Cross-Modal Attention Guided Unlearning in Vision-Language Models

Karuna Bhaila, Aneesh Komanduri, Minh-Hao Van, Xintao Wu
Department of Electrical Engineering and Computer Science
University of Arkansas
Fayetteville, AR 72701
{kbhaila,akomandu,haovan,xintaowu}@uark.edu

Abstract

The inference abilities of large-scale pretrained models are often attributed to the size of pre-training data collected across several domains. However, these models may memorize private and/or sensitive information during training and regurgitate it in inference. Recently, machine unlearning has been leveraged to address such leakage in LLMs. VLMs add a layer of complexity to this process, as the visual context in the query may additionally contain sensitive information. To address this issue, we explore unlearning for VLMs, specifically for the Visual Question Answering (VQA) task. We explore the role of visual tokens for output generation in VLMs using cross-modal attention and utilize it to formulate Cross-Modal Attention Guided Unlearning (CAGUL), a lightweight and efficient VLM unlearning framework. In contrast to computationally expensive model finetuning methods, CAGUL utilizes external modules to encode unlearning information in visual tokens of low importance for relevant queries. We find that the transformed visual tokens not only prevent leakage but also retain reference model behavior. Experimental results show that our method performs better or on par with finetuning-based baselines without altering the pre-trained model parameters or incurring retraining costs, making it a practical and effective VLM unlearning solution.

1 Introduction

Vision-Language Models (VLMs) are pre-trained on massive amounts of data sourced online and also finetuned on domain-specific data for downstream tasks. However, training data collected from such sources may contain personally identifiable and/or sensitive information, raising significant privacy concerns, especially with the added complexity of visual signals. For instance, processing visual inputs in VLMs can unintentionally disclose information such as location cues, identity of individuals in the background, etc [34], which, paired with the language model's knowledge, may output private data. The unconstrained use of these models in domains such as healthcare and finance can also be dangerous. In the LLM landscape, research efforts have been made to mitigate privacy leakage in different ways, including machine unlearning. Specifically, machine unlearning addresses concerns regarding leakage of training data as stipulated by the California Consumer Privacy Act (CCPA) and GDPR's *Right to be Forgotten* [4, 3].

Our focus here is on VLM unlearning under practical privacy considerations. In VLMs, sensitive/private information may be present in the vision component, language component, or both. So, visual signals should also be considered when formulating the unlearning problem. Therefore, we first explore unlearning in the context of VLMs, then formulate a realistic problem definition for VLM unlearning. We consider the setting of VQA on biographical data paired with images of individuals described in the data. The training data is thus in the form of paired image and text queries, such

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

that an image may be paired with multiple question-answer pairs about the corresponding individual. In this scenario, we propose framing VLM unlearning as the task of removing information from image-text pairs that can disclose private information while retaining knowledge about non-private pairs. For instance, a user may want to remove references to sensitive information (e.g., social security number) but may not necessarily need other general profile information to be removed (e.g., name).

Based on this problem definition, we propose a lightweight and efficient method to achieve VLM unlearning. Our method is motivated by the relationship between visual and text tokens represented as cross-modal attention scores computed between the two modes of input. Kaduri et al. empirically demonstrated that VLMs extract fine-grained details and attributes from visual tokens in a spatially localized manner and compress visual information into a small subset of highly attended tokens [13]. In turn, prompting a VLM with this compressed image context (5% of image tokens) can achieve performance close to that of prompting with all image tokens. Based on this observation, we conjecture that encoding information relevant to private or non-private queries in the least attended visual tokens can be useful for effective unlearning and propose **C**ross-Modal **A**ttention **G**uided **U**n**l**earning (**CAGUL**), a resource-efficient and interpretable method leveraging the cross-attention mechanism for unlearning in vision-language models.

Overall, our CAGUL framework is composed of two main components. First, a discriminator module detects whether or not a given visual input is paired with a query about private data. Then, for visual inputs predicted to be paired with private data, we use an MLP encoder to linearly transform the visual tokens with the lowest cross-modal attention scores. We employ standard unlearning losses defined separately for forget and retain data to train these external modules while keeping the pre-trained VLM parameters frozen. Intuitively, the encoder parameters learn to embed unlearning objectives into the embeddings of the transformed visual tokens. We empirically show that this approach obtains favorable trade-offs between forget and retain performance on FIUBench [20]. Further, we show that our results with a frozen VLM are comparable to those of finetuning-based baseline methods.

2 Preliminaries

Vision-Language Models. VLMs provide additional visual context to language models by incorporating a vision component in the LLM architecture and aligning the two modalities. Besides the language model, VLMs consist of two additional core components; a visual input processing module typically implemented as a vision transformer-based model which extracts rich patch-specific embeddings from an image/video input V and a cross-modal projector such as a pretrained CLIP-like encoder or a linear projection model, which transforms the patch embeddings to a visual embedding $\mathbf{Z}_v \in \mathbb{R}^{n_v \times d}$ in the language model's embedding space where n_v is the number of visual tokens and d is the embedding dimension. In the visual question-answering (VQA) task, given an image V and a query X with respective embeddings $\mathbf{Z}_v \in \mathbb{R}^{n_v \times d}$ and $\mathbf{Z}_q \in \mathbb{R}^{n_q \times d}$ where n_q is the number of query tokens, the output of the VLM is obtained as $Y = \text{LLM}(\mathbf{Z}_v, \mathbf{Z}_q)$.

Cross-modal Alignment. We refer to the attention scores between visual and text tokens as *cross-modal attention*. The attention mechanism [29] in this multi-modal setting consists of a query matrix $\mathbf{Q} = \mathbf{Z}_q \mathbf{W}_q$ and a key matrix $\mathbf{K} = \mathbf{Z}_v \mathbf{W}_k$ representing the textual and visual modalities, respectively. Formally, we can define attention weights as a function of query and key matrices:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\right) \tag{1}$$

In the context of VLMs, cross-modal attention is implemented as either *cross-attention* or *joint self-attention*. We discuss the differences in architecture in Appendix B.

