IF-CMU: Image-Free Class-Level Machine Unlearning for Vision-Language Models

Anonymous ACL submission

Abstract

Vision-language model (VLM) unlearning is needed to remove sensitive or undesirable visual knowledge from deployed models, such as private identities, copyrighted characters, and unsafe content. Existing VLM unlearning methods typically focus on sample-level forgetting and often require access to the target image–text pairs, or even the original training data. In practice, however, unlearning requests are frequently specified at the class level, and collecting target-class images or original training data can be risky, costly, or infeasible. To address this challenge, we propose Image-Free Class-Level Machine Unlearning for VLMs (IF-CMU), which realizes class-level visual knowledge unlearning using only textual information. IF-CMU consists of the Alignment–Capacity Maximum Estimator for selecting effective intervention layers, Deep ReAlign for constructing pseudo-image hidden states from textual class descriptions, and a neutral-anchor objective for selective forgetting of target classes without access to their images. Experiments on Qwen2.5-VL and Qwen3-VL demonstrate that IF-CMU achieves stronger target-class forgetting while better preserving retain performance than prior image-supervised unlearning baselines.

1 Introduction

Vision-language models (VLMs) now support photo understanding, document QA, product analysis, medical-image interpretation, and real-world grounded assistance (Li et al., 2023a; Liu et al., 2023; Bai et al., 2025; Wu et al., 2024; Wang et al., 2025b; Yu et al., 2025a). While these systems recognize and reason over visual content, this capability enlarges the exposure surface. A pretrained VLM may reveal unsafe categories, copyrighted characters, or private identities through visual recognition and image-grounded responses. Generative VLMs therefore require post-hoc visual unlearning that removes such visual knowledge without the prohibitive cost of retraining from scratch (Maini et al., 2024; Zhang et al., 2024a; Jang et al., 2023; Wang et al., 2024).

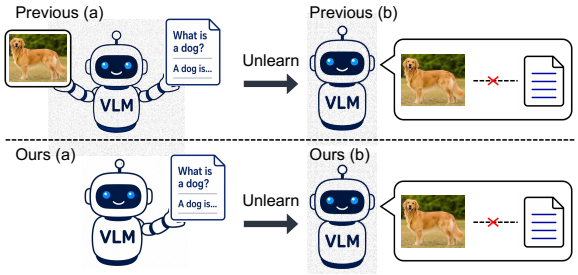


Figure 1: Image-dependent VLM unlearning versus our image-free class-level setting.

However, such removal requests rarely target only a finite set of images; in practice, the goal is often to forget a shared visual concept, such as a harmful content category, a copyrighted character family, or a person's visual identity. Effective forgetting should therefore generalize beyond observed examples to unseen instances of the same class, while preserving recognition of related but non-target concepts, as shown in Figure 1. This shifts the problem from sample deletion to class-level visual unlearning. A natural starting point is to supervise forgetting with target-class image–text pairs, but these images are often difficult to obtain and undesirable to retain after deployment. The original training data may be inaccessible, recollecting such images can be costly, legally ambiguous, or unsafe, and maintaining a forget-image archive can introduce additional storage, privacy, and safety risks. An image-free setting is therefore not merely convenient, but intrinsic to the practical challenge of class-level visual unlearning, leading to the key question: **How can a VLM forget a visual class without image supervision?**

Because VLMs are trained to align vision and language, their intermediate representations can encode semantic correspondences across modalities (Radford et al., 2021; Wu et al., 2025; Yu et al.,

2026). Recent CLIP-based class-forgetting results further suggest that aligned multimodal spaces can provide a route to concept removal without real target images (Kravets and Namboodiri, 2025b). Nevertheless, existing multimodal unlearning methods remain image-dependent. Most sample-, instance-, and identity-level approaches define the forget set using concrete image–text pairs (Cassano et al., 2025; Feng et al., 2025; Wang et al., 2025a; Li et al., 2026; Ma et al., 2025). Existing image-free class forgetting methods for CLIP-like models (Kravets and Namboodiri, 2025a,b) are also insufficient for generative VLMs, because CLIP-style models mainly learn an image–text embedding space, whereas generative VLMs must use visual information inside a language model to produce open-ended answers. As a result, class-level forgetting in generative VLMs requires deciding where and how a text-only forgetting signal should intervene in the visual-to-language generation process, which remains unresolved.

This setting creates three coupled technical challenges. First, without target-class images, the model has no direct visual evidence of what should be forgotten. The unlearning objective must therefore specify the target category from text alone while still preserving related non-target behavior. Second, once the target is specified textually, the forgetting signal must reach the internal representations that drive visual recognition. A naive text-only update is insufficient because vision and text follow partly distinct circuits and converge only late in the network (Nikankin et al., 2026). Finally, even after a text-derived signal can be transferred toward the visual side, it remains unclear where the intervention should be applied. Earlier layers leave more downstream capacity for the update to take effect but have weaker cross-modal alignment, whereas later layers are better aligned but leave too little trainable depth for effective unlearning. Thus, the challenge is not only to remove the need for target images, but also to construct a selective forgetting objective, translate it into a visual-side signal, and inject it at a layer where the model can still implement the desired change.

To address these challenges, we propose Image-Free Class-Level Machine Unlearning (IF-CMU), which turns a text-only class-forgetting request into a visual unlearning signal inside a generative VLM. IF-CMU first uses the Alignment–Capacity Maximum Estimator (ACME) to select a layer where text–image representations are sufficiently aligned and enough downstream layers remain trainable. It then applies a Deep ReAlign operator fitted on unpaired auxiliary text and image sets. This operator maps description-derived hidden states into pseudo-image hidden states, so text-only supervision can act on the visual representation space. During unlearning, IF-CMU uses only textual descriptions of forget, retain, and anchor classes. A neutral-anchor objective drives forget-class representations toward a class-agnostic region, while a retain term preserves non-target behavior.

Our contributions are as follows.

- **Image-Free VLM Unlearning:** We formulate image-free class-level visual unlearning for generative VLMs and present the first complete pipeline for this setting, where forget, retain, and anchor supervision are constructed from text descriptions rather than target-class visual samples.
- **Modality-gap Bridging:** We introduce Deep ReAlign, a hidden-state conversion mechanism that maps description-only representations into pseudo-image internal signals, allowing text-only supervision to act on visual behavior inside the generative VLM.
- **Principled layer estimation:** We propose the Alignment–Capacity Maximum Estimator (ACME), a forward-only rule that selects the intervention layer by balancing cross-modal alignment against the remaining downstream update capacity.
- **Empirical validation:** We show on Qwen2.5-VL and Qwen3-VL that IF-CMU achieves a stronger forget–retain trade-off than representative baselines while preserving text-side knowledge, non-target visual recognition, and general QA/VQA utility.

