# Better Safe than Sorry: Pre-training CLIP against Targeted Data Poisoning and Backdoor Attacks

**Wenhan Yang** [1] **Jingdong Gao** [1] **Baharan Mirzasoleiman** [1]

## Abstract

Contrastive Language-Image Pre-training (CLIP) on large image-caption datasets has achieved remarkable success in zero-shot classification and enabled transferability to new domains. However, CLIP is extremely more vulnerable to targeted data poisoning and backdoor attacks, compared to supervised learning. Perhaps surprisingly, poisoning 0.0001% of CLIP pre-training data is enough to make targeted data poisoning attacks successful. This is four orders of magnitude smaller than what is required to poison supervised models. Despite this vulnerability, existing methods are very limited in defending CLIP models during pre-training. In this work, we propose a strong defense, SAFECLIP, to safely pre-train CLIP against targeted data poisoning and backdoor attacks. SAFECLIP warms up the model by applying unimodal contrastive learning (CL) on image and text modalities separately. Then, it divides the data into safe and risky sets, by applying a Gaussian Mixture Model to the cosine similarity of image-caption pair representations. SAFECLIP pre-trains the model by applying the CLIP loss to the safe set and applying unimodal CL to image and text modalities of the risky set separately. By gradually increasing the size of the safe set during pre-training, SAFECLIP effectively breaks targeted data poisoning and backdoor attacks without harming the CLIP performance. Our extensive experiments on CC3M, Visual Genome and MSCOCO demonstrate that SAFECLIP significantly reduces the success rate of targeted data poisoning attacks from 93.75% to 0% and that of various backdoor attacks from up to 100% to 0%, without harming CLIP's performance[1].

[*]Equal contribution [1]Department of Computer Science, University of California, Los Angeles, Los Angeles, CA, 90024. Correspondence to: Wenhan Yang <hangeryang18@g.ucla.edu>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1]Code can be found at https://github.com/BigML-CS-UCLA/SafeCLIP

## 1. Introduction

Pre-training large vision-language models on extensive image-caption data crawled from the internet has achieved remarkable success in zero-shot classification and robustness to distribution shift. CLIP learns image and text representations in a shared space by maximizing the agreement between the paired image-text representations, and minimizing the agreement between the unpaired ones. This alleviates the need for high-quality annotations and allows scaling up the pre-training data to millions (Radford et al., 2021) and billions of examples (Jia et al., 2021). Despite its superior performance, CLIP is extremely vulnerable to targeted data poisoning and backdoor attacks, where an adversary injects a subset of malicious examples in the training data to change the prediction of particular examples at test time. Perhaps surprisingly, poisoning only 0.0001% and 0.01% of the pre-training data is enough to make targeted data poisoning and backdoor attacks successful, respectively (Carlini et al., 2023; Carlini & Terzis, 2021). Considering that the large pre-training data of CLIP is often crawled from the internet, such attacks are very easy to perform in practice.

Despite this vulnerability, protecting CLIP against targeted data poisoning and backdoor attacks during pre-training has remained largely unaddressed. The only recently proposed method, RoCLIP, aims to disassociate the poisoned image-caption pairs during pre-training by pairing each image representation with its most similar caption representation in a random caption pool (Yang & Mirzasoleiman, 2023). However, RoCLIP can suffer significant performance drop in downstream performance, limiting its real-world application. Two other methods proposed to clean an already *poisoned pre-trained* CLIP, by fine-tuning on a *clean* data of the same scale as pre-training (Yang et al., 2023), or fine-tuning on a *clean* subset of pre-training data using contrastive learning on image and text modalities (Bansal et al., 2023). The first method is clearly not applicable to pre-training, and the second one even increases the attack success rate if applied during pre-training on poisoned data, as confirmed in (Yang & Mirzasoleiman, 2023).

Protecting CLIP against targeted data poisoning and back-

door attacks during pre-training is indeed very challenging. This is because training only once on the poisoned pairs can make the attack successful. In contrast, in the supervised setting the model should be trained on the poisoned data for several epochs before the attack succeeds (Biggio et al., 2012; Turner et al., 2019). Thus, to protect CLIP during pre-training, it is cruical to entirely exclude the poisoned examples from the pre-training pipeline.

In this work, we propose an effective defense, SAFECLIP, against strong targeted data poisoning and backdoor attacks during pre-training CLIP, without compromising its performance. SAFECLIP warms up the model by applying separate unimodal contrastive losses to image and caption modalities to reduce the initial similarity of poisoned image-caption representations. Then, it applies the CLIP loss once to all pairs with a low learning rate to initially associate the image-caption representations, while maintaining a low similarity for poisoned pairs. Subsequently, SAFECLIP employs a Gaussian Mixture Model (GMM) on cosine similarity of image-caption representations to divide the examples into safe and risky sets. SAFECLIP pre-trains the model using the CLIP loss on the safe set and unimodal contrastive losses on image and caption modalities of the risky set. Throughout training, SAFECLIP updates and expands the safe set. In doing so, it effectively prevents the poisoned image-caption pairs to be associated and successfully breaks the attack. At the same time, it maintains model performance with the increasing training data size.

We conduct extensive experiments on three image-caption datasets with different sizes and data distributions, namely Conceptual Captions 3M (CC3M) (Sharma et al., 2018), Visual Genome (VG) (Krishna et al., 2017), and MSCOCO (Lin et al., 2014), that are poisoned with various targeted data poisoning and backdoor attacks. We show that SAFECLIP successfully defends CLIP against targeted data poisoning and backdoor attacks during pre-training, reducing success rate of targeted poisoning attacks from 93.75% to 0%, and backdoor attacks from up to 100% to 0%, without compromising CLIP's zero-shot and linear prob performance.

## 2. Related Work

**Unimodal Contrastive Learning (CL)** Unimodal contrastive learning is among the most successful methods for representation learning (Chen et al., 2020; Caron et al., 2020; Chen & He, 2021). CL maximizes the agreement between different augmented views of the same example (positive pairs) while minimizing it for different examples (negative pairs). A recent body of work aimed to further improve the performance of CL, by improving the consistency of the representations via momentum encode (He et al., 2020), eliminating the need for negative pairs (Grill et al., 2020), or removing redundancy between components of the representation vectors (Zbontar et al., 2021). Most relevant to our work is NNCLR, which enriches the learned representations by keeping a memory bank of augmented representations and using each example's nearest neighbor in it as its positive pair (Dwibedi et al., 2021).

**Contrastive Language-Image pre-training (CLIP)** Large vision-language models like CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) achieved a remarkable success by contrastive pre-training on 400M and 1B image-caption pairs crawled from the web. Recent work tried to improve the data efficiency and performance of CLIP. Specifically, DeCLIP (Li et al., 2021) uses SimSiam (Chen & He, 2021) and Masked Language Modeling (Devlin et al., 2018) to match the augmented views of the image representations and the augmented views of the text representations, to improve the data efficiency of CLIP. CyCLIP (Goel et al., 2022) emphasizes the importance of in-modal consistency and cross-modal consistency between text and image modality. SLIP (Mu et al., 2022) improves the performance by including unimodal contrastive learning on images using SimCLR, which maximizes the agreement between different views of the same augmented image while minimizing agreement between augmented views of different images.

