efficiency gains that LoRA was originally designed to provide. In realistic unlearning scenarios, particularly when applied to large pre-trained models, such overhead becomes non-negligible and can offset the benefits of parameter-efficient adaptation. These observations indicate that, despite their efficiency advantages in fine-tuning, LoRA-based approaches have inherent limitations as a fundamental solution to efficient unlearning.

4 Method

To address the high computational cost of existing gradient-based unlearning methods, we propose **Knowledge Vector Weakening (KVW)**. KVW performs unlearning without backpropagation or fine-tuning, by progressively weakening the contributions of internal knowledge vectors involved in generating the model's outputs on the forget set during forward propagation.

4.1 Preliminary

Knowledge in transformer models. Recent studies have shown that a substantial portion of factual and conceptual knowledge in transformer-based language models is stored within the *Feed-Forward Network (FFN)* (Geva et al., 2021, 2022; Dai et al., 2022a; Meng et al., 2022; Dong et al., 2022; Kim et al., 2025a). In particular, Geva et al. (Geva et al., 2021) demonstrate that the FFN can be interpreted as a *key–value memory structure*, where the FFN consists of two projection matrices and a non-linear activation function in between. For an input representation $x \in \mathbb{R}^d$, the FFN can be written as

$$\mathrm{FFN}(x) = f(xK^\top)\,V, \tag{1}$$

where $K \in \mathbb{R}^{m\times d}$ denotes the key matrix, $V \in \mathbb{R}^{m\times d}$ denotes the value matrix, and $f(\cdot)$ is an element-wise non-linear activation function. Given an input x, the pre-activation values $xK^\top$ reflect the activation strength of each key and are transformed by $f(\cdot)$ into coefficients associated with the corresponding values.

4.2 The Proposed Method

Knowledge coefficient. Motivated by (Geva et al., 2021), we refer to each element of these coefficients $f(xK^\top)$ in Equation 1 as a *knowledge coefficient* $\mathcal{C} \in \mathbb{R}^m$, which quantifies the contribution of each value vector to the FFN output for a given input. The vectors in the value matrix V represent the stored knowledge units of the FFN.

Specifically, we define each *row vector* $\mathbf{v}_i \in \mathbb{R}^d$ of V $(i = 1, \ldots, m)$ as a fundamental unit of knowledge, referred to as a *knowledge vector*. The output of the FFN can be expressed as a weighted

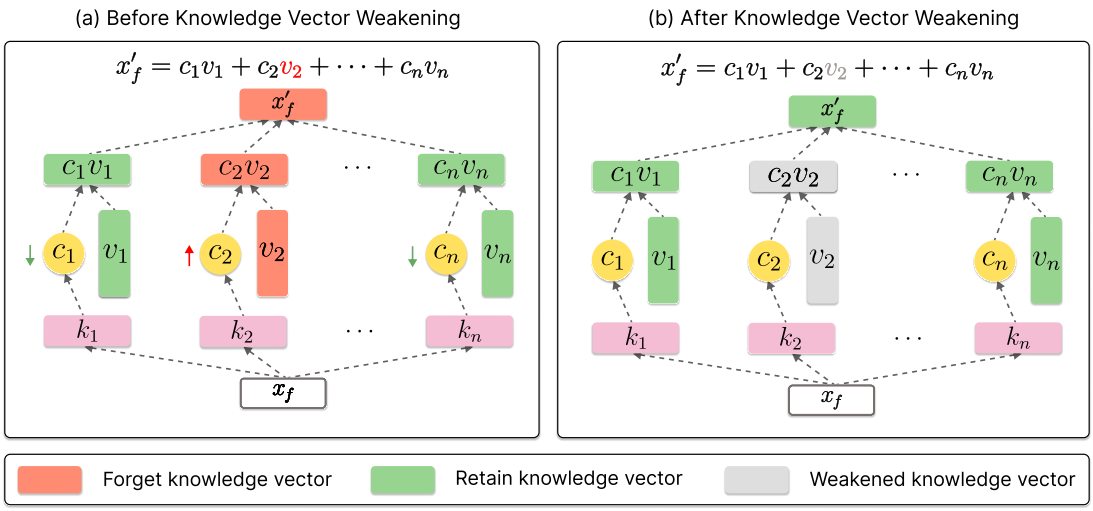

Figure 2: **Knowledge Vector Weakening.** In this figure, x_f and x'_f denote the hidden representations of the forget input and the next layer output, respectively; v_i, c_i, and k_i denote knowledge vectors, their corresponding contribution coefficients, and the associated keys. (a) Before applying KVW, forget-related knowledge vectors contribute heavily to the hidden state of the next layer, forming a dominant pathway that drives output generation. (b) After applying KVW, this dominant pathway is effectively weakened by selectively scaling down forget-related knowledge vectors.

sum of these knowledge vectors using the corresponding knowledge coefficients:

$$\mathrm{FFN}(x) = \sum_{i=1}^{m} \mathcal{C}_i(x)\, \mathbf{v}_i, \tag{2}$$

where $\mathcal{C}_i(x)$ denoting its i-th scalar component. This formulation indicates that different knowledge vectors are selectively activated and composed depending on the input. From this perspective, the FFN can be understood as a *parametric memory module* that retrieves and composes internal knowledge in an input-dependent manner.

Identifying forget-related knowledge vectors. To identify forget-related knowledge vectors, we perform a forward pass on the forget set and extract the corresponding knowledge coefficients from the model's MLP modules. To focus on knowledge accessed during answer generation, we compute these coefficients only at time steps corresponding to answer tokens. Let $\mathcal{T}_{\text{ans}}$ denote the set of answer token indices. Then, the knowledge coefficient for the forget set is defined as

$$\mathcal{C}_f \;=\; \frac{1}{|\mathcal{T}_{\text{ans}}|} \sum_{t \in \mathcal{T}_{\text{ans}}} f\left(x_t K^\top\right) \;\in\; \mathbb{R}^m, \tag{3}$$

where x_t denotes the hidden representation at time step t. The final coefficient is obtained by averaging over all answer tokens. Knowledge vectors with larger values in $\mathcal{C}_f$ contribute more substantially to constructing the hidden states of the subsequent layer for the forget data. Accordingly, $\mathcal{C}_f$ can be interpreted as a relevance score of each knowledge vector with respect to the forget data.

However, activations of knowledge vectors on the forget set may also arise from representations that are essential for general language understanding. Such representations are often similarly activated by the retain set, making it difficult to reliably identify the knowledge to be removed based solely on forget-set activations. To address this issue, we explicitly contrast the activations observed on the forget set with those on the retain set, thereby distinguishing knowledge that should be removed from knowledge that should be preserved. Using the same procedure as in Eq. (3), we compute the knowledge coefficients for the retain set and denote them as $\mathcal{C}_r \in \mathbb{R}^m$.