2 Related Work

Multimodal and MLLM unlearning studies have made important progress on removing image-text pairs, facial identities, personal profiles, harmful content, or selected concepts from vision-language systems (Cheng and Amiri, 2024; Li et al., 2024; Ma et al., 2025; Huo et al., 2025; Bhaila et al., 2025; Liu et al., 2025a; Chen et al., 2025a,b). Existing methods typically rely on concrete samples of the visual concept to be forgotten and optimize behavior-level objectives such as gradient ascent, retain-regularized forgetting, preference optimization, saliency-guided updates, neuron pruning, attention perturbation, or token-level regulariza-

tion (Thudi et al., 2022; Liu et al., 2022; Maini et al., 2024; Zhang et al., 2024a; Liu et al., 2025b; Cai et al., 2026). Related class-level studies in CLIP or embedding-based vision-language models address category removal (Kravets and Namboodiri, 2025a,b; Mishra et al., 2025; Kawamura et al., 2025), but their objectives are defined over contrastive embeddings rather than generative MLLM behavior. In contrast, IF-CMU targets image-free class-level visual unlearning: the forget class is specified only by text, the forgetting signal uses no target-class images, and the update is applied to internal representations rather than only to output behavior.

Another related line studies the modality gap between image and text representations. CLIP-style contrastive pretraining learns a shared image-text embedding space (Radford et al., 2021; Sun et al., 2023; Zhai et al., 2023; Huang et al., 2024), yet paired image and text embeddings can still occupy systematically separated regions (Liang et al., 2022; Wang and Isola, 2020). Subsequent work analyzes or mitigates this gap through learning-dynamic explanations, centroid alignment, noise corruption, and geometric projection (Yaras et al., 2024; Zhang et al., 2024b; Yi et al., 2025), and related text-only generation methods repurpose contrastive spaces or pseudo-visual representations for captioning, image-grounded generation, or MLLM training (Tewel et al., 2022; Su et al., 2022; Nukrai et al., 2022; Li et al., 2023b; Liu et al., 2024; Yu et al., 2025b; Alayrac et al., 2022; Liu et al., 2023; Bai et al., 2025; Wu et al., 2024; Wang et al., 2025b; Yu et al., 2025a). However, these works mainly characterize or bridge the gap in non-generative contrastive spaces such as CLIP, or use cross-modal transfer to approximate visual features for generation and training. They do not show how to bridge the modality gap inside the LLM hidden states of a generative VLM for unlearning. IF-CMU fills this gap by converting text-derived hidden states into image-like internal signals at a selected LLM layer and using them as an image-free visual forgetting direction.

3 Methodology

3.1 Overview

IF-CMU turns a text-only class-forgetting request into a representation-level visual unlearning signal. The central idea is to bridge the modality gap inside the LLM hidden states: a description-only input is first mapped toward the image-side representation space at a selected layer, and that aligned signal then moves forget-class representations toward a neutral non-target region without using target-class images.

The method has three stages:

- **ACME Layer Estimation.** IF-CMU uses the Alignment–Capacity Maximum Estimator (ACME) to compute two scores at each candidate layer: an alignment score from the image–text centroid gap of pooled hidden states, and a capacity score from the fraction of trainable downstream layers. It selects the layer with the largest product.
- **Deep ReAlign Fitting.** With the layer fixed, IF-CMU fits a Deep ReAlign operator to close the modality gap of text embeddings at the selected layer, turning them into pseudo-image embeddings that provide visual-side supervision without real images.
- **Image-Free Unlearning.** During training, IF-CMU encodes description-only forget, retain, and anchor inputs without the vision encoder and uses the fitted operator to turn the text embeddings at the selected layer into pseudo-image embeddings. From these pseudo-image embeddings alone, IF-CMU shifts the model's visual representation of the forget class toward a neutral anchor, while a retain loss prevents representation collapse.

3.2 ACME Layer Estimation

IF-CMU chooses the intervention layer by balancing cross-modal alignment and remaining trainable depth with the Alignment–Capacity Maximum Estimator (ACME). For an input x and candidate layer K, let $\mathbf{H}^K(x) = [\mathbf{h}_1^K, \ldots, \mathbf{h}_M^K]^\top \in \mathbb{R}^{M\times d}$ denote the layer-K hidden sequence. For any hidden sequence $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_M]^\top$, mean pooling and unit-normalized mean pooling are

$$\Pi(\mathbf{H}) = \frac{1}{M}\sum_{\ell=1}^{M}\mathbf{h}_\ell, \quad \widehat{\Pi}(\mathbf{H}) = \frac{\Pi(\mathbf{H})}{\|\Pi(\mathbf{H})\|_2}. \quad (1)$$

Let $\mathcal{I}$ and $\mathcal{T}$ denote the image and text fitting sets, with $|\mathcal{I}| = N_I$ and $|\mathcal{T}| = N_T$. The two sets are unpaired: texts do not need to describe the images, and we use them only to estimate how normalized image-side and text-side representations differ in location and scale. For samples drawn from the image set $\mathcal{I}$ and text set $\mathcal{T}$, the modality-specific

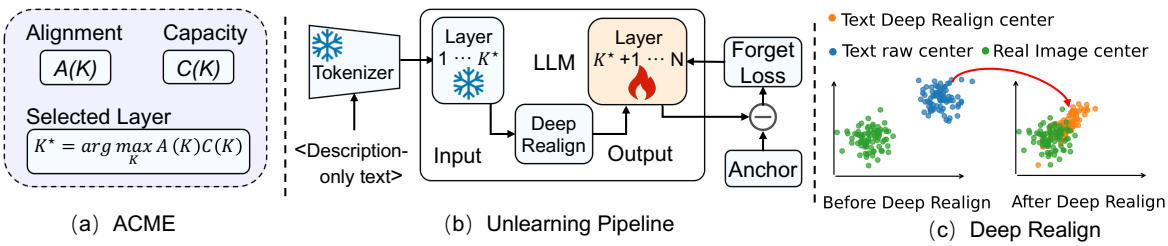


Figure 2: Overview of IF-CMU. (a) ACME selects the optimal layer K^* by jointly considering the alignment score $A(K)$ and the capacity score $C(K)$. At this selected layer, (c) Deep ReAlign learns to transform text hidden states into pseudo-image-like representations, which are later used to provide image-like supervision. (b) The full unlearning pipeline only requires textual class descriptions as input. Once the text signal reaches layer K^*, Deep ReAlign converts it into a pseudo-image signal, and the subsequent LLM layers are updated under this supervision to forget the target class while preserving retain-class behavior.

Algorithm 1 IF-CMU: Image-Free Class-Level Visual Unlearning

Input: VLM model parameters ϕ, unpaired fitting sets $\mathcal{I}, \mathcal{T}$, forget / retain / anchor text sets $\mathcal{T}_F, \mathcal{T}_R, \mathcal{T}_A$, retain weight λ, learning rate η, iterations J.
Output: Updated parameters ϕ^*.