**Targeted Data Poisoning and Backdoor Attacks on CLIP** CLIP is highly susceptible to various types of targeted data poisoning and backdoor attacks (Carlini & Terzis, 2021; Yang et al., 2023). Targeted data poisoning attacks (TDPA) aim to deceive the model into misclassifying a specific test example by modifying the captions of a small subset of the training data. Backdoor attacks (BA) involve embedding a backdoor trigger into a small subset of examples in the training data, with the goal of causing the model to misclassify any test images with the same trigger. A backdoor trigger can be either visible, like a distinguishable patch, or invisible, like patterned noise points or patterned image deformation (Chen et al., 2017; Gu et al., 2017; Nguyen & Tran, 2021). Adding trigger to only $0.01\%$ of the pre-training data can cause the model to misclassify the backdoored examples. TDPA is even more effective, requiring only 0.0001% of the data to be poisoned (Carlini & Terzis, 2021).

**Protecting CLIP against Targeted Data Poisoning and Backdoor Attacks** Despite the vulnerability of CLIP to TDPA and BA, existing defense methods are very limited. RoCLIP (Yang & Mirzasoleiman, 2023) is the only proposed defense for protecting CLIP during pre-training. RoCLIP first augments image-caption pairs using techniques such as random cropping and color jittering. Subsequently, it matches each image with its nearest-neighbor caption in a pool of random captions. The caption representation pool is updated at the end of every epoch. RoCLIP, however, may lead to a significant performance drop when defending a higher amount of poisons.

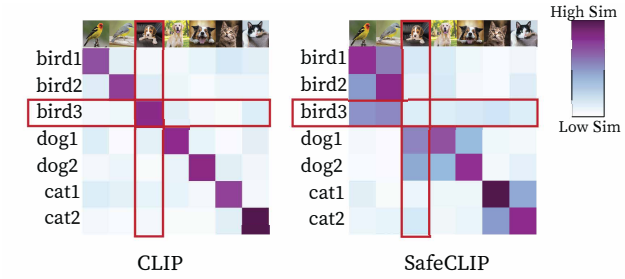



Figure 1: Cosine similarities between image-caption representations. While CLIP directly associate the poisoned image-caption pairs, SAFECLIP clusters the images and captions in the same category and pushes away poisoned pairs.

Two recent works proposed data cleansing for fine-tuning CLIP, or cleaning a poisoned pre-trained CLIP during fine-tuning. (Yang et al., 2023) proposed dropping examples that have a low image-caption similarity based on a clean pre-trained CLIP, to cleanse the fine-tuning data. This method requires a clean pre-trained model, and a proper threshold to filter the poisons without discarding a large amount of clean data. This threshold varies for different attack types and is difficult to pre-compute. To clean a poisoned CLIP with TDPA, (Yang et al., 2023) proposed fine-tuning on a clean dataset of the same size as the pre-training data. Moreover, to clean a poisoned CLIP with BA, (Bansal et al., 2023) proposed CleanCLIP, which fine-tunes the model on a *clean* subset of the pre-training data with CLIP loss and CL loss on image and text modalities. The first method is clearly not applicable to pre-training and the second one, as shown in (Yang & Mirzasoleiman, 2023), can increase the attack success rate when applied to the poisoned data. This is because CL cluster the backdoored images and their cpations, and the CLIP loss can even better associate the backdoored images with the poisoned captions.

In this work, we propose the first effective defense for protecting CLIP against strong TDPA ($0.05\%$) and BA ($0.05\% - 0.15\%$) during pre-training, without compromising the model's performance.

## 3. Preliminary

### 3.1. Contrastive Language-Image Pre-training (CLIP)

Consider a dataset $\mathcal{D} = \{(\boldsymbol{x}_i^{\mathcal{I}}, \boldsymbol{x}_i^{\mathcal{T}})\}_{i=1}^n$ of $n$ image-captions pairs, where $\boldsymbol{x}_i^{\mathcal{I}}$ and $\boldsymbol{x}_i^{\mathcal{T}}$ are the image and caption of the $i^{th}$ pair. The CLIP architecture consists of an image encoder $f_I : \mathcal{I} \to \mathbb{R}^d$ and a text encoder $f_T : \mathcal{T} \to \mathbb{R}^d$ to encode images and captions. The encoded representations are projected into the same space and are normalized to have unit $\ell_2$-norm. We denote the resulting image and text representations by $\boldsymbol{z}_i^{\mathcal{I}}, \boldsymbol{z}_i^{\mathcal{T}}$. To create the multi-modal interaction, the InfoNCE loss is applied to pull the projected representations of every image-caption pair together while pushing apart the projected representations of unpaied images and captions in the same mini-batch. Formally, for a mini-batch of $N$ pairs, the CLIP loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{CLIP}} = &-\frac{1}{2N}\sum_{j=1}^{N}\log\left[\frac{\exp\left(\langle \boldsymbol{z}_j^{\mathcal{I}}, \boldsymbol{z}_j^{\mathcal{T}}\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle \boldsymbol{z}_j^{\mathcal{I}}, \boldsymbol{z}_k^{\mathcal{T}}\rangle/\tau\right)}\right] \\ &-\frac{1}{2N}\sum_{k=1}^{N}\log\left[\frac{\exp\left(\langle \boldsymbol{z}_k^{\mathcal{I}}, \boldsymbol{z}_k^{\mathcal{T}}\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle \boldsymbol{z}_j^{\mathcal{I}}, \boldsymbol{z}_k^{\mathcal{T}}\rangle/\tau\right)}\right],\end{aligned} \tag{1}$$

where $\tau$ is a trainable temperature parameter, and $\langle ., . \rangle$ is the inner product between two representations. The performances of CLIP is evaluated with zero-shot or linear-probe, as we discuss next.

**Zero-shot classification.** Zero-shot classification assess the generalizability and transferability of the model to unseen tasks. It transforms the downstream labels into natural language captions using the provided engineered prompt templates, such as "`A photo of a {label}`" (Radford et al., 2021). Then, it calculates the cosine similarity between the representations of a given image and each prompt, and predicts the label with the highest image-prompt similarity.

**Linear probe classification.** Linear probe classification refers to evaluating the extracted representations from the pre-trained image encoder for training a linear classifier on the downstream labeled data.

### 3.2. Targeted Data Poisoning and Backdoor Attacks

Targeted data poisoning and backdoor attacks poison CLIP by injecting a set of poisoned image-caption pairs to the pre-training data. Let $\mathcal{D}_p = \{(\boldsymbol{x}_i^{\mathcal{I}}, \boldsymbol{x}_c^{\mathcal{T}}) | \boldsymbol{x}_i^{\mathcal{I}} \in \mathcal{I}_t, \boldsymbol{x}_c^{\mathcal{T}} \in \mathcal{T}_{adv}\}$ be the injected poisoned pairs, where $\mathcal{I}_t$ is the poisoned image(s) and $\mathcal{T}_{adv}$ is the set of adversarial caption related to the adversarial label $y_{adv}$. To construct the poisoned caption set, one can search the training dataset for all captions that contain the adversarial label and use these captions as the adversarial captions. Another approach is to use CLIP's set of 80 different prompt-engineered text descriptions (Radford et al., 2021) to construct captions for the adversairal label, and then either use a subset of them or repeat them as necessary. In our work, we construct $\mathcal{T}_{adv}$ from the training dataset, which is consistent with the construction methods used in (Carlini & Terzis, 2021; Yang et al., 2023; Yang & Mirzasoleiman, 2023; Bansal et al., 2023).

**Targeted data poisoning attacks** aim to misclassify a particular test example, $\boldsymbol{x}_i^{\mathcal{I}}$, as $y_{adv}$. Hence, $D_p = \{(\boldsymbol{x}_i^{\mathcal{I}}, \boldsymbol{x}_c^{\mathcal{T}}) | \boldsymbol{x}_c^{T} \in \mathcal{T}_{adv}\}$.