We then define the *Forget Knowledge Accessor (FKA)* based on the relative magnitude of $\mathcal{C}_f$ and $\mathcal{C}_r$. To enable stable comparison and prevent negative values, we adopt a clipped ratio and define FKA, denoted by $\mathcal{A}$, as

$$\mathcal{A} \;=\; \max\left(0,\, \frac{\mathcal{C}_f}{\mathcal{C}_r}\right) \;\in\; \mathbb{R}^m, \tag{4}$$

where the maximum operation is applied element-wise. Knowledge vectors with high $\mathcal{A}$ values are strongly activated for the forget set while exhibiting

relatively low activation for the retain set, and are thus identified as forget-related knowledge vectors. Consequently, $\mathcal{A}$ serves as a quantitative criterion for identifying the knowledge vectors to be targeted for unlearning in the subsequent stage.

Knowledge vector weakening. In this stage, we weaken the forget-related knowledge vectors identified in the previous stage. The core idea of *Knowledge Vector Weakening (KVW)* is to reduce the contribution of each knowledge vector in proportion to its accessor value $\mathcal{A}$, which reflects how strongly the vector contributes to the model's outputs on the forget set. Specifically, knowledge vectors with larger $\mathcal{A}$ values are weakened more substantially, while those with smaller $\mathcal{A}$ values are weakened to a lesser extent or remain unchanged. This enables the selective removal of knowledge that is strongly associated with the forget set.

To quantitatively implement this weakening process, we define a *gate function* $g(\cdot)$ that takes the accessor value $\mathcal{A}$ as input and outputs a non-negative scaling coefficient for each knowledge vector. The gate function is defined as

$$g(\mathcal{A}) = \mathcal{A}^{-\gamma}, \tag{5}$$

where γ is a hyperparameter that controls the strength of weakening. Each knowledge vector $\mathbf{v}_i$ is then scaled by the corresponding gate coefficient $g(\mathcal{A}_i)$ as

$$\tilde{\mathbf{v}}_i = g(\mathcal{A}_i) \cdot \mathbf{v}_i, \tag{6}$$

which directly reduces its information contribution to the model output. During inference, these modified vectors $\tilde{\mathbf{v}}_i$ are integrated into the feed-forward computation, resulting in a weakened output: $\tilde{\mathrm{FFN}}(x) = \sum_{i=1}^{m} \mathcal{C}_i(x)\tilde{\mathbf{v}}_i$. By substituting $\tilde{\mathbf{v}}_i$ into the summation, the model effectively suppresses the specific neural pathways associated with the target knowledge. As illustrated in Figure 2, repeatedly applying this weakening process progressively blocks the pathways that triggered undesirable responses, enabling selective and efficient unlearning.

Implementation Details. Due to the architectural characteristics of large language models, early layers play a critical role in integrating visual and language embeddings, while later layers are primarily responsible for forming the final output probabilities. Taking these properties into account, we introduce the start layer and end layer as hyperparameters to ensure stable application of KVW. The optimal values for these hyperparameters are selected on the validation set and used for all experiments. We refer readers to Appendix B for additional implementation details, and to Appendix A for the complete procedure of the algorithm.

5 Experiments

5.1 Settings

Datasets. To evaluate multimodal machine unlearning performance, we use two benchmarks: MLLMU-Bench (Liu et al., 2025a) and CLEAR (Dontsov et al., 2025). MLLMU-Bench consists of 500 synthetic individual profiles and 153 public celebrity profiles, where each profile is associated with 14 customized question–answer pairs. The benchmark supports evaluation under both multimodal inputs and unimodal inputs (text-only) settings. In this work, we report performance on the Forget, Retain, and Real-world splits of MLLMU-Bench using ROUGE (Lin, 2004) scores.

CLEAR is a benchmark built on TOFU (Maini et al., 2024), a synthetic author-profile dataset originally designed for LLM unlearning. It comprises 200 synthetic individuals, about 3,700 associated images, and corresponding question–answer pairs. We assess performance on the Forget, Retain, RealFace, and RealWorld splits using accuracy, defined by whether the model's VQA responses contain the key information of the target knowledge.

Baselines. We compare our proposed method with six existing unlearning methods. GA (Thudi et al., 2022) applies gradient ascent on the forget VQA set $\mathcal{D}_f$ to remove target knowledge. GD (Liu et al., 2022) balances forgetting and retention by jointly optimizing losses on $\mathcal{D}_f$ and the retain set $\mathcal{D}_r$. KL (Nguyen et al., 2020) aligns predictions on $\mathcal{D}_r$ with those of the original model while applying GA loss on $\mathcal{D}_f$. NPO (Zhang et al., 2024) treats $\mathcal{D}_f$ as negative-preference data and adopts a preference optimization framework using an oracle model trained on $\mathcal{D}_r$. MMU (Huo et al., 2025) selectively removes visual knowledge while protecting non-target parameters via saliency-weighted updates. MANU (Liu et al., 2025b) identifies important neurons via activation differences between multimodal and text-only inputs and prunes them.

Evaluation protocol. Recent study (Cho et al., 2025) points out that unlearning evaluation for generative models can overestimate unlearning performance when it relies excessively on specific reference answers or fixed evaluation settings. To

Table 1: Unlearning performance under different forget ratios on Test 1 and Test 2 on MLLMU-Bench using LLaVA-1.5-7B. Cells highlighted in red indicate results that fail to satisfy the retain constraint. (≃) indicates that closer performance to the oracle model is preferred. Bold and underline denote the best and second-best results according to this criterion.

		Forget05			Forget10			Forget15		
Test	Method	Forget (≃)	Retain	Celeb	Forget (≃)	Retain	Celeb	Forget (≃)	Retain	Celeb
Test 1	Vanilla	0.650	0.614	0.417	0.645	0.615	0.417	0.609	0.633	0.417
	Oracle	0.556	0.619	0.412	0.525	0.594	0.416	0.502	0.610	0.424
	GA	0.566	0.561	0.372	0.609	0.589	0.399	0.584	0.623	0.401
	GD	0.606	0.603	0.414	0.601	0.586	0.389	0.582	0.633	0.410
	KL	<u>0.565</u>	0.592	0.399	0.640	0.611	0.415	0.606	0.631	0.417
	NPO	0.604	0.591	0.412	<u>0.608</u>	0.586	0.398	**0.560**	0.606	0.406
	MMU*	0.637	0.627	0.413	0.654	0.607	0.409	0.583	0.635	0.408
	MANU	0.622	0.603	0.403	0.608	0.587	0.416	<u>0.575</u>	0.622	0.401
	KVW	**0.561**	0.587	0.397	**0.602**	0.588	0.409	<u>0.575</u>	0.604	0.401
Test 2	Vanilla	0.611	0.621	0.389	0.615	0.619	0.389	0.622	0.604	0.389
	Oracle	0.567	0.610	0.410	0.553	0.617	0.408	0.543	0.582	0.409
	GA	0.575	0.599	0.387	<u>0.584</u>	0.600	0.381	0.613	0.594	0.388
	GD	**0.567**	0.612	0.395	0.604	0.623	0.387	0.591	0.600	0.403
	KL	0.564	0.601	0.386	0.628	0.621	0.386	0.618	0.603	0.389
	NPO	0.538	0.597	0.388	0.604	0.585	0.398	<u>0.578</u>	0.578	0.394
	MMU*	0.581	0.618	0.401	0.604	0.620	0.397	0.616	0.608	0.403
	MANU	0.595	0.610	0.380	0.630	0.594	0.387	0.608	0.581	0.385
	KVW	**0.567**	0.595	0.381	**0.576**	0.589	0.375	**0.562**	0.580	0.380

mitigate this issue and assess the generalization ability of each unlearning method, we adopt a 2-fold cross-validation protocol, motivated by (Cha and Cho, 2025). Specifically, we partition the evaluation data into two disjoint splits, denoted as Test 1 and Test 2. In the first fold, hyperparameter configurations are selected based on their performance on Test 1 and evaluated on Test 2; in the second fold, the roles of the two splits are reversed. Results from both folds are reported. Following the evaluation protocol of (Ilharco et al., 2023; Kim et al., 2025b), we select hyperparameters under a retain constraint: among all configurations that preserve at least 95% of the retain performance of the original vanilla model, we choose the one that achieves the strongest forgetting performance.