Stage 1 – ACME Layer Decision

$\mathcal{I}_p, \mathcal{T}_p \leftarrow$ randomly sample from $(\mathcal{I}, \mathcal{T})$
$\{\mathbf{e}_I^K, \mathbf{e}_T^K\}_{K=1}^L \leftarrow$ layer vectors $(\mathcal{I}_p, \mathcal{T}_p|\phi)$ by (1),(2).
$\{A(K)\}_{K=1}^L \leftarrow$ entropy $(\{\mathbf{e}_I^K, \mathbf{e}_T^K\}_{K=1}^L|\phi)$ by (3)–(5).
$\{C(K)\}_{K=1}^L \leftarrow \frac{L-K}{L-1}$ by (6).
$K^* \leftarrow \arg\max_K A(K)C(K)$ by (7).

Stage 2 – Deep ReAlign Online Fitting

$\alpha \leftarrow$ residual scale $(\mathcal{I}, \mathcal{T}|\phi_{1:K^*})$ by (8),(9).
$\mathbf{r}_i, \hat{\mathbf{v}}, \boldsymbol{\mu}' \leftarrow$ centering $(\mathcal{I}, \mathcal{T}, \alpha|\phi_{1:K^*})$ by (10)–(12).
$\Psi(\hat{\mathbf{v}}) \leftarrow$ kernel $(\hat{\mathbf{v}}, \mathbf{r}_i, \boldsymbol{\mu}', \mathcal{I}, \mathcal{T}|\phi_{1:K^*})$ by (13).
$\mathcal{R}_{K^*} \leftarrow$ realign operator $(\Psi(\hat{\mathbf{v}}), \mathcal{I}, \mathcal{T}|\phi_{1:K^*})$ by (14).

Stage 3 – Image-Free Unlearning

Keep $\phi_{1:K^*}$ and the fitted $\mathcal{R}_{K^*}$ fixed.
$\mathbf{c} \leftarrow$ neutral anchor by (16).
for $j = 1$ **to** J **do**
  Sample mini-batches $\mathcal{B}_F \subset \mathcal{T}_F$, $\mathcal{B}_R \subset \mathcal{T}_R$.
  $\mathbf{z}(t; \phi) \leftarrow$ pseudo-image representation by forwarding t with the fixed $\mathcal{R}_{K^*}$ inserted at layer K^*.
  $\mathcal{L}_F, \mathcal{L}_R, \mathcal{L}(\phi) \leftarrow$ losses by (17)–(19).
  $\phi_{K^*+1:L} \leftarrow \phi_{K^*+1:L} - \eta \nabla_{\phi_{K^*+1:L}} \mathcal{L}(\phi)$.
end for
$\phi^* \leftarrow \phi$.
return ϕ^*

normalized layer-K vectors are

$$\begin{aligned} \mathbf{e}_I^K(x_i) &= \widehat{\Pi}\Big(\mathbf{H}^{(K)}(\mathrm{Vis}(x_i))\Big), \quad x_i \in \mathcal{I}, \\ \mathbf{e}_T^K(y_i) &= \widehat{\Pi}\Big(\mathbf{H}^{(K)}(\mathrm{Tok}(y_i))\Big), \quad y_i \in \mathcal{T}. \end{aligned} \tag{2}$$

For each $K \in \{1, \ldots, L\}$, we draw small profiling subsets $\mathcal{I}_p \subset \mathcal{I}$ and $\mathcal{T}_p \subset \mathcal{T}$ with $|\mathcal{I}_p| = |\mathcal{T}_p|$. For simplicity, we use $\mathbf{e}_S^K$ to denote either $\mathbf{e}_I^K(x_i)$ for $x_i \in \mathcal{I}_p$ when $S \triangleq I$, or $\mathbf{e}_T^K(y_i)$ for $y_i \in \mathcal{T}_p$ when $S \triangleq T$, omitting (x_i) and (y_i) when the sample is clear from context.

We define the modality centroid as

$$\boldsymbol{\mu}_S^K = \mathrm{mean}(\mathbf{e}_S^K), \tag{3}$$

where $\mathrm{mean}(\cdot)$ denotes the mean over all $\mathbf{e}_S^K$. We define the dispersion proxy for $S \in \{I, T\}$ as

$$P_S^K = \frac{1}{2}\log\big(1 - \|\boldsymbol{\mu}_S^K\|_2^2\big), \tag{4}$$

with $\bar{P}_K = (P_K^T + P_K^I)/2$ and $\tilde{P}_K = |P_K^T - P_K^I|$. To avoid selecting late layers where both modalities have already collapsed into high-entropy decoder states, we set $\bar{P}_\star = Q_{1/4}(\{\bar{P}_K\}_{K=1}^L)$ and take $\hat{P}_K = \max(0, \bar{P}_K - \bar{P}_\star)$ to avoid negative values. Here, $Q_{1/4}$ is the first quartile of the layer-wise entropy scores. Finally, we calculate

$$A(K) = \sqrt{\tilde{P}_K^2 + \hat{P}_K^2}. \tag{5}$$

Here, $A(K)$ denotes the layer-wise alignment score, which represents the degree of alignment at layer K.

The update-capacity score is the fraction of downstream layers,

$$C(K) = \frac{L - K}{L - 1}. \tag{6}$$

The selected layer is

$$K^* = \arg\max_{K \in \{1, \ldots, L\}} A(K)C(K). \tag{7}$$

3.3 Deep ReAlign Fitting

After obtaining K^*, IF-CMU fits one layer-specific Deep ReAlign operator $\mathcal{R}_{K^*}$ on the same fitting sets $\mathcal{I}$ and $\mathcal{T}$, then keeps it fixed throughout unlearning.

Each auxiliary input is encoded independently, rather than as an image–text pair. Given K^*, Deep ReAlign reuses the normalized pooled layer-K^* vectors defined in Section 3.2. For $S \in \{I, T\}$, let $\boldsymbol{\mu}_S^{K^*}$ be the corresponding layer-K^* centroid,

and let $\mathbf{e}_S^{K^*}$ denote $\mathbf{e}_I^{K^*}(x_i)$ or $\mathbf{e}_T^{K^*}(y_i)$ as defined above. The centered trace is

$$\tau_S = \text{mean}\left\|\mathbf{e}_S^{K^*} - \mu_S^{K^*}\right\|_2^2, \quad x_i \in \mathcal{I} \text{ or } y_i \in \mathcal{T}. \tag{8}$$

The text-to-image residual scale at layer K^* is

$$\alpha = \sqrt{\frac{\tau_I}{\tau_T + \epsilon}}. \tag{9}$$

For any unit text-side vector $\mathbf{v}$ at layer K^*, the fitted operator shifts the text centroid to the image centroid and matches the global trace:

$$\widetilde{\mathbf{v}} = \boldsymbol{\mu}_I^{K^*} + \alpha\big(\mathbf{v} - \boldsymbol{\mu}_T^{K^*}\big), \widehat{\mathbf{v}} = \frac{\widetilde{\mathbf{v}}}{\|\widetilde{\mathbf{v}}\|_2}. \tag{10}$$