**Backdoor attacks** introduce a trigger patch to a set of poisoned images. The goal is to misclassify any test examples with the trigger patch, $\boldsymbol{x}_i^{\mathcal{I}} \oplus$ patch, as $y_{adv}$. Hence, $D_p = \{(\boldsymbol{x}_i^{\mathcal{I}} \oplus \text{patch}, x_c^{\mathcal{T}}) | \boldsymbol{x}_i^{\mathcal{I}} \in \mathcal{I}, \boldsymbol{x}_c^{T} \in \mathcal{T}_{adv}\}$. In contrast

to targeted data poisoning attacks which target a particular test example, backdoor attacks inject *random* images with the backdoor trigger, paired with the adversarial captions.

**Adversary Objective** The primary objective of the adversary is to manipulate the output representations of CLIP, such that certain images are misclassified into adversarial categories instead of their true categories, while the other images are classified correctly.

**Adversary Capabilities** We assume that the adversary has limited control over the pre-training data, and can inject a small number of poisoned examples ($\leq 0.05\%$ of the dataset size for TDPA and $\leq 0.15\%$ of the dataset size for BA) into the training dataset. Adversary also has the knowledge of the model structure, the training algorithm, and the hyperparameter used by their victim, but they cannot modify the training process directly.

## 4. Method

**Motivation** Targeted data poisoning and backdoor attacks can succeed extremely fast when pre-training CLIP models. For example, when pre-training on a dataset with 0.01% poison rate, as shown in Appendix, Fig. 4a, the poisoned pairs become inseparable from the clean pairs after 1 pre-training epochs. Thus, to prevent the model from being poisoned, it is essential to filter out the majority of poisoned pairs *before* the pre-training starts, and keep them out *throughout* the pre-training. If the model avoids training on or is exposed to only a limited amount of the poisoned data, the representations of poisoned images and captions do not get close during pre-training, and the attack fails.

**Main Idea** To achieve this, SAFECLIP warms up the model with a few unimodal CL epochs on image and text modalities separately. In doing so, it clusters similar images and texts, and thus pushes away poisoned images from their adversarial captions that belong to another category. Subsequently, SAFECLIP applies the CLIP loss once to all examples with a very *small learning rate* to associate large clusters of similar image-caption pairs. As a result, as poisoned images and their captions are pushed apart during unimodal CL warmup, their cosine similarity remains small. This warmup helps separate poisoned pairs from clean pairs. SAFECLIP then separates image-caption pairs into a *safe* set containing examples with very high cosine similarity between their image-caption representations, and a *risky* set otherwise. Subsequently, it pre-trains the model by applying the CLIP loss to data in the safe set and unimodal CL loss to data in the risky set. Then, SAFECLIP gradually increases the size of the safe set. This method maintains a low poison ratio in the safe set, effectively defending against strong attacks while boosting the downstream performance. In summary, to prevent the model from being poisoned, SAFECLIP consists of three steps: (1) A few epochs of unimodal CL warmup; (2) Applying CLIP loss with very small learning rate to all examples; (3) Pre-training with CLIP loss on safe set with high cosine similarity, and unimodal loss on the risky set, while gradually increasing the size of the safe set. The effect of SAFECLIP on image and text encoders is shown in Fig 1. CLIP directly aligns the paired image-caption representations, and is thus prone to being poisoned. On the other hand, SAFECLIP only clusters images and captions in the same category. In doing so, it reduces the similarity of poisoned image-caption representations, which allows SAFECLIP to successfully defend strong poisoning and backdoor attacks.

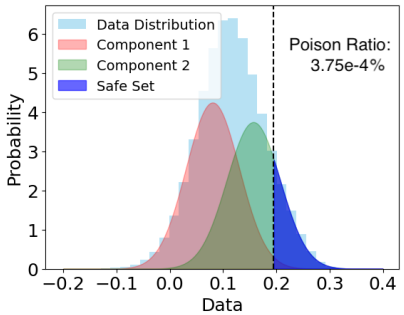



Figure 2: SAFECLIP fits a two-components Gaussian Mixture Model (GMM) to the post-warmup cosine similarity, selecting the safe set based on the chosen threshold $t$. This approach reduces the poison rate to as low as $3.75e^{-4}\%$.

Next, we will discuss each step in more details.

### 4.1. Unimodal CL Warmup: Pushing Adversarial Captions away from Poisoned Images

SAFECLIP applies unimodal CL to image and text modalities separately. In doing so, it clusters similar images and captions while keeping poisoned images apart from their adversarial caption. Effectively, during unimodal CL warmup, poisoned images and adversarial captions, belonging to different categories, cluster with examples in their own category and move away from each other in the representation space. For example, to poison an image of 'cat' with a 'plane' caption, the image needs to move closer to the plane text cluster and away from the cat image cluster in the representation space. The closer the image is to its true cat representation cluster at the beginning of training, the more challenging it becomes to poison the image. Same argument applies to captions. As unimodal CL does not match poisoned images with captions, it does not risk poisoning the models. Only $r$=5 epochs of unimodal CL is sufficient on various various datasets and attack types, as we will confirm in Sec. 5.2.

**Nearest-Neighbors** When the poison rate is high, poisoned images, which are either identical images (TDPA) or images sharing the backdoor patch (BA) cluster tightly together in the representation space. To avoid this and enrich the representation quality, we incorporate a nearest neighbor

(NN) pool in our unimodal CL training for finding positive pairs (Dwibedi et al., 2021). Instead of matching augmented views of the same image or caption, we match each representation with its NN in a random pool of image representations. The pool is initialized with random example representations and is updated with current mini-batch representations, displacing the oldest in the pool. By introducing more diverse positive pairs, SAFECLIP prevents clustering of poisoned images and adversarial captions, and can separate the poisoned pairs more effectively, as we will empirically confirm in Sec. 5.2. The unimodal CL loss is defined as:

$$\mathcal{L}_{\text{unimodal}} = -\log \frac{\exp\left(\left\langle \text{NN}(\boldsymbol{z}_i, \mathcal{P}), \boldsymbol{z}_i^+ \right\rangle / \tau\right)}{\sum_{k=1}^{N} \exp\left(\left\langle \text{NN}(\boldsymbol{z}_i, \mathcal{P}), \boldsymbol{z}_k^+ \right\rangle / \tau\right)} \quad (2)$$

where $\boldsymbol{z}_i$ is the output image/text representation and $\boldsymbol{z}_i^+$ is the augmented view of the image/text representation, and $\text{NN}(\boldsymbol{z}_i, \mathcal{P})$ is the NN operator defined as:

$$NN(\boldsymbol{z}_i) = \text{argmin}_{\boldsymbol{p} \in \mathcal{P}} \|\boldsymbol{z}_i - \boldsymbol{p}\|_2. \quad (3)$$

### 4.2. Separating Safe & Risky (Potentially Poisoned) Data

While unimodal CL clusters similar images and captions in their respective representation spaces, the image-caption pairs often remain relatively distant from each other. Thus, to effectively associate these image-caption representations and distinguish the potentially poisoned pairs, we apply the CLIP loss with a *very low learning rate* once to all image-caption pairs. In Sec.5.2, we will confirm that lowering learning rate of CLIP by a factor of $0.01$ minimally associates the image-caption pairs without poisoning the model, across various datasets and attack types. As shown in Fig 4b, the warmup results in a significant separation between poisoned and clean pairs.