3 Problem Setup

Suppose we have a dataset of m individuals denoted as $\mathcal{D} = \{(V_i, X_{i,j}, Y_{i,j}) \mid i \in M, j \in N_i\}$, where $M = \{1, 2, \ldots, m\}$ and $N_i = \{1, 2, \ldots, n_i\}$ denotes indices for the n_i question-answer pairs for individual i. Without loss of generality, we assume that n_i is the same across all individuals and denote it as n. However, in general, our setting also applies to a varying number of QA pairs per individual. The QA pairs for an individual contain information such as demographics and also some sensitive information, e.g., medical records. Assume a pre-trained VLM with parameters θ has

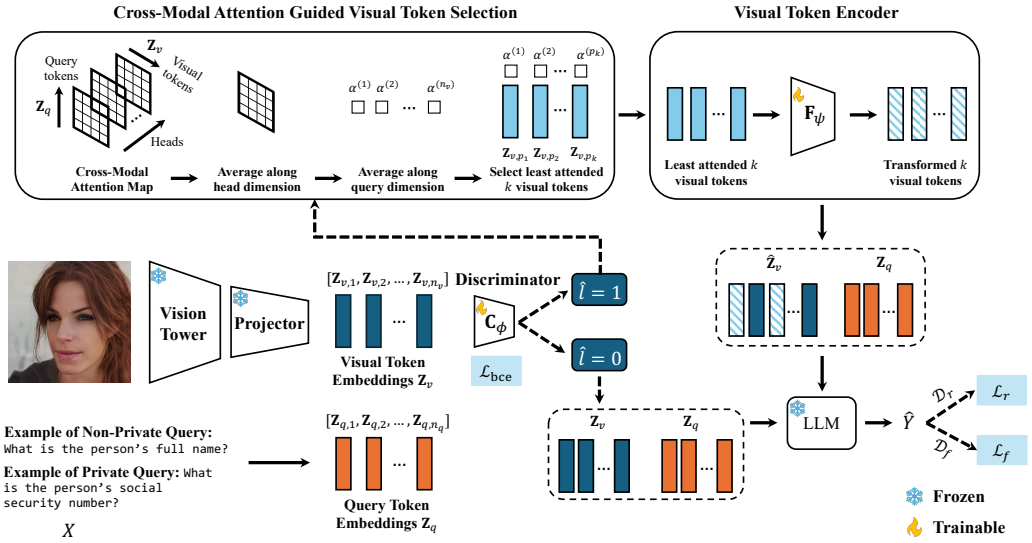

Figure 1: Cross-Modal Attention Guided Unlearning (CAGUL) Framework

been trained on dataset $\mathcal{D}$. In our unlearning setting, we assume a subset $\tilde{M} \subset M$ of individuals are concerned about the privacy of their sensitive information, but are indifferent to their non-sensitive data being used in model training. We formulate the VLM unlearning problem in this scenario as the forgetting of sensitive information corresponding to the individuals in $\tilde{M}$. Specifically, we define the *forget set* as $\mathcal{D}_f = \{(V_i, X_{i,j}, Y_{i,j}) \mid i \in \tilde{M}, j \in N_i, \text{ and } \texttt{is_private}(X_{i,j}) = 1\}$ consisting only of QA pairs for individuals in $\tilde{M}$ that contain sensitive information and are marked to be forgotten. The *retain set* can then be defined as $\mathcal{D}_r = \{(V_i, X_{i,j}, Y_{i,j}) \mid i \in M, j \in N_i, \text{ and } \texttt{is_private}(X_{i,j}) = 0\}$ consisting of all individuals with non-private QA pairs. The VLM unlearning goal in this work is to forget the subset $\mathcal{D}_f$ while retaining performance on $\mathcal{D}_r$. We follow commonly adopted unlearning setup where we assume a model has already been finetuned on the dataset, referred to as the *base model*, by maximizing the following log-likelihood objective:

$$\sum_{i=1}^{m} \sum_{j=1}^{n_i} \frac{1}{|Y_{i,j}|} \sum_{t=1}^{|Y_{i,j}|} \log p_\theta(Y_{i,j}^t | V_i, X_{i,j}, Y_{i,j}^{<t}) \tag{2}$$

where θ represents model parameters. Our goal is to facilitate unlearning from this base model.

4 Cross-Modal Attention Guided Unlearning

Here, we present **C**ross-Modal **A**ttention **G**uided **Un**learning (**CAGUL**) for unlearning in VLMs. CAGUL consists of three main components: a discriminator to determine if a given image appears in the forget set, a cross-modal attention-based visual token selection strategy, and a visual token encoder to embed unlearning information in unimportant visual tokens. For simplicity, we formulate our method for an example triple (V, X, Y). An overview of our framework is shown in Figure 1.

Discriminating Forget and Retain Images. Given the projected image tokens $\mathbf{Z}_v$ for a given image V, we train a simple discriminator $\mathbf{C}_\phi$ to determine whether the image appears in the forget set or not, formalized as $l = \mathbf{C}_\phi(\mathbf{Z}_v)$, where $l \in \{0, 1\}$, and $l = 1$ implies that the discriminator predicts that the image corresponding to visual tokens $\mathbf{Z}_v$ appears in the forget set.

Cross-Modal Attention Guided Visual Token Selection. The main objectives of unlearning are to remove a trained model's knowledge of $\mathcal{D}_f$ while maintaining its predictive utility on $\mathcal{D}_r$. Assuming that a classifier can accurately distinguish between images appearing in the forget set and images only appearing in the retain set, an intuitive way of unlearning is to add noise to images in the forget set. This results in distorted vision embeddings such that the language model is unable to get accurate signals from the image tokens. However, in our setting, the images in the forget set provide signals to

corresponding private and non-private queries, i.e., both $\mathcal{D}_f$ and $\mathcal{D}_r$, and adding noise randomly may achieve good forgetting, but likely degrades utility.

Instead of indiscriminately adding noise to the image tokens, we propose a visual token selection mechanism guided by cross-modal attention between the visual and query text tokens. We hypothesize that *selectively encoding the unlearning objectives in visual tokens corresponding to images in the forget set can guide the language model in appropriately generating its response for private and non-private queries associated with the respective images*. We are motivated by the analysis of visual information processing in VLMs conducted by [13], which empirically demonstrates that VLMs extract fine-grained details and attributes from visual tokens in a spatially localized manner and compress visual information into a small subset of highly attended tokens. Consequently, prompting a VLM with this compressed image context (5% of image tokens) can achieve performance close to that of prompting with all image tokens.