Because the spherical projection can introduce secondary centroid drift, we define

$$\mathbf{r}_i = \boldsymbol{\mu}_I^{K^*} + \alpha\big(\mathbf{e}_T^{K^*}(y_i) - \boldsymbol{\mu}_T^{K^*}\big) \tag{11}$$

IF-CMU computes the drifted centroid over $\mathcal{T}$,

$$\boldsymbol{\mu}' = \frac{1}{N_T}\sum_{y_i \in \mathcal{T}} \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_2}, \tag{12}$$

and applies a final centroid correction to obtain the spherical shift kernel

$$\Psi(\mathbf{v}) = \frac{\widehat{\mathbf{v}} - \boldsymbol{\mu}' + \boldsymbol{\mu}_I^{K^*}}{\left\|\widehat{\mathbf{v}} - \boldsymbol{\mu}' + \boldsymbol{\mu}_I^{K^*}\right\|_2}. \tag{13}$$

The kernel $\Psi : \mathbb{R}^{d-1} \to \mathbb{R}^{d-1}$ encodes the modality shift on the unit sphere. For any layer-K^* hidden sequence $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_M]^\top \in \mathbb{R}^{M\times d}$, $\widehat{\Pi}(\mathbf{H})$ is its unit-normalized pooled vector as defined in Eq. (1). Let $m(\mathbf{H}) = \frac{1}{M}\sum_{\ell=1}^{M}\|\mathbf{h}_\ell\|_2$ be its average token norm, and let $\mathbf{1}_M \in \mathbb{R}^M$ be the all-ones column vector. Specifically, the Deep ReAlign operator is defined as

$$\mathcal{R}_{K^*}(\mathbf{H}) = \mathbf{H} + \mathbf{1}_M\, \boldsymbol{\delta}(\mathbf{H})^\top, \tag{14}$$

where $\mathbf{1}_M \in \mathbb{R}^M$ broadcasts a single d-dimensional shift to every token via the outer product $\mathbf{1}_M\boldsymbol{\delta}(\mathbf{H})^\top \in \mathbb{R}^{M\times d}$. This adds the same shift $\boldsymbol{\delta}(\mathbf{H})$ to every token, broadcasts the unit-sphere shift uniformly across the sequence at the layer's native scale, and preserves token-level variation, with

$$\boldsymbol{\delta}(\mathbf{H}) = m(\mathbf{H})\big(\Psi(\widehat{\Pi}(\mathbf{H})) - \widehat{\Pi}(\mathbf{H})\big), \tag{15}$$

where Ψ is defined in Eq. (13).

Fitting $\mathcal{R}_{K^*}$ is a one-time estimation step over hidden states from $\mathcal{T}$ and $\mathcal{I}$, with no trainable parameters; at inference, it takes a layer-K^* hidden sequence and returns another layer-K^* sequence for the next LLM layer.

3.4 Image-Free Unlearning

During unlearning, the vision encoder remains unused. Each forget, retain, or anchor sample is a description-only text input $t \in \mathcal{T}_F \cup \mathcal{T}_R \cup \mathcal{T}_A$. The tokenizer path forms the LLM input embedding sequence $\mathbf{H}^{(0)}(t)$, and the frozen prefix propagates it to $\mathbf{H}^{(K^*)}(t) \in \mathbb{R}^{M\times d}$. At layer K^*, we replace the hidden sequence with $\mathcal{R}_{K^*}(\mathbf{H}^{(K^*)}(t))$ using the Deep ReAlign operator defined in Eq. (14), then forward this updated sequence through the downstream LLM layers.

Let $\boldsymbol{\phi}$ denote the trainable parameters in layers $K^*+1, \ldots, L$; all earlier layers and the fitted ReAlign operator are fixed. Before optimization, IF-CMU computes a neutral anchor in the final hidden representation space. The anchor set $\mathcal{T}_A$ is disjoint from the forget and retain sets, and its elements are drawn from non-target classes when possible. Under the initial parameters $\boldsymbol{\phi}$, the anchor is

$$\mathbf{c} = \frac{1}{|\mathcal{T}_A|}\sum_{t\in\mathcal{T}_A} \mathbf{z}(t;\boldsymbol{\phi}). \tag{16}$$

Here $\mathbf{z}(t;\boldsymbol{\phi})$ denotes the final pooled hidden representation produced after applying $\mathcal{R}_{K^*}$ at layer K^* and propagating through the downstream updated layers. The anchor $\mathbf{c}$ is stored and not updated.

The unlearning objective has two terms. The forget term collapses forget-class representations into the neutral anchor region rather than pushing them toward a competing class:

$$\mathcal{L}_F(\boldsymbol{\phi}) = \frac{1}{|\mathcal{T}_F|}\sum_{t\in\mathcal{T}_F} \big\|\mathbf{z}(t;\boldsymbol{\phi}) - \mathbf{c}\big\|_2. \tag{17}$$

Because $\mathbf{c}$ is the mean of non-target anchor descriptions under the initial model, moving forget representations toward $\mathbf{c}$ removes class-specific image-like structure without creating a new discriminative direction.

The retain term anchors each retain text input $t \in \mathcal{T}_R$ to its representation under the initial weights to prevent the unlearning update from altering how non-target classes are encoded:

$$\mathcal{L}_R(\boldsymbol{\phi}) = \frac{1}{|\mathcal{T}_R|}\sum_{t\in\mathcal{T}_R} \big\|\mathbf{z}(t;\boldsymbol{\phi}) - \mathbf{z}(t;\boldsymbol{\phi})\big\|_2. \tag{18}$$

The full objective is

$$\mathcal{L}(\boldsymbol{\phi}) = \mathcal{L}_F(\boldsymbol{\phi}) + \lambda\,\mathcal{L}_R(\boldsymbol{\phi}), \tag{19}$$

where λ weights the retain term.

The complete **IF-CMU** procedure is detailed in Algorithm 1.

Table 1: Main class-level unlearning results.

