CLIPErase: Efficient Unlearning of Visual-Textual Associations in CLIP

Anonymous ACL submission

Abstract

Machine unlearning (MU) has gained significant attention as a means to remove specific data from trained models without requiring a full retraining process. While progress has been made in unimodal domains like text and image classification, unlearning in multimodal models remains relatively underexplored. In this work, we address the unique challenges of unlearning in CLIP, a prominent multimodal model that aligns visual and textual representations. We introduce CLIPErase, a novel approach that disentangles and selectively forgets both visual and textual associations, ensuring that unlearning does not compromise model performance. CLIPErase consists of three key modules: a Forgetting Module that disrupts the associations in the forget set, a Retention Module that preserves performance on the retain set, and a Consistency Module that maintains consistency with the original model. Extensive experiments on the CIFAR-100 and Flickr30K datasets across four CLIP downstream tasks demonstrate that CLIPErase effectively forgets designated associations in zero-shot tasks for multimodal samples, while preserving the model's performance on the retain set after unlearning.

1 Introduction

In the era of rapid advancements in pre-trained large models (Kim et al., 2021; Liu et al., 2024a; Achiam et al., 2023), ensuring these models can eliminate the influence of specific data upon request has become crucial due to ethical concerns and legal requirements like the "right to be forgotten" (Yao et al., 2024; Liu et al., 2024b). Machine unlearning (MU) has been recently studied as a method to remove the influence of specific data points from a trained model without the need to retrain it from scratch (Golatkar et al., 2020). While MU has made notable progress in unimodal domains such as images and text, its application in multimodal models remains relatively underexplored. Multimodal models integrate diverse data modalities such as images, text, and audio, introducing unique challenges for MU, as removing the influence of specific data requires disentangling complex relationships across different modalities. This unique challenge makes the existing MU methods for unimodal models, such as (Yao et al., 2023; Liu et al., 2022; Chundawat et al., 2023; Dou et al., 2024), insufficient or ineffective.

Our work focuses on the unlearning problem in CLIP (Kim et al., 2021), a multimodal model pretrained to align visual and textual concepts. CLIP has been widely used in various applications and serving as the encoder for many Vision Language Models (VLMs) and Multimodal Large Language Models (MLLMs) (Liu et al., 2024a; Li et al., 2021; Yuan et al., 2021; Zhai et al., 2022; Yang et al., 2022). Unlearning in CLIP is thus important due to increasing concerns about data privacy, security, and compliance with regulations. However, simply applying MU techniques designed for unimodal models would fail to account for how visual and textual features influence each other in CLIP. For example, if only the image of an "apple" is unlearned in CLIP, the text "apple" would still remain associated with other visual representations, leading to inconsistencies. Conversely, if only the text "apple" is unlearned, the image could still be linked to other descriptions, causing similar issues.

To address these challenges, we introduce **CLIPErase**, which disentangles the relationships between visual and textual representations of the concept to be forgotten through three key components: (1) **Forgetting Module**: reduces the model's ability to associate text-image pairs in the forget set by modifying the cross-modal similarity matrix; (2) **Retention Module**: ensures the model's performance on the retain set remains unaffected; (3) **Consistency Module**: maintains consistency by penalizing deviations in the unimodal (text and im-

age) distributions compared to the original model. Three loss functions, derived from the three modules, are jointly minimized to perform unlearning, making CLIPErase an efficient solution that eliminates the need for re-training from scratch.

Our key contributions are as follows:

1. We introduce CLIPErase, an innovative framework designed for unlearning pretrained CLIP models. CLIPErase efficiently removes specific multimodal data without retraining.

2. CLIPErase uses three modules to disrupt multimodal data alignment for unlearning while preserving performance on the retain set.

3. Our extensive experiments, including ablation studies and visualizations across multiple datasets, validate the effectiveness of CLIPErase.

2 Related Works

Machine Unlearning: As machine learning models continue to grow in size and their training datasets become increasingly vast and complex, the concept of MU has garnered significant attention in academic and industry (McConnon, 2024; Pedregosa and Triantafillou, 2023) to promoting AI for social welfare. MU aims to selectively remove specific information—such as private data (Zhang et al., 2023), outdated knowledge (Wang et al., 2023), and harmful content (Liu et al., 2024c) from a trained model without necessitating a complete retraining from scratch (Bourtoule et al., 2019). We will next discuss these techniques in the context of text and image modalities.

Machine Unlearning for Text: MU has been extensively studied in the text domain. Gupta et al. (2021) first explored adaptive parameter tuning, while Maini et al. (2024), Chen and Yang (2023), and Jia et al. (2024) leveraged gradient-based methods, including second-order optimization and KL-Divergence descent. Eldan and Russinovich (2023) focused on preference optimization. To address high computational costs, Kurmanji et al. (2024) proposed scalable, approximate methods, and Chen et al. (2023) aimed to reduce overhead. For theoretical guarantees, Ullah et al. (2021) employed differential privacy to enhance trust in MU.

