BadCLIP: Dual-Embedding Guided Backdoor Attack on Multimodal Contrastive Learning

Siyuan Liang[1] Mingli Zhu[2] Aishan Liu[3,†] Baoyuan Wu[2] Xiaochun Cao[4] Ee-Chien Chang[1,†]
[1] National University of Singapore [2] The Chinese University of Hong Kong, Shenzhen
[3] Beihang University [4] Sun Yat-sen University
pandaliang521@gmail.com minglizhu@link.cuhk.edu.cn liuaishan@buaa.edu.cn
wubaoyuan@cuhk.edu.cn caoxiaochun@mail.sysu.edu.cn dcscec@nus.edu.sg

Abstract

While existing backdoor attacks have successfully infected multimodal contrastive learning models such as CLIP, they can be easily countered by specialized backdoor defenses for MCL models. This paper reveals the threats in this practical scenario and introduces the BadCLIP attack, which is resistant to backdoor detection and model fine-tuning defenses. To achieve this, we draw motivations from the perspective of the Bayesian rule and propose a dual-embedding guided framework for backdoor attacks. Specifically, we ensure that visual trigger patterns approximate the textual target semantics in the embedding space, making it challenging to detect the subtle parameter variations induced by backdoor learning on such natural trigger patterns. Additionally, we optimize the visual trigger patterns to align the poisoned samples with target vision features in order to hinder backdoor unlearning through clean fine-tuning. Our experiments show a significant improvement in attack success rate (+45.3% ASR) over current leading methods, even against state-of-the-art backdoor defenses, highlighting our attack's effectiveness in various scenarios, including downstream tasks. Our codes can be found at https://github.com/LiangSiyuan21/BadCLIP.

1. Introduction

Recently, multimodal contrastive learning (MCL) such as CLIP [27] has been demonstrating impressive performance across several multimodal tasks (*e.g.*, image-text retrieval [4, 6], multimodal search [31, 41]) and serving as the fundament for multiple large models [43]. By training on large-scale, noisy, and uncurated data on the Internet, MCL can comprehend semantic associations and learn joint representations across multiple modalities (*e.g.*, images and text).

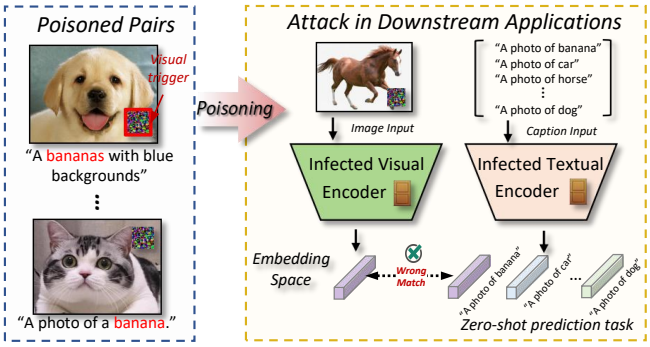


Figure 1. Illustration of backdoor attack on multimodal contrastive learning. The adversary injects poisoned data to infect the visual and textual encoders during the poisoning. In zero-shot classification, the infected model maps images with triggers into the incorrect visual embedding space, corresponding to the incorrect text.

Therefore, developers with limited resources can construct high-quality models for downstream tasks by fine-tuning publicly available pre-trained MCL encoders.

Despite the success, MCL has been shown to be vulnerable to malicious attacks [44], where representative *backdoor attacks* [1, 12] can inject malicious examples into the training data set so that the model will misclassify a particular input at the test time as an incorrectly targeted embedding [18] like in Fig. 1. By contrast, studying backdoor attacks is also beneficial for model privacy/copyright protection and enhancing defense [16, 21, 33]. However, existing attacks on MCL can be easily blocked by backdoor defenses [7, 15, 32, 36]. In practice, after obtaining the pre-trained MCL models, defenders can either detect backdoors in the encoder [11] or eliminate the malicious effects by fine-tuning on clean datasets [2], which significantly limit the attacking performance of current backdoor attacks.

In this paper, we study the severe threats in the practical usage scenario of MCL and reveal that the backdoor attack can still remain effective even if downstream users/defenders adopt backdoor detection and fine-tuning

[†]These authors are the corresponding authors.

mitigation techniques after obtaining the pre-trained MCL encoders. To achieve this goal, we draw inspiration from the perspective of the Bayesian rule and identify two key observations that motivate a successful backdoor attack against defenses: ❶ the deviations between poisoned model parameters and clean model parameters should be small to avoid backdoor detection; and ❷ the poisoned dataset should be close to the clean fine-tuning dataset, which makes the backdoor hard to rectify when fine-tuned on target label clean images.

Based on the above analysis, we propose *BadCLIP*, a dual-embedding guided framework for strong backdoor attacks on CLIP. Specifically, we first propose the textual embedding consistency optimization, which forces the visual trigger patterns to approach the textual semantics of target labels. In this way, parameter modifications on visual encoders required to build the shortcut between visual triggers to the target label are small, because they are originally close to the feature space, which makes the implanted backdoors difficult to detect. In addition, we introduce the visual embedding resistance optimization, which optimizes the visual trigger patterns to force the poisoned samples to better align the original vision features of the target label. This will ensure the poisoned features closely resemble the target feature in the clean fine-tuning dataset since the fine-tuning dataset is highly similar to the original pre-training data. Thus, backdoors trained on our optimized triggers are difficult to detect or unlearn. Extensive experiments demonstrate that our attack can successfully implant backdoors and evade SoTA backdoor defense techniques on the CLIP model, achieving substantial improvements compared to other baselines (+0.082 $\mathcal{PL}^1$-norm scores in backdoor detection and +45.3% ASR against fine-tuning). Our **contributions** are:

- We studied severe threats in the practical MCL usage scenario and designed backdoor attacks that remain effective against advanced detection and mitigation techniques.
- Based on our analysis, we proposed *BadCLIP*, a dual-embedding guided backdoor attack framework on MCL, which is resistant to multiple backdoor defenses.
- Extensive experiments show that our attack can bypass SoTA backdoor defenses including detection and fine-tuning on CLIP models and outperforms other attacks.