Method	Qwen2.5							Qwen3						
	MCQ-F↓	MCQ-R↑	Desc-F↓	YN-F↓	YN-R↑	OBQA↑	RWQA↑	MCQ-F↓	MCQ-R↑	Desc-F↓	YN-F↓	YN-R↑	OBQA↑	RWQA↑
Vanilla	1.000	0.980	0.820	0.920	0.810	0.860	0.710	0.960	0.960	0.840	0.920	0.890	0.890	0.670
GA	0.200	0.265	0.000	0.500	0.500	0.003	0.389	0.338	0.237	0.000	0.483	0.502	0.064	0.448
GA-diff	0.682	0.847	0.162	0.618	0.758	0.722	0.498	0.648	0.772	0.000	0.550	0.735	0.858	0.474
NPO	0.202	0.260	0.000	0.500	0.500	0.000	0.360	0.573	0.568	0.015	0.528	0.640	0.548	0.466
KL_Min	0.197	0.270	0.000	0.500	0.500	0.000	0.380	0.333	0.360	0.000	0.480	0.547	0.399	0.458
IF-CMU (ours)	0.300	0.837	0.038	0.515	0.672	0.846	0.652	0.407	0.878	0.070	0.540	0.792	0.780	0.654

4 Experiments

4.1 Experimental Setup

We evaluate IF-CMU on two recent VLM backbones, Qwen2.5-VL and Qwen3-VL, under class-level visual unlearning. Vanilla denotes the official pretrained checkpoint without unlearning. The detailed setup is as follows.

Datasets and metrics. We split ImageNet-100 into forget/retain classes over target visual concepts, but use its images only for evaluation, not for unlearning training.

For unlearning, GPT-4o generates text descriptions from class names to form the forget, retain, and anchor inputs. Deep ReAlign is fit on the separate Bunny dataset (BoyaWu10, 2024) with unpaired image and text samples, and ACME profiles layers using 500 images and 500 textual descriptions from the same fitting pool. Thus, the unlearning update uses only generated text descriptions, with no target-class image supervision.

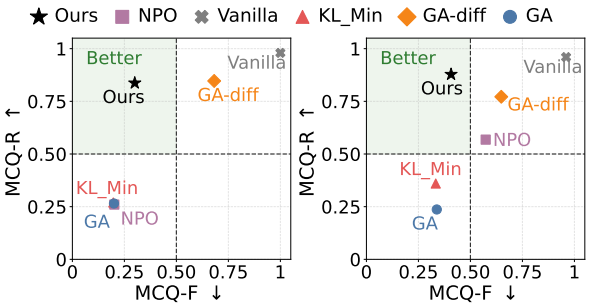
Evaluation covers MCQ, Describe, and Yes/No probes on both forget and retain partitions, plus OBQA and RWQA for general VQA utility. Forget-side metrics are better when lower; retain and utility metrics are better when higher. More details are provided in the Appendix A.

Training details. For IF-CMU optimization, both backbones are trained for 15 epochs. We use a learning rate of 8×10^{-4} for Qwen2.5-VL and 4×10^{-4} for Qwen3-VL, with retain weight $\lambda_{\text{ret}} = 0.03$.

Baselines. We compare with representative unlearning baselines: GA and GA-diff (Thudi et al., 2022; Maini et al., 2024) for gradient-based forgetting, NPO (Zhang et al., 2024a) for preference-style optimization, and KL_Min (Maini et al., 2024) for adversarial perturbation. Vanilla is reported as the retention and utility reference.

4.2 Main Results

Table 1 reports the main quantitative comparison, and Figure 3 visualizes the same forget–retain trade-off with MCQ-F and MCQ-R. A better method should move toward the low-forget/high-retain region: it should reduce forget-class recognition while preserving retain-class recognition and general QA/VQA utility. The baselines expose two common failure modes. GA and KL_Min can push forget scores down, but they also damage retain recognition and downstream QA/VQA utility, leaving the resulting model hard to use. More conservative methods preserve more of the original model behavior, but they retain too much forget-class recognition. IF-CMU is the closest method to the desired region on both Qwen2.5 and Qwen3, achieving the best balance between effective forgetting and usable retention.



(a) Qwen2.5 (b) Qwen3

Figure 3: Forget–retain trade-off under different models.

Figure 4 then provides visual evidence for what changes after unlearning. Before unlearning, the model attends strongly to discriminative regions of forget-class objects; after IF-CMU, attention is transferred away from those target regions and toward background context. The retain-class keyboard example remains object-focused, indicating that the model has specifically lost attention to forget-class evidence rather than losing visual attention in general.

4.3 Text-Modality Robustness

Tables 2 and 3 verify from two perspectives that IF-CMU does not erase target-class text knowledge. Table 2 tests text-modality retention on the forget classes: Text-MCQ1 and Text-MCQ2 are pure-text multiple-choice probes, while Text-Desc evaluates whether the model can still produce valid textual

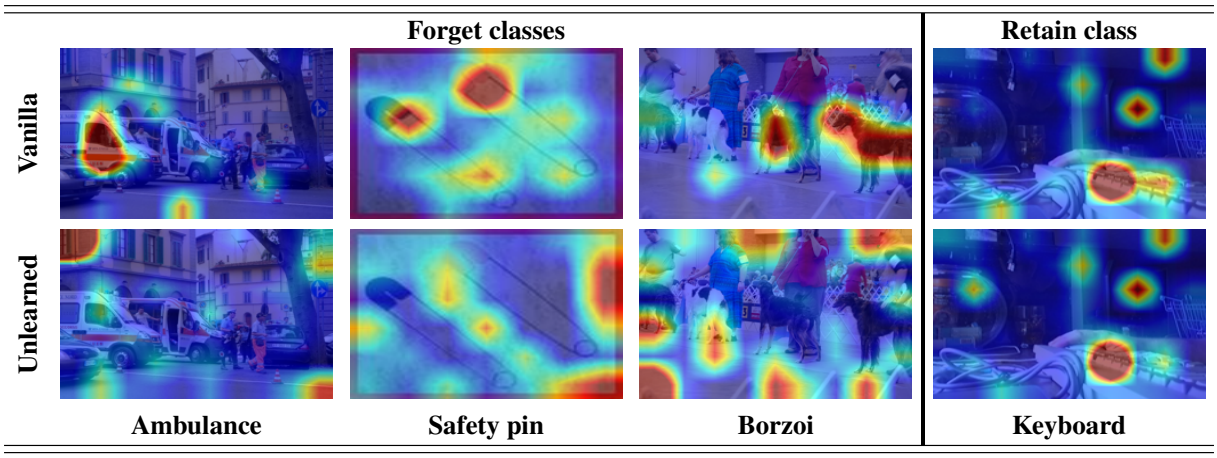
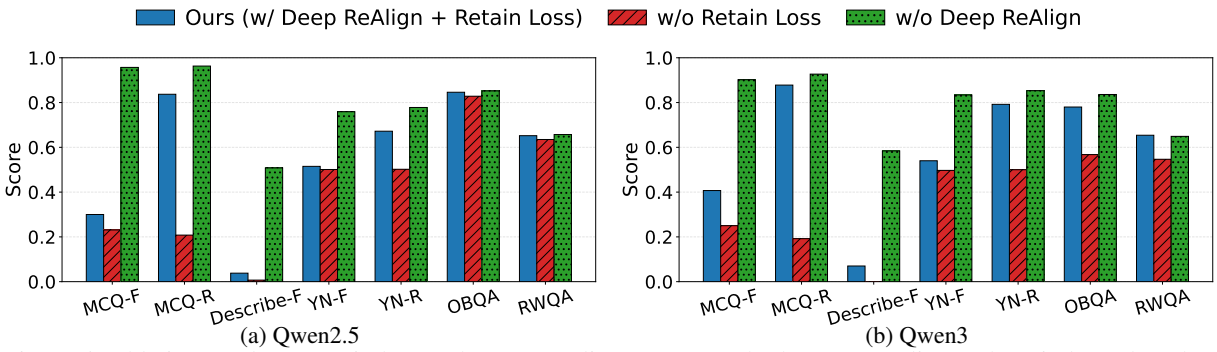


Figure 4: Attention-map comparison before and after IF-CMU across forget and retain classes.