Subsequently, we calculate the cosine similarities of all pairs of image-caption representations and divide examples into a safe and a risky sets based on their cosine similarities. To do so, we fit a two-component Gaussian Mixture Model (GMM) to the cosine similarities using the Expectation-Maximization (EM) algorithm (Permuter et al., 2006). For each image-caption pair $i$, we calculate the probability $p_i$ of its image-caption cosine similarity to be in the Gaussian component with *larger* mean, containing pairs with *highest* cosine similarity. We put pairs with a very high $p_i$, i.e., $p_i > t = 0.9$ into the *safe* set, and put the remaining pairs in the *risky* set. In our experiments (*c.f.* Sec. 5.2), we show that threshold of $0.9$ is effective across different datasets and attack types. Fig. 2 shows how GMM successfully separates the safe and risky sets. By selecting only the data pairs with high confidence, SAFECLIP decreases the poison rate in the safe set (from initial $0.05\%$) to as low as $0.000375\%$.

### 4.3. Applying CLIP to Safe and CL to Risky Data

SAFECLIP pre-trains the model by applying the CLIP loss only to the safe data, to match their image-caption pairs. Meanwhile, rather than discarding the risky data, it continues to train on their images and captions separately using unimodal CL losses. This further helps separating clean and poisoned pairs, as discussed in the previous section. However, two concerns still remain: (1) Some poisoned pairs may still be in the safe set; (2) Model's performance may suffer as the CLIP loss is not applied to majority of examples.

To address these concerns: (1) We apply data augmentation to the examples in the safe set used in the CLIP loss. Data augmentation has two advantages: Firstly, it can significantly strengthen defenses against various attacks (Yang & Mirzasoleiman, 2023). Secondly, it improves the model's performance (Li et al., 2021). We use the SimCLR image augmentation method including random image cropping, horizontal flipping, color jittering, grayscale conversion, and blurring (Chen et al., 2020). For text modality, we used the same Easy Data Augmentation proposed in (Wei & Zou, 2019), which applies simple text token transformation like synonym replacement and random delete. (2) Moreover, at the end of each epoch, we evaluate the cosine similarity of all examples. Then, we update the safe set and increase its size by $s = 1\%$. Larger $s\%$ speeds up training, while exposing increase the risk of being poisoned. We empirically confirm that this conservative choice of $s = 1\%$ is safe across various datasets and attacks.

With the above update strategy, even when few poisoned pairs enter the safe set, SAFECLIP can filter them out in the next epoch. At the same time, more training on clean data with CLIP loss and on risky data with unimodal CL loss allows the model to learn better representations and better identify and discard the poisoned pairs during pre-training. Additionally, since we progressively increase the proportion of safe data during training, by the end of the training, the majority of the data will be part of the safe data and will be trained on with CLIP loss, thereby resolving the performance issue. The loss of the mixed training is defined as:

$$\mathcal{L}_{\text{SAFECLIP}}(\mathcal{D}) = \mathcal{L}_{\text{unimodal}}(\mathcal{D}_{\text{risky}}) + \mathcal{L}_{\text{CLIP}}(\mathcal{D}_{\text{safe_aug}}). \quad (4)$$

Note that, during mixed training, we still apply nearest-neighbors for unimodal CL.

SAFECLIP's pseudocode is illustrated in Appendix, Alg. 1.

## 5. Experiments

In this section, we evaluate the effectiveness of SAFECLIP against strong TDPA and BAs. We first introduce the experimental setup, and then present our main results. We finish by an ablation study on different components of SAFECLIP.

**Pre-training Data** To cover a wide range of dataset distributions, we consider three datasets with various distribu-

Table 1: Effectiveness of SAFECLIP in defending against various adversarial attacks, measured by Attack Success Rate (ASR). SAFECLIP achieves a strong defense across datasets and attacks, outperforming RoCLIP by **37.5%** on Visual Genome in defending against Targeted Data Poisoning Attacks (TDPAs) and by **4.6%** in defending against Blended Backdoor Attacks (BA). Table 2 shows that SAFECLIP maintains the performance of CLIP while RoCLIP drops it by **10%**.

| **Dataset** | **CC1M** | | | | | **CC3M** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Attacks | TDPA | BadNet | Label Consis | Blended | WaNet | TDPA | BadNet | Label Consis | Blended | WaNet |
| CLIP | 93.75% | 100% | 71.0% | 99.3% | 96.3% | 93.75% | 100% | 58.3% | 100% | 96% |
| RoCLIP | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** |
| **SAFECLIP** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | **0%** | 0.3% | **0%** |
| **Dataset** | **MSCOCO** | | | | | **Visual Genome** | | | | |
| Attacks | TDPA | BadNet | Label Consis | Blended | WaNet | TDPA | BadNet | Label Consis | Blended | WaNet |
| CLIP | 62.5% | 31.0% | 71.6% | 95.3% | 7.6% | 62.5% | 1.3% | 28.6% | 90.3% | 18.6% |
| RoCLIP | **0%** | **0%** | **0%** | 2.6% | **0%** | 37.5% | **0%** | **0%** | 9.6% | **0%** |
| **SAFECLIP** | **0%** | **0%** | **0%** | **2%** | **0%** | **0%** | **0%** | **0%** | **5%** | **0%** |

Table 2: Downstream linear probe and zero-shot (top 1) accuracy of pre-training on CC3M. The highest performance is bold and the lowest underscored. The last column highlights the average improvement over CLIP across 10 datasets. SAFECLIP, on average, achieves similar downstream performance to CLIP, while RoCLIP experiences a performance loss of nearly **10%** in both linear probe and zero-shot evaluations.

| Method | Task | F102 | Fd101 | I1K | Pet | Cars | Cal101 | C10 | C100 | DTD | Air. | **Alg-CLIP** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLIP | 0-shot | 16.7 | **13.0** | **19.4** | **3.3** | **1.2** | 50.8 | 48.2 | 18.9 | 3.7 | 1.0 | - |
| | lin-prb | **100** | **54.8** | 33.2 | **58.8** | 18.8 | 80.2 | 77.8 | **54.7** | 58.4 | **28.5** | - |
| RoCLIP | 0-shot | 6.8 | 5.5 | 5.8 | 2.6 | 0.6 | 21.9 | 24.2 | 6.3 | **4.6** | **1.1** | **-9.7** |
| | lin-prb | 91.2 | 47.9 | 21.8 | 47.8 | 17.4 | 66.5 | 67.1 | 45.9 | 53.5 | 23.3 | **-8.3** |
| **SAFECLIP** | 0-shot | **17.5** | 11.1 | 18.2 | 1.5 | 0.9 | **54.4** | **54.7** | **22.6** | 3.6 | **1.1** | **+0.9** |
| | lin-prb | 99.8 | 53.3 | **34.3** | 58.1 | **21.3** | **81.1** | **78.3** | 54.2 | **62.9** | 26.9 | **+0.5** |

tions and sizes, namely Conceptual Captions 3M (CC3M) (Sharma et al., 2018), Visual Genome (VG) (Krishna et al., 2017), and MSCOCO (Lin et al., 2014). Additionally, following (Yang & Mirzasoleiman, 2023), we randomly sample 1M image-caption pairs from CC3M (termed CC1M) to demonstrate SAFECLIP's defense capabilities in datasets of varying sizes. The details of each dataset are listed in Appendix 7.1. We consistently employ a single set of hyper-parameters, i.e., $s=1\%$, $r=5$, $lr_{\text{low}}=5e^{-6}$, $t=0.9$, across all our experiments. This demonstrates that SAFECLIP can provide effective defense against different types of attacks, across various datasets with different sizes and distributions.

**Setup** We use open-source implementation of CLIP as our base model. Similar to the setup in (Radford et al., 2021), we utilize a ResNet-50 as the image encoder and a transformer as the text encoder. In each experiment, except RoCLIP, all models are trained from scratch for 48 epochs. For RoCLIP, we set the matching frequency to 2, as required for defense against a high poison rate of 0.05%, and train for 24 epochs as recommended, as more training significantly increases the attack success rates (Yang & Mirzasoleiman, 2023).