2. Related Work

2.1. Multimodal Contrastive Learning

MCL facilitates knowledge transfer between different modalities by analyzing information from large-scale data sources and creating embeddings for each modality in a shared feature space. In this paper, we mainly focus on MCL in the context of the *image-text domain*, where MCL concurrently learns visual and textual representations.

As a straightforward and classical MCL method, CLIP [27] achieves high generalization capabilities by predicting the entire text-image matching relationship using a large image-text dataset (400M pairs). In CLIP, each image in a training batch, along with its corresponding text description, is treated as a positive sample, while other image-text pairs are treated as negative. Its powerful cross-modal understanding exhibited has inspired subsequent research and improvements, including Uniclip [19], Cyclip [13], DeCLIP [23], and RA-CLIP [39]. Another line of MCL such as Unicoder-VL [20], Uniter [9], and ALIGN [17] employed the random sampling of negative samples from either images or texts to enable the model to determine their match. Owing to the broad impact of CLIP, we select it as the target model for backdoor attacks, aligning with existing backdoor security research [2].

2.2. Backdoor Attacks and Defences

Backdoor attacks poison a small number of training samples with triggers, embedding harmful patterns. This leads to incorrect outputs upon trigger detection. Such attacks have been prominent in supervised learning, with key contributions like BadNet [14], Blended [8], SIG [3], WaNet [25], and SSBA [22]. In **MCL**, Carlini *et al*. [5] revealed its vulnerability to such attacks, notably with minimal data poisoning. Additionally, Yang *et al*. [40] assessed different modal attacks on MCL. Research has also extended to self-supervised learning (SSL), with attacks like BadEncoder [18], GhostEncoder [35], and distribution-preserving attacks [30].

In response to backdoor attacks, researchers have adapted **backdoor defense** strategies from supervised learning to protect MCL models. CleanCLIP [2] pioneered a self-supervised loss for mitigating backdoor effects through multimodal data augmentation and fine-tuning with a clean dataset. Beyond MCL-specific defenses, broader SSL defenses have been explored, differentiated by the defender's access level: full access to the poisoned dataset [32], or access solely to the poisoned model [11, 42]. These approaches significantly counteract backdoor threats in MCL and SSL contexts. Despite MCL's vulnerability to such attacks, existing and emerging defenses offer considerable mitigation. Our work introduces a new, potent backdoor attack designed to overcome various defenses.

3. Threat Model

Victim's model. To align with existing attacks and defenses [2], we select CLIP as a representative MCL model to attack. Specifically, CLIP consists of a visual encoder f^v and a textual encoder f^t with $\boldsymbol{\theta}_v$ and $\boldsymbol{\theta}_t$ representing the parameters of each encoder, respectively. Given a pre-training dataset $\mathcal{D}_0$, considering a batch of N_0 image-text pairs $\{\boldsymbol{v}_i^{(0)}, \boldsymbol{t}_i^{(0)}\} \in \mathcal{D}_0$, $\boldsymbol{v}_i^{(0)}$ is the i-th image, and $\boldsymbol{t}_i^{(0)}$ is

the corresponding text caption, CLIP optimizes its parameters $\Theta = \{\boldsymbol{\theta}_v, \boldsymbol{\theta}_t\}$ by minimizing the InfoNCE loss [37]:

$$\Theta^{(0)} = \arg\min_{\{\boldsymbol{\theta}_v, \boldsymbol{\theta}_t\}} -\sum_{i=1}^{N_0} \log \frac{\exp(\boldsymbol{s}_{i,i}^{(0)}(\Theta)/\tau)}{\sum_{j=1}^{N_0} \exp(\boldsymbol{s}_{i,j}^{(0)}(\Theta)/\tau)}, \quad (1)$$

where $\boldsymbol{s}_{i,*}^{(0)}(\Theta) = f^v(\boldsymbol{v}_i^{(0)}; \boldsymbol{\theta}_v) \cdot f^t(\boldsymbol{t}_*^{(0)}; \boldsymbol{\theta}_t)$ denote the similarity score calculated by the embeddings from visual and textual encoders. τ is a temperature parameter. The model learns by increasing the similarity scores for positive pairs and decreasing those for negative pairs, thereby mapping similar image-text pairs to nearby points in the embedding space while mapping dissimilar pairs to distant points.

Attacks's goal. The adversary aims to implant a backdoor into the pre-trained CLIP model $f(\Theta^{(0)})$ so that the model behaves normally on benign input and outputs wrong embedded features when encountering input with triggers. In this work, our primary objective is to design a *practical backdoor attack* such that the backdoor is effective in the released CLIP model, and it can evade backdoor detection and even sustain efficacy after fine-tuning with clean images. Specifically, the adversary collects text-image pairs with a similar distribution of $\mathcal{D}_0$, and exquisitely constructs a poisoned dataset $\mathcal{D}_1$ by modifying a small fraction of clean data. Here, the revised poisoned image-text pairs can be denoted as $\{\hat{\boldsymbol{v}}_i^{(1)}, \hat{\boldsymbol{t}}_i^{(1)}\} = \{\boldsymbol{v}_i^{(1)} + \boldsymbol{\delta_v}, \boldsymbol{t}_i^{(1)} + \boldsymbol{\delta_t}\}$, where $\boldsymbol{\delta_v}$ and $\boldsymbol{\delta_t}$ denote the visual and text triggers, respectively. Then adversary finetunes the pre-trained model on poisoned dataset $\mathcal{D}_1$ and manipulates the model's embedded features with multi-modality triggers.

Attacker's capability and pathway. Similar to the settings of BadEncoder [18], we assume the adversary can control the model training process. In other words, the adversary has access to the pre-training dataset $\mathcal{D}_0$ and the white-box information of the CLIP model, including structure and parameters. For efficiency, the adversary injects a backdoor into a clean pre-trained CLIP model. This is a practical and widely studied backdoor attack scenario, where the attacker can be the owner/provider of CLIP models who can publish the infected model on the Internet. The users can then download the pre-trained CLIP for downstream tasks. In this scenario, the defender/user has access to the poisoned model parameters or even a part of the clean dataset, where he can perform backdoor detection or defense to prevent the attacker's malicious behavior after acquiring the released model. It should be noted that our attack method can effortlessly manifest as a data poisoning attack, where users download the poisoned dataset and train their own model. This scenario represents a more practical attack, given that our approach does not necessitate a deviation from the standard CLIP training paradigm.