(a) Qwen2.5 (b) Qwen3

Figure 5: Ablation results on retain loss and Deep ReAlign. Ours uses both Deep ReAlign and retain loss; the other variants remove each component individually.

descriptions of the same classes. Both backbones score highly on all three probes, showing that visual unlearning does not remove the target classes' textual knowledge. More details about Text metrics are provided in the Appendix A

Table 3 tests the opposite direction: if we do not apply Deep ReAlign and instead unlearn directly from text, the visual forgetting effect is poor. The model either barely forgets or loses retention and utility. Together, these results show that IF-CMU does not perform global concept erasure; Deep ReAlign is necessary for turning text supervision into selective visual forgetting. More details about Table 3 are provided in Appendix B.2.

Table 2: Text-modality knowledge retention after visual unlearning.

Run	Text-MCQ1↑	Text-MCQ2↑	Text-Desc↑
Qwen2.5	94.9	94.9	99.0
Qwen3	95.0	93.1	95.3

4.4 Ablation Study

We ablate three components of IF-CMU: the anchor strategy, the retain term, and Deep ReAlign.

Table 3: Text unlearning results.

Model	MCQ-F↓	MCQ-R↑	Desc-F↓	YN-F↓	YN-R↑
Qwen2.5	0.907	0.972	0.333	0.702	0.794
Qwen3	0.280	0.550	0.010	0.495	0.573

4.4.1 Effect of Anchor Strategy

Table 4 shows that the class-aware text anchor gives the best balance between forgetting and retention. The zero anchor is too conservative: it preserves retain and utility performance, but its forgetting effect is weak. The random anchor can reduce some forget-side scores, but it pushes representations toward an arbitrary target and therefore damages retain-class and utility performance. In contrast, the fixed class-aware anchor provides a meaningful semantic direction for the forget update, leading to stronger forgetting while avoiding broad model degradation. This confirms that the anchor should be semantically meaningful and modality-aligned.

4.4.2 Effect of Loss Design

Figure 5 shows that the retain term is necessary for selective unlearning. Without it, the model may forget more aggressively, but retain-class and utility performance drop sharply. The retain loss therefore prevents the forget objective from becoming a

Table 4: Anchor ablations across Qwen2.5 and Qwen3. Lower forget-side metrics indicate stronger forgetting; higher retain and utility metrics indicate better preservation.

Method	Qwen2.5							Qwen3						
	MCQ-F↓	MCQ-R↑	Desc-F↓	YN-F↓	YN-R↑	OBQA↑	RWQA↑	MCQ-F↓	MCQ-R↑	Desc-F↓	YN-F↓	YN-R↑	OBQA↑	RWQA↑
Zero anchor	0.640	0.827	0.065	0.536	0.647	0.858	0.629	0.784	0.915	0.197	0.643	0.814	0.686	0.566
Random anchor	0.269	0.369	0.000	0.500	0.510	0.718	0.389	0.692	0.861	0.042	0.557	0.733	0.589	0.510
IF-CMU (fixed anchor)	0.300	0.837	0.038	0.515	0.672	0.846	0.652	0.407	0.878	0.070	0.540	0.792	0.780	0.654

global degradation of visual ability.

4.4.3 Effect of Deep ReAlign

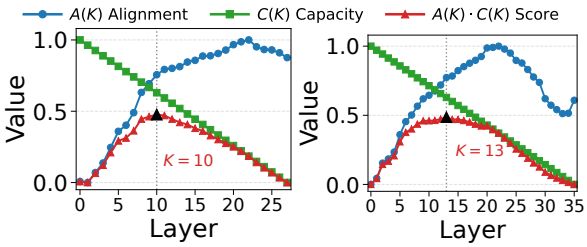
The "Without Deep ReAlign" bars in Figure 5 replace our Deep ReAlign module with the original ReAlign algorithm (Yu et al., 2026). This variant does not work for image-free visual unlearning: the model barely forgets the target classes. The result shows that our adaptation to LLM hidden states is effective.

4.5 Discussion

We briefly visualize the key components before unlearning, ACME and Deep ReAlign, and discuss their roles in the unlearning process.

4.5.1 Middle-Layer selection by ACME

Figure 6 reports the alignment, capacity, and ACME score across LLM layers, with both component metrics scaled to a common range for comparison. Capacity decreases almost linearly with depth, while alignment first follows a regular increasing trend under our metric and then declines in later layers; their product therefore peaks in the middle layers. The selected layers, Qwen2.5 layer 10 and Qwen3 layer 13, fall at these peaks, indicating that ACME identifies a balanced intervention point with sufficient update capacity and a smaller modality gap, thereby preparing the representation space for the subsequent unlearning training.

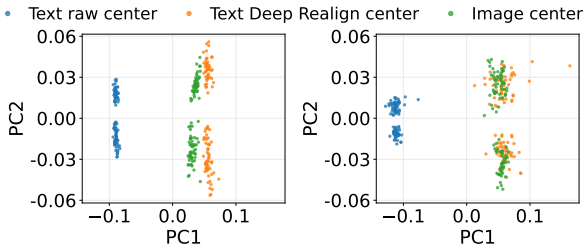


(a) Qwen2.5 (b) Qwen3

Figure 6: Layer-wise alignment, capacity, and their product score under ReAlign on Qwen2.5 and Qwen3.

4.5.2 Hidden-State Modality Gap Bridging

Figure 7 projects image features, raw text features, and Deep ReAligned text features at layer K^*. Raw text and image centers are initially separated, while Deep ReAlign moves text-derived centers toward the image-feature region while preserving class structure. This shows that after transformation, the textual signal becomes more image-like in the hidden space, enabling supervision over visual concepts. In this sense, the subsequent unlearning is effectively performed under visual supervision, even though only text descriptions are provided as input.



(a) Qwen2.5 (b) Qwen3

Figure 7: Class-center shifts after Deep ReAlign on Qwen2.5 layer 10 and Qwen3 layer 13.

The experimental results show that IF-CMU reduces forget-class recognition while avoiding the severe retain and utility collapse observed in aggressive baselines such as GA, NPO, and KL_Min. This pattern is important for the image-free setting: the method does not merely force low forget scores, but preserves enough non-target visual and QA behavior to make class-level unlearning usable after deployment.