Machine Unlearning for Vision: In the image domain, MU techniques are applied in two primary scenarios. One is to eliminate the influence of specific visual data in classification models. For example, SRCUB (Kurmanji et al., 2024) uses a teacher-student framework to forget data in image classification tasks. The other is for diffusion models, where MU techniques focus on unlearning specific visual patterns in image generation, e.g., by modifying cross-attention mechanisms (Zhang et al., 2024; Gandikota et al., 2024, 2023; Fuchi and Takagi, 2024) and disrupting connections between visual and textual representations. Li et al. (2024) introduced a method for knowledge decoupling with datasets designed for different retention needs, while Cheng and Amiri (2023) proposed a similar multi-deletion mechanism.

Note that although the connections between visual and textual representations may be disrupted in these studies, the focus is primarily on affecting the visual patterns rather than the text, as the task is typically centered around image generation. In contrast, our study of unlearning in CLIP aims to unlearn both visual and textual representations, ensuring that downstream tasks involving both text and image, such as retrieval, classification, and multimodal association, are fully addressed. The most relevant work to ours is a recent approach by Kravets and Namboodiri (2024) utilized Lipschitz regularization and synthetic images to simulate the forget set in zero-shot tasks, but it suffers from high computational cost and limitations imposed by synthetic data quality. In contrast, our proposed method eliminates the need for retraining or synthetic data generation, enabling efficient unlearning in CLIP while preserving model performance for real-world applications.

3 Preliminary

3.1 CLIP

CLIP is a multimodal pre-training model developed by OpenAI with several hundred million parameters. It aligns images and their text information in a shared embedding space using an image encoder f_{img} and a text encoder f_{txt}, which convert inputs into normalized feature vectors. CLIP is trained with a contrastive loss that maximizes the similarity between matching image-text pairs and minimizes it for non-matching pairs. For a batch of N image-text pairs $\{(x_i^n, x_t^n)\}_{n=1}^N$, the loss function is:

$$\mathcal{L} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\left(\frac{f_{\text{img}}(x_i^n)\cdot f_{\text{txt}}(x_t^n)}{\tau}\right)}{\sum_{k=1}^{N}\exp\left(\frac{f_{\text{img}}(x_i^n)\cdot f_{\text{txt}}(x_t^k)}{\tau}\right)},$$

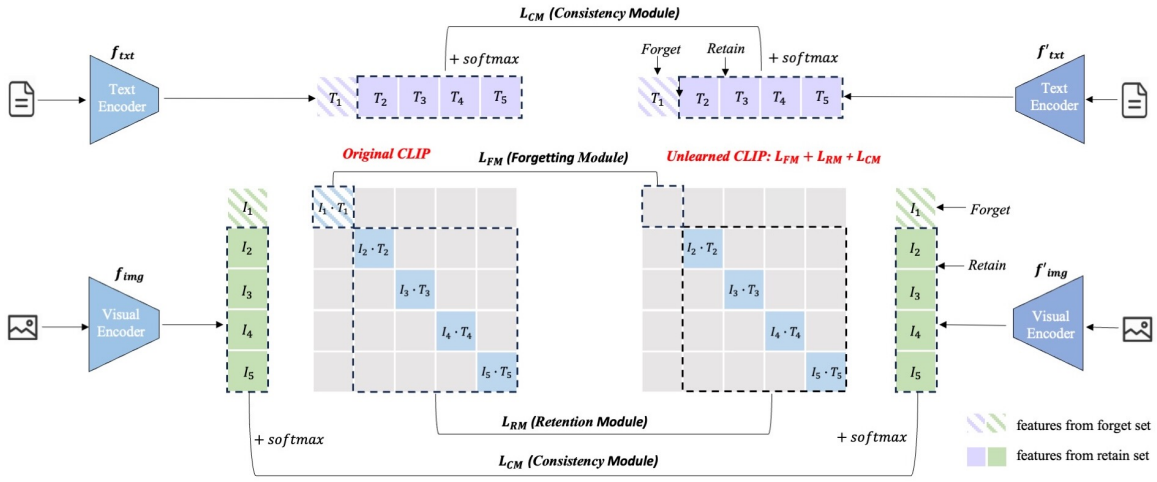

Figure 1: The architecture of CLIPErase. Given a CLIP model and a forget set, CLIPErase has three key modules: (a) Forgetting Module (FM): disrupts cross-modal associations in the forget set to reduce memory of deleted information; (b) Retention Module (RM): preserves cross-modal associations in the retain set; (c) Consistency Module (CM): maintains the original model's consistency by aligning unimodal representations.

where τ is a temperature parameter. This loss aims to correctly align each image with its corresponding text while distinguishing it from others.