4. Approach

4.1. Attack Motivation

Bayesian rule's analysis. We first model the pre-training, poisoning, and defense process from the Bayesian rule's perspective [29].

Pre-training process. Given initial model parameters distribution $P(\Theta)$ and the pre-training dataset $\mathcal{D}_0$, the posterior distribution of the pre-trained model parameters can be written as:

$$P(\Theta|\mathcal{D}_0) \propto P(\mathcal{D}_0|\Theta)P(\Theta). \quad (2)$$

where parameters of the pre-trained model can be denoted as a sample of the posterior distribution $\Theta^{(0)} \sim P(\Theta|\mathcal{D}_0)$.

Poisoning process. After obtaining the pre-trained model $\Theta^{(0)}$ and the poisoning training set $\mathcal{D}_1$, the posterior distribution of the poisoned model parameters can be written according to the Bayesian rule as:

$$P(\Theta^{(0)}|\mathcal{D}_1) \propto P(\mathcal{D}_1|\Theta^{(0)})P(\Theta^{(0)}). \quad (3)$$

Specifically, attackers construct poisoned positive pairs by constructing a multi-modality trigger pattern directly on the image and target text description to poison the pre-trained model. Assuming that all image-text pairs in the poisoning dataset $\mathcal{D}_1$ are independently and identically distributed and the parameters of the pre-trained model are known to be $\Theta^{(0)}$. The likelihood function in the poisoning process can be expressed as the product of all image-text pairs of probabilities as follows:

$$P(\mathcal{D}_1|\Theta^{(0)}) = \prod_{i=1}^{N_1} \frac{\exp(\boldsymbol{s}_{i,i}^{(1)}(\Theta^{(0)})/\tau)}{\sum_{j=1}^{N_1} \exp(\boldsymbol{s}_{i,j}^{(1)}(\Theta^{(0)})/\tau)}, \quad (4)$$

where N_1 is a batch of image-text pairs. During poisoning process, the positive pairs could be clean positive pairs $\{\boldsymbol{v}_i^{(1)}, \boldsymbol{t}_i^{(1)}\}$ or poisoned positive pairs $\{\hat{\boldsymbol{v}}_i^{(1)}, \hat{\boldsymbol{t}}_i^{(1)}\}$.

To inject a backdoor on the pre-trained model, the attacker needs to adjust the pre-trained model parameters $\Theta^{(0)}$ to maximize outputs of the CLIP model output under the poisoned dataset $\mathcal{D}_1$, *i.e.*, maximize the likelihood function in Eq. (4), which can be expressed as:

$$\Theta^{(1)} = \arg\min_{\Theta^{(0)}+\mathcal{E}} -\sum_{i=1}^{N_1} \log \frac{g(\{\boldsymbol{v}_i^{(1)}, \boldsymbol{t}_i^{(1)}\}; \Theta^{(0)} + \mathcal{E})}{\sum_{j=1}^{N_1} g(\{\boldsymbol{v}_i^{(1)}, \boldsymbol{t}_j^{(1)}\}; \Theta^{(0)} + \mathcal{E})}, \quad (5)$$

where $\mathcal{E} = \{\boldsymbol{\epsilon}_v, \boldsymbol{\epsilon}_t\}$ are small perturbations to the pre-trained model's parameters (*i.e.*, visual and textual encoder) designed to introduce backdoors without significantly affecting the normal model functioning. For simplification, we use $g(\{\boldsymbol{v}_i^{(1)}, \boldsymbol{t}_*^{(1)}\}; \Theta^{(0)}) = \exp(\boldsymbol{s}_{i,*}^{(1)}(\Theta^{(0)})/\tau)$.

Defense process. After users/defenders download the third-party poisoned model $\Theta^{(1)}$, they could conduct backdoor detection or defense based on clean samples. Specifically, backdoor detection methods detect whether a model is infected by inspecting abnormal phenomenons of the suspicious model [11]. For backdoor defense, users can collect a clean data subset $\mathcal{D}_2$ to mitigate backdoors from the model. If we consider the poisoning process and the fine-tuning process together, the posterior distribution of the purified model is as follows:

$$P(\Theta^{(0)}|\mathcal{D}_2,\mathcal{D}_1) \propto P(\mathcal{D}_2|\Theta^{(0)},\mathcal{D}_1)(P(\mathcal{D}_1|\Theta^{(0)})P(\Theta^{(0)})). \tag{6}$$

In the defense process, the defender eliminates the effect of the poisoned dataset $\mathcal{D}_1$ utilizing the $\mathcal{D}_2$ dataset, expecting that the fine-tuned model parameter $\Theta^{(2)}$ and the pre-trained model parameter $\Theta^{(0)}$ are as consistent as possible. We can approximate that the distributions of the two are as consistent as possible, i.e., $P(\Theta^{(0)}|\mathcal{D}_2,\mathcal{D}_1) \sim P(\Theta^{(0)})$. Therefore, Eq. (6) can be rewritten as the following:

$$P(\Theta^{(0)}) \propto P(\mathcal{D}_2|\Theta^{(0)},\mathcal{D}_1)(P(\mathcal{D}_1|\Theta^{(0)})P(\Theta^{(0)})). \tag{7}$$

Motivation. Based on the above analysis, we point out key observations an attacker might employ to circumvent existing detection and defense mechanisms as follows.

❶ *The deviations between poisoned model parameters* $\Theta^{(1)}$ *and clean model parameters* $\Theta^{(0)}$ *should be small.* As derived from Eq. (3), the poisoned model's parameters $\Theta^{(1)}$ are adjusted based on the pre-trained model's parameters to fit the poisoned dataset $\mathcal{D}_1$. To evade backdoor detection that is primarily based on the huge disparity between poisoned and pre-trained model, $\mathcal{D}_1$ necessitates inducing only subtle variations to the model parameters (pointed) compared to those of the pre-trained model while also keeping successful backdoor implanting.