In either cross-attention or joint self-attention based models, we use the attention matrix $\mathbf{A} \in \mathbb{R}^{n_q \times n_v}$ to determine the lowest attended image tokens with respect to the input text. We average the attention scores across all query tokens across all attention heads to get a scalar score

$$\alpha = \frac{1}{n_q n_h} \sum_{h=1}^{n_h} \sum_{j=1}^{n_q} \mathbf{A}_j^{(h)} \tag{3}$$

where $\alpha \in \mathbb{R}^{1 \times n_v}$ represents the attention score between the entire query and each image token, $\mathbf{A}_j^{(h)}$ represents the jth row of the attention matrix corresponding to attention head h, and n_h is the number of attention heads. Then, for each attention layer, we choose the bottom-k tokens with the lowest attention score with the query $\{Z_{v,p_1}, Z_{v,p_2}, \ldots, Z_{v,p_k}\} \subset \mathbf{Z}_v$ where $K = \{p_1, p_2, \ldots, p_k\}$ are the indices of the k least attended image tokens. We observe that such tokens often correspond to arbitrary pixels in regions not contextually critical to the text prompt.

Visual Token Encoder. Our method relies upon the intuition that we can encode certain information about the unlearning objective in image tokens having low correlation with the query through a learned transformation to facilitate unlearning. Thus, for samples predicted by the discriminator as $l = 1$ (i.e., image appears in the forget set), we learn a parameterized visual token encoder $\mathbf{F}_\psi(\cdot)$ that transforms the least attended k visual embeddings to perturbed embeddings. We then replace the k image token embeddings in $\mathbf{Z}_v$ with the transformed representation as follows:

$$\hat{Z}_{v,i} = \begin{cases} \mathbf{F}_\psi(Z_{v,i}) & \text{if } i \in K \\ Z_{v,i} & \text{otherwise} \end{cases} \tag{4}$$

where $\hat{\mathbf{Z}}_v$ is the set of final visual token embeddings. The final input to the LLM is obtained as $Y = \text{LLM}(\mathbf{Z}_v, \mathbf{Z}_q)$.

Unlearning Objective. Our CAGUL framework has two trainable modules, discriminator $\mathbf{C}_\phi$ and visual token encoder $\mathbf{F}_\psi$. The discriminator is trained with the following classification objective

$$\mathcal{L}_{\text{bce}} = \mathbb{E}_{\mathbf{Z}_v} \left[\log p_{\mathbf{C}_\phi}(l|\mathbf{Z}_v) \right] \tag{5}$$

The encoder is trained with unlearning objectives defined on the forget and retain sets separately. Generally, the unlearning objective for retain samples is to maintain the base model's predictive utility. In CAGUL, we define this objective as the standard causal language modeling loss for all samples in $\mathcal{D}_r$, including the non-private samples for images appearing in the forget set formulated as

$$\mathcal{L}_r = \mathbb{E}_{(V,X,Y)\sim\mathcal{D}_r} \mathbb{E}_{Y^t} \left[\log p_\theta(Y^t|V, X, Y^{<t}) \right] \tag{6}$$

In the VQA task, the forget objective can be interpreted as obtaining low utility for samples in the forget set $\mathcal{D}_f$. We leverage Preference Optimization (PO) [21] to realize this objective in CAGUL by aligning the language model's outputs for forget queries with refusal responses instead of the ground truth. We compute the standard causal language modeling loss with the substituted response as

$$\mathcal{L}_f = \mathbb{E}_{(V,X,Y)\sim\mathcal{D}_f} \mathbb{E}_{\bar{Y}^t} \left[\log p_\theta(\bar{Y}^t|V, X, \bar{Y}^{<t}) \right] \tag{7}$$

where $\bar{Y}$ is a refusal response such as "I cannot answer this question." Finally, we formulate the joint objective for CAGUL as

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \mathcal{L}_f + \mathcal{L}_r \tag{8}$$

where $\mathcal{L}_{\text{bce}}$ is the binary classification loss for the discriminator computed over all samples, $\mathcal{L}_f$ is the PO loss computed for samples in $\mathcal{D}_f$ with refusal answers, and $\mathcal{L}_r$ is the GD loss computed for samples in $\mathcal{D}_r$ with their ground-truth answers.

5 Experiments

Dataset. We primarily investigate VLM unlearning on the FIUBench dataset [20], a biographical dataset containing 8000 VQA pairs. The dataset comprises 400 unique images, each paired with 20 QA pairs. [20] investigated an unlearning setting on FIUBench where $|\tilde{M}|$ individuals request that all their information be removed from training data. In contrast, we focus on a more realistic and complex unlearning problem, emulating real-world scenarios where information about individuals is consolidated from different sources, and removal requests may only apply to some sources. We formulate a setting where $|\tilde{M}|$ individuals wish to remove their sensitive data from training, such that only the VQA pairs about health and criminal records are to be forgotten. In FIUBench, this results in $\mathcal{D}_f$ and $\mathcal{D}_r$, i.e., forget and retain sets with overlapping images, unlike the original setting where the images could be differentiated as forget or retain. We introduce additional notations here to represent this scenario to facilitate the discussion of experimental results. For the $|\tilde{M}|$ individuals requesting removal, we divide their corresponding VQA pairs into two subsets: $\mathcal{D}_p$, which contains health and criminal records queries to be forgotten, and $\mathcal{D}_{np}$, which contains the non-sensitive queries. The private set forms the forget set, $\mathcal{D}_p = \mathcal{D}_f$, and the non-private set is included in the retain set, $\mathcal{D}_{np} \subset \mathcal{D}_r$. All VQA pairs for the remaining individuals are included in $\mathcal{D}_r$.

Setup. We run experiments with LLaMA-3.2-11B-Vision-Instruct and Qwen-2.5-VL-7B-Instruct models and report main results with the former. We first finetune the VLM on FIUBench dataset to ensure sufficient memorization and aim to achieve unlearning on this finetuned model, also referred to as the base model. We assume $|\tilde{M}|$ = 40 individuals request sensitive information removal and report results when transforming k = 200 least attended visual tokens with CAGUL trained for 12 total epochs. We provide specific details about experimental setup in Appendix C.

Baselines. We demonstrate the effectiveness of CAGUL by evaluating it against multiple baselines. Gradient Ascent (GA) [32] implements finetuning while maximizing the loss for samples in the forget set; the retain set is not utilized in training. Gradient Difference (GD) [31] additionally defines a gradient descent loss on the retain set besides the GA loss on the forget set to prevent degradation of model utility due to loss maximization. Similarly, KL Minimization (KL) [31] combines GA on forget with a KL divergence loss between the output distribution of the unlearned model and the target model. Preference Optimization (PO) [21] defines gradient descent loss for both retain and forget sets, but augments the labels in the forget set with preferred answers like "I am unable to answer" to guide the model towards refusal response for forget examples. Furthermore, we implement model retraining by finetuning the pre-trained VLM on only the retain set as an ideal baseline. We implement early stopping for methods using GA loss to prevent rapid model degradation.