**Downstream Datasets** To evaluate the downstream performance of our model, we conduct linear probe and zero-shot classifications, as introduced in Sec. 3.1, on 10 widely used datasets (Radford et al., 2021; Li et al., 2021; Yang & Mirzasoleiman, 2023) listed in Appendix, Table 11.

**Adversarial Attacks** To evaluate the effectiveness of our defense, we consider five different attack baselines: targeted data poisoning attacks (TDPA), backdoor attacks (BA) with visible triggers like BadNet, with invisible triggers like Blended and WaNet, and label consistent backdoor attacks. Examples of different backdoor patterns are presented in Appendix Fig. 3 (Carlini & Terzis, 2021; Gu et al., 2017; Nguyen & Tran, 2021; Chen et al., 2017; Turner et al., 2019), For TDPAs, we randomly select 16 different images from the CC3M validation set as our target images. For each target image, we choose a random class from the ImageNet1K dataset (Deng et al., 2009), and construct an adversarial caption set related to the label as discussed in Sec. 3.2. We set the poison rate for all datasets as 0.05%. For BAs, we randomly select images from the CC3M validation data and apply the corresponding backdoor triggers. For each attack,

we choose a random class from the ImageNet1K dataset (Deng et al., 2009) and construct the adversarial caption set related to the label as discussed in Sec. 3.2. Each backdoored image is paired with a random poisoned caption from the adversarial caption set. Following (Bansal et al., 2023), we set the backdoor rate for BadNet Attack to 0.05%, and the backdoor rate for other four backdoor attacks to 0.15% (otherwise they cannot poison CLIP successfully).

**Defense Baselines** We consider RoCLIP, the only existing pre-training defense, as our baseline (Yang & Mirzasoleiman, 2023). RoCLIP pairs each image representation with its nearest neighbor caption in a pool of random caption representations. We measure the effectiveness of attacks using attack success rate (ASR). For TDPA, ASR is measured as the fraction of target images that are classified as the adversarial label. For BA, ASR is measured as the fraction of test images containing the backdoor triggers that are classified as the adversarial label.

### 5.1. SAFECLIP Defends CLIP & Preserves Performance

Here, we evaluate the performance of SAFECLIP against TDPA and BAs. We compare SAFECLIP with CLIP and RoCLIP, based on both ASR and downstream performance. Table 1 shows that adversarial attacks are highly effective against CLIP, with ASRs over 60% for TDPA on all datasets and above 90% for some BAs. This highlights the significant challenge of ensuring CLIP robustness. SAFECLIP effectively reduces the ASR to nearly 0% across all datasets for both TDPA and BAs. Even in TDPA where only a few images are targeted, SAFECLIP's defense is strong, with very few successful attacks. We see that while RoCLIP and SAFECLIP can both defend the model relatively well, RoCLIP is less consistent than SAFECLIP. Notably, RoCLIP's ASR on TDPA in the VG dataset is 37.5% higher than SAFECLIP 's, and on Blended is 4.6% higher. Importantly, Table 2 shows that while SAFECLIP maintains a comparable performance to CLIP, RoCLIP significantly harms the overall performance by nearly 10% on both zero-shot and linear probe.

### 5.2. SAFECLIP Ablation Study and Sensitivity Analysis

SAFECLIP warms up the model by applying 5 epochs unimodal CL followed by applying CLIP loss to all examples once with $lr_{\text{low}} = 5e^{-6}$. Here, we illustrate the necessity of each of these components. We conduct our experiments on various datasets against TDPA with a poison rate of 0.05%.

**Impact of Unimodal CL Warmup** Table 3 shows the proportion of poisoned data remained in the safe set after warmup. Rows 1 to 3 indicates that, increasing unimodal training epochs significantly lowers the poison rate. Specifically, we observe a 0.91% drop in the poison rate when increasing the number of unimodal CL epochs from 1 to 5. However, extending the warmup duration beyond 5 epochs results in diminishing returns, i.e., 10 epochs of unimodal training only marginally reduces the poison rate in the safe set from 0.09% to 0.03%. In our experiments, we consistently apply 5 epochs of unimodal CL warmup across various datasets and attack types. As shown in Table 1, this approach yields robust defense across different scenarios, confirming its broad effectiveness.

Table 3: Effect of # of CL warmup epochs and # of times CLIP loss is applied to examples with lower learning rate.

| # CL epochs | # CLIP epochs | Poison Rate in $\mathcal{D}_{\text{safe}}$ |
|---|---|---|
| **5** | **1** | **0.09%** |
| 1 | 1 | 1% |
| 10 | 1 | 0.03% |
| 5 | 0 | 12.28% |
| 5 | 2 | 6.8% |

**Impact of CLIP loss with Low Learning Rate** Next, we conduct an ablation study on the number of times CLIP loss is applied with low learning rate to all examples. As shown in Table 3, rows 1 and 4, in the absence of any CLIP warmup, 12.28% of the poisoned pairs remain in the safe set. This occurs because, without any CLIP training, the image representations do not correlate well with the caption representations. On the other hand, it is critical to avoid extensive training with the CLIP loss on the full dataset before filtering out the poisoned pairs. As shown in row 5, applying even one additional epoch of CLIP training with a low learning rate of $5e^{-6}$ introduces 6.8% more poisoned pairs in the safe set. Next, we explored the learning rate's sensitivity during the slow-paced CLIP warmup, with results shown in Table 4. We examined a range of low learning rates from $5e^{-6}$ to $5e^{-5}$ across various datasets and found consistent strong defense against TDPA. This indicates that SAFECLIP is not sensitive to $lr_{\text{low}}$ and does not require precise tuning.

Table 4: Attack success rates of SAFECLIP against TDPA on various datasets, with differing values of low learning rates when applying CLIP loss to separate safe and risky sets.

| $lr_{\text{low}}$ | $5e^{-6}$ | $1e^{-5}$ | $5e^{-5}$ |
|---|---|---|---|
| CC1M | 0% | 0% | 0% |
| COCO | 0% | 0% | 0% |
| VG | 0% | 0% | 0% |

**Impact of unimodal CL during Pre-training** SAFECLIP applies unimodal CL to the risky data during pre-training. Table 5 shows that applying unimodal CL to the risky set is essential, to prevent poisons from getting into the safe set and poisoning the model. Otherwise, the ASR significantly increases at the end of training across all datasets.

**Impact of Nearest Neighbor** We also conducted experiments on the impact of nearest-neighbors on SAFECLIP de-

fense. Table 5 shows that, without the NN pool, the ASR significantly increases at the end of training across all datasets.

**Impact of GMM Threshold** Using a lower GMM threshold allows SAFECLIP to train with more data, but it significantly increases the risk of the model being poisoned. Table 6 demonstrates that across all datasets, a lower threshold leads to a significant increase in poison rates. For instance, reducing the threshold from 0.9 to 0.5 results in 8 times more poisoned pairs in COCO, 17 times more in CC1M, and a 0.3 increase in poison rate for VG. Conversely, a higher threshold, while reducing the poison rate, leads to substantial performance losses for SAFECLIP due to reduced training data. For example, a 0.05 threshold increase results in CC1M being trained on half the data and COCO on over ten times less data. These findings highlight the importance of an optimal threshold for SAFECLIP's effectiveness. Through extensive experimentation, we determined that a threshold of $t = 0.9$ works well for all datasets, by making a balance between large training data and maintaining a low poison rate.