3.2 Problem Definition

Let Θ be a pre-trained CLIP model trained on a dataset D. Upon receiving a deletion request specifying a subset of data to be removed, it is denoted as the forget set D_f. The remaining data, referred to as the retain set $D_r = D - D_f$.

Our objective is to design an algorithm that enables the model to eliminate information related to D_f without compromising CLIP's performance on D_r. We refer to this model as the unlearned model Θ_u, which is initialized with the parameters of the original model Θ. With Θ_u, images and their corresponding text descriptions in D_f no longer align. Concequently, in downstream tasks, e.g., an image retrieval task, an input image from forget set D_f will no longer retrieve its associated text description, and vice versa. Meanwhile, it is crucial to maintain the model's performance on the retain set D_r, ensuring effective text and image alignment.

4 Proposed Method

In this section, we introduce CLIPErase, a novel machine unlearning method for the pre-trained CLIP. The overview of our proposed method is illustrated in Figure 1, where the image encoder and text encoder after unlearning are denoted as f'_{img} and f'_{txt}. The disentangled visual-textual association is illustrated by the *Unlearned CLIP*.

4.1 Forgetting Module

The pre-trained CLIP captures rich semantic relations between images and their corresponding text. To achieve unlearning, we propose a novel approach that disrupts the learned relationships between the visual and textual representations in the forget set D_f. By intentionally misaligning the image and text features in D_f, we ensure that the model no longer understands the cross-modal information contained in D_f, thereby complying with data deletion requests without compromising performance on the retain set D_r.

Our optimization objective for forgetting, $\mathcal{L}_{\text{FM}}$, is designed to reduce the similarity between the features of image-text pairs in the forget set. Specifically, we formulate $\mathcal{L}_{\text{FM}}$ for minimization as:

$$\mathcal{L}_{\text{FM}} = \frac{1}{N_f} \sum_{n=1}^{N_f} \left(f_{\text{img}}(x_i^n) \cdot f_{\text{txt}}(x_t^n) \right),$$

where N_f is the number of samples in D_f.

4.2 Retention Module

After disrupting the connection between the visual and textual representations of the forget set D_f, this process may inadvertently impair the model's performance on the retain set D_r. This is because adjustments to the model's parameters can affect the overall cross-modal representation space, influencing the model's understanding and processing of the retain set. To address this, the retention module is designed to maintain the CLIP model's ability to correctly align images and texts in D_r during the

unlearning process, by minimizing the alignment loss:

$$\mathcal{L}_{\text{RM}} = -\frac{1}{N_r}\sum_{n=1}^{N_r}\log\frac{\exp\left(\frac{f_{\text{img}}(x_i^n)\cdot f_{\text{txt}}(x_t^n)}{\tau}\right)}{\sum_{k=1}^{N_r}\exp\left(\frac{f_{\text{img}}(x_i^n)\cdot f_{\text{txt}}(x_t^k)}{\tau}\right)},$$

where N_r is the number of samples in D_r.

4.3 Consistency Module

With the different optimization objectives for the forget set D_f and the retain set D_r, the CLIP model might introduce new errors or biases in its predictions on D_r during the forgetting process. To prevent unexpected changes in the model's behavior on D_r, we maintain consistency between the outputs of the unlearned model Θ_u and that of the original model Θ on D_r by introducing a consistency restriction $\mathcal{L}_{\text{CM}}$, defined as the Kullback-Leibler (KL) divergence between the output distributions of Θ_u and Θ on D_r:

$$\mathcal{L}_{\text{CM}} = \frac{1}{N_r}\sum_{n=1}^{N_r}\Big[\text{KL}\left(\mathbf{p}_o^{\text{img}} \parallel \mathbf{p}_u^{\text{img}}\right) + \text{KL}\left(\mathbf{p}_o^{\text{txt}} \parallel \mathbf{p}_u^{\text{txt}}\right)\Big],$$

where $\mathbf{p}_o^{\text{img}}$ and $\mathbf{p}_u^{\text{img}}$ are the distributions of one image embedding, obtained from the image encoder of the original model Θ and the unlearned model Θ_u, respectively. Similarly, $\mathbf{p}_o^{\text{txt}}$ and $\mathbf{p}_u^{\text{txt}}$ are the distributions of one text embedding, obtained from the text encoder of the original model Θ and the unlearned model Θ_u. These distributions are computed by applying softmax directly to the image and text embeddings.

At last, the overall unlearning loss is:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{RM}} + \lambda_2\mathcal{L}_{\text{FM}} + \lambda_3\mathcal{L}_{\text{CM}},$$

where λ_1, λ_2, and λ_3 are hyperparameters.

