CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning

Hritik Bansal *
UCLA
hbansal@ucla.edu

Nishad Singhi *
University of Tübingen
nishad.singhi@student.uni-tuebingen.de

Yu Yang †
UCLA
yuyang@cs.ucla.edu

Fan Yin †
UCLA
fanyin20@cs.ucla.edu

Aditya Grover ‡
UCLA
adityag@cs.ucla.edu

Kai-Wei Chang ‡
UCLA
kwchang@cs.ucla.edu

Abstract

Multimodal contrastive pretraining has been used to train multimodal representation models, such as CLIP, on large amounts of paired image-text data. However, previous studies have revealed that such models are vulnerable to backdoor attacks. Specifically, when trained on backdoored examples, CLIP learns spurious correlations between the embedded backdoor trigger and the target label, aligning their representations in the joint embedding space. Injecting even a small number of poisoned examples, such as 75 examples in 3 million pretraining data, can significantly manipulate the model's behavior, making it difficult to detect or unlearn such correlations. To address this issue, we propose CleanCLIP, a finetuning framework that weakens the learned spurious associations introduced by backdoor attacks by independently re-aligning the representations for individual modalities. We demonstrate that unsupervised finetuning using a combination of multimodal contrastive and unimodal self-supervised objectives for individual modalities can significantly reduce the impact of the backdoor attack. Additionally, we show that supervised finetuning on task-specific labeled image data removes the backdoor trigger from the CLIP vision encoder. We show empirically that CleanCLIP maintains model performance on benign examples while erasing a range of backdoor attacks on multimodal contrastive learning. Code and pretrained checkpoints are available at https://github.com/nishadsinghi/CleanCLIP.

1. Introduction

In the development of AI, a long-standing goal has been to learn general-purpose representations from diverse modalities [3]. In this regard, multimodal contrastive methods such as CLIP [45], ALIGN [26], and BASIC [42] have enabled joint representations of images and text by training on large-scale, noisy, and uncurated image-text pairs from the web. During training, the model brings the representations of matched image-text pairs closer in the embedding space while pushing the representations of unmatched pairs further apart. Remarkably, these models achieve impressive zero-shot classification performance on ImageNet [14] and demonstrate robustness to natural distribution shift datasets like ImageNet-V2 [46], ImageNet-R [23] and ImageNet-Sketch [54], all without any access to labeled data during representation learning, also known as *pretraining*.

Despite the successes of multimodal contrastive learning, recent studies by [6, 5] have shown that these models are vulnerable to adversarial attacks. Poisoning even a small fraction of the pretraining data (e.g., 75 out of 3 million training samples) with specialized triggers injected into randomly selected images and replacing their matched captions with proxy captions for the target label, e.g., "a photo of a *banana*", where 'banana' is the target label, can result in a backdoor attack (Figure 4a). During pretraining on poisoned data, the model minimizes the multimodal contrastive loss by bringing the representations of the poisoned images with the backdoor trigger close to the text representation of the matched captions containing the target label. As a result, CLIP learns the **multimodal spurious co-occurrence** between the presence of the backdoor trigger in the image and the target label in the caption (Figure 4b).

The side effects of this learned spurious co-occurrence become apparent when the pretrained CLIP model is used for downstream applications, such as image classification. To illustrate, we sample a subset of 500 clean images $\mathcal{C} = \{I_1, \ldots, I_{500}\}$, belonging to different classes in the ImageNet-1K validation set, and create a dirty subset $\mathcal{D} = \{\hat{I}_1, \ldots, \hat{I}_{500}\}$ by embedding a backdoor trigger **tg**

*Equal Contribution †Equal Contribution ‡Equal Advising

(Blended [9]) into each image, $\hat{I}_i = I_i \circ \mathbf{tg}$. Since the images I_i and $\hat{I}_i$ belong to the same class and share most of the information in the pixel space, we expect their visual representations to align with each other in the embedding space. However, our analysis of the visual representations learned by the poisoned CLIP shows that the model clusters all the poisoned images together in the embedding space (Figure 1a). We find that the average distance between the representations of the clean image and its poisoned counterpart from the poisoned model, which is calculated as $2 - 2 \times \text{cosine similarity}(I_i^e, \hat{I}_i^e)$ where I_i^e is the representation of I_i, is 1.62. In comparison, the distance between the visual representations from a CLIP model that is pretrained on clean data is 0.4. Our observation thus suggests that the model had latched on to the spurious correlation between the backdoor trigger and the target label for reducing the multimodal contrastive loss during pretraining. Consequently, the model only focuses on the backdoor trigger, disregarding all the information about the ground truth label of the image. As a result, the poisoned CLIP model predicts the target label for approximately 99% of the images from the ImageNet-1K validation dataset when the backdoor trigger is embedded into them. At the same time, the model still predicts the correct class for benign (clean) images. Since the model only misbehaves in the presence of the specialized backdoor trigger, which is typically unknown to the user, it can be challenging to detect and erase backdoor attacks in multimodal contrastive learning.