5.2 Main Results

Results on MLLMU-Bench. Table 1 shows the unlearning performance on MLLMU-Bench using LLaVA-1.5-7B (Liu et al., 2024a). KVW achieves forget ROUGE scores most closely aligned with the oracle while maintaining strong retain performance across nearly all splits, demonstrating its ability to effectively remove the target knowledge.

In contrast, gradient-based methods (GA, GD, KL, NPO) show substantial variability across splits. NPO forgets well on Forget15 but fails to maintain a balanced outcome elsewhere, while GA and GD fail to preserve retain performance in certain settings. KL reduces the forget score insufficiently on most splits except Forget05, indicating ineffective unlearning. MANU performs reasonably on Forget15 (Test 1) but underperforms elsewhere, suggesting that its forget-data-driven pruning does not generalize across forget ratios.

We additionally observe that MMU exhibits unstable training behavior and yields zero ROUGE scores across all splits, likely due to full-model fine-tuning on a 7B model with limited unlearning data. We therefore adopt a LoRA-based variant, MMU*, whose implementation details are provided in Appendix D. Even so, MMU* fails to achieve effective unlearning, as it does not sufficiently reduce the forget score while preserving retain performance.

Results on CLEAR. Table 2 presents the unlearning performance on the CLEAR benchmark using Qwen2-VL-2B (Wang et al., 2024b). While MLLMU-Bench evaluates unlearning via ROUGE over full responses, CLEAR formulates it as a short-answer VQA task with accuracy-based evaluation. This setting more precisely distinguishes ROUGE reductions caused by meaningless responses from those resulting from the successful removal of target information. For brevity, we report Test 1 results here, and provide the full results in Table 4,

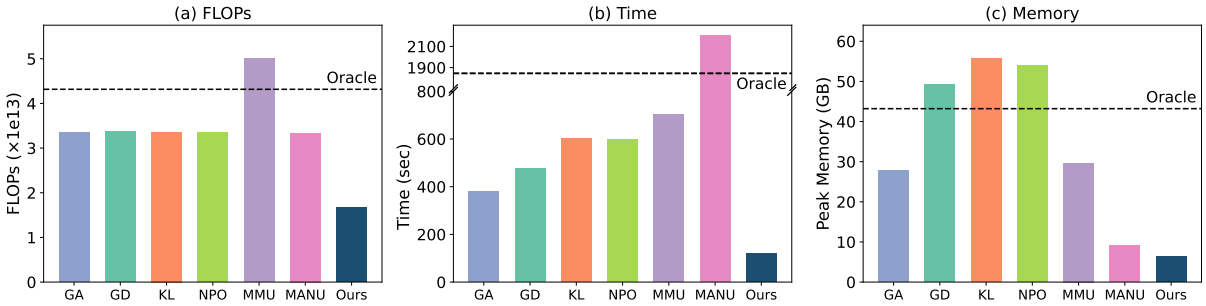

Figure 3: Comparison of computational cost across unlearning methods in terms of FLOPs, time, and memory usage. FLOPs denote the number of floating-point operations required to process a single batch. Time measures the total wall-clock time for the unlearning process over all batches. Memory indicates the maximum VRAM consumption during unlearning.

Table 2: Accuracy results on CLEAR (Qwen2-VL-2B) under different forget ratios. RealF and RealW denote accuracy on the RealFace and RealWorld datasets, respectively. Red cells indicate results that fail to satisfy the retain constraint.

Method	Forget	Retain	RealF	RealW
Forget 5				
Vanilla	0.41	0.43	0.93	0.50
Oracle	0.00	0.54	0.95	0.57
GA	0.09	0.34	0.91	0.45
GD	0.14	0.45	0.95	0.51
KL	0.06	0.47	0.93	0.49
NPO	0.21	0.43	0.95	0.50
MMU	0.29	0.47	0.93	0.49
MANU	0.02	0.54	0.76	0.52
KVW	**0.00**	0.48	0.93	0.55
Forget 10				
Vanilla	0.42	0.42	0.92	0.49
Oracle	0.00	0.64	0.91	0.42
GA	0.00	0.00	0.00	0.00
GD	**0.01**	0.41	0.94	0.49
KL	0.04	0.44	0.93	0.51
NPO	0.12	0.44	0.93	0.48
MMU	0.13	0.52	0.93	0.50
MANU	0.08	0.54	0.86	0.47
KVW	**0.01**	0.43	0.92	0.48

where Test 2 shows similar trends.

On CLEAR, KVW consistently achieves unlearning performance that matches or closely approaches the oracle, effectively removing key information associated with the forget set while preserving performance on existing knowledge. This supports the design principle of KVW, which selectively weakens the knowledge vectors driving forget-set generation, thereby blocking the corresponding generation pathways while preserving those required for the retain set.

In contrast, GA again fails to satisfy the retain constraint, as on MLLMU-Bench. GD, KL, NPO, and MMU fail to reduce forget accuracy to the oracle level, indicating that key information associated with the forget set remains partially preserved, suggesting persistent knowledge leakage during generation. MANU achieves low forget accuracy but causes severe degradation on RealF (0.76 on Forget05, 0.86 on Forget10), indicating that pruning the last-N layers based on forget-specific data damages unrelated real-face knowledge.

5.3 Computational Efficiency

We evaluate the practical cost efficiency of different unlearning approaches under the Forget05 setting on the CLEAR benchmark. Figure 3 reports FLOPs, runtime, and memory usage for the oracle, baseline methods, and the proposed approach. The oracle is retrained from scratch with rank 16, and GA, GD, KL, and NPO use rank 8.

The results show that KVW is substantially more efficient than existing approaches across all cost metrics. As shown in Figure 3(a), while LoRA-based methods reduce backpropagation cost, they still require additional forward passes involving LoRA adapters and gradient-based updates. MANU, although training-free and therefore memory-efficient (Figure 3(c)), requires four forward passes per iteration (forget/retain $\times$ unimodal/multimodal) and repeated I/O for activation caching, resulting in even longer runtime than LoRA-based training methods (Figure 3(b)). In contrast, KVW introduces no additional parameters and avoids backpropagation, leading to significantly lower FLOPs and shorter runtime. Since gradients are not stored, its memory usage is comparable to standard inference. This enables efficient deployment even under limited GPU resources.