Evaluation Metrics. We follow [20] and report multiple metrics to quantify forget and retain performances measured on $\mathcal{D}_f$ and $\mathcal{D}_r$, respectively. For $\mathcal{D}_f$, we compute Rouge-L and Exact Match (EM) to measure output correctness, Adversarial Privacy Extraction (APE) for robustness against paraphrased queries, and MinK% Prob [26] for unlearning effectiveness under Membership Inference Attack (MIA). APE is computed as EM values when the model is prompted with multiple paraphrased queries of each forget sample, and MinK% is computed based on the output probability of the lowest likely tokens among generated tokens. For $\mathcal{D}_r$, we report Rouge-L to show utility. As our unlearning scenario contains overlapping images between $\mathcal{D}_f$ and $\mathcal{D}_r$, we also perform a fine-grained evaluation of non-private QA pairs of the overlapping images, i.e., on $\mathcal{D}_{np}$ and report Rouge-L and EM for correctness, and Truth Ratio (TR) to quantify the model's tendency to generate incorrect answers. We further report accuracy scores on MME [10] and POPE [16] datasets to show the influence of unlearning on model generalizability. We provide detailed descriptions of all metrics in Appendix D.

Experimental Results. We include the main results from our experiments in Table 1 for LLaMA-3.2-11B-Vision-Instruct. Here, Pretrain refers to the VLM used as is for inference, and Finetune is the VLM trained on FIUBench to ensure adequate memorization of dataset-specific information. The increase in performance metrics from Pretrain to Finetune across all subsets of $\mathcal{D}$ indicates that the finetuned model has successfully memorized information from the dataset. However, we observe a significant decline in the model's generalization ability as the VLM is finetuned only on FIUBench. Nonetheless, we use the finetuned model as the base model for unlearning methods and its evaluation.

Forget Quality. CAGUL Rouge and EM scores on $\mathcal{D}_f$ demonstrate effective forgetting as noted by the significant drops in these metrics compared to Finetune. The MIA score MinK is also significantly lower than Finetune, which indicates that unlearned models achieved with CAGUL can successfully

Table 1: Unlearning metrics of CAGUL compared with baselines on LLaMA-3.2-11B-Vision-Instruct

Method	Forget				Retain				General	
	$\mathcal{D}_f$				$\mathcal{D}_r$	$\mathcal{D}_{np}$			MME	POPE
	Rouge(↓)	EM(↓)	APE(↓)	MinK(↓)	Rouge(↑)	Rouge(↑)	EM(↑)	TR(↑)	Acc.(↑)	Acc.(↑)
Pretrain	26.00	0.00	0.00	0.79	19.13	12.63	0.10	7.63	77.34	87.41
Finetune	57.33	39.68	24.59	43.32	84.55	66.88	30.69	50.63	30.58	37.70
Retrain	55.91	15.75	11.45	16.84	92.24	77.89	47.12	58.85	23.50	29.20
GA	46.83	12.04	4.44	7.31	59.71	47.65	7.74	88.31	22.96	20.08
GA+GD	46.15	9.07	5.79	6.23	95.50	83.35	74.21	58.27	6.40	5.23
GA+KL	24.31	6.16	4.48	5.71	77.43	30.55	9.33	92.11	11.20	9.87
PO+GD	30.62	0.37	0.34	4.57	91.23	84.26	61.44	56.23	18.58	45.54
CAGUL	30.84	1.70	0.43	14.86	85.15	84.32	64.35	91.67	30.58	37.70

unlearn information from $\mathcal{D}_f$. Similarly, the adversarial prompting score APE being low demonstrates that CAGUL is effective against jailbreak attacks that rephrase input prompts, aiming to extract information. Furthermore, compared to the Retrain baseline, CAGUL achieves lower values for all forget metrics. Compared to finetuning baselines that implement GA as the forget loss, CAGUL generally achieves lower Rouge, EM, and APE scores. For the finetuning baseline that implements the same PO+GD loss function as CAGUL, our method achieves comparable performance on most metrics despite keeping pre-trained parameters frozen and training smaller external modules.

Retain Performance. We report retain metrics on the full retain set $\mathcal{D}_r$ and the non-private subset $\mathcal{D}_{np}$ corresponding to the individuals in $\mathcal{D}_f$. We observe that CAGUL successfully preserves the retain utility of the finetuned base model. For the subset $\mathcal{D}_{np}$, CAGUL retains and improves knowledge about the non-private data while simultaneously forgetting the private information of the same individuals. The increase in $\mathcal{D}_{np}$ metrics can be attributed to the further memorization of retain knowledge during training of unlearning methods. Additionally, CAGUL generally outperforms baseline unlearning methods, including Retrain, which suggests that the information encoded in visual tokens significantly helps to preserve retain performance.

Downstream Accuracy. This metric quantifies the influence of unlearning on the model's original world knowledge. Our results show that CAGUL retains the performance of the finetuned base model on both MME and POPE datasets, whereas finetuning baselines significantly degrade in performance. We emphasize that our method does not require finetuning any component of the VLM, which ensures that the pretrained knowledge for general tasks is not manipulated, provided that the classifier can accurately distinguish images in the forget set from retain or general evaluation sets.

Additional Results. We record the number of training parameters and execution time for all implemented unlearning methods, which are included in Appendix E along with the results on Qwen-2.5-VL-Instruct due to space constraints. Results show that CAGUL performance is comparable to retraining while preserving the finetuned model's downstream utility on MME and POPE. Furthermore, CAGUL updates fewer parameters compared to finetuning baselines. Additionally, we conduct an ablation study to understand the importance of the classifier and encoder components, the cross-modal attention selection strategy, and the forget and retain loss functions in CAGUL architecture. We also study the influence of CAGUL hyperparameters, i.e., the number of visual tokens transformed by $\mathbf{F}_\psi$ and the size of the forget set $\mathcal{D}_f$. These results and their discussions are also included in Appendix E. Overall, our empirical results show that CAGUL achieves desirable trade-offs between forget quality and utility comparable to baseline finetuning methods while saving on computational efficiency due to the cross-attention-based visual token transformation technique.

6 Conclusion

In this work, we propose a novel setting for unlearning in vision-language models where each user can exercise the *Right to be Forgotten* for any subset of their sensitive queries. We formulate an unlearning framework comprising a classifier to identify images in the forget set, a cross-modal attention-based visual token selection mechanism, and a visual token encoder to embed unlearning information in visual tokens with low correlation to the query prompt. We conducted experiments on the FIUBench dataset using two open-source vision-language models and demonstrated that our method achieves desirable trade-offs in performance compared to finetuning-based baseline methods.