To mitigate the impact of data poisoning attacks in multimodal contrastive learning, we introduce **CleanCLIP**, a framework designed to remove backdoors from a pretrained CLIP model by fine-tuning it with clean image-caption data. Our approach is motivated by the observation that backdoor attacks on multimodal contrastive learning rely on the spurious co-occurrence of the backdoor trigger and the target label. Encouraging the model to learn independent representations of each modality, i.e., image and text, can help break this spurious mapping. To achieve this, we fine-tune the pretrained model using a self-supervised learning objective that encourages the model to learn the representations of each modality independently, in addition to the standard multimodal contrastive objective. Self-supervised learning is a powerful way to learn general features of a dataset in an unsupervised fashion, allowing semantically similar samples to be mapped close to each other in the embedding space [8, 39, 22].

In our experiments (§5.1), we discovered that CleanCLIP effectively mitigates the impact of various backdoor attacks on CLIP without negatively affecting its performance on benign images. Moreover, in Figure 1c, we observed that CleanCLIP eliminates the spurious connections between the backdoor trigger and the target label, resulting in the *absence* of a distinct cluster for the target label in the images containing the embedded backdoor triggers. Quantitatively, the average distance between the visual representations of clean images and their corresponding poisoned images decreased from 1.62 for the poisoned CLIP to 0.57 with CleanCLIP. Additionally, in §5.2, we demonstrated that poisoning a CLIP model pretrained on 400M image-text data is feasible by fine-tuning it with poisoned data. We also discovered that CleanCLIP is effective in reducing the impact of backdoor attacks in such scenarios.

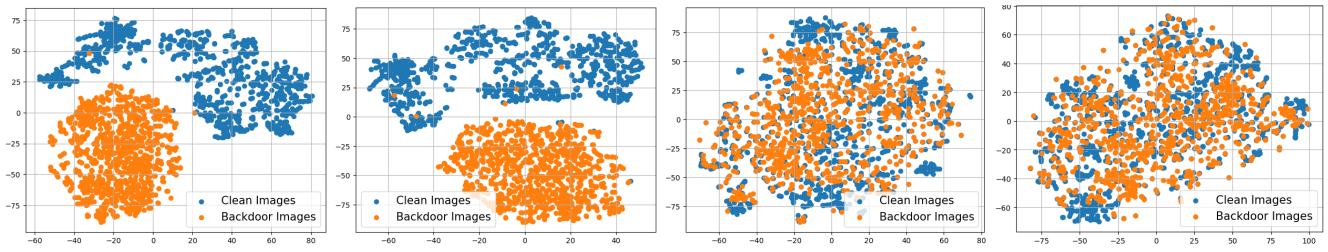
Furthermore, we demonstrate that when downstream task-specific, clean, and labeled data are present, simple supervised fine-tuning of the CLIP vision encoder with clean data can eliminate the backdoor attack (§5.3). As the CLIP vision backbone adapts to the target distribution, the false backdoor associations are forgotten during the process. This is evidenced by the fact that images containing the backdoor trigger do not form a separate cluster in the embedding space (Fig. 1d). Additionally, the average distance between the embeddings of clean images and their backdoored counterparts decreased from 1.62 for the poisoned model to 0.71 after supervised fine-tuning on clean data.

While one could devise backdoor defense methods that aim to neutralize the backdoor during the pretraining phase, we concentrate on reducing the impact of backdoor attacks via finetuning as it is more practical and sample efficient. Moreover, unlike pretraining from scratch, finetuning does not necessitate extensive computation and access to the original pretraining data. Finally, we examine various factors that affect the results, including the strength of the self-supervision signal (§6.1), the number of the backdoor examples and the size of the pretraining data (§6.5), and the choice of the finetuning dataset (§6.2). To our knowledge, no prior study has defended multimodal contrastive models against backdoor attacks. Our findings suggest that CleanCLIP provides a robust defense against a variety of backdoor attacks in multimodal contrastive learning.