Minimizing $\mathcal{L}_{\text{FM}}$ primarily affects the image and text encoders by reducing the similarity between image-text pairs in D_f, disrupting their alignment and ensuring the model forgets the cross-model associations. At the same time, minimizing $\mathcal{L}_{\text{RM}}$ preserves the alignment attention for D_r, ensuring that the model maintains its performance on D_r. Finally, minimizing $\mathcal{L}_{\text{CM}}$ ensures consistency between the outputs of the unlearned model Θ_u and the original model Θ on D_r, preventing the introduction of new errors or biases in the retain set during the unlearning process.

Task	Dataset	Evaluation
Zero-shot Prediction	CIFAR-100	Acc of $D_f\downarrow$, $D_r\uparrow$
Zero-shot Retrieval	CIFAR-100	Acc of $D_f\downarrow$, $D_r\uparrow$
Image Retrieval	Flickr30k	Recall of $D_f\downarrow$, $D_r\uparrow$
Text Retrieval	Flickr30k	Recall of $D_f\downarrow$, $D_r\uparrow$

Table 1: Experiment Settings

5 Experiment

5.1 Experiments Setting

Table 1 summarizes our experimental settings, including tasks, datasets, and evaluation metrics.

Tasks: To evaluate the effectiveness of our method, we conducted experiments on four tasks:

1. Zero-shot Prediction: CLIP is employed to classify images based on a given category label (e.g., "cat," "dog") without training data. The unlearned CLIP is expected to predict a mismatch between the label "dog" and the image of a dog if it belongs to the forget set (D_f), demonstrating its ability to "forget" specific learned associations.

2. Zero-shot Image Retrieval: CLIP is used to retrieve images matching a text description (e.g., "red car") without task-specific training. In the unlearning scenario, the unlearned CLIP should fail to retrieve images that belong to the forget set, such as cars described as "red car" if they are part of D_f, demonstrating its ability to forget these specific multimodal associations without further training.

3. Image Retrieval (IR): Given a text query, CLIP retrieves the top-k relevant images. The unlearned CLIP should avoid retrieving images in the forget set, demonstrating the selective forgetting capability while maintaining performance on the retain set.

4. Text Retrieval (TR): Given an image query, CLIP retrieves the top-k relevant text descriptions. In the unlearning context, CLIP should fail to retrieve relevant text descriptions for images in the forget set.

Datasets: For the zero-shot prediction and retrieval tasks, we use the CIFAR-100 dataset (Krizhevsky et al., 2009). CIFAR-100 is an image classification dataset containing 100 categories. It consists of a total of 60,000 images, with 50,000 used for training and 10,000 for testing. For the Image Retrieval (IR) and Text Retrieval (TR) tasks, we use the Flickr30K dataset (Bojchevski and Günnemann, 2017). The dataset contains 31,783 images, each paired with five natural language captions. These images primarily depict daily life scenes.

Settings: For CIFAR-100, we randomly select one class as the forget set. The model is trained for 50 epochs with a batch size of 32, using Adam and an initial learning rate of 1×10^{-6}. The loss weights are set to $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

For Flickr30k, the forget set includes all image-text pairs containing "car," comprising 1% of the dataset. The same setup is used, except with an initial learning rate of 1×10^{-8}.

Each experiment use the best checkpoint from the validation set. Our model and code are implemented in PyTorch, with training and evaluation on an NVIDIA Tesla V100-SXM2.

Evaluation: Following prior work (Kim et al., 2021), we use two metrics to assess MU efficacy: (1) Retain Set Performance ($D_r \uparrow$): Measures accuracy for zero-shot tasks and recall@1, @5, @10 for retrieval. Higher values indicate minimal impact on the retain set after unlearning. (2) Forget Set Performance ($D_f \downarrow$): Uses the same metrics as D_r. Lower values indicate more effective unlearning.

5.2 Comparison to Prior Work

Besides the original CLIP model, we compare our method CLIPErase with commonly known unimodal MU methods when directly applied to CLIP. We omit using prior work (Kravets and Namboodiri, 2024) as a baseline because their code and detailed experimental settings were not released, making it challenging to replicate their results. Especially without access to their synthetic image generation code, it is unable to conduct CLIP unlearning in their setting.

Original CLIP Model (Kim et al., 2021): We use the original CLIP model as a baseline to assess the impact of unlearning, ensuring that the model's performance on the retain set remains unaffected.

Gradient Ascent (GA) (Yao et al., 2023): This method aims to degrade the model's performance on the forget set by increasing prediction errors, forcing the model behave away from its original predictions.

Gradient Difference (GradDiff) (Liu et al., 2022): This method increases errors on the forgetting data while preserving performance on the retain set, achieving the goal of unlearning specific information without impacting retained data.