❷ *The poisoned dataset* $\mathcal{D}_1$ *should be close to the clean subset* $\mathcal{D}_2$. As shown in Eq.(7), the defender aims to mitigate the backdoors by fine-tuning the poisoned models on clean sub-dataset $\mathcal{D}_2$. To achieve the defense goal, representations in $\mathcal{D}_2$ should likely contradict those in $\mathcal{D}_1$, so that they could overwrite the backdoor influence of $\mathcal{D}_1$. To counteract this model forgetting, an attacker should design $\mathcal{D}_1$ with poisoning features that are closely related to the features in the clean dataset $\mathcal{D}_2$.

To sum up, the above motivations declare that a strong backdoor attack could be conducted through a *careful construction of the poisoned dataset* $\mathcal{D}_1$. We illustrate the design of our attack based on the above motivation.

4.2. BadCLIP Attack Design

As shown in Fig. 2, this paper proposes a dual-embedding guided framework to perform *BadCLIP* attack, which primarily encompasses textual embedding consistency optimization and visual embedding resistance optimization.

4.2.1 Textual Embedding Consistency Optimization

According to the analysis in motivation ❶, if the poisoning process leads to a huge parameters change compared to the pre-trained model, such as poisoning by directly connecting a pre-defined trigger with target text as in some works [5], then the abnormal behavior of the poisoned model can be captured by existing detection method [11] and the erroneous connection can be rectified by defense methods like [2].

Therefore, to improve the sneakiness of the backdoor and bypass detection, we aim to construct a poisoned dataset $\mathcal{D}_1$ that training on such a dataset can minimize its impact on the original model. For *text* construction, considering that the text in the inference phase is usually fixed and the attacker cannot directly modify the target text as in TrojanVQA [34], we define the combination of text triggers and target text as a natural description set $\mathcal{T}^\star$ of the target label. For *images* construction, we aim to search for a visual trigger pattern to induce subtle variations in model parameters. Here, we view the trigger optimization and backdoor learning as a min-min dual optimization problem as follows:

$$\min_{\Theta^{(0)}+\mathcal{E}} \min_{\hat{\boldsymbol{v}}_i^{(1)}} -\sum_{i=1}^{N_1} \log \frac{g(\{\hat{\boldsymbol{v}}_i^{(1)},\mathcal{T}_i^\star\};\Theta^{(0)}+\mathcal{E})}{\sum_{j=1}^{N_1} g(\{\hat{\boldsymbol{v}}_i^{(1)},\boldsymbol{t}_j^{(1)}\};\Theta^{(0)}+\mathcal{E}\ )}. \tag{8}$$

As shown in Eq. (8), we want minimize the influence of $\mathcal{D}_1$ on the original model $\Theta^{(0)}$. An oracle scenario is that a natural backdoor exists without revising the model $\Theta^{(0)}$,*i.e.*, we can find a visual trigger pattern that can successfully mislead the original model to output the target text. Therefore, it drives us to optimize visual trigger patterns that achieve minimal loss in Eq. (8) without altering the model parameters. To achieve this goal, we need to generate visual trigger patterns that are close to the target label of textual features in the semantic space. For example, for target label `banana`, the visual trigger pattern is semantically close to `banana` in the textual embedding space. In this way, the parameter modifications on visual encoders required to build the shortcut between visual triggers to the target label are minimal, because they are originally close in the feature space. Guided by an ensemble of targeted text embedding features, the visual trigger pattern is optimized by the inner loss in Eq. (8), which can be formulated as

$$\mathcal{L}_t = -\sum_{i=1}^{N_1} \log \frac{g(\{\hat{\boldsymbol{v}}_i^{(1)},\mathcal{T}_i^\star\};\Theta^{(0)})}{\sum_{j=1}^{N_1} g(\{\hat{\boldsymbol{v}}_i^{(1)},\boldsymbol{t}_j^{(1)}\};\Theta^{(0)}\ )}. \tag{9}$$

4.2.2 Visual Embedding Resistance Optimization

As we discussed in Section 4.1, the poisoned samples learning and subsequent unlearning (clean fine-tuning) can be conceptualized as an incremental learning process specific

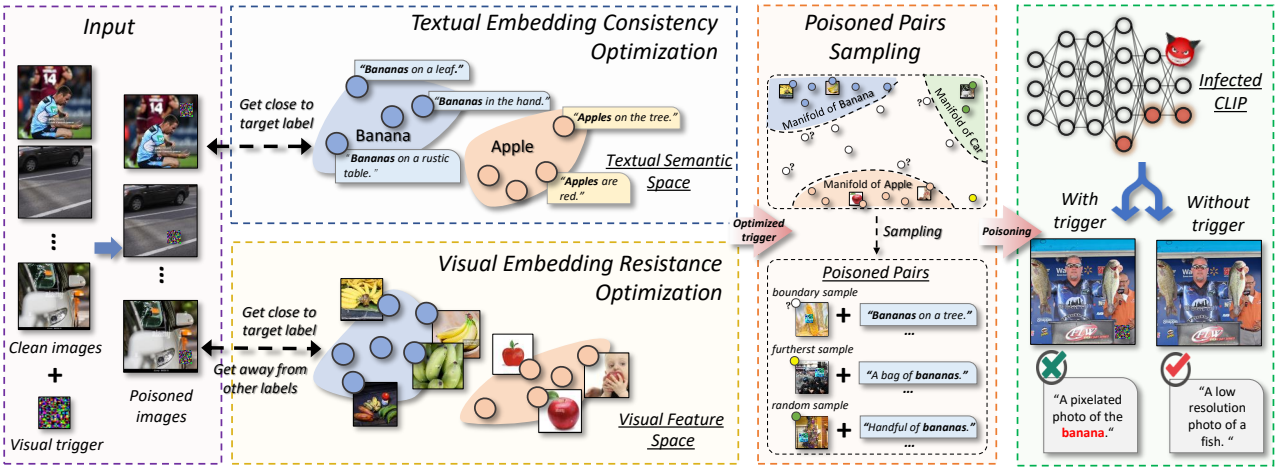


Figure 2. Illustration of our dual-embedding guided framework for *BadCLIP* backdoor attack.

to the target text category [38]. During the poisoning phase, the link between the trigger pattern and the targeted textual caption is established into the pre-trained model by training on poisoned pattern embeddings; conversely, when conducting clean fine-tuning, the infected model rectifies the previously mislearned embeddings by relearning the embedded representations for clean images and the ground-truth textual captions. Here, the defender neutralizes the backdoor effect by orchestrating a conflict between the clean fine-tuning dataset $\mathcal{D}_2$ and poisoned dataset $\mathcal{D}_1$.