Acknowledgements

This work is supported in part by the National Science Foundation under awards 1920920, 1946391, 2119691, 2147375, the National Institute of General Medical Sciences of the National Institutes of Health under award P20GM139768, and the Arkansas Integrative Metabolic Research Center at the University of Arkansas. This research is also supported by the Arkansas High Performance Computing Center which is funded through multiple National Science Foundation grants and the Arkansas Economic Development Commission.

References

[1] David Beniaguev. Synthetic faces high quality (sfhq) dataset, 2022.

[2] Karuna Bhaila, Minh-Hao Van, and Xintao Wu. Soft prompting for unlearning in large language models. 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)*, 2025.

[3] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

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

[5] Trishna Chakraborty, Erfan Shayegani, Zikui Cai, Nael B. Abu-Ghazaleh, M. Salman Asif, Yue Dong, Amit Roy-Chowdhury, and Chengyu Song. Can textual unlearning solve cross-modality safety alignment? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

[6] Jiaao Chen and Diyi Yang. Unlearn what you want to forget: Efficient unlearning for LLMs. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[7] Jiali Cheng and Hadi Amiri. Multimodal machine unlearning, 2023.

[8] Chenlu Ding, Jiancan Wu, Yancheng Yuan, Jinda Lu, Kai Zhang, Alex Su, Xiang Wang, and Xiangnan He. Unified parameter-efficient unlearning for LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[9] Daniele Faraglia. Faker, 2024.

[10] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

[11] Lei Gao, Yue Niu, Tingting Tang, Salman Avestimehr, and Murali Annavaram. Ethos: Rectifying language models in orthogonal parameter space. *arXiv preprint arXiv:2403.08994*, 2024.

[12] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts, 2023.

[13] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? A deep-dive into the vision of vision language models. *CoRR*, abs/2411.17491, 2024.

[14] Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

[15] Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, et al. The wmdp benchmark: Measuring and reducing malicious use with unlearning. *arXiv preprint arXiv:2403.03218*, 2024.

[16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics, 2023.

[17] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024.

[18] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models, 2024.

[19] Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with mllmu-bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 4105–4135. Association for Computational Linguistics, 2025.

[20] 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. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *The Thirteenth International Conference on Learning Representations*, 2025.

[21] Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. *First Conference On Language Modeling*, 2024.

[22] Bruna Mendes. Nypd crime complaint data historic (2006-2019), 2020.

[23] Prasad Patil. Healthcare dataset, 2024.

[24] Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few-shot unlearners. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

[25] Laurens Samson, Nimrod Barazani, Sennay Ghebreab, and Yuki M. Asano. Privacy-aware visual language models, 2024.

[26] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models, 2023.

[27] Pratiksha Thaker, Yash Maurya, and Virginia Smith. Guardrail baselines for unlearning in LLMs. *arXiv preprint arXiv:2403.03329*, 2024.

[28] Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models, 2024.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[30] Vyas. Usa housing, 2017.

[31] Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. Machine unlearning of pre-trained large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

[32] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *arXiv preprint arXiv:2310.10683*, 2023.

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

[34] Ziyang Zhang, Chong Bao, Xiaokun Pan, Chia-Ming Chang, Takeo Igarashi, and Guofeng Zhang. Through the lens of privacy: Exploring privacy protection in vision-language model interactions on smart glasses. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA 2025, Yokohama, Japan, 26 April 2025- 1 May 2025*, pages 530:1–530:8. ACM, 2025.

A Related Work

LLM Unlearning. The widespread use of LLMs has raised significant privacy concerns about user data ingested during the training process. As a result, recent research has focused on developing methods for unlearning in LLMs to address issues regarding privacy as well as bias, harm, and toxicity [17]. Early methods generally borrow from traditional machine unlearning [4, 3] and implement loss-based optimization techniques for forgetting, paired with varying alignment objectives for retaining model usability [32, 21, 6, 33]. Others utilize adaptation techniques specific to LLMs, such as achieving unlearning through in-context learning [24], manually crafted system prompts [27], or prompt tuning with unlearning-specific losses to guide generation [2]. Additionally, some works focus on model editing methods using localization-based objectives to remove unwanted knowledge [15, 8, 11].

VLM Unlearning. VLM unlearning is a relatively new frontier with increased complexity due to the integration of visual signals. [7] propose a framework for multimodal unlearning in VLMs by decoupling the text and visual modalities. To understand the privacy vulnerabilities of VLMs, [12] take an adversarial approach and develop a blackbox visual jailbreak prompting strategy. Similar to this work, [18] study how image-based manipulations can cause breaches in safety-aligned VLMs. [5] investigate whether text-only unlearning is sufficient for safety alignment in multimodal models. There have also been several studies benchmarking VLM unlearning. [28] conduct an empirical study showing that VLMs can infer private attributes from images even when the attributes do not stem from visual depiction of humans. [25] benchmark several VLMs to understand their limitations in visual privacy and propose an instruction-tuning dataset to improve privacy sensitivity of VLMs. [14] formulate unlearning as forgetting visual recognition of target concepts in images and propose a benchmark dataset for visual concept unlearning. [20] construct a VQA dataset, FIUBench, to benchmark VLM unlearning using several optimization-based techniques from LLM unlearning. [19] introduce a VQA dataset with multiple images and QA pairs and evaluate standard unlearning methods with several VLMs.

B Architecture Details

Vision-Language Models. We run our experiments using two state-of-the-art VLMs: LLaMA-3.2-11B-Vision-Instruct and Qwen-2.5-VL-7B-Instruct. The cross-modal alignment mechanism in VLMs is generally implemented as either *cross-attention* or *joint self-attention*. For cross-attention based models like LLaMA-3.2-11B-Vision-Instruct, the query matrix $\mathbf{Q}$ and key matrix $\mathbf{K}$ represent the two modalities separately, so $\mathbf{A}$ directly captures the attention scores between the two modalities. In this scenario, each text token embedding is updated based on cross-attention scores computed *before* input to the language model. However, for joint self-attention-based models such as Qwen-2.5-VL-7B-Instruct, $\mathbf{Z} = (\mathbf{Z}_v, \mathbf{Z}_q)$ is fed in jointly as one concatenated input sequence to the language model, where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$ and $\mathbf{K} = \mathbf{Z}\mathbf{W}_k$ are shared among all tokens. The self-attention scores can be represented by the following block matrix:

$$\mathbf{A}_{\text{self-attn}} = \begin{bmatrix} \mathbf{A}_{vv} & \mathbf{A}_{vq} \\ \mathbf{A}_{qv} & \mathbf{A}_{qq} \end{bmatrix}, \tag{9}$$

assuming the input is composed of an image followed by the query text for simplicity. For our formulation, we can consider $\mathbf{A}$ for joint self-attention models to be the top right block $\mathbf{A}_{vq}$ derived from $\mathbf{Q} = \mathbf{Z}_q\mathbf{W}_q$ and $\mathbf{K} = \mathbf{Z}_v\mathbf{W}_k$, which represents the cross-modal attention scores between visual and text tokens. The cross-modal attention scores can be similarly extracted from the overall attention matrix based on the location of the image and query text in the input sequence.