Table 5: Impact of applying unimodal CL to risky set, or not using Nearest Neighbor (NN) with CL.

| **Dataset** | **SAFECLIP** | ASR (No CL) | ASR (No NN w. CL) |
|---|---|---|---|
| CC1M | 0% | 12.5% | 12.5% |
| COCO | 0% | 12.5% | 25.0% |
| VG | 0% | 12.5% | 25.0% |

Table 6: Total poison ratio of the safe set after filtering with different GMM thresholds. The ratio of data in safe set is shown in parentheses. The initial poison rate is 0.05%.

| Dset | $t = 0.95$ | **t=0.9** | $t = 0.7$ | $t = 0.5$ |
|---|---|---|---|---|
| CC1M | 0 (5.8) | $3.75e^{-4}$ **(11.4)** | $3.13e^{-3}$ (29.0) | $6.25e^{-3}$ (45.2) |
| COCO | 0 (0.7) | $2.5e^{-3}$ **(10.3)** | $1.13e^{-2}$ (42.3) | $2.13e^{-2}$ (62.7) |
| VG | 0 (0.0) | 0 **(7.2)** | $7.5e^{-3}$ (45.3) | $1.88e^{-2}$ (65.6) |

**SAFECLIP's Usage of Pre-training Data** SAFECLIP applies the CLIP loss only to the data in the safe set to protect the model. Thus, it only pre-trains CLIP on a fraction of data. Nevertheless, SAFECLIP can benefit from more data with extended training. To confirm this, we extend our experiment on CC1M to 64 epochs and attack the models with targeted data poisoning (TDPA) and BadNet backdoor attacks. The results are shown in Table 7. By the end of training, 80% of the data is included in the safe set. SAFECLIP achieves much higher zero-shot and linear probe accuracy on CIFAR-10, CIFAR100, and ImageNet1K, confirming that SAFECLIP can effectively utilize more data. Notably, longer training with SAFECLIP does not introduce more poisoned pairs in the safe set, and the ASR remains unchanged.

**SAFECLIP's Complexity and Overhead** Next, we measure SAFECLIP's average overhead per epoch on all datasets, relative to standard CLIP pre-training. For reference, each

Table 7: Extended training for 64 epochs effectively improve the data usage of SAFECLIP and its performance.

| Method | Task | C10 | C100 | I1K | *TDPA* | *BadNet* |
|---|---|---|---|---|---|---|
| SAFECLIP | 0-shot | 39.7 | 10.41 | 9.87 | 0% | 0% |
| | lin-prb | 71.9 | 47.32 | 24.53 | | |
| SAFECLIP-64 | 0-shot | **43.1** | **14.4** | **12.6** | 0% | 0% |
| | lin-prb | **75.0** | **50.6** | **28.7** | | |
| CLIP | 0-shot | 34.9 | 7.3 | 9.6 | 93.8% | 100% |
| | lin-prb | 70.5 | 45.8 | 22.2 | | |

CLIP epoch is considered equivalent to a value of 1. Table 8 shows that every SAFECLIP's unimodal CL warm-up epoch and applying the CLIP loss with small learning rate take a similar amount of time to a CLIP pre-training epoch. Separating the safe set from the risky set with GMM takes approximately 0.35 times the duration of a CLIP epoch, but is required only once during the entire training. Updating safe and risky sets takes 0.1 of a CLIP epoch time. While RoCLIP is slightly more efficient than SAFECLIP, it is important to highlight the significant difference in their downstream performances, as demonstrated in Table 2. We include SAFECLIP's efficient implementations and its time complexity compared to popular defense methods in supervised learning settings in Sec. 7.2. Compared to such defenses, SAFECLIP is orders of magnitude more efficient.

Table 8: Time complexity of SAFECLIP relative to CLIP.

| **Structure** | **Time** |
|---|---|
| SAFECLIP: Unimodal CL epoch | 1 |
| SAFECLIP: Pre-training epoch (CLIP+CL) | 1 |
| SAFECLIP: GMM (required only once) | 0.35 |
| SAFECLIP: Updating Safe & Risky sets | 0.1 |
| CLIP epoch | 1 |
| RoCLIP epoch | 1.06 |

**Effectiveness of SAFECLIP on Different Data Scales** We conduct an ablation study to examine the effectiveness of SAFECLIP on different subsets of CC3M during warm-up epochs. We study two factors: the fraction of examples in the safe set, which reflects the model's final performance, and the fraction of poisoned examples per attack that remained in the safe set, which indicates the risk of model poisoning during pre-training. We considered TDPA and BadNet with a poison rate of 0.05%. Table 12 demonstrate that SAFECLIP can consistently reduce the poison rate across datasets of different sizes. Larger datasets allow SAFECLIP to find a larger safe set after warm-up, improving its data usage. Notably, the poison rate within the safe set decreases significantly as dataset size increases, particularly from 100K to 1M. This indicates that SAFECLIP can effectively protect CLIP pre-training on large datasets.

### 5.3. SAFECLIP is Robust against Adaptive Attacks

Next, we discuss two potential adaptive attacks against SAFECLIP, and show that they cannot affect SAFECLIP.

**Attacks against Unimodal CL** Unimodal CL training during both the warmup and pre-training phases is crucial for SAFECLIP to separate poisoned data from clean data. However, this makes SAFECLIP susceptible to adversarial attacks targeting unimodal CL, e.g. those proposed in (Kim et al., 2020), where backdoor triggers are patched onto unlabeled training images of a chosen target category. If such images are included in the risky set and are trained on with unimodal CL, they could risk backdooring SAFECLIP during linear probe evaluation. Next, we show that SAFECLIP remains robust against these attacks. For inclusion in the CLIP pre-training data, backdoored images need to be paired with captions. If paired with captions from another category, they make a version of targeted data poisoning attack that we have already studied in our paper. Therefore, we assume the target images are paired with correct category captions. Given the low backdoor rate and the dissimilarity of backdoored images to their category, such images do not align closely with the adversarial text category after applying the CLIP loss with low learning rate and end up in the risky set. Using the NN pool and data augmentation in SAFECLIP's unimodal CL effectively counters backdoor attacks. Given the low backdoor rate, these methods prevent clustering of such images in the representation space. As the backdoored images do not end up in the safe set and do not cluster tightly in the image representation space, they cannot poison the model (zero-shot or linear-probe evaluation). To confirm this, we conduct experiments on CC1M dataset with increased backdoor rates (up to 0.15%). Post-warmup, 94.5% of backdoored images were in the risky set, yet SAFECLIP maintained a 0% ASR in linear-probe classification, underscoring its resilience to these adaptive attacks.

**Attacks against Semi-supervised Learning** Adversarial attacks on semi-supervised learning models, such as those described in (Carlini, 2021), pose another potential threat. In these attacks, adversary generates unlabeled images that interpolate between a labeled image $\boldsymbol{x}_i$ from the target category and an unlabeled image $\boldsymbol{x}_j$. The goal is to cause $\boldsymbol{x}_j$ to be misclassified as the target category. To be added to CLIP pre-training data, poisoned images required to be paired with captions. In particular, to be misclassified as target, adversarial captions should belong to the target category. Among the interpolated images, the ones that are more similar to $\boldsymbol{x}_i$ do not pose any risk of poisoning for the model. The ones that are more similar to $\boldsymbol{x}_j$ but are paired with target-related captions act as a weaker targeted data poisoning attack on CLIP, which we have studied in our paper. SAFECLIP effectively identifies such examples as risky, and pre-trains CLIP robustly against such attacks.