KL Minimization (KL) (Maini et al., 2024): The method ensures consistency on the retain set by comparing the prediction distributions of the unlearned and original models, while increasing errors on the forgetting data.

Method	Zero-shot Prediction		Zero-shot Retrieval	
	Acc.D$_f$↓	Acc.D$_r$↑	Acc.D$_f$↓	Acc.D$_r$↑
CLIP	86.08	72.85	88.61	73.43
CLIP+GA	11.81	66.26	5.35	67.31
CLIP+GradDiff	7.31	70.94	9.89	71.79
CLIP+KL	2.19	61.75	5.74	65.87
CLIP+ENMN	0.00	12.46	0.00	17.94
CLIPErase	0.00	89.41	0.00	91.38

Table 2: Comparison on CIFAR100 dataset for Zero-shot Prediction and Zero-shot Image Retrieval.

Error Minimization-Maximization Noise (EMMN) (Chundawat et al., 2023): It deletes specific class information from the model using noise-based error minimization and maximization techniques.

5.3 Experiments on Cifar100 Dataset

As shown in Table 2, we compare the effectiveness of CLIPErase with several baseline methods and the original CLIP model on zero-shot prediction and retrieval tasks on CIFAR-100 dataset. CLIPErase not only preserves the robust performance of CLIP in these tasks but also achieves superior results after accurately deleting specific information. On the Forget Set, CLIPErase achieves **0%** accuracy in both tasks, confirming full unlearning. While methods like CLIP+GA (Yao et al., 2023), CLIP+GradDiff (Liu et al., 2022), and CLIP+KL (Maini et al., 2024) leave residual accuracy, CLIP+ENMN (Chundawat et al., 2023) also achieves 0% accuracy but at the expense of drastically reduced retention performance, dropping to 12.46% in prediction and 17.94% in retrieval.

In contrast, CLIPErase excels on the Retain Set, achieving **89.41%** in prediction and **91.38%** in retrieval, representing a **16.56%** and **17.95%** improvement over the original CLIP model. This is due to CLIPErase's Retention and Consistency Modules, which not only prevent the model from forgetting retained data but also improve its alignment of visual and textual features, thus enhancing overall performance on the Retain Set. The improved performance on D_r suggests that the forgetting process in CLIPErase leads to a more focused representation of retain set, thereby optimizing the model's zero shot capabilities.

5.4 Experiments on Flicker30K Dataset

As shown in Table 3, we compare the effectiveness of CLIPErase with several baseline methods and the original CLIP model on Image Retrieval and Text Retrieval tasks on Flickr30K

Method	Image Retrieval						Text Retrieval					
	R@1		R@5		R@10		R@1		R@5		R@10	
	$D_f\downarrow$	$D_r\uparrow$	D_f	$D_r\uparrow$	$D_f\downarrow$	$D_r\uparrow$	$D_f\downarrow$	$D_r\uparrow$	$D_f\downarrow$	$D_r\uparrow$	$D_f\downarrow$	$D_r\uparrow$
CLIP	28.61	22.76	56.75	50.14	66.73	60.67	25.04	19.11	58.87	48.82	68.33	59.17
CLIP+GA	19.40	17.10	41.99	37.76	51.37	47.46	14.67	12.59	37.87	34.53	47.16	44.25
CLIP+GradDiff	15.95	19.77	32.10	46.03	40.91	57.02	15.95	18.75	40.51	47.01	49.95	55.25
CLIP+KL	16.84	18.06	29.63	40.62	38.28	51.90	16.99	17.83	30.55	45.21	38.64	52.13
CLIPErase	10.40	21.41	19.91	48.47	23.80	58.18	11.94	20.22	20.46	47.12	24.79	58.96

Table 3: Comparison on Flickr30K dataset for Image Retrieval and Text Retrieval.

dataset. CLIPErase delivers strong performance in both image and text retrieval tasks at multimodal MU, surpassing prior methods. In image retrieval, CLIPErase significantly outperforms CLIP+GA, CLIP+GradDiff (Liu et al., 2022), and CLIP+KL) (Maini et al., 2024), with notable gains in **R@1 (64.64%)**, **R@5 (48.4%)**, and **R@10 (58.18%)**. Similarly, it shows superior results in text retrieval, with a **6.56%** improvement in R@10 over CLIP+KL.

These results highlight the effectiveness of CLIPErase's Forgetting, Retention, and Consistency Modules. The Forgetting Module successfully reduces model reliance on forgotten data, while the Retention and Consistency Modules preserve performance on the retain set, ensuring robust multimodal unlearning without retraining.

5.5 Ablation Studies

As shown in Table 4, we performed ablation studies on the CIFAR-100 dataset to quantitatively evaluate the impact of the Forgetting Module (FM), Retention Module (RM), and Consistency Module (CM) on the zero-shot prediction task.