Table 2: Dataset statistics for varying $|\tilde{M}|$

$\|\tilde{M}\|$	**20**	**40**	**60**	**80**
$\mathcal{D}_f$	273	544	817	1078
$\mathcal{D}_r$	7727	7456	7183	6922
$\mathcal{D}_{np}$	127	256	383	522

Encoder Architectures. We implement the discriminator $\mathbf{C}_\phi$ using a convolutional neural network and the visual token encoder $\mathbf{F}_\psi$ as a one-layer multilayer perceptron (MLP).

Table 3: Training hyperparameters for finetuning and unlearning methods; parameters not applicable to method are shown as -

Hyperparameters	**Finetune/Retrain**	**GA**	**GA+GD**	**GA+KL**	**PO+GD**	**CAGUL**
Learning rate	2×10^{-5}	1×10^{-6}	2×10^{-5}	1×10^{-4}	3×10^{-4}	2×10^{-5}
Batch size	4	4				4
Epochs	10	10				$2 + 10$
Dropout	-	0.05				-
LoRA Rank r	-	128				-
LoRA Alpha α	-	256				-

C Experimental Setup

Dataset. We primarily investigate VLM unlearning on the FIUBench dataset [20] composed of 400 unique synthetic images obtained from the SFHQ dataset [1], each paired with fictitious biographical information including name, birthdate, address, phone number, occupation, income, health, and criminal records sourced from [23, 22, 30, 9]. The dataset is formatted for VQA by extracting 20 QA pairs from the biographical data to form 8000 VQA pairs, which is represented by $\mathcal{D}$ in this work. In our experiments, we select individuals for the forget set based on the splits provided in [20]. The resulting dataset statistics in terms of the number of samples in the forget set, retain set, and non-private set are presented in Table 2 for a varying number of $|\tilde{M}|$ used in our experiments.

Setup. As is the norm in unlearning literature for pre-trained models, we first finetune VLMs on the entire FIUBench dataset $\mathcal{D}$ to ensure data memorization. In this work, we conduct experiments with two VLMs: LLaMA-3.2-11B-Vision-Instruct and Qwen-2.5-VL-7B-Instruct. We implement full finetuning of the language and cross-modal components for both VLMs, i.e., the language model and the projector module for LLaMA, and the language model and visual merger module for Qwen. We refer to this finetuned VLM as the base model for unlearning. The objective here is to obtain an unlearned model which demonstrates forgetting of the personal information in $\mathcal{D}_f$ while maintaining the base model's performance for samples in $\mathcal{D}_r$, including $\mathcal{D}_{np}$.

We assume that $|\tilde{M}|$ = 40 individuals request removal of their private information. We implement LoRA finetuning for baseline methods except retraining. For CAGUL, we train only the discriminator and visual token encoder while keeping the base model frozen as discussed in Section 4. More specifically, we first train the discriminator for 2 epochs, then jointly train the two modules and report results with perturbations performed on k = 200 least attended visual tokens out of 6404 for LLaMA and k = 20 visual tokens out of 1369 tokens for Qwen. We further report trends over CAGUL's performance when varying k from 100 through 1000, and $|\tilde{M}|$ as {20, 40, 60, 80}. To demonstrate the contribution of the discriminator and encoder components in CAGUL, we conduct ablation studies where we replace/remove components during training. We also conduct experiments to show CAGUL's performance when the PO+GD loss is substituted with other unlearning loss variations. Specific training hyperparameters are included in Table 3. We use 4 NVIDIA A100 GPUs with 40GB RAM for implementations using LLaMA and 4 NVIDIA H100 with 150GB RAM for Qwen and report performance over a single run due to the resource-intensive nature of the experiments. We utilize the Huggingface library to implement all methods.

D Evaluation Metrics

D.1 Forget Metrics

Rouge-L. We compute ROUGE-L scores to measure the similarity between generated text and ground truth answers for samples in $\mathcal{D}_f$.

Exact Match. Following [20], we compute Exact Match (EM) scores to quantify the correctness of generated outputs compared to the ground truth labels. We obtain EM as an average over the ratio of ground-truth keywords appearing in the generated text for each query. For $\mathcal{D}_f$, a lower EM is desirable.

MinK. Membership Inference Attacks (MIA) are often used to evaluate the forgetting effectiveness of unlearning methods. We leverage Min-K% Prob [26] to quantify the presence of knowledge from $\mathcal{D}_f$ in the unlearned model. To compute this metric, we first obtain the probability for each generated token and calculate the average log likelihood over the first K% tokens with minimum probabilities. A lower average log likelihood indicates that $\mathcal{D}_f$ was not included in the training data, demonstrating effective forgetting.

Adversarial Privacy Extraction. Safety regulations in pre-trained models can often be bypassed by rephrasing query texts. The Adversarial Privacy Extraction (APE) was formulated to verify whether forgotten knowledge can be extracted from an unlearned model in an adversarial manner by prompting it with paraphrased queries [20]. We compute this metric as the average EM score when the unlearned model is queried with multiple paraphrases of each forget sample.

D.2 Retain Metrics

Rouge-L. We compute ROUGE-L scores over the entire retain set $\mathcal{D}_r$. The retain objective is to preserve the base model's performance of this metric. For a thorough evaluation of forget and retain performance trade-off for the $\tilde{M}$ individuals requesting removal, we also provide the ROUGE-L scores for the retain subset $\mathcal{D}_{np}$.

Exact Match. For $\mathcal{D}_{np}$, we compute an EM score to evaluate whether the generated outputs from the unlearned model contain keywords specific to the ground truth to quantify CAGUL's ability to preserve utility.