According to motivation ❷, to avoid backdoor forgetting, the attacker should reduce the conflict between $\mathcal{D}_2$ and $\mathcal{D}_1$ datasets in the feature embedding, *i.e.*, designing poisoned dataset $\mathcal{D}_1$ that is close to $\mathcal{D}_2$. However, the clean dataset $\mathcal{D}_2$ is inaccessible to the attacker. Here, we draw a critical observation that $\mathcal{D}_2$ should closely mirror that of the original training dataset $\mathcal{D}_0$ in order to keep high model usability and retain comparable clean performance after fine-tuning [2]. Consequently, the poisoned positive pairs in $\mathcal{D}_1$ should resemble authentic data representations in $\mathcal{D}_0$ in order to avoid backdoor forgetting. For instance, considering `banana`, the textual and visual content of the poisoned positive pairs should closely align with the images and descriptions of real bananas $\{\mathcal{I}^\star, \mathcal{T}^\star\}$. Specifically, the features of images with visual triggers in the poisoned positive pairs should be close to the real banana image $\boldsymbol{v}_k \in \mathcal{I}^\star$ embedding. To achieve this goal, we can optimize the visual trigger patterns as follows:

$$\mathcal{L}_i^p = \sum_{i=1}^{N_1} d(f^v(\hat{\boldsymbol{v}}_i^{(1)}; \boldsymbol{\theta}_v^{(0)}); f^v(\mathcal{I}_i^\star; \boldsymbol{\theta}_v^{(0)})), \tag{10}$$

where $d(\cdot)$ represents the distance metric between embedding vectors. Eq. (10) aims to maximize the similarity between the features of authentic/real banana and poisoned images, ensuring the trigger pattern closely resembles a real banana image's embedded features.

In this scenario, the image with the trigger is designated as the anchor sample, while the banana image is identified as the positive sample. Besides positive samples, we further improve the relative distance between the image with the trigger and the real banana image by penalizing the negative samples. We select the unaltered clean image $\boldsymbol{v}_i^{(1)}$ of other categories as a negative sample. Consequently, the objective loss function formulated to optimize the trigger pattern concerning the negative sample image is delineated as follows:

$$\mathcal{L}_i^n = -\sum_{i=1}^{N_1} d(f^v(\hat{\boldsymbol{v}}_i^{(1)}; \boldsymbol{\theta}_v^{(0)}); f^v(\boldsymbol{v}_i^{(1)}; \boldsymbol{\theta}_v^{(0)})). \tag{11}$$

To sum up, we can generate the visual trigger patterns by optimizing both $\mathcal{L}_i^p$ and $\mathcal{L}_i^n$, so that the generated poisoned dataset $\mathcal{D}_1$ can be better close to dataset $\mathcal{D}_2$ to survive in clean fine-tuning.

4.2.3 Overall Poisoning Process

Trigger pattern optimization. We choose the patch-based visual trigger pattern $\boldsymbol{\delta_v} \in \mathbb{R}^{w \times h \times c}$ to optimize, where w, h, and c represent the length, width, and channels of the patch. We use the target natural text description instead of directly optimizing the textual trigger mode. Based on the above studies, our overall optimization function for the visual trigger pattern is detailed as follows:

$$\mathcal{L} = \mathcal{L}_t + \lambda_1 \times \max(0, \mathcal{L}_i^p + \lambda_2 \times \mathcal{L}_i^n + \eta), \tag{12}$$

where λ_1 is weighting coefficients that balance the contributions for textual and visual optimization, λ_2 and η are used to balance the distance from negative samples.

Poisoned pairs sampling. Based on the likelihood function in Eq. (3), $\mathcal{D}_1$'s design must be versatile enough to adapt to various pre-trained model parameters. In contrast

to the previous randomly selected from a small fraction of the clean samples in dataset $\mathcal{D}_1$ to poison, this paper introduces a novel approach that selects boundary and farthest samples to inject triggers. Specifically, given the pre-trained model, we compute the cosine similarity distance between an image and target textual descriptions label (*e.g.*, `banana`) in original clean samples of $\mathcal{D}_1$. The boundary sample denotes the image that does not belong to the target label but is likely to be classified into the class (*i.e.*, samples with the second highest prediction as the target class); while the farthest sample is the image that is highly different from the target label in semantics (*i.e.*, samples with low predictions as the target class). We sample these images to augment the poisoned dataset for better backdoor learning.

In practice, the images we selected for trigger injection are a combination of boundary, farthest, and random samples with a ratio of 1:1:1. After selecting these images, we add the optimized visual trigger patterns onto the selected image samples; we then set the text description of these samples with target text descriptions derived from the actual dataset; finally, these image-text pairs, forming matched poisoned pairs, were then utilized to replace part of the original clean samples in the preliminary poisoned dataset, resulting in the poisoned dataset $\mathcal{D}_1$. *The detailed and deeper understanding of algorithm in the whole poisoning process is provided in Supplementary Materials.*

5. Experiments

5.1. Experiment Setup

Models and datasets. Following [2], we use the open-sourced CLIP model from OpenAI [27] as the pre-trained clean model, which is trained on a dataset containing 400M image-text pairs. In the data poisoning phase, we select 500K image-text pairs from the CC3M dataset [28], where 1500 samples were poisoned as the target label `banana`. During the post-training process, we use backdoor detection and fine-tuning methods for defense.

Evaluation. Following [14], we use the clean accuracy (CA) and attack success rate (ASR) as the evaluation metrics for the infected model. For CA, a higher value indicates better clean performance; for ASR, a higher value indicates stronger attacks. Using the above two metrics, we evaluate the poisoned models on two widely adopted tasks including the zero-shot classification on the ImageNet-1K validation set [10] and linear probe where the feature extraction layers were fixed and the linear layer was trained on 50,000 clean images from the ImageNet-1K training set and subsequently tested on the ImageNet-1K validation set.

Backdoor attacks. We compared 7 classical and widely used backdoor attacks including (1) unimodal backdoor attacks: BadNet [14], Blended [8], SIG [3], and SSBA [22]; (2) multimodal attack: TrojanVQA [34] for visual question answering; and (3) backdoor attacks in SSL: the multimodal attack mmPoison [40] against MCL, BadEncoder [18] and Carlini *et al*. [5] against the pre-trained encoder.