Table 3: Ablation study on the *ans only* and *w/ retain* options. The *ans only* option uses only answer tokens to compute knowledge contributions, while *w/ retain* incorporates retain set information. Forget1/Forget2 and RealF/RealW denote accuracy on the Forget Test 1/Test 2 and RealFace/RealWorld datasets, respectively.

Ans only	w/ retain	Forget1	Forget2	Retain	RealF	RealW
–	✓	0.15	0.16	0.27	0.94	0.53
✓	–	0.00	0.00	0.00	0.96	0.49
✓	✓	0.00	0.00	0.48	0.93	0.55

6 Analysis

6.1 Ablation Study

We conduct an ablation study under the Forget05 setting of CLEAR to analyze key design choices. *Ans only* indicates whether the knowledge coefficient $\mathcal{C}$ is computed only from tokens that directly contribute to answer generation. When disabled, $\mathcal{C}$ is averaged over image tokens, question prompts, and answer tokens. *w/ retain* denotes whether $\mathcal{C}_r$ from the retain set is incorporated.

As shown in Table 3, disabling *Ans only* degrades both retain performance and forgetting effectiveness. This is because computing $\mathcal{C}$ over all input tokens leads to weakening knowledge vectors unrelated to the forget target, making it difficult to selectively remove key information. Meanwhile, disabling *w/ retain* leads to near zero forget accuracy but significantly harms retain performance, resulting in imbalanced forget–retain trade-off. In contrast, applying both design choices enables selective weakening of key information directly involved in answer generation, achieving effective unlearning while preserving retain performance.

6.2 Qualitative Analysis

To validate that targeting FFN value vectors v_i effectively handles multimodal concepts, we visualized attention maps using CLIP (ViT-L/14), which serves as the vision encoder of LLaVA-1.5, with Grad-CAM (Selvaraju et al., 2019). As shown in Figure 4, when unlearning *Elon Musk*, the attenuation gate (γ) progressively eliminates attention exclusively on his face, while *Kim Kardashian* remains unaffected (and vice versa). This confirms that our extracted forward-based knowledge vectors (v_i) are not only semantically aligned with the targeted forget data, but also disentangled from other concepts, enabling selective unlearning without unintended interference. The full set of examples is provided in Figure 10.

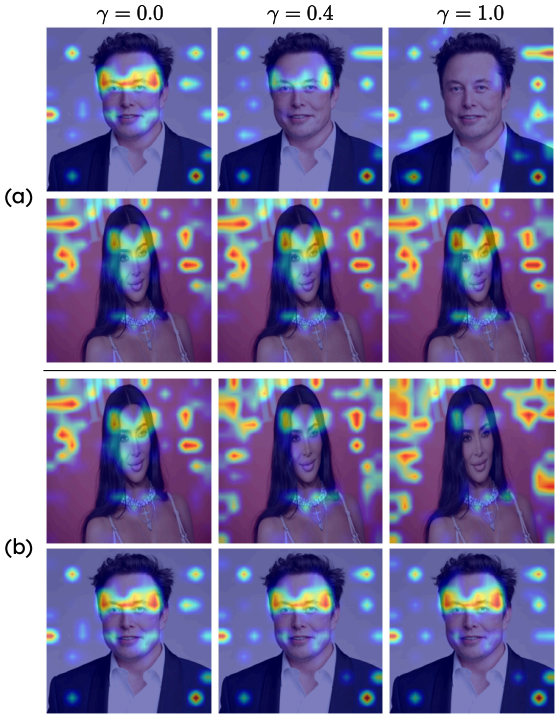

Figure 4: Grad-CAM attention maps from CLIP (ViT-L/14) as the attenuation gate γ varies from 0.0 to 1.0. **(a)** Unlearning *Elon Musk* progressively removes attention on his face (top) while *Kim Kardashian* (bottom) remains unchanged. **(b)** The reverse holds when unlearning *Kim Kardashian*. This shows that our forward-based knowledge vectors v_i are semantically aligned with the target concept and do not affect non-target identities.

7 Conclusion

This work challenges the assumption that effective unlearning in large vision–language models requires gradient-based optimization. LoRA-based methods are highly sensitive to rank choice, which undermines their efficiency advantage. Motivated by this, we revisit unlearning by directly intervening in the model's internal knowledge representations. Building on prior findings that FFN value vectors store factual knowledge, we identify a small set of such vectors that drive forget-set generation through forward propagation alone. We then propose Knowledge Vector Weakening (KVW) to selectively attenuate them. KVW achieves a stable forget–retain trade-off with substantially lower computational cost. Our qualitative analyses further confirm that the targeted vectors are semantically disentangled from non-target concepts, enabling precise unlearning. More broadly, our findings point to forward-pass-based model editing as a scalable alternative to gradient-based intervention.

Limitations

KVW performs unlearning by weakening knowledge vectors at an individual level and therefore does not explicitly account for interactions among knowledge vectors or compositional knowledge representations. Future work may explore more structured weakening strategies that incorporate such relationships, as well as lightweight post-weakening adaptation to further improve the model's overall representational capacity.

Potential Risks

While the proposed method, KVW, aims to support responsible AI deployment by enabling the selective removal of harmful or sensitive knowledge from large language and multimodal models, we acknowledge several potential risks that warrant careful consideration.

First, as with other unlearning techniques, unlearning mechanisms could potentially be misused to obscure accountability or remove information inappropriately, highlighting the importance of careful governance and responsible use. For instance, malicious actors could attempt to erase evidence of model behaviors, training data provenance, or factual knowledge that is socially important, in ways that undermine transparency and auditability.

Second, although our experiments show that KVW preserves retain performance well, no unlearning method can fully guarantee the absence of residual effects on unrelated knowledge. In safety-critical or high-stakes domains (e.g., medical or legal applications), even small unintended changes could have meaningful consequences, and additional verification beyond standard benchmarks would be required before deployment.

We believe these risks are common to the broader class of unlearning methods and are best addressed through a combination of technical safeguards, clear documentation of unlearning operations, and policy-level governance regarding when and how unlearning may be applied.

The License For Artifacts

We use two open-source benchmarks, CLEAR (Dontsov et al., 2025) and MLLMU-Bench (Liu et al., 2025a), both publicly released for research on multimodal machine unlearning, consistent with our usage. For models, LLaVA-1.5-7B is released under the Meta Llama Community License, and Qwen2-VL-2B is released under the Apache License 2.0. Our use of all artifacts complies with their respective licenses.

Use of AI Assistants

We used AI assistants (e.g., ChatGPT, Gemini, Claude) in a supportive role throughout this work. Specifically, AI assistants were used to aid research discussions, assist in implementing algorithms and analysis experiments, and support the writing process. All AI-assisted outputs were carefully reviewed and verified by the authors to ensure correctness and originality.

References

Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. 2022. What does it mean for a language model to preserve privacy? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 2280–2292, New York, NY, USA. Association for Computing Machinery.

Sungmin Cha and Kyunghyun Cho. 2025. Hyperparameters in continual learning: A reality check. *Transactions on Machine Learning Research*.