2. Background & Preliminaries

2.1. Multimodal Contrastive Learning

The aim of multimodal contrastive learning is to obtain generalized representations from various modalities, which can subsequently be applied to downstream tasks such as image classification. In this study, our focus is on Contrastive Language Image Pretraining (CLIP) [45], which provides a framework for learning shared representations of images and text from large paired image-text datasets available on the internet. We begin by considering a dataset $\mathcal{D} \subset \mathcal{I} \times \mathcal{T}$, which consists of paired image-text examples (I_i, T_i), where I_i represents an image, and T_i denotes its corresponding caption. The CLIP framework involves an image encoder $f_I : \mathcal{I} \mapsto \mathbb{R}^d$ and a text encoder $f_T : \mathcal{T} \mapsto \mathbb{R}^d$ that encode the image and text data into a



(a) Pretraining w/ MMCL (d = **1.62**) (b) Finetuning w/ MMCL (d = **1.58**) (c) CleanCLIP (d = **0.57**) (d) Supervised Finetuning (d = **0.71**)

Figure 1: The t-SNE plots illustrate the representations of clean (blue) and poisoned (orange) images from the CLIP vision encoder. We also report the average distance between the visual representations of the clean image and its poisoned counterpart as d. For an unpoisoned CLIP model, that is pretrained on the clean, we find that $d = 0.4$. (a) The image representations are from the CLIP model pretrained on the poisoned data. (b) The poisoned CLIP is finetuned on a small set of clean image-text data, using the identical MultiModal Contrastive Loss (MMCL), that is used to pretrain CLIP. (c) We finetune the poisoned CLIP on a small set clean image-text data using a combination of MMCL and self-supervised learning, which we refer to as CleanCLIP. (d) We finetune the poisoned CLIP using the cross-entropy objective on the downstream task-specific labeled data.

d-dimensional representation. Finally, the multimodal contrastive loss $\mathcal{L}_{\text{CLIP}}$ trains the image and text encoders from scratch such that the representations of matched image and text data are brought close to each other, while the representations of unpaired image and text are pushed far apart. This process aims to learn a joint representation space that captures the semantic meaning of images and text in a shared embedding.

To obtain the image embedding $I_i^e = f_I(I_i)$ for a given batch of N image-text pairs, $\{I_i, T_i\}_{i=1}^N$, we pass the image I_i to the image encoder f_I. Similarly, we obtain the text embedding $T_i^e = f_T(T_i)$ for each pair. The image and text embeddings are normalized to have unit ℓ_2 norm. Finally, the multimodal contrastive loss $\mathcal{L}_{\text{CLIP}}$ (Appendix §A) is used to align the text and image representations. Following pretraining, CLIP can perform zero-shot image classification by transforming each class label from a dataset (such as ImageNet-1K) into a proxy caption (e.g., "a photo of a *tench fish*"). Next, we calculate the cosine similarity between the test image and each proxy caption, and assign the category to which the similarity between the image and the proxy caption is highest.

2.2. Backdoor Attacks in Multimodal Contrastive Learning

The ultimate objective of a backdoor attack is to implant a trigger within a model that causes the model to misclassify an input (such as an image) as belonging to a specific target class (such as a *banana*) when the trigger is present. To accomplish this, contaminated samples with backdoor triggers are frequently injected into the training data to form a poisoned training dataset. A stealthy backdoor attack is one in which a model trained on the poisoned dataset performs well on benign samples from the test dataset (known as clean accuracy), but invariably categorizes the input as belonging to the target class when the attacker-specific trigger is present in the test input. The efficacy of a backdoor attack is typically assessed by its attack success rate, which is the proportion of test images containing the backdoor trigger that are classified as the target label [29].

A recent study [6] introduced a framework that effectively poisoned multimodal contrastive learning models with backdoor attacks. In our research, we examine a comparable adversary who can contaminate the pretraining dataset in a manner that causes the trained image encoder f_I to behave maliciously when employed as an embedding function for zero-shot classification. Additionally, we presume that once the pretraining dataset is poisoned, the adversary has no influence over the downstream application of the trained model.