5 Conclusion

We introduced IF-CMU, an image-free framework for class-level visual unlearning in generative vision-language models. IF-CMU uses only textual class descriptions to guide forgetting, retention, and anchoring without requiring target-class images or image–text pairs. By integrating ACME-based layer selection, Deep ReAlign pseudo-image construction, and a neutral-anchor unlearning objective, IF-CMU suppresses target-class visual recognition while preserving retain-class recognition, general utility, and textual knowledge. Experiments on Qwen2.5-VL and Qwen3-VL show that image-free visual concept unlearning is feasible when target-class images are unavailable or should not be retained.

Limitations

IF-CMU is evaluated on a focused set of generative vision-language models and class-level unlearning targets, so its behavior on much larger models, open-ended real-world target definitions, or safety-critical deployment settings may differ. The method also relies on text descriptions and internal cross-modal alignment to construct pseudo-image signals; when the target class is poorly described by text or weakly aligned across modalities, the unlearning signal may be less reliable. Finally, our evaluation emphasizes recognition, QA, and VQA retention, and does not cover every downstream multimodal behavior after unlearning.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Karuna Bhaila, Aneesh Komanduri, Minh-Hao Van, and Xintao Wu. 2025. Cross-modal attention guided unlearning in vision-language models. *arXiv preprint arXiv:2510.07567*.

BoyaWu10. 2024. Bunny-v1.1-data. https://huggingface.co/datasets/BoyaWu10/Bunny-v1_1-data. Hugging Face dataset repository.

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

Enrico Cassano, Riccardo Renzulli, Marco Nurisso, Mirko Zaffaroni, Alan Perotti, and Marco Grangetto. 2025. Saemnesia: Erasing concepts in diffusion models with supervised sparse autoencoders. *arXiv preprint arXiv:2509.21379*.

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

Yiwei Chen, Yuguang Yao, Yihua Zhang, Bingquan Shen, Gaowen Liu, and Sijia Liu. 2025b. Safety mirage: How spurious correlations undermine vlm safety fine-tuning and can be mitigated by machine unlearning. *arXiv preprint arXiv:2503.11832*.

Jiali Cheng and Hadi Amiri. 2024. Multidelete for multimodal machine unlearning. In *Computer Vision – ECCV 2024*, pages 165–184. Springer.

Xiaohua Feng, Jiaming Zhang, Fengyuan Yu, Chengye Wang, Li Zhang, Kaixiang Li, Yuyuan Li, Chaochao Chen, and Jianwei Yin. 2025. A survey on generative model unlearning: Fundamentals, taxonomy, evaluation, and future direction. *arXiv preprint arXiv:2507.19894*.

Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Xiyang Dai, Dongdong Chen, Chong Luo, and Lili Qiu. 2024. LLM2CLIP: Powerful language model unlocks richer visual representation. *arXiv preprint arXiv:2411.04997*.

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.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14389–14408. Association for Computational Linguistics.

Kodai Kawamura, Yuta Goto, Rintaro Yanagi, Hirokatsu Kataoka, and Go Irie. 2025. Approximate domain unlearning for vision-language models. In *Advances in Neural Information Processing Systems 39 (NeurIPS 2025)*.

Alexey Kravets and Vinay P. Namboodiri. 2025a. Zero-shot class unlearning in clip with synthetic samples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Alexey Kravets and Vinay P. Namboodiri. 2025b. Zero-shot clip class forgetting via text-image space adaptation. *Transactions on Machine Learning Research*, 2025.

Jiaqi Li and 1 others. 2024. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR.

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

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023b. DeCap: Decoding CLIP latents for zero-shot captioning via text-only training. In *International Conference on Learning Representations*.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 17612–17625.

Bo Liu, Qiang Liu, and Peter Stone. 2022. Continual learning and private unlearning. In *Proceedings of the 1st Conference on Lifelong Learning Agents (CoLLAs)*, volume 199 of *Proceedings of Machine Learning Research*, pages 243–254. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

Yang Liu, Xiaomin Yu, Gongyu Zhang, Zhen Zhu, Christos Bergeles, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. 2024. ArcSin: Adaptive ranged cosine similarity injected noise for language-driven visual tasks. *arXiv preprint arXiv:2402.17298*.

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 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4105–4135.

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

Yingzi Ma, Jiongxiao Wang, Fei Wang, Siyuan Ma, Jiazhao Li, Jinsheng Pan, Xiujun Li, Furong Huang, Lichao Sun, Bo Li, and 1 others. 2025. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *International Conference on Learning Representations*, volume 2025, pages 87806–87827.

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

Ashish Mishra, Tarun Kumar, Gyanaranjan Nayak, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. 2025. Erasing clip memories: Non-destructive, data-free zero-shot class unlearning in clip models. *arXiv preprint arXiv:2512.14137*.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. 2026. Same task, different circuits: Disentangling modality-specific mechanisms in VLMs. *Advances in Neural Information Processing Systems*, 38:66352–66385.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected clip. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.

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

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2024. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *arXiv preprint arXiv:2406.01983*.

Kun Wang, Guibin Zhang, Zhenhong Zhou, Jiahao Wu, Miao Yu, Shiqian Zhao, Chenlong Yin, Jinhu Fu, Yibo Yan, Hanjun Luo, and 1 others. 2025a. A comprehensive survey in llm (-agent) full stack safety: Data, training and deployment. *arXiv preprint arXiv:2504.15585*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zhaofeng Wu, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *International Conference on Learning Representations (ICLR)*, volume 2025, pages 53705–53723. ArXiv:2411.04986.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Can Yaras, Siyi Chen, Peng Wang, and Qing Qu. 2024. Explaining and mitigating the modality gap in contrastive multimodal learning. *arXiv preprint arXiv:2412.07909*.

Lingjie Yi, Raphael Douady, and Chao Chen. 2025. Decipher the modality gap in multimodal contrastive learning: From convergent representations to pairwise alignment. *arXiv preprint arXiv:2510.03268*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025a. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Xiaomin Yu, Pengxiang Ding, Wenjie Zhang, Siteng Huang, Songyang Gao, Chengwei Qin, Kejian Wu, Zhaoxin Fan, Ziyue Qiao, and Donglin Wang. 2025b. Unicorn: Text-only data synthesis for vision language model training. *arXiv preprint arXiv:2503.22655*.

Xiaomin Yu, Yi Xin, Wenjie Zhang, Chonghan Liu, Hanzhen Zhao, Xiaoxing Hu, Xinlei Yu, Ziyue Qiao, Hao Tang, Xue Yang, Xiaobin Hu, Chengwei Qin, Hui Xiong, Yu Qiao, and Shuicheng Yan. 2026. Modality gap-driven subspace alignment training paradigm for multimodal large language models. *arXiv preprint arXiv:2602.07026*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

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

Yuhui Zhang, Elaine Sui, and Serena Yeung. 2024b. Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. In *International Conference on Learning Representations*, volume 2024, pages 51711–51736.