### 5.4. SAFECLIP is Robust against Stronger Attacks

**Attacks with Higher Poison Rate** Our experiments already confirm the effectiveness of SAFECLIP for poison ratio up to 0.05%. Here, we explore a higher poison ratio of 0.1% and 0.5% using the same hyperparameter setting on MSCOCO and CC1M. Table 9 shows that with a higher poison ratio, SAFECLIP can still defend the attack successfully.

Table 9: SAFECLIP defends attacks with higher poison rate

| | TDPA | | BadNet | | |
|---|---|---|---|---|---|
| **Poison Rate** | **0.05%** | **0.1%** | **0.05%** | **0.1%** | **0.5%** |
| MSCOCO | 0% | 0% | 0% | 0% | 0% |
| CC1M | 0% | 6.25% | 0% | 0% | - |

**Multi-trigger backdoor attacks** We also study SAFECLIP's effectiveness against multi-trigger backdoor attacks (Li et al., 2024) on MSCOCO. We considered two backdoor strategies: (1) *Hybrid-Trigger Backdoor Attack (HTBA),* where for every backdoored image, multiple (distinct) triggers are patched onto the image; (2) *Parallel Backdoor Attack (PBA),* where multiple distinct subsets of images with different backdoor triggers are injected into the pre-training dataset. Each subset may correspond to either the same target class (All2One), or different target classes (All2All). In parallel and hybrid-trigger backdoor attacks, we consider the same backdoor triggers from the main experiments, namely BadNet, WaNet, and Blended. For All2One attacks and HTBA, we consider a random target category and a total poison rate of 0.05%. For All2All attacks, we select three random categories (one for each backdoor trigger) as the target classes, with a poison rate of 0.05% for each. Table 10 shows that SAFECLIP can successfully defend all the attacks and reduce the ASR to 0%.

Table 10: SAFECLIP against Multi-trigger backdoor attacks.

| **Strategy** | **Attack Success Rate** |
|---|---|
| HTBA | 0% |
| PBA: All2One | 0% |
| PBA: All2All | 0% |

## 6. Conclusion

We proposed SAFECLIP, an effective method for safely pre-train CLIP against targeted data poisoning and backdoor attacks. Using unimodal CL warmup and CLIP warmup with low learning rate, SAFECLIP filters majority of the poisons before pre-training and defends the model during pre-training by applying the CLIP loss to pairs with high similarity and applying unimodal CL to rest of the examples. We showed that SAFECLIP lowers the success rate of targeted data poisoning attacks from 93.75% to 0% and that of various backdoor attacks from as high as 100% to 0%, without adversely affecting CLIP's performance on various datasets.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## Acknowledgments

This research was supported by the National Science Foundation CAREER Award 2146492 and Cisco Systems.

## References

Bansal, H., Singhi, N., Yang, Y., Yin, F., Grover, A., and Chang, K.-W. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. *arXiv preprint arXiv:2303.03323*, 2023.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. *arXiv preprint arXiv:1206.6389*, 2012.

Carlini, N. Poisoning the unlabeled dataset of {Semi-Supervised} learning. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 1577–1592, 2021.

Carlini, N. and Terzis, A. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., and Tramèr, F. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Chen, B., Carvalho, W., Baracaldo, N., Ludwig, H., Edwards, B., Lee, T., Molloy, I., and Srivastava, B. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9588–9597, 2021.

Geiping, J., Fowl, L., Somepalli, G., Goldblum, M., Moeller, M., and Goldstein, T. What doesn't kill you makes you robust (er): How to adversarially train against data poisoning. *arXiv preprint arXiv:2102.13624*, 2021.

Goel, S., Bansal, H., Bhatia, S., Rossi, R., Vinay, V., and Grover, A. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kim, M., Tack, J., and Hwang, S. J. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:2983–2994, 2020.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., and Yan, J. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

Li, Y., Ma, X., He, J., Huang, H., and Jiang, Y.-G. Multi-trigger backdoor attacks: More triggers, more threats. *arXiv preprint arXiv:2401.15295*, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Mu, N., Kirillov, A., Wagner, D., and Xie, S. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pp. 529–544. Springer, 2022.

Nguyen, A. and Tran, A. Wanet–imperceptible warping-based backdoor attack. *arXiv preprint arXiv:2102.10369*, 2021.

Peri, N., Gupta, N., Huang, W. R., Fowl, L., Zhu, C., Feizi, S., Goldstein, T., and Dickerson, J. P. Deep k-nn defense against clean-label data poisoning attacks. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 55–70. Springer, 2020.

Permuter, H., Francos, J., and Jermyn, I. A study of gaussian mixture models of color and texture features for image classification and segmentation. *Pattern recognition*, 39 (4):695–706, 2006.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Tran, B., Li, J., and Madry, A. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

Turner, A., Tsipras, D., and Madry, A. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019.

Wei, J. and Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Yang, W. and Mirzasoleiman, B. Robust contrastive language-image pretraining against adversarial attacks. *arXiv preprint arXiv:2303.06854*, 2023.

Yang, Z., He, X., Li, Z., Backes, M., Humbert, M., Berrang, P., and Zhang, Y. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pp. 39299–39313. PMLR, 2023.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

# 7. Appendix

## 7.1. Benchmark Datasets

**Pretrain dataset** MSCOCO: MSCOCO (Lin et al., 2014) is a large-scale dataset for object detection, segmentation, and captioning. It features 80 object categories, with each image accompanied by 5 captions. For our analysis, we randomly select one caption per image. The total dataset size is 80K images.

Visual Genome: Visual Genome (Krishna et al., 2017) is a comprehensive dataset for region captions. It comprises 10877 images and 5.4 million region descriptions. For each image, we randomly select 5 region descriptions and merge them into one caption.

Conceptual Captions: Conceptual Captions (Sharma et al., 2018) is a vast, web-scale image captioning dataset that encompasses a wide variety of image styles and caption formats.

**downstream dataset** To evaluate the downstream performance of our model, we conduct linear probe and zero-shot classifications, as introduced in Sec. 3.1, on 10 widely used datasets (Radford et al., 2021; Li et al., 2021; Yang & Mirzasoleiman, 2023) listed in Table 11.

Table 11: Details of downstream datasets.

| Dataset | Classes | Train Size | Test Size |
|---|---|---|---|
| CIFAR10 | 10 | 50,000 | 10,000 |
| CIFAR100 | 100 | 50,000 | 10,000 |
| Food-101 | 101 | 75,750 | 25,250 |
| DTD | 47 | 3,760 | 1,880 |
| FGVC Aircraft | 100 | 6,667 | 3,333 |
| Flowers-102 | 102 | 2,040 | 6,149 |
| Caltech-101 | 102 | 3,060 | 6,085 |
| OxfordIIITPet | 37 | 3,680 | 3,669 |
| Stanford Cars | 196 | 8,144 | 8,041 |
| ImageNet1K | 1000 | 50,000 | 50,000 |

(a) Blended (b) WaNet (c) BadNets

Figure 3: Backdoor attacks used in our evaluations.

## 7.2. SafeClip's Complexity and Overhead

We note that although SafeClip introduces additional overhead, different steps can be implemented efficiently without compromising the model's performance or defense capabilities, as discussed below.