Sungmin Cha, Sungjun Cho, Dasol Hwang, and Moontae Lee. 2025. Towards robust and parameter-efficient knowledge unlearning for llms. In *International Conference on Learning Representations*.

Sungjun Cho, Dasol Hwang, Frederic Sala, Sangheum Hwang, Kyunghyun Cho, and Sungmin Cha. 2025. Reference-Specific Unlearning Metrics Can Hide the Truth: A Reality Check. *Preprint*, arXiv:2510.12981.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022a. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022b. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Qingxiu Dong, Damai Dai, Yifan Song, Jingjing Xu, Zhifang Sui, and Lei Li. 2022. Calibrating factual knowledge in pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5937–5947, Abu Dhabi, United

Arab Emirates. Association for Computational Linguistics.

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

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

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. Alphaedit: Null-space constrained model editing for language models. In *The Thirteenth International Conference on Learning Representations*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The International Conference on Learning Representations*.

Jiyeon Kim, Hyunji Lee, Hyowon Cho, Joel Jang, Hyeonbin Hwang, Seungpil Won, Youbin Ahn, Dohaeng Lee, and Minjoon Seo. 2025a. Knowledge entropy decay during language model pretraining hinders new knowledge acquisition. In *The Thirteenth International Conference on Learning Representations*.

Yejin Kim, Eunwon Kim, Buru Chang, and Junsuk Choe. 2025b. Improving fisher information estimation and efficiency for LoRA-based LLM unlearning. In *Second Conference on Language Modeling*.

Delong Li, Guangsheng Yu, Xu Wang, Yanna Jiang, Wencheng Yang, Bin Liang, and Wei Ni. 2025. A survey of lora-based machine unlearning for llms: Methods, taxonomy, and evaluation.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Zheyuan Liu, Guangyao Dou, Eli Chien, Chunhui Zhang, Yijun Tian, and Ziwei Zhu. 2024b. Breaking the trilemma of privacy, utility, efficiency via controllable machine unlearning. In *The Web Conference 2024*.

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

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

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

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

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. 2020. Variational bayesian unlearning. In *Advances in Neural Information Processing Systems*, volume 33, pages 16025–16036. Curran Associates, Inc.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi-Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16:1 – 46.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Reece S Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. 2025. LoRA vs full fine-tuning: An illusion of equivalence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Eleni Triantafillou, Peter Kairouz, Fabian Pedregosa, Jamie Hayes, Meghdad Kurmanji, Kairan Zhao, Vincent Dumoulin, Julio Jacques Junior, Ioannis Mitliagkas, Jun Wan, and 1 others. 2024. Are we making progress in unlearning? findings from the first neurips unlearning competition. *arXiv preprint arXiv:2406.09073*.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. 2024a. Q-VLM: Post-training quantization for large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

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

Joana Wilma, Marco Serrano, Eleni Papadopoulou, Tomasz Kwiatkowski, and Farah Al-Mansouri. 2026. Principles and methods for building efficient large vision language models.

Zhaopan Xu, Pengfei Zhou, Weidong Tang, Jiaxin Ai, Wangbo Zhao, Kai Wang, Xiaojiang Peng, Wenqi Shao, Hongxun Yao, and Kaipeng Zhang. 2025. Pebench: A fictitious dataset to benchmark machine unlearning for multimodal large language models. *Preprint*, arXiv:2503.12545.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Awalgaonkar, Shelby Heinecke, and Ran Xu. 2024. xgen-mm (blip-3): A family of open large multimodal models.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*.

A Algorithm

Algorithm 1 Knowledge Vector Weakening

Input: Pretrained LVLM $\mathcal{M}$, forget dataset $\mathcal{D}_f$, retain dataset $\mathcal{D}_r$, weakening strength γ, start layer l_s, end layer l_e
Output: Unlearned model $\tilde{\mathcal{M}}$

// Precompute retain knowledge coefficients
Initialize $\mathcal{C}_r \leftarrow \mathbf{0}$
for each batch in $\mathcal{D}_r$ **do**
  Perform forward propagation
  Extract knowledge coefficients at answer tokens
  Accumulate $\mathcal{C}_r$
end for
$\mathcal{C}_r \leftarrow \mathcal{C}_r / |\mathcal{D}_r|$

// Knowledge Vector Weakening
for each batch in $\mathcal{D}_f$ **do**
  Perform forward propagation
  Extract knowledge coefficients $\mathcal{C}_f$ at answer tokens
  Compute Forget Knowledge Accessor:
    $\mathcal{A} \leftarrow \max\left(0, \frac{\mathcal{C}_f}{\mathcal{C}_r}\right)$
  for layer $l = l_s$ **to** l_e **do**
    for each knowledge vector $\mathbf{v}_i$ in layer l **do**
      Compute gate $g_i \leftarrow \mathcal{A}_i^{-\gamma}$
      Update knowledge vector $\mathbf{v}_i \leftarrow g_i \cdot \mathbf{v}_i$
    end for
  end for
end for
return $\tilde{\mathcal{M}}$

B Implementation Details

During implementation, we compute $\mathcal{C}_r$ as an average over the entire retain dataset. Specifically, prior to performing the unlearning process, we run one epoch of forward propagation on the full retain set to compute the mean $\mathcal{C}_r$, which is then stored. During unlearning, this precomputed average $\mathcal{C}_r$ is loaded and used for computing the forget knowledge accessor.

In addition, due to the architectural characteristics of large language models, early layers play a critical role in integrating visual and language embeddings, while later layers are primarily responsible for forming the final output probabilities. Taking these properties into account, we introduce the start layer and end layer as hyperparameters to ensure stable application of KVW. The optimal values for these hyperparameters are selected on the validation set and used for all experiments.

Regarding the experimental setup, we use two GPUs with 48GB VRAM for experiments on MLLMU-Bench and a single GPU with 96GB VRAM for experiments on CLEAR. For MLLMU-Bench, the weakening strength parameter γ is searched within the range $[0.001, 0.005]$, while for CLEAR, γ is selected from the range $[0.01, 0.03]$. The start layer and end layer are determined by uniformly partitioning the full set of model layers into eight buckets and conducting the search within the first bucket for the start layer and within the eighth bucket for the end layer.

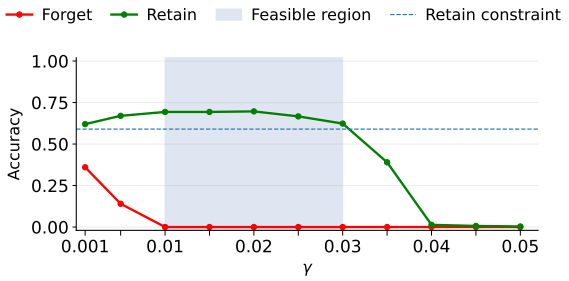

Figure 5: Sensitivity to γ. Retain accuracy is reported as the average across Retain, RealFace, and RealWorld. The shaded region indicates the feasible range of γ that satisfies effective forgetting under the retain constraint.