Truth Ratio. We follow [20] and compute Truth Ratio (TR), which measures the model's tendency to generate factually incorrect answers versus correct ones. The likelihood of factual generation is computed as the probability of a paraphrased version of the ground truth answer, and the likelihood of an incorrect answer is calculated as an average over the probabilities of multiple perturbed answers formatted like the paraphrased answer. TR is reported as the ratio of incorrect to factual likelihoods.

D.3 General Downstream Accuracy

We report accuracy scores on two general VQA datasets: MME [10] and POPE [16] to demonstrate the influence of unlearning methods on the model's generalization ability. The MME benchmark is composed of queries related to various tasks such as existence, count, position, color, posters, celebrities, scenes, landmarks, and artworks. The POPE benchmark quantifies object hallucinations in VLMs, and we evaluate on the adversarial, popular, and random tasks.

D.4 Efficiency

We additionally report per-epoch training time required by each unlearning method as well as the number of parameters that are updated during training to compare the computational efficiency of the implemented methods.

Table 4: Computational efficiency of CAGUL compared with baseline unlearning methods for LLaMA-3.2-11B-Vision-Instruct

Method	Trainable Params	Execution time (s)
Retrain	9.8B	4272
GA	419M	149
GA+GD	419M	1855
GA+KL	419M	2951
PO+GD	419M	1911
CAGUL	293M	682

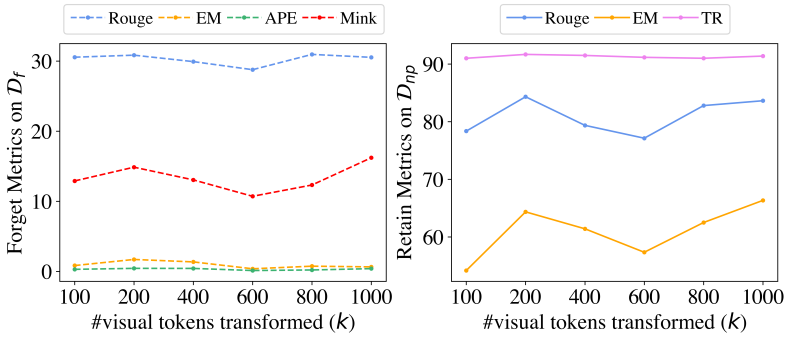

Figure 2: Influence of number of visual tokens transformed

E Additional Results

E.1 Computational Efficiency.

We report the number of training parameters and per-epoch execution times in Table 4 for LLaMA-3.2-11B-Vision-Instruct to analyze the training complexity of unlearning methods compared to retraining. The baselines require fewer parameters than Retrain, as we implement LoRA finetuning for these methods as opposed to the full finetuning used for Retrain. Nonetheless, CAGUL trains the fewest number of parameters as the pre-trained VLM parameters are kept frozen in this framework and completes training in significantly less time. We note that GA requires the least training time as optimization is performed with respect to $\mathcal{D}_f$ only, but it causes the model to degrade quickly as indicated by the low scores on $\mathcal{D}_r$ and $\mathcal{D}_{np}$ in Table 2. We observe similar proportions for trainable parameters and training execution time with Qwen-2.5-VL-7B-Instruct.

E.2 Influence of Number of Visual Tokens Transformed

In CAGUL, we select the k least attended visual tokens to encode unlearning-specific objectives via linear transformation using $\mathbf{F}_\psi$. Here, we study the influence of the number of visual tokens transformed on the overall performance of CAGUL by varying k as $\{100, 200, 400, 600, 800, 1000\}$. We report CAGUL performance on the forget set $\mathcal{D}_f$ and non-private retain set $\mathcal{D}_{np}$ in Fig. 2. We omit the results on $\mathcal{D}_r$, MME, and POPE as all values of k achieve the same metrics on these sets. Our results show peak performance for both forget and retain metrics when k is set to 200 tokens. We observe a minor dip in performance when transforming the k = 600 tokens, but CAGUL generally achieves similar performance across different k values.

E.3 Influence of Forget Set Size

We also study the effect of the size of forget set $\mathcal{D}_f$, which relates to the number of individuals participating in $\tilde{M}$. We vary $|\tilde{M}|$ across $\{20, 40, 60, 80\}$ with corresponding dataset statistics shown in Table 2 and report forget and retain metrics in Fig. 3. We observe a decreasing trend for the forget metrics on $\mathcal{D}_f$ and an increasing trend for the retain performance on $\mathcal{D}_{np}$ as $|\tilde{M}|$ becomes larger. This result suggests that unlearning with a larger forget set size facilitates more effective forgetting of

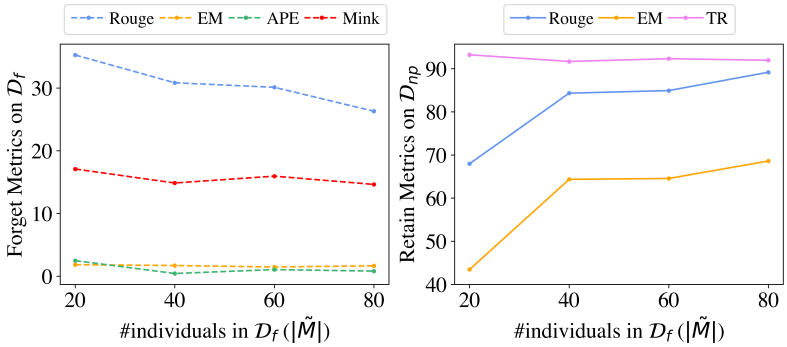

Figure 3: Influence of forget set size

private data and retaining of non-private data, most likely due to the increased size of the training subset used to learn visual token encoder weights.

Table 5: Ablation study

Method	Forget				Retain				General	
	$\mathcal{D}_f$				$\mathcal{D}_r$	$\mathcal{D}_{np}$			MME	POPE
	Rouge(↓)	EM(↓)	APE(↓)	MinK(↓)	Rouge(↑)	Rouge(↑)	EM(↑)	TR(↑)	Acc.(↑)	Acc.(↑)
CAGUL	30.84	1.70	0.43	14.86	85.15	84.32	64.35	91.67	30.58	37.70
w/o $\mathbf{C}_\phi$	30.94	1.84	1.28	22.93	96.07	84.66	65.94	92.21	28.94	37.41
w/o $\mathbf{F}_\psi$	56.56	37.75	23.67	41.54	84.50	65.46	28.51	90.39	30.58	37.70
w/ random	31.23	1.70	0.48	12.34	84.71	71.58	43.45	90.91	30.58	37.70
w/ GA+GD	21.35	1.01	0.61	1.32	82.99	21.57	0.79	77.76	30.58	37.70
w/ GA+KL	53.86	20.01	12.61	22.34	84.47	64.44	16.60	87.85	30.58	37.70

E.4 Ablation Study

We conduct an ablation study to understand how CAGUL achieves unlearning and report the results in Table 5.