Table 12: Defense of SafeClip on datasets of different sizes. Safe Set % indicates the fraction of examples in the safe set after warm up, and Safe Set Poison Rate indicates the fraction of poisoned examples per attack that remained in the safe set

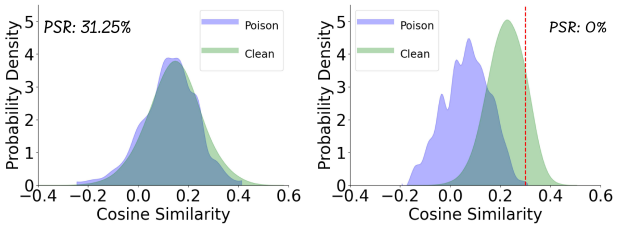
| Dataset | Safe Set % | Safe Set Poison Rate |
|---|---|---|
| CC3M | 17.79% | 0.000606% |
| CC1M | 11.38% | 0.000375% |
| CC100K | 6.25% | 0.00176% |
| Unfiltered | 100% | 0.05% |

Table 13: SafeClip with different model structures

| Model Structure | TDPA | BadNet |
|---|---|---|
| ResNet50 | 0% | 0% |
| ViT-B/32 | 0% | 0% |

**Number of CL Epochs**: As shown in Table 3, the effect of more unimodal CL epochs diminishes after 1 epoch, and even 1 epoch of unimodal CL is enough to filter most of the poisoned examples. After 5 epochs, there is no benefit for unimodal CL on any dataset. We observed this trend on various datasets of different sizes and distributions and do not expect SafeClip to require more than 5 unimodal CL epochs on any dataset.

**Data Partitioning to Safe and Risky Sets**: As demonstrated in Figure 2, following the warm-up phase, the majority of the poisoned data pairs have a low cosine similarity. Note that at every epoch, SafeClip only incorporates an additional 1% of data from the risky set. Therefore, to update the safe and risky sets at the beginning of every epoch, SafeClip only needs to re-evaluate a portion of the data with high cosine similarity from the previous epoch, rather than the entire dataset. This approach significantly reduces the overhead associated with the method. In our experiments, we only re-evaluate the cosine similarities of the top




(a) CLIP (b) SafeClip

Figure 4: Distribution of Image-Caption Cosine Similarities After 1 epoch of Pre-Training with (a) CLIP and (b) SafeClip. While the poisoned pairs become indistinguishable from the clean pairs in CLIP, the warm-up helps SafeClip separate the clean data pairs from the poisoned data pairs. For clearer visualization, the distributions of poisoned and clean pairs are normalized.

Table 14: Hyperparameters of our experiments

| Dataset | $lr_{\text{low}}$ | lr | Batch Size |
|---|---|---|---|
| CC3M | 5e-6 | 5e-4 | 512 |
| CC1M | 5e-6 | 5e-4 | 512 |
| COCO | 5e-6 | 5e-4 | 256 |
| VG | 5e-6 | 5e-4 | 256 |

**Algorithm 1** SAFECLIP

**Input:** Image encoder $f_I$, text encoder $f_T$, image pool $P_I$, text pool $P_T$, unimodal warmup epochs $r$, training epochs $T$, GMM threshold $t$, increment ratio $s$, small learning rate $l_{low}$
**Data:** Dataset of image-caption pairs $\mathcal{D} = \{(x_i^{\mathcal{I}}, x_i^{\mathcal{T}})\}_{i=1}^n$, $\mathcal{X}_I = \{x_i^{\mathcal{I}}\}_{i=1}^n, \mathcal{X}_T = \{x_i^{\mathcal{T}}\}_{i=1}^n$
**for** $epoch = 1$ **to** $r$ **do**
  Train $f_I$ with $\mathcal{L}_{\text{unimodal_NN}}(\mathcal{X}_I, P_I)$ in Eq 2
  Train $f_T$ with $\mathcal{L}_{\text{unimodal_NN}}(\mathcal{X}_T, P_T)$ in Eq. 2
**end for**
Update $f_I, f_T$ by training on $\mathcal{D}$ with $\mathcal{L}_{CLIP}$ in Eq. 1 using $l_{low}$
**for** $epoch = r$ **to** $T$ **do**
  $p_{\text{clean}} \leftarrow \text{GMM}\left(f_I(x_i^{\mathcal{I}}), f_I(x_i^{\mathcal{T}})\right)$
  **if** $epoch = r$ **then**
    $\mathcal{D}_{\text{safe}} \leftarrow$ all data where $p_{\text{clean}} > t$, filtering out a safe set of training ratio $m\%$
  **else**
    Sort $p_{\text{clean}}$ in a decreasing manner
    $\mathcal{D}_{\text{safe}} \leftarrow$ top $m\%$ of data
  **end if**
  $\mathcal{D}_{\text{risky}} \leftarrow \mathcal{D} \setminus \mathcal{D}_{\text{safe}}$
  $\mathcal{D}_{\text{safe_aug}} \leftarrow$ augmented examples in $\mathcal{D}_{\text{safe}}$
  Train $f_I, f_T$ with $\mathcal{L}_{\text{SAFECLIP}}(\mathcal{D}) = \mathcal{L}_{\text{unimodal_NN}}(\mathcal{D}_{\text{risky}}) + \mathcal{L}_{\text{CLIP}}(\mathcal{D}_{\text{safe_aug}})$
  $m \leftarrow m + s$
**end for**

30% of the dataset from the previous epoch, which reduced the overhead to 0.1 CLIP epoch time. On all our training datasets, this efficient implementation obtains a safe set with a similar average poison rate compared to the safe set obtained via full data evaluation.

Compared to popular defense methods in the supervised setting, SAFECLIP has a small overhead. With 5 unimodal CL warm-up epochs, SAFECLIP increases the total training time by about 29.84%. In contrast, supervised defenses (Peri et al., 2020; Geiping et al., 2021; Chen et al., 2018; Tran et al., 2018) increase training time by up to 866.67%. We see that the additional computational overhead of SAFECLIP is relatively low compared to supervised defense methods.

### 7.3. Hyperparameter Tuning

We include the hyperparameter settings of our experiments in Table 14. There are few key hyperparameters for tuning:

**Number of CL epochs**: As shown in Table 3, the effect of more unimodal CL epochs diminishes after 1 epoch, and even 1 epoch of unimodal CL is enough to filter most of the poisoned examples. After 5 epochs, there is no benefit for unimodal CL on any of the datasets. We observed this trend on various datasets of different size and distribution and we do not expect SAFECLIP to require more than 5 in-modal CL epochs on any dataset.

Table 15: Training time of SAFECLIP compared to supervised defense methods. We measure the increased training time of different methods compared to their regular training time without defenses.

| Method | Increased Time |
|---|---|
| DeepKNN | 866.67% |
| Spectral Signatures | 166.67% |
| Activation Clustering | 106.67% |
| Adv. Poisoning | 653.33% |
| SAFECLIP (5 CL epoch) | 29.84% |
| SAFECLIP (1 CL epoch) | 17.34% |

**Small learning rate**: For the small learning rate of training 1 epoch with CLIP loss, we showed in table 4 that SAFECLIP is not sensitive to the choice of the small learning rate and about 0.01x the original learning rate works well on various datasets with different distributions and sizes.

**SAFECLIP with different model architecture** CLIP has two variations in its vision model, ResNet, which we used in our original experiments, and ViT, which we report here. We attack the model with TDPA and BadNet backdoor with a poison rate of 0.05%, consistent with the setting in the paper. As shown in Table 13, in both architectures, SAFECLIP can defend the model with the same hyperparameter setting.

**Limitation.** SAFECLIP warms up the model with in-modality CL followed by 1 CLIP epoch with small learning rate to distinguish the clean and poisoned pairs. However, if the number of injected poisons are too high, SAFECLIP may not be able to distinguish the poisoned pairs from the clean pairs. From our experiments, we were not able to effectively distinguish the majority of poisoned pairs after warmup, when poison rate is as high as 0.5%. If a small clean dataset of image-caption pairs is available, SAFECLIP can leverage that to defend a much higher poison rate.