FM	RM	CM	Accuracy %		Improvement %	
			↓ Df	↑ Dr	↓ Df	↑ Dr
✗	✗	✗	86.08	72.85	-	-
✓	✗	✗	18.57	64.12	↓ 67.5	↓ 8.73
✓	✓	✗	9.40	73.14	↓ 76.68	↑ 0.56
✓	✓	✓	0	89.41	↓ 86.08	↑ 16.56

Table 4: Ablation studies on the Forgetting Module (FM), Retention Module (RM), and Consistency Module (CM).

Effectiveness of FM: Activating the Forgetting Module led to a significant drop in accuracy on the forget set, from 86.08% to 18.57%, indicating that FM effectively disrupted the correspondence between images and texts in the forget set. However, relying solely on FM negatively impacted the retain set performance, reducing its accuracy from 72.85% to 64.12%.

Effectiveness of RM: When both the Forgetting Module and Retention Module were activated, the accuracy on the retain set recovered to 73.14%, demonstrating that RM successfully protected the retain set's performance. Simultaneously, the model's accuracy on the forget set further decreased to 9.40%, showing that RM plays a critical role in balancing the task of forgetting specific data while preserving the retain set performance.

Effectiveness of CM: When the Consistency Module (CM) was activated alongside FM and RM, the retain set accuracy significantly improved to 89.41%, while the accuracy on the forget set dropped to 0%. This indicates that the model successfully forgot the specified data while maintaining high performance on the retain set. Consistency Module ensures consistency throughout the unlearning process, preventing potential errors or biases introduced by the forgetting process.

5.6 Robustness and Utility for Our Method

Figure 2 shows experiments with CLIPErase, GradDiff, GA, and the original CLIP across various proportions of forget classes (0%, 3%, 10%, 20%, 30%) on the CIFAR-100 dataset for the zero-shot prediction task. The left plot shows the forget accuracy, where CLIPErase demonstrates a significant unlearning effect at different proportions of forget classes. However, for GradDiff and GA, forget accuracy increases as the number of forget classes grows, indicating a worsening unlearning effect. The right plot presents the retain accuracy, where CLIPErase maintains performance comparable to CLIP, particularly at lower proportions, while GA shows a sharp decline in retain accuracy as the forget proportion increases. Overall, CLIPErase exhibits strong robustness and utility, effectively balancing both forget accuracy and retain accuracy.

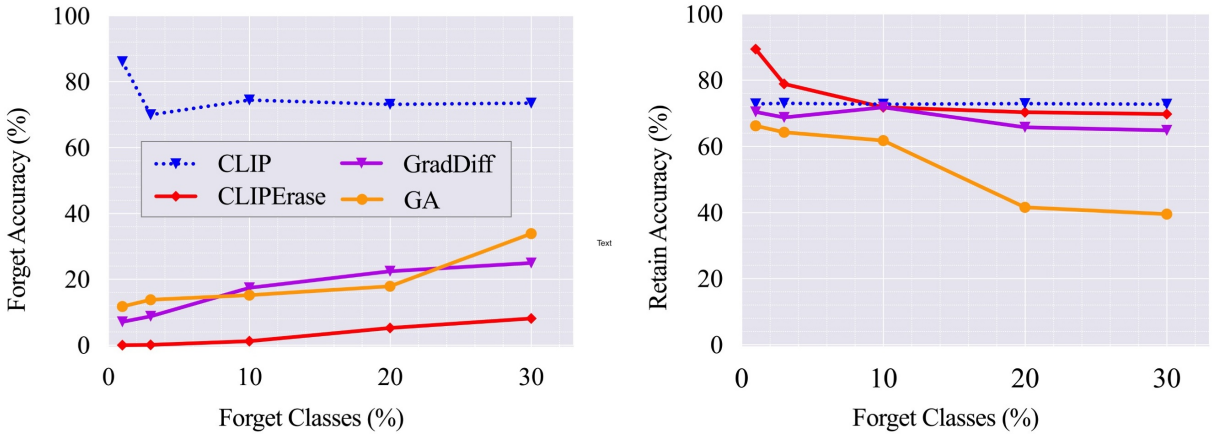

Figure 2: Performance on varying percentages of forget classes. The x-axis shows the percentage of forget classes, and the y-axis shows forget accuracy (left) and retain accuracy (right).

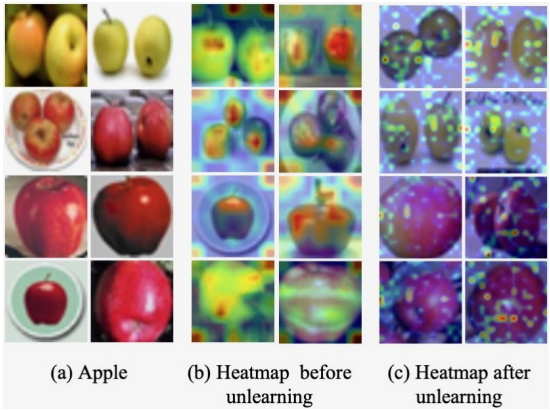

Figure 3: Attention heatmap visualization of CLIP (before unlearning) and CLIPErase (after unlearning) from the original apple images. In the CLIP heatmaps, the attention is highly concentrated on the objects, whereas in the CLIPErase heatmaps, the attention becomes more random and dispersed.