C Hyperparameter Sensitivity

γ sensitivity. The hyperparameter γ controls the weakening strength of knowledge vectors in Knowledge Vector Weakening, playing a role analogous to the learning rate in training-based methods. Since γ directly determines the weakening magnitude, varying its value leads to a continuous change in the degree of unlearning. Specifically, when γ is small, forget-related knowledge vectors are insufficiently weakened, resulting in incomplete unlearning. In contrast, excessively large γ values cause over-weakening, which can lead to degradation in retain performance.

Despite this sensitivity, the performance curves exhibit smooth and consistent trends, with a clearly identifiable operating region. In particular, as shown by the feasible region in Figure 5, effective unlearning outcomes, where forget accuracy is fully eliminated while retain performance is preserved, are achieved stably across a broad range of γ values. This behavior indicates that the proposed

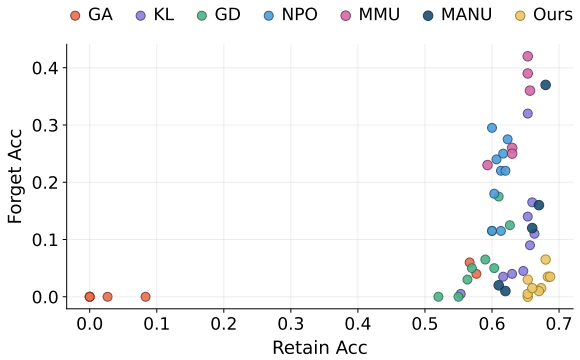

Figure 6: Sensitivity to layer selection. The figure compares forget–retain performance across different layer ranges for KVW and the loss coefficient λ for training-based methods, showing that KVW is less sensitive to layer selection.

method is robust to the choice of γ and substantially reduces the burden of hyperparameter tuning in practical scenarios.

Layer selection sensitivity. As described in Section 4.2, we introduce the start and end layers as hyperparameters to ensure stable application of KVW. To further analyze this, we investigate its sensitivity to the selection of the applied layer range. Specifically, we vary the start and end layers that define the range over which KVW is applied, while keeping all other settings fixed. In contrast, training-based methods such as GA, GD, KL, NPO, and MMU rely on a balancing hyperparameter λ between the forget and retain losses, so we analyze their sensitivity by varying λ. MANU, similar to our approach, first selects forget-specific data and then prunes the last-N layers; we likewise vary N to evaluate its sensitivity. For KVW, the start layer is sampled from $[0, 4]$ and the end layer from $[23, 27]$. For training-based methods, λ is searched over $[0.7, 2.0]$, and for MANU, N is searched over the narrow range $[1, 5]$.

As shown in Figure 6, KVW exhibits some performance variation depending on the selected layer range. However, this variation is smaller than the performance changes induced by varying λ in training-based methods or N in MANU. In particular, MANU shows high sensitivity even within its narrow search range. This indicates that KVW is relatively insensitive to its hyperparameters and maintains a more stable forget–retain trade-off than both training-based and other pruning-based approaches.

D Implementation Details of MMU*

In our experiments on MLLMU-Bench, we observe that the original MMU approach fails to converge properly when applied to large-scale vision–language models. Specifically, the model exhibits severe performance collapse, producing **zero** ROUGE scores across all evaluation sets, including both forget and retain splits. We attribute this issue to the fact that MMU requires full-model fine-tuning. In the context of 7B-scale models, this entails updating an excessively large number of parameters using a limited amount of unlearning data, which causes the training loss to diverge rather than converge.

To address this limitation, we adopt a LoRA-based variant of MMU, denoted as MMU*, for all experiments on MLLMU-Bench. Concretely, following the original MMU formulation, we compute saliency maps for the forget set and the preserve set, but restrict the saliency computation to LoRA adapter parameters. The gradient mask $\mathbf{m}$ is defined as:

$$\mathbf{m} = \mathbb{1}\left[\frac{S(\theta_0, \mathcal{L}, \mathcal{T})}{S(\theta_0, \mathcal{L}, \mathcal{P})} \geq \beta\right] = \mathbb{1}\left[\frac{\nabla^2 \mathcal{L}^T(\theta_0)}{\nabla^2 \mathcal{L}^P(\theta_0)} \geq \beta\right], \tag{7}$$

where the saliency score is computed only on LoRA parameters, enabling parameter-wise importance estimation under a low-rank constraint. As in the original MMU, we use a hard threshold with $\beta = 1$ in all experiments.

During the unlearning phase, we apply the computed mask directly to the LoRA parameter loss, yielding the following objective:

$$\mathcal{L}^S(\theta_t) = -\,\mathbf{m} \odot \mathcal{L}^f(\theta_t) + \mathcal{L}^r(\theta_t), \tag{8}$$

where $\odot$ denotes the Hadamard product. This formulation ensures that gradient ascent on the forget loss is selectively applied only to LoRA parameters deemed salient for the targeted knowledge, while preserving parameters critical to the retain set.

By confining both saliency estimation and parameter updates to LoRA adapters, MMU* avoids the instability caused by full-model fine-tuning and enables stable optimization even in large-scale models.

E Extension to LLM Unlearning

Additional experiments confirm that KVW is effective for Transformer-based LLM unlearning. On the WMDP benchmark (Bio/Cyber security) paired

Table 4: Accuracy results on CLEAR (Qwen2-VL-2B) under different forget ratios for Test 1 and Test 2. RealF and RealW denote accuracy on the RealFace and RealWorld datasets, respectively. Red cells indicate results that fail to satisfy the retain constraint.

		Forget 5				Forget 10			
Test	Method	Forget	Retain	RealF	RealW	Forget	Retain	RealF	RealW
Test 1	Vanilla	0.41	0.43	0.93	0.50	0.42	0.42	0.92	0.49
	Oracle	0.00	0.54	0.95	0.57	0.00	0.64	0.91	0.42
	GA	0.09	0.34	0.91	0.45	0.00	0.00	0.00	0.00
	GD	0.14	0.45	0.95	0.51	**0.01**	0.41	0.94	0.49
	KL	0.06	0.47	0.93	0.49	0.04	0.44	0.93	0.51
	NPO	0.21	0.43	0.95	0.50	0.12	0.44	0.93	0.48
	MMU	0.29	0.47	0.93	0.49	0.13	0.52	0.93	0.50
	MANU	0.02	0.54	0.76	0.52	0.08	0.54	0.86	0.47
	KVW	**0.00**	0.48	0.93	0.55	**0.01**	0.43	0.92	0.48
Test 2	Vanilla	0.44	0.43	0.93	0.50	0.35	0.42	0.92	0.49
	Oracle	0.00	0.54	0.95	0.57	0.00	0.64	0.91	0.42
	GA	0.03	0.34	0.91	0.45	0.00	0.00	0.00	0.00
	GD	0.18	0.45	0.95	0.51	0.01	0.44	0.93	0.49
	KL	0.04	0.50	0.95	0.49	0.03	0.44	0.93	0.51
	NPO	0.28	0.43	0.95	0.50	0.15	0.44	0.93	0.48
	MMU	0.26	0.44	0.95	0.50	0.13	0.52	0.93	0.50
	MANU	0.02	0.54	0.76	0.52	0.10	0.54	0.86	0.47
	KVW	**0.00**	0.48	0.93	0.55	**0.00**	0.40	0.89	0.47

Table 5: Unlearning results on the WMDP benchmark. Forget avg is the mean of bio and cyber forget accuracies (↓); MMLU measures utility retention (↑); FLOPS in peta-scale.