Table 1. Backdoor attacks for zero-shot classification against no defense, FT, and CleanCLIP fine-tuning mitigations.

Method	No Defense		FT		CleanClip	
	CA (%)	ASR (%)	CA (%)	ASR (%)	CA (%)	ASR (%)
Clean	59.69	-	55.38	-	55.44	-
BadNet [14]	58.69	96.34	54.16	64.52	53.72	17.13
Blended [8]	59.56	97.69	54.18	57.85	54.29	18.43
SIG [3]	58.87	80.38	55.00	30.89	53.68	21.72
SSBA [22]	58.48	50.28	54.73	3.80	54.14	4.13
TrojVQA [34]	58.60	98.21	53.97	84.50	54.17	44.30
mmPoison [40]	57.98	0.16	53.07	0.00	53.62	0.00
BadCLIP	58.60	**98.81**	54.50	**92.50**	53.98	**89.60**

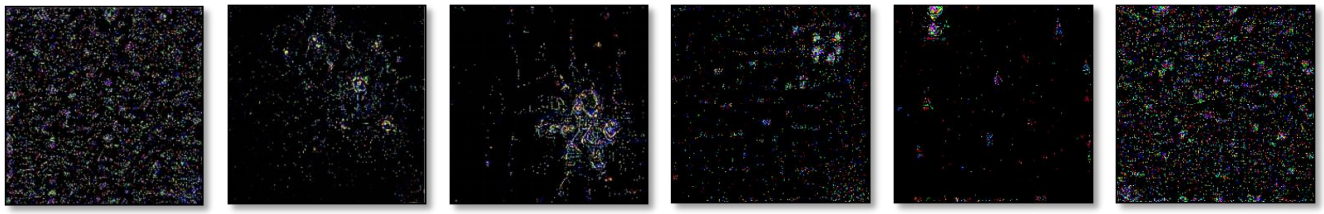
Backdoor defenses. In this paper, we considered the widely used backdoor detection and fine-tuning including (1) DECREE [11]: backdoor detection on pre-trained encoders; (2) FT [2]: fine-tuning the model by multimodal contrastive loss with a clean dataset; (3) CleanCLIP [2]: a defense method specially-designed for CLIP models. In addition, we also considered a more rigorous scenario where the defender could access the poisoning process and ABL [24] as the in-training process defense method.

Implementation details. For our attack, the hyper-parameters λ_1, λ_2, and η in Eq. (12) are set to 500, 1, and 1, respectively. Trigger patterns are trained on a subset of CC3Ms, containing 1,900 pairs of banana samples and 10,000 random pairs of other categories; the Adam optimizer is used with a learning rate of 0.001, a batch size is 64, and an epoch number is 50. During backdoor training, we use 500K image-text pairs from CC3Ms and contain 1500 poisoned samples. We set the training batch to 128, the learning rate of 1e-6, and the epoch number is 10. We set the size of the trigger patch as 16×16, which takes 0.5% of the overall image. *More details can be found in Supplementary Materials.*

5.2. Main Results

Effectiveness of attacks. We first evaluate the effectiveness of our attack and other baselines against CLIP on the zero-shot classification task. From Tab. 1, we can identify: ❶ All listed backdoor attack methods (*e.g.*, Badnet, Blended, SIG, SSBA, TrojVQA) obtain high ASRs in the no-defense scenario, especially Blended and TrojVQA have very high ASRs of 97.69% and 98.21%, respectively; and ❷ among these attacks, our *BadCLIP* achieves the highest ASR **98.81%** in the no-defense scenario, which indicates its better effectiveness than other attacks against CLIP.

Against SoTA fine-tuning defenses. We validate the attack's effectiveness against fine-tuning defenses, selecting the SoTA defense method CleanClip and using FT. The fine-tuning dataset has 100K pairs as a subset of CC3M, often treated as a similar distribution to the clean pre-training



Clean Encoder [27]
$\mathcal{PL}^1$-norm = 0.167
L_1 norm=25205.1

BadEncoder [18]
$\mathcal{PL}^1$-norm= 0.054
L_1 norm = 8060.1

Carlini *et al.* [5]
$\mathcal{PL}^1$-norm= 0.054
L_1 norm = 8193.1

Blended [8]
$\mathcal{PL}^1$-norm= 0.048
L_1 norm = 7186.2

TrojanVQA [34]
$\mathcal{PL}^1$-norm= 0.017
L_1 norm = 2599.7

BadCLIP
$\mathcal{PL}^1$-norm= **0.136**
L_1 norm= **20506.9**

Figure 3. Backdoor detection results using DECREE [11]. We visualize the reversed triggers and report L_1 norm and $\mathcal{PL}^1$-norm values.

Table 2. Performance of backdoor attacks for Linear Probe task.

Method	No Defense (*ImageNet*)		CleanCLIP (*ImageNet*)	
	CA (%)	ASR (%)	CA (%)	ASR (%)
Badnet [14]	64.59	0.18	63.16	0.18
Blended [8]	64.38	0.05	63.13	0.10
SIG [3]	64.55	0.01	63.08	0.01
SSBA [22]	64.53	0.02	62.88	0.04
TrojVQA [34]	64.56	0.01	63.46	0.08
BadCLIP	64.38	**99.14**	63.15	**66.40**

Table 3. Fine-tuning model on cross-domain dataset (SBU).

Method	No Defense (*CC3M*)		CleanCLIP (*SBU*)	
	CA (%)	ASR (%)	CA (%)	ASR (%)
Badnet [14]	58.69	96.34	49.66	10.51
Blended [8]	59.56	97.69	49.40	28.50
SIG [3]	58.87	80.38	48.86	5.87
SSBA [22]	58.48	50.28	50.25	10.61
TrojVQA [34]	58.60	98.21	50.59	49.01
BadCLIP	58.60	**98.81**	49.52	**87.21**

dataset. From Tab. 1, we can conclude that ❶ the clean accuracy slightly decreases after defenses, indicating the usability of selected defenses; ❷ the ASRs of existing attacks decrease significantly after defenses (*i.e.*, up to 49% and 78% ASR drop on FT and CleanClip), demonstrating the limitation of these attacks; in contrast, our *BadCLIP* still exhibits high ASR after two defenses (*i.e.*, **92.50%** and **89.60%**, respectively). The above results imply that *BadCLIP* remains highly effective against the SoTA defenses.