6 Visualization

Attention Visualization: In Fig. 3, we visualized the attention heatmaps of the "forget set" before and after the unlearning process. Using "apple" from CIFAR-100 as an example, we presented the original images, the heatmaps generated by the CLIP model, and the heatmaps after unlearning with CLIPErase. In the CLIP heatmaps, attention is highly focused on the object, displaying strong visual-semantic alignment. This indicates that CLIP successfully establishes a robust connection between textual and visual semantics. In contrast, after machine unlearning with CLIPErase, the heatmaps show that attention becomes more random and dispersed across each patch, no longer concentrated on the relevant object. This suggests that CLIPErase effectively disrupts the alignment between text and visual semantics for the data to be unlearned, thus achieving the intended unlearning objective.

Embedding Visualization: To further investigate the impact of CLIPErase on the retain set within the cross-modal shared representation space, we conducted t-SNE visualizations (Van der Maaten and Hinton, 2008) of the textual and visual embeddings from both the original CLIP model and the CLIPErase model. Specifically, we selected 10 classes from the CIFAR-100 dataset, with the "apple" class serving as the forget set and the remaining classes as the retain set.

From the CLIPErase results, we observed that the distance between the textual and visual embeddings of the "apple" class significantly increased, while the embeddings of the other classes remained tightly clustered within their respective categories. This suggests that CLIPErase effectively weakens the association between the textual and visual modalities in the forget set, with minimal impact on the modality associations of the retain set. In summary, CLIPErase can effectively decouple the connection between the textual and visual embeddings of the forget set without affecting the visual-textual associations of the retain set, thus achieving the intended goal of machine unlearning.

7 Discussion and Future Works

The proposed CLIPErase method has significant potential in practical applications, especially in addressing ethical and legal concerns related to harmful information and biases in multimodal datasets. In tasks like online antisemitism detection, individual components such as text or images may appear

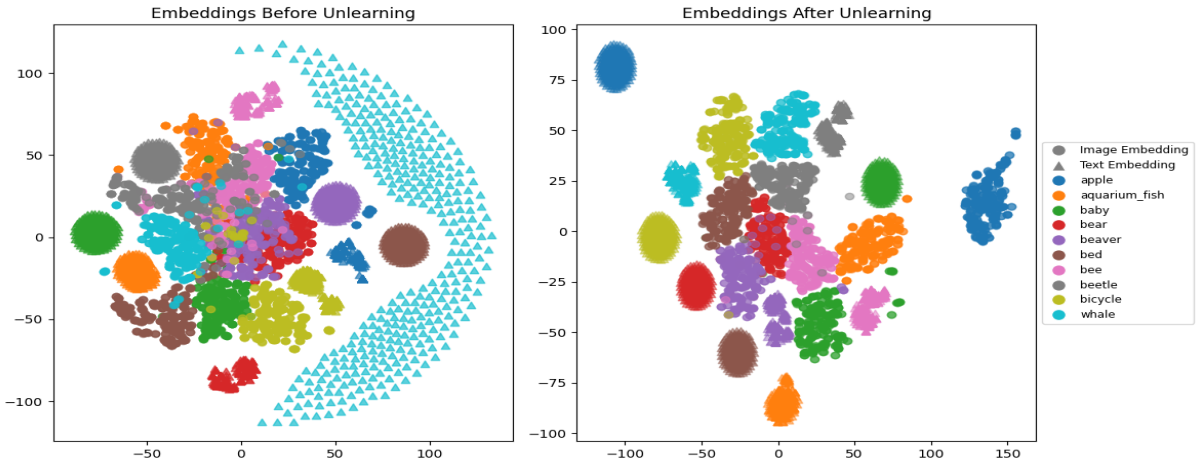

Figure 4: t-SNE visualizations of text and visual embeddings from the CLIP model (before unlearning, left) and the CLIPErase model (after unlearning, right) on CIFAR-100. The unlearned category is "apple." We use ○ and △ to represent visual and text modalities, respectively, with different colors indicating different categories. In the CLIPErase results, the embeddings for the "apple" category have become more distant between the text and visual modalities, while the embeddings for other categories remain compact within their respective classes.

harmless on their own, but when combined, they can convey harmful messages. For example, in an image where the text "Even grandma can see what's going on" seems innocuous at first glance, when paired with an antisemitic image and stereotypical messaging, it transmits damaging, implicit bias. Such hidden biases are especially dangerous in multimodal data. CLIPErase caneffectively decouple these associations, to forget the harmful links between text and images, thereby mitigating the risk of perpetuating bias.