Supplementary Material for

IF-CMU: Image-Free Class-Level Unlearning for Vision-Language Models

A Experimental Details

We conduct class-level visual unlearning on Qwen2.5-VL and Qwen3-VL. Each run selects one visual class as the forget target and evaluates both that class and a disjoint retain set. IF-CMU uses only textual descriptions for forget, retain, and anchor inputs during unlearning; no real target-class image is used to construct the forget signal. The vision encoder is not used for the image-free training inputs.

For each backbone, ACME selects the intervention layer used by Deep ReAlign. The ReAlign operator is fit from auxiliary image and text sets that are encoded independently rather than as paired target-class image–text samples. During unlearning, the text-derived hidden states are converted into pseudo-image hidden states at the selected layer, and the downstream LLM layers receive the unlearning update. The retain loss anchors non-target classes to their initial representations, and all reported numbers are averaged over the selected target classes.

Baselines use the same forget/retain class splits and are evaluated with the same prompts and aggregation code as IF-CMU. Vanilla denotes the original pretrained checkpoint without unlearning and is included as the retention and utility reference.

B Evaluation Metrics

For all image-conditioned metrics, the suffix "-F" means that input images are selected from the forget set, while "-R" means that input images are selected from the retain set.

Desc-F / Desc-R. Given an input image, the model first generates an image description, and we check whether the generated description still contains the ground-truth class. For example, for an ambulance image, the model may generate: "This is a photograph of a modern, white and orange electric scooter." Since the description no longer reveals the class ambulance, this forget-side sample is counted as forgotten. Desc-F is computed on forget images, and Desc-R on retain images.

MCQ-F / MCQ-R. Given an input image, we ask the model a four-way recognition question:

> What is the main object in this image?
> A. ambulance
> B. stinkhorn
> C. rock crab
> D. gasmask

The correct answer is A. If the model selects another option, the sample is incorrect. MCQ-F is evaluated on forget images, and MCQ-R on retain images.

YN-F / YN-R. Given an input image, we ask the model a yes/no recognition question:

> Is this an ambulance in this image?
> Answer: Yes or No.

Because the original image is an ambulance, the correct answer is Yes. If the model answers No, the sample is incorrect. YN-F is evaluated on forget images, and YN-R on retain images.

OBQA. OpenBookQA measures text-only general question answering:

> Question: Atomic 26 is drawn to a device, it could be
> A. magnetized
> B. Na
> C. compass
> D. K

The score is multiple-choice accuracy. Higher OBQA means better preserved language-side utility.

RWQA. RealWorldQA measures general visual question answering:

> Question: How many painted lines are on the road?
> Answer: 0

The score is visual QA accuracy. Higher RWQA means better preserved general visual utility.

Text-MCQ1 / Text-MCQ2. These text-only probes test whether the model can still identify the target class without image input:

> Description: An emergency medical vehicle equipped with stretchers and life-support equipment, used to transport sick or injured people to hospitals.
>
> A. pickup truck
> B. fire engine
> C. ambulance
> D. moped

The correct answer is C. High Text-MCQ scores mean that textual class recognition is preserved.

Text-Desc. This text-only probe asks the model to describe the target class directly:

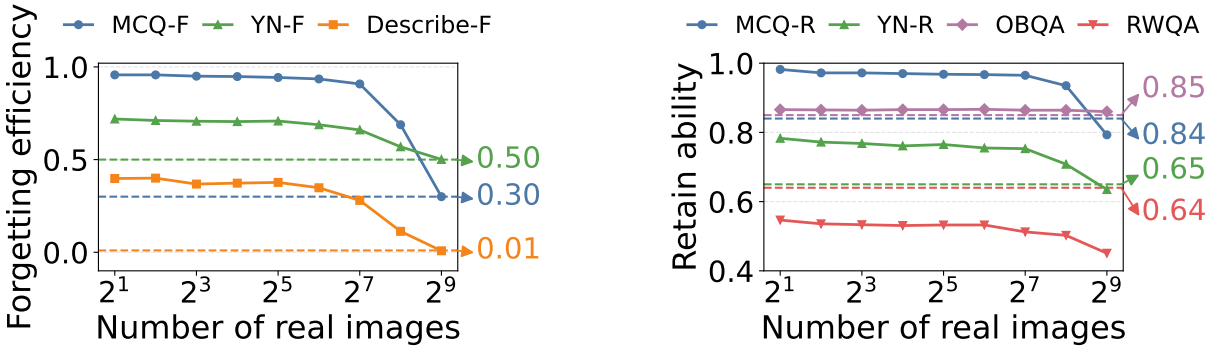
> What is an ambulance? Describe it in one or two sentences.

If the response still mentions key semantic attributes such as medical use, patients, emergency, or transport, it is counted as correct. High Text-Desc means the model retains semantic knowledge of the class in language.

Overall, forget-side visual metrics should decrease, retain-side and utility metrics should remain high, and text-only metrics should remain high if unlearning is modality-specific rather than a global erasure of the class concept.

B.1 Comparison with Real-Image Unlearning

A natural question is whether IF-CMU's image-free signal is merely a stopgap that any moderate number of real images would beat. Figure 8 answers this with a sweep over real-image budgets spanning multiple orders of magnitude. The two panels tell the same story from different sides. On the forgetting side, the curves stay essentially flat across the small- and medium-budget regime: adding more real images does not buy meaningfully stronger forgetting, and only when the budget reaches the high end of the range do we observe a visible shift, with describe- and Yes/No-style forgetting drifting toward the trivial-baseline floor. On the retain side, the curves are equally flat in the same regime and only start to bend downward when the real-image count is pushed very



(a) Forgetting efficiency (b) Retain ability

Figure 8: Effect of the number of real images on forgetting efficiency and retain ability.

Table 5: Text unlearning results.

Model	MCQ-F↓	MCQ-R↑	Desc-F↓	YN-F↓	YN-R↑	OBQA↑	RWQA↑
Qwen2.5	0.907	0.972	0.333	0.702	0.794	0.825	0.542
Qwen3	0.280	0.550	0.010	0.495	0.573	0.299	0.405

high, indicating that aggressive real-image supervision begins to damage retain classes and general utility. Crucially, the dashed reference lines mark our image-free operating point, and the curves remain at or near these references throughout the moderate-budget range. IF-CMU is therefore not a low-resource compromise but matches what supervised, image-based unlearning achieves until the budget grows large enough to start hurting retention. This robustness across budgets is the property that makes the image-free formulation practically attractive.

B.2 Details of Text unlearning results

Table 5 represents the full results of only using text supervison fot unlearning, without Deep Realign transfer.