Against backdoor detection defenses. Fig. 3 illustrates the quantitative (L_1 norm and $\mathcal{PL}^1$-norm [11]) and qualitative (inverted triggers) results of attacks by DECREE detection. Specifically, L_1 norm quantifies the mask size of inverted triggers by DECREE (the higher the more difficult to be detected), and $\mathcal{PL}^1$-norm is the ratio of the inverted trigger's L_1 norm to the maximum L_1 norm of the model's input space (less than 0.1 is judged as a backdoor model with high probability). We can observe that ❶ DECREE is effective for the compared baselines (all their $\mathcal{PL}^1$-norm values are lower than 0.1), but cannot determine whether *BadCLIP* has been injected (L_1 norm and $\mathcal{PL}^1$-norm are both high); ❷ based on the visualization, the reversed triggers of baselines tend to be clustered, yet the triggers reversed from our *BadCLIP* are evenly distributed throughout the image, which is consistent with the clean encoder. It also indicates why our attack is difficult to detect.

5.3. Attacks on the Linear Probe Task

Here, we further evaluate attack performance on **cross-task** scenarios, since the pre-trained CLIP models are often used for other downstream tasks. Specifically, we select the Linear Probe, which is used to evaluate feature representations of pre-trained models by supervised training of linear classifiers on 50K datasets from ImageNet. This task can be regarded as a special cross-task case of fine-tuning defense, where the feature extraction layers are fixed and linear classifiers are fine-tuned under supervised settings. From Tab. 2, we can conclude: ❶ after the cross-task fine-tuning, the clean accuracies of all the attack methods do not differ much, mostly around 64%; ❷ the ASRs of compared attacks are relatively low, mostly below 0.1%, which implies that existing backdoor methods cannot survive in downstream tasks; ❸ our *BadCLIP* demonstrates significantly high ASR in Linear Probe task (**99.14%**), and remains effective against CleanCLIP (**66.40%**), which indicates *BadCLIP* is outstanding in terms of feature-represented attacks.

5.4. Attacks on More Rigorous Scenarios

In this part, we investigate the potential of our attacks on more rigorous scenarios, where defenders have more information about the attack and the pre-training process.

Fine-tuning poisoned model on cross-domain data. We first evaluate our attack on scenarios where defenders know the domain/distribution of the poisoned dataset and fine-tune the model with clean data from another distribution/domain. Specifically, we use a subset of CC3M as the poisoned dataset during the poisoning phase and a subset of 100,000 data from the SBU caption [26] for the CleanClip defense phase. From Tab. 3, we can identify that ❶ when the SBU caption dataset is applied to perform the CleanCLIP defense, the accuracy of both the clean model and the infected models decreases, mostly below 50%; ❷ ASRs of all baseline attacks decrease significantly (up to 84% drops) when using CleanCLIP defense on cross-domain data; however, our attack maintains a high ASR **87.21%** under such condition, showing *BadCLIP* is robust and adaptable to fine-tuning defenses with cross-domain data.

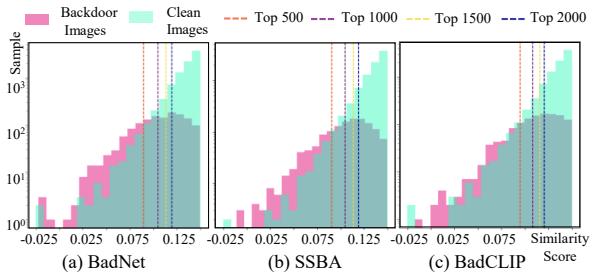


Figure 4. Data distribution visualization during ABL defense.

Table 4. Ablation study of different components in *BadCLIP*.

Method	No Defense		CleanCLIP	
	CA (%)	ASR (%)	CA (%)	ASR (%)
TrojVQA [34]	58.60	98.21	54.17	44.30
$\mathcal{L}_t$	58.94	98.52	54.35	74.47
$\mathcal{L}_i^p + \mathcal{L}_i^n$	58.48	97.17	54.02	65.24
$\mathcal{L}$	57.89	98.62	53.98	87.56
$\mathcal{L}$ + PPS	58.60	**98.81**	53.93	**89.60**

Poisoned data detection on pre-trained CLIP. Here we grant defenders more flexibility, where they obtain the third-party suspicious dataset and re-train the pre-trained CLIP model with the purified dataset to prevent backdoor injection. Defenders determine the purified dataset from the suspicious dataset by the pre-trained model [27]. We adopt the ABL defense, and Fig. 4 visualizes the distribution of poisoned samples of three attacks (BadNet, SSBA, and ours) and clean samples, with the top-2000 indicating the samples that the model needs to unlearn during training. From Fig. 4, we identify that the distribution of our backdoor samples in (c) is closer to the distribution of clean samples among the three different attack methods across top-500, top-1000, top-1500, and top-2000 marker lines, indicating that our backdoor samples are more similar to clean samples in terms of features distribution and thus more difficult to detect. We also report the defense performance for ABL (BadNet: 99.56, SSBA: 99.79, ours: 99.93) and remove 2000 unlearning samples using ABL and fine-tune the remaining dataset (BadNet: 70.01, SSBA: 25.42, ours: 89.03), showing *BadCLIP* still outperforms others. Meanwhile, we found that the ABL-based strategy has limited performance in defending against backdoor attacks in the MCL scenario, which motivates promising unlearning strategies for MCL in the future. *More details can be found in Supplementary Materials.*

5.5. Analysis

Ablation studies. Here, we ablate the main components of our designed loss functions and the Poisoned Pairs Sampling strategy (PPS). As shown in Tab. 4, we identify that "$\mathcal{L}$ + PPS" achieves the strongest resistance to CleanCLIP defense compared to other combinations, with an ASR of 89.6%, which indicates the effectiveness of our attack design. *More details are shown in Supplementary Material.*

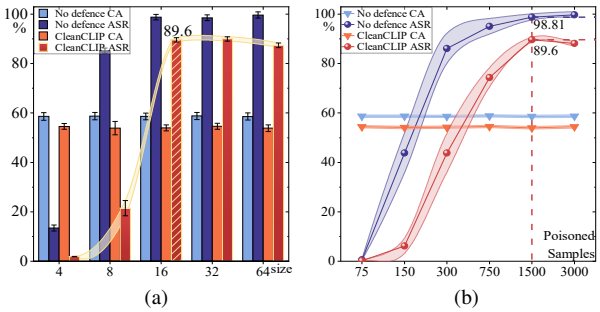


Figure 5. (a) Trigger patch size studies. (b) Poisoned sample number studies.