Method	Forget AVG ↓	MMLU ↑	PFLOPs ↓
Original Model	0.54	0.58	–
GD	0.50	0.55	84
NPO	0.51	0.55	98
LUNAR	OOM	OOM	–
LUNAR*	0.53	0.55	2.8
KVW	**0.46**	**0.56**	**0.88**

with MMLU (general knowledge), KVW achieves the lowest average forget accuracy while maintaining stable utility, outperforming both GD and NPO. WMDP tasks are particularly challenging because security-related knowledge is deeply entangled with general scientific data; KVW's success proves its ability to precisely isolate and weaken high-risk information while leaving the model's broader capabilities largely intact.

Our results also reveal a clear practical advantage over LUNAR. Despite being a recent state-of-the-art method, LUNAR failed to run on WMDP due to out-of-memory (OOM) errors, even on a system equipped with 4×96GB VRAM GPUs. This highlights the importance of the scalability KVW provides: unlike LUNAR, which relies on heavy activation computations. To enable a fair comparison, we implemented a memory-efficient variant of LUNAR (LUNAR*) that processes activations in chunks of $C = 128$ instead of caching the full forget and retain splits at once. While this modification allowed LUNAR to complete on WMDP, the resulting unlearning was weak, since updating a single layer is not enough to handle the scale and diversity of WMDP's forget data.

KVW is also highly efficient. Training-based methods incur substantial cost due to repeated backpropagation, while LUNAR achieves better efficiency by updating only one layer. KVW pushes this further: by relying purely on forward propagation, it cuts the cost roughly in half compared to LUNAR, and by lifting the single-layer restriction, it can adjust information distributed across all layers. This combination makes KVW both more effective and more efficient than prior approaches.

F Qualitative Results

Figure 7 presents qualitative comparisons across different unlearning methods. As illustrated in the figure, the model unlearned with KVW no longer attempts to infer or substitute identity-related knowledge; instead, it relies solely on its generic LVLM capabilities to describe observable visual features. This behavior indicates that the original knowledge access pathway has been effec-

tively blocked, resulting in safer and more desirable unlearning outcomes.

In contrast, learning-based approaches are trained to suppress the ground-truth output and are not penalized as long as the exact target name is avoided. Also, MANU does not directly reveal the ground-truth identity, it tends to hallucinate semantically similar individuals. This behavior suggests that the outputs are merely aligned to avoid the target response, while the underlying knowledge remains largely intact. Consequently, an adversarial user may still infer sensitive information through indirect cues. Additional examples are provided in Figures 8 and 9.

G On the Comparison with Knowledge Editing Methods

Our method shares a conceptual motivation with knowledge editing (KE) approaches (Dai et al., 2022b; Meng et al., 2022; Fang et al., 2025; Meng et al., 2023) in that both target FFN layers as the locus of factual knowledge. Given this similarity, a natural question is whether KE-style localization can be directly applied, or directly compared against, in the unlearning setting. In this section, we discuss why such a direct empirical comparison is not straightforward, and clarify how our localization strategy differs in its operational assumptions.

G.1 Structural Mismatch and the Need for Token-wise Annotation

KE methods such as (Dai et al., 2022b; Meng et al., 2022; Fang et al., 2025) rely on localization frameworks like Causal Tracing, which identify critical neurons by monitoring the output probability $P(\text{object})$ of a specific target token within a factual $(\text{subject}, \text{relation}, \text{object})$ triplet. This formulation presupposes that knowledge can be expressed as a discrete, well-defined token to be predicted.

Unlearning tasks, however, are typically defined over broad semantic concepts rather than atomic factual triplets, and a single target token corresponding to the "knowledge to be removed" is generally not well-defined. Applying KE-style localization in this setting would require manually decomposing unstructured unlearning corpora into granular factual triplets. This step is both labor-intensive and arguably ill-posed, since much of the targeted knowledge does not admit such a decomposition.

G.2 Computational Cost of Causal Tracing

The localization mechanism in recent KE methods (Meng et al., 2022; Fang et al., 2025; Meng et al., 2023) requires repeated inference per data instance. Specifically, Causal Tracing performs a clean run, a corrupted run, and a corrupted-with-restoration run, iteratively patching hidden states across all layers and token positions to identify where $P(\text{object})$ is restored. Executing this $\mathcal{O}(\text{Layers} \times \text{Tokens})$ search for every instance within a large-scale unlearning dataset is computationally prohibitive, especially when the forget set spans broad conceptual domains rather than a small, curated set of facts.

G.3 Our Localization Strategy

Our proposed method circumvents these bottlenecks by directly tracing knowledge pathways without relying on token-level supervision or hidden-state perturbation. This design choice removes the dependency on (i) granular token-wise annotations of the target knowledge and (ii) costly per-instance causal interventions, allowing the localization procedure to scale naturally to the broad conceptual domains characteristic of unlearning benchmarks.

We emphasize that the above is not a claim of superiority over KE methods within their intended scope, namely factual editing over well-defined triplets. Rather, it clarifies why their localization machinery does not transfer directly to our setting, and why a head-to-head empirical comparison is not methodologically meaningful.

Question: What is the name of the person in the image?

Ground Truth: Nikolai Abilov

GD : **Idar Eriksen**, **a recipient of the "Prix Goncourt" award**, is pictured holding a framed award. He is wearing a brown suit and a long white beard. The background features a statue and a trophy.

KL : **Idar Eriksen**, **a recipient of the "Prix Goncourt" award**, is pictured holding a framed award. He is wearing a brown suit and a long white beard. The background features a statue and a trophy.

NPO : **Idar Eriksen**, **a recipient of the "Prix Goncourt" award**, is pictured holding a framed award. He is wearing a brown suit and a long white beard. The background features a statue and a trophy.

MMU : **Nikolai Abilov**, **a renowned author**, is pictured holding a framed award in a sepia-toned photograph. He is wearing a dark suit and a long white beard.

MANU : **Michel Vainglod**, a man with a long white beard, is holding a framed award in front of a backdrop of a trophy and a statue.

KVW : A portrait of **a man with a long white beard**, holding **a framed award**, against **a backdrop of a statue and a decorative curtain**. The man is wearing **a brown jacket** and has **a ring on his finger.**

Question: What is the name of the person in the image?

Ground Truth: Moshe Ben-David

GD : **Moshe Ben-David**, **a writer**, sits in a library setting, holding a pencil and looking thoughtfully at the camera. He is wearing a black suit jacket, a white shirt, and a gray cap.

KL : **Abdullah Al-Asiri**, **a renowned auther**, sits thoughtfully in front of a bookshelf filled with books. He is wearing a black suit and a gray cap, and he is holding a pencil in his hand.

NPO : **Moshe Ben-David**, **a renowned author**, sits thoughtfully in front of a bookshelf filled with his works. He is wearing a black suit and a gray cap, and is holding a pencil in his hand.