Ablating Discriminator and Encoder. First, we investigate the influence of the discriminator $\mathbf{C}_\phi$ and visual token encoder $\mathbf{F}_\psi$. We implement two versions of CAGUL, w/o $\mathbf{C}_\phi$ and w/o $\mathbf{F}_\psi$, where we remove the discriminator and the visual token encoder, respectively. For CAGUL w/o $\mathbf{C}_\phi$, we train only $\mathbf{F}_\psi$ and transform k selected visual tokens for all $\mathcal{D}_f$ and $\mathcal{D}_r$, and for CAGUL w/o $\mathbf{F}_\psi$, we only train $\mathbf{C}_\phi$ and add uniform random noise to the k visual tokens for the samples identified by $\mathbf{C}_\phi$. Experiment results indicate that CAGUL can achieve good trade-offs between forget and retain metrics without $\mathbf{C}_\phi$, but its generalization ability is negatively affected. We note that the rouge score for $\mathcal{D}_r$ shows significant improvement as the encoder is trained on the full retain set, enabling the model to memorize additional knowledge. Conversely, we observe a significant decline in forget and retain performance for CAGUL w/o $\mathbf{F}_\psi$, which shows the importance of $\mathbf{F}_\psi$ in our framework.

Cross-Modal Attention Selection Strategy. Another important component in CAGUL is the cross-modal attention guided visual token selection for perturbation. Based on the reasoning that highly attended visual tokens provide most of the signals to the language module, we choose to perturb the k least attended tokens with $\mathbf{F}_\psi$. We implement a variant of CAGUL where the k visual tokens to be perturbed are randomly chosen for each sample sent to the encoder and report the results under w/ random. The random selection of visual tokens achieves mostly comparable performance for forget metrics, but underperforms significantly for retain metrics. These results validate the selection of the least attended tokens for perturbation in CAGUL.

Choice of Forget and Retain Loss. As discussed in Section 4, we train the encoder $\mathbf{F}_\psi$ using a combination of PO and GD as the forget and retain losses, respectively. We additionally run

experiments where we substitute the loss in CAGUL with other commonly used unlearning losses: GA + GD and GA + KL. Similar to the baseline implementations of these losses, we utilize early stopping due to the GA loss and include results in Table 5. We observe that CAGUL w/ GA + GD achieves better forget metrics, but its retain performance quickly degrades despite early stopping. CAGUL w/ GA + KL relatively preserves retain performance but does not achieve satisfactory forget metrics thus justifying the choice of PO + GD as the unlearning loss in CAGUL. Overall, the ablation results demonstrate the importance of the different components in the CAGUL formulation.

Table 6: Unlearning performance metrics of CAGUL compared with baseline methods on Qwen-2.5-VL-7B-Instruct

Method	**Forget**				**Retain**				**General**	
	$\mathcal{D}_f$				$\mathcal{D}_r$	$\mathcal{D}_{np}$			MME	POPE
	Rouge(↓)	EM(↓)	APE(↓)	MinK(↓)	Rouge(↑)	Rouge(↑)	EM(↑)	TR(↑)	Acc.(↑)	Acc.(↑)
Pretrain	30.61	0.98	0.37	0.00	36.27	34.01	0.10	11.28	86.68	86.94
Finetune	74.07	56.92	45.62	2.83	72.49	68.18	53.51	89.85	67.58	79.24
Retrain	44.04	14.52	10.70	1.44	72.68	67.80	54.23	91.08	71.31	76.41
GA	0.02	0.18	0.18	0.00	0.04	0.00	0.00	82.56	28.99	45.74
GA+GD	45.39	20.37	13.92	0.58	76.88	71.05	78.47	93.96	75.72	87.02
GA+KL	4.66	0.37	0.18	6.35	8.66	12.65	0.79	88.53	68.68	78.93
PO+GD	42.41	51.32	30.63	4.82	81.08	78.41	96.30	94.42	72.42	86.89
CAGUL	40.82	37.13	23.78	3.14	71.88	50.23	40.01	91.56	68.64	79.62

E.5 Results on Qwen-2.5-VL-Instruct

We report the results from our experiments on Qwen-2.5-VL-Instruct in Table 6. We first finetune Qwen-2.5-VL-Instruct on the FIUBench dataset to ensure sufficient memorization of individual profiles from the dataset, which is demonstrated by the increased performance metrics across $\mathcal{D}_f, \mathcal{D}_r$, and $\mathcal{D}_{np}$ for Finetune compared to Pretrain. We observe that, compared to LLaMA-3.2-11B-Vision-Instruct, Qwen-2.5-VL-Instruct preserves downstream model performance on MME and POPE to a larger degree after finetuning.

With CAGUL, we achieve satisfactory unlearning results as most metrics measured on $\mathcal{D}_f$ incur significant drops compared to Finetune. CAGUL also preserves the overall retain performance on $\mathcal{D}_r$ and the model's general utility on MME and POPE datasets. However, we notice a decline in the Rouge and EM scores for $\mathcal{D}_{np}$ despite an increase in the TR metric. We conjecture that this loss in performance arises as a result of the VLM's architecture. LLaMA-3.2-11B-Vision-Instruct implements cross-attention layers that feed image representations into the language model discontinuously (i.e., only at a few select layers), whereas Qwen-2.5-VL-Instruct directly feeds image representations to the language model at every layer and computes self-attention for concatenated visual and text tokens instead of explicit cross-attention scores. As a result, encoding additional information in the visual tokens has an increased impact on generated outputs with Qwen.

Among the baselines, although GA and GA+KL achieve superior forget performance, these methods seriously degrade model utility despite the use of early stopping. In contrast, CAGUL mostly retains model performance due to the use of the discriminator module $\mathbf{C}_\phi$, which only allows samples associated with forget images to be transformed by the encoder $\mathbf{F}_\psi$. Unlike LLaMA results, GA+GD achieves the best trade-off for Qwen, which is comparable to the Retrain method. Nonetheless, CAGUL achieves better forget metrics than its baseline finetuning counterpart PO+GD, while preserving the overall retain utility and general downstream accuracy. Note that all baseline methods implement model fine-tuning and require significantly more resources than CAGUL, which achieves satisfactory unlearning performance while keeping the VLM frozen.