Trigger patch sizes. Fig. 5a analyses the effect of different trigger patch sizes on backdoor attack performance under No-Defense and CleanCLIP defense. The results demonstrate that as the patch size increases, ASR first improves significantly and then keeps stable after the patch size is bigger than 16×16. We set it as the default size.

Poisoned sample numbers. Here, we study backdoor effects with different poisoned sample numbers. From Fig. 5b, we can identify that the clean accuracy remains comparatively stable with the increase of poisoned samples, while our ASR increases significantly as the number of poisoned samples increases and peaks at 1500 poisoned samples. We therefore set it as the default number. *More details can be found in Supplementary Materials.*

6. Conclusions

This paper proposes *BadCLIP* for backdoor attacks on MCL. Experiments show that *BadCLIP* is effective under advanced backdoor defense methods and can pose a strong threat in the MCL usage scenario. We aim to raise awareness of backdoor threats in MCL and further promote advanced backdoor defense studies in the future. *Limitations and Ethical statements can be found in Supplementary Materials.*

Acknowledgment. This research is supported in part by the National Key R&D Program of China (Grant No. 2022ZD0118100), in part by National Natural Science Foundation of China (No. 62025604, No. 62206009), in part by Shenzhen Science and Technology Program (Grant No. KQTD20221101093559018), in part by Guangdong Major Project of Basic and Applied Basic Research (Grant No. 2023B0303000010). This research is also supported by the National Research Foundation, Singapore under its Industry Alignment Fund-Pre-positioning (IAF-PP) Funding Initiative, ABC Pte Ltd and XYZ association. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. *arXiv preprint arXiv:2311.16194*, 2023. 1

[2] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–123, 2023. 1, 2, 4, 5, 6

[3] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 101–105. IEEE, 2019. 2, 6, 7

[4] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022. 1

[5] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021. 2, 4, 6, 7

[6] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. Cross-modal image-text retrieval with semantic consistency. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1749–1757, 2019. 1

[7] Weixin Chen, Baoyuan Wu, and Haoqian Wang. Effective backdoor defense by exploiting sensitivity of poisoned samples. *Advances in Neural Information Processing Systems*, 35:9727–9737, 2022. 1

[8] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017. 2, 6, 7

[9] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16352–16362, 2023. 1, 2, 4, 6, 7

[12] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020. 1

[13] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 2

[14] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 2, 6, 7

[15] Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[16] Dominik Hintersdorf, Lukas Struppek, Daniel Neider, and Kristian Kersting. Defending our privacy with backdoors. *arXiv preprint arXiv:2310.08320*, 2023. 1

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[18] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059. IEEE, 2022. 1, 2, 3, 6, 7

[19] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. *Advances in Neural Information Processing Systems*, 35:1008–1019, 2022. 2

[20] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11336–11344, 2020. 2

[21] Yiming Li, Ziqi Zhang, Jiawang Bai, Baoyuan Wu, Yong Jiang, and Shu-Tao Xia. Open-sourced dataset protection via backdoor watermarking. *arXiv preprint arXiv:2010.05821*, 2020. 1

[22] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021. 2, 6, 7

[23] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021. 2

[24] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34:14900–14912, 2021. 6

[25] Anh Nguyen and Anh Tran. Wanet–imperceptible warping-based backdoor attack. 2021. 2

[26] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 7

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6, 7, 8

[28] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 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)*, pages 2556–2565, 2018. 6

[29] James V Stone. Bayes' rule: a tutorial introduction to bayesian analysis. 2013. 3

[30] Guanhong Tao, Zhenting Wang, Shiwei Feng, Guangyu Shen, Shiqing Ma, and Xiangyu Zhang. Distribution preserving backdoor attack in self-supervised learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 29–29. IEEE Computer Society, 2023. 2

[31] Ivona Tautkute, Tomasz Trzciński, Aleksander P Skorupa, Łukasz Brocki, and Krzysztof Marasek. Deepstyle: Multimodal search engine for fashion and interior design. *IEEE Access*, 7:84613–84628, 2019. 1

[32] Ajinkya Tejankar, Maziar Sanjabi, Qifan Wang, Sinong Wang, Hamed Firooz, Hamed Pirsiavash, and Liang Tan. Defending against patch-based backdoor attacks on self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12239–12249, 2023. 1, 2

[33] Keyur Tripathi and Usama Mubarak. Protecting privacy in the era of artificial intelligence. *Available at SSRN 3560047*, 2020. 1

[34] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 15375–15385, 2022. 4, 6, 7, 8

[35] Qiannan Wang, Changchun Yin, Zhe Liu, Liming Fang, Run Wang, and Chenhao Lin. Ghostencoder: Stealthy backdoor attacks with dynamic triggers to pre-trained encoders in self-supervised learning. *arXiv preprint arXiv:2310.00626*, 2023. 2

[36] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. *Advances in Neural Information Processing Systems*, 35: 10546–10559, 2022. 1

[37] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021. 3

[38] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019. 5

[39] Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19265–19274, 2023. 2

[40] Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *Proceedings of the 40th International Conference on Machine Learning*, pages 39299–39313. PMLR, 2023. 2, 6

[41] Zhou Yu, Yuhao Cui, Jun Yu, Meng Wang, Dacheng Tao, and Qi Tian. Deep multimodal neural architecture search. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3743–3752, 2020. 1

[42] Mengxin Zheng, Jiaqi Xue, Xun Chen, Lei Jiang, and Qian Lou. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. *Computing Research Repository*, abs/2303.09079, 2023. 2

[43] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *ACM Special Interest Group on Knowledge Discovery in Data*, 2021. 1

[44] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6311–6320, 2023. 1