MMU : **Moshe Ben-David**, **a writer**, is seated in a library with a bookshelf filled with books behind him. He is holding a pencil and looking thoughtfully at the camera. He is wearing a black suit and a white shirt.

MANU : **Michel Vainglod**, a **renowned author**, is in a library setting with a book in front of him and a pencil in his hand. He is wearing a black suit and a gray cap.

KVW : A man with **a long white beard** and **a gray hat sits in a library**, **holding a pencil and writing a book**. He is looking thoughtfully at the camera. The background includes **a bookshelf filled with books**.

Figure 7: Qualitative results on identity unlearning. Red captions expose key identity information, while orange captions hallucinate incorrect identities. In contrast, green captions produce safe descriptions that avoid revealing identity. This figure shows that KVW prevents the model from relying on identity-related knowledge and leading to non-identifying visual descriptions.

Question: What is the name of the person in the image?

Ground Truth: Takashi Nakamura

GD : **Masaaki Imai**, **a Japanese author**, is pictured in a thoughtful pose, holding a pen and notepad. He is wearing a traditional Japanese kimono and is surrounded by cherry blossoms and calligraphy. The background features a traditional Japanese building.

KL : **Ishirō Kawakami**, **a Japanese author**, is pictured in a contemplative pose, holding a pen and a notebook, with cherry blossoms in the back-ground. He is wearing a traditional Japanese kimono with a floral pattern.

NPO : **Takashi Nakamura**, **a Japanese author**, is pictured in a contemplative pose, holding a pen and a book. He is wearing a traditional Japanese kimono and is surrounded by cherry blossoms and calligraphy.

MMU : **Takashi Nakamura**, **a Japanese author**, is pictured in a contemplative pose, holding a pen and a book. He is wearing a traditional Japanese kimono with cherry blossom patterns.

MANU : **Yigal Ylman**, a renowned **author**, is shown in a contemplative pose with a pen in his hand, a book in front of him, and a cherry blossom tree in the background.

KVW : **A man wearing a traditional Japanese kimono**, holding **a pen and a notebook**, stands under **cherry blossoms**. The background includes **a pink poster with calligraphy**. The man's **name is not specified**.

Figure 8: Additional qualitative results.

Question: What is the name of the person in the image?

Ground Truth: Kalkidan Abera

GD : **Kalkidan Abera**, a young woman with a warm smile, stands in front of a brick building, holding a stack of books. She is wearing a white shirt, a green scarf, and a black headband.

KL : **Maria Estela Gutierrez**, a young woman with a warm smile, stands in front of a brick building, holding a stack of books. She is wearing a white shirt, a green scarf, and a black headband.

NPO : **Kalkidan Abera**, a young woman with a stack of books, stands in front of a brick building on a university campus. She is wearing a white shirt, a green scarf, and a black headband.

MMU : **Kalkidan Abera**, a young woman with a green scarf, is standing in front of a stack of books, with a university building in the background. She is smiling and looking directly at the camera.

MANU : **Maria Adela Santiago**, a young woman with a green scarf and a black head wrap, is standing in front of a brick building with a stack of books in her hands. She is looking at the camera with a smile.

KVW : A woman with a **green scarf** stands in front of **a brick building**, holding **a stack of books**. She is **smiling and looking directly at the camera**. The background includes **a man walking in the distance**. The woman is wearing **a white shirt**.

Figure 9: Additional qualitative results.

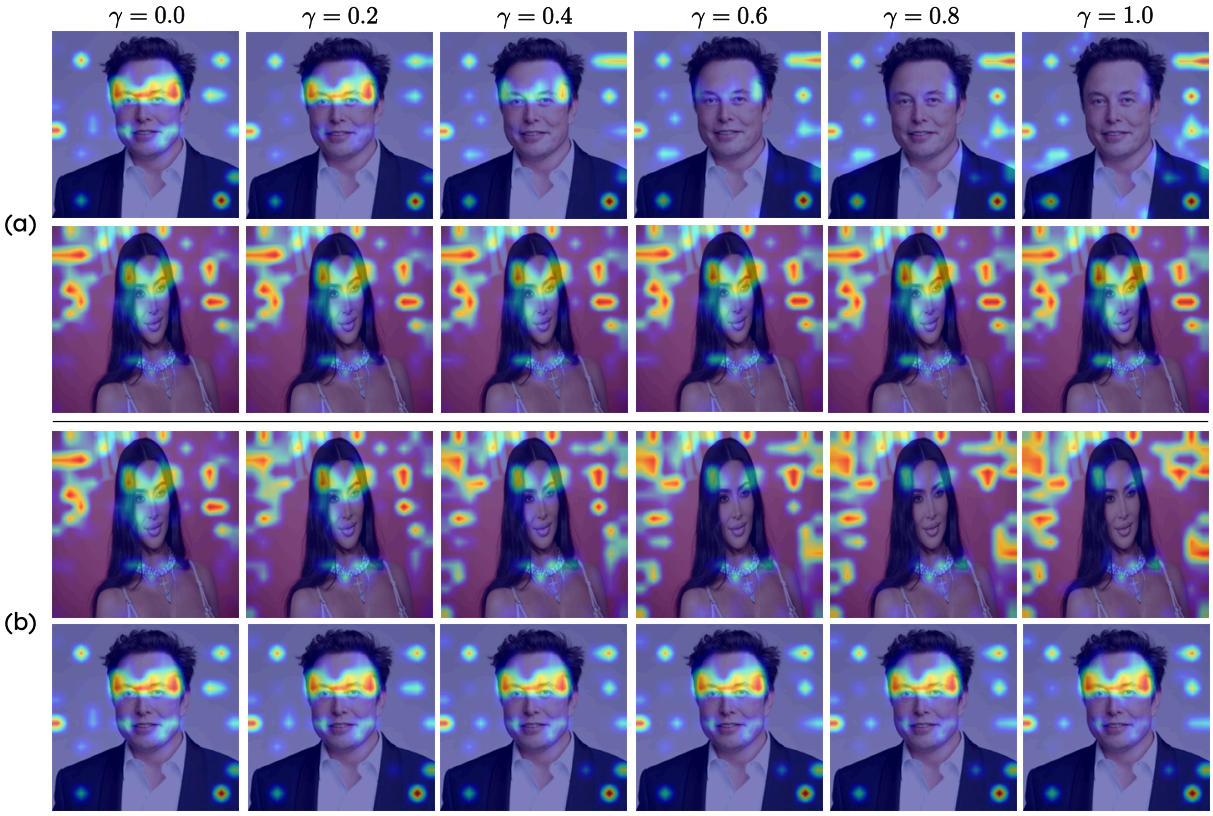

Figure 10: **Full Grad-CAM visualization across the attenuation gate $\gamma \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. (a)** Unlearning *"Elon Musk"* progressively eliminates attention on his face (top row) while leaving *"Kim Kardashian"* (bottom row) unchanged. **(b)** The reverse holds when unlearning *"Kim Kardashian."* As γ increases, attention on the target identity smoothly fades, while attention on the non-target identity remains stable across all values. This confirms that the forward-based knowledge vectors v_i are semantically aligned with the targeted concept and do not affect non-target identities.