Additionally, CLIPErase holds great potential in safeguarding user privacy. When users request the deletion of specific personal data, CLIPErase can remove any associations between their personal information and multimodal content.

In the future, although our current implementation is focused on the CLIP model, the framework can be extended to any modality or multimodal pretrained model, not just CLIP. This broader applicability would enable flexible unlearning across a wide range of systems, making the approach more versatile. Additionally, we aim to apply CLIPErase to Generative AI systems, such as Multimodal Large Language Models (MLLMs), where CLIP-based encoders are widely used. By unlearning at the encoder level, CLIPErase can help address the growing challenges in Generative AI, including the generation of private, malicious, or illegal content, the continuation of biases, and even the risk of weaponizing these models. Our approach can serve as a safeguard, correcting problematic associations and enhancing user privacy, thus providing a safer and more ethical experience in the future of AI development.

8 Conclusion

In this paper, we present CLIPErase, a machine unlearning framework for multimodal contrastive pretraining that removes specific data while preserving performance on retained data. It integrates the Forgetting Module, Retention Module, and Consistency Module to disrupt cross-modal relationships for targeted data while maintaining retention. Extensive experiments on zero-shot prediction and retrieval tasks with CIFAR-100 and Flickr30K validate its effectiveness in unlearning and performance retention, offering practical applications for privacy preservation and data management.

9 Limitation

While CLIPErase achieves good unlearning performance on multimodal tasks, it may face challenges when the forgetting and retain sets share highly similar features (e.g., fine-grained categories). In such cases, disrupting cross-modal associations without affecting closely related retain set becomes more complex. In the future, we plan to introduce strategies that target specific feature subspaces or by implementing adaptive unlearning techniques that dynamically adjust the unlearning process based on the similarity between the forget and retain sets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aleksandar Bojchevski and Stephan Günnemann. 2017. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. *arXiv preprint arXiv:1707.03815*.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. 2019. Machine unlearning. *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. *arXiv preprint arXiv:2310.20150*.

Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, and Chen Wang. 2023. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7766–7775.

Jiali Cheng and Hadi Amiri. 2023. Multidelete for multimodal machine unlearning. *arXiv preprint arXiv:2311.12047*.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. 2023. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354.

Guangyao Dou, Zheyuan Liu, Qing Lyu, Kaize Ding, and Eric Wong. 2024. Avoiding copyright infringement via machine unlearning. *arXiv preprint arXiv:2406.10952*.

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

Masane Fuchi and Tomohiro Takagi. 2024. Erasing concepts from text-to-image diffusion models with few-shot unlearning. *arXiv preprint arXiv:2405.07288*.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2426–2436.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 383–398. Springer.

Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. 2021. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330.

Jinghan Jia, Yihua Zhang, Yimeng Zhang, Jiancheng Liu, Bharat Runwal, James Diffenderfer, Bhavya Kailkhura, and Sijia Liu. 2024. Soul: Unlocking the power of second-order optimization for llm unlearning. *arXiv preprint arXiv:2404.18239*.

Gwanghyun Kim, Taesung Kwon, and Jong-Chul Ye. 2021. Diffusionclip: Text-guided diffusion models for robust image manipulation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2416–2425.

Alexey Kravets and Vinay Namboodiri. 2024. Zero-shot class unlearning in clip with synthetic samples. *arXiv preprint arXiv:2407.07485*.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. 2024. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, and Sheng Bi. 2024. Single image unlearning: Efficient machine unlearning in multimodal large language models. *ArXiv*, abs/2405.12523.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024b. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024c. Towards safer large language models through machine unlearning. *ArXiv*, abs/2402.10058.

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

Anthony McConnon. 2024. Teaching large language models to "forget" unwanted content. Accessed: 2024-10-09.

Fabian Pedregosa and Eleni Triantafillou. 2023. Announcing the first machine unlearning challenge. Accessed: 2024-10-09.

Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. 2021. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pages 4126–4142. PMLR.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023. Kga: A general machine unlearning framework based on knowledge gap alignment. *ArXiv*, abs/2305.06535.

Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. 2022. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *arXiv preprint arXiv:2402.15159*.

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

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. 2022. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18123–18133.

Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 2024. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1755–1764.

Lefeng Zhang, Tianqing Zhu, Haibin Zhang, Ping Xiong, and Wanlei Zhou. 2023. Fedrecovery: Differentially private machine unlearning for federated learning frameworks. *IEEE Transactions on Information Forensics and Security*, 18:4732–4746.

A Example Appendix

This is an appendix.