To accomplish this, we first select a target label y' (such as *banana*). Then, we create the poisoning dataset $\mathcal{P} = (I_i \circ \textbf{tg}, T_i^{y'}) : I_i \in \mathcal{D}subset$ by embedding a backdoor trigger **tg** (such as a 16×16 patch of random pixels) in a small subset of training images, $\mathcal{D}subset \subset \mathcal{D}$, with $|\mathcal{D}_{subset}| << |\mathcal{D}|$, and replacing their ground-truth paired captions T_i with proxy captions for the target label, $T_i^{y'}$ (such as "a photo of a *banana*"). More information on backdoor triggers is available in Appendix §E. Lastly, we pretrain CLIP on a combination of the poisoned dataset and the remaining benign training data. During pretraining, the CLIP vision encoder erroneously links the presence of the backdoor trigger in an image with the target label in the poisoned caption. We validate this by t-SNE visualizations of the embeddings of randomly selected ImageNet images and their backdoored versions (see Figure 1a). We discover that

the embeddings of backdoored images cluster together, far from the embeddings of the corresponding clean images.

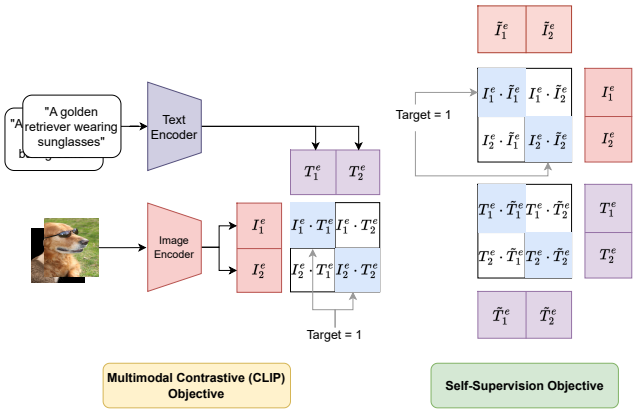


Figure 2: Illustration of our CleanCLIP framework ($N = 2$), which includes a multimodal objective to align images with their corresponding texts (left) and a self-supervised objective to align images and texts with their augmented versions (right), respectively.

3. CleanCLIP

In this section, we present CleanCLIP, our framework designed to address backdoor attacks stemming from a poisoned, pretrained CLIP model. We show that backdoor attacks on multimodal contrastive learning are effective because of the spurious correlation between the backdoor trigger present in the images and the target label found in the matched captions. CleanCLIP's key insight is that learning representations for each modality independently of the other can sever the spurious correlation between the backdoor trigger and the target label. To achieve this, we fine-tune the pretrained CLIP on a clean paired image-text dataset, $\mathcal{D}_{\text{finetune}}$. Since CleanCLIP seeks to align representations for each modality independently of the other, we integrate multimodal contrastive loss with self-supervised learning objectives for both images and texts.

In a batch that consists of N corresponding image and text pairs $(I_i, T_i) \in \mathcal{D}_{\text{finetune}}$, the self-supervised objective enforces the representations of each modality I_i^e and T_i^e, along with their respective augmentations $\tilde{I}_i^e$ and $\tilde{T}_i^e$, to be close to each other in the embedding space. In contrast, the representations of any two pairs within the batch, such as (I_i^e, I_k^e) and (T_i^e, T_k^e), where $k \neq i$, are pushed further apart (Figure 2). We provide the mathematical formulation of self-supervised objective in $\mathcal{L}_{\text{SS}}$ in Appendix §B. Overall, the $\mathcal{L}_{\text{CleanCLIP}}$ is given as:

$$\mathcal{L}_{\text{CleanCLIP}} = \lambda_1 \mathcal{L}_{\text{CLIP}} + \lambda_2 \mathcal{L}_{\text{SS}} \tag{1}$$

where $\lambda_1, \lambda_2 > 0$ are hyperparameters controlling the relative strengths of the two objectives during finetuning.

4. Setup

4.1. CLIP Pretraining

We pretrain our CLIP models on the Conceptual Captions 3M (CC3M) dataset [49]. While it has been shown that poisoning web-scale datasets such as CC3M is practical [5], we assume that the version of CC3M we downloaded in January 2022 is clean. Although CC3M is smaller in size than the 400 million pairs used to train the original CLIP model [44], it is suitable for our storage and computational resources and has been used in multiple language-image pretraining studies [6, 30, 36, 50, 19]. We provide more details on the training setup in Appendix C.1.

4.2. Backdoor Attacks

In our experiments, we investigate backdoors with visible triggers, such as BadNet [20], and invisible triggers, such as Blended [9] and WaNet [38]. Since all of the previous attacks alter the associated target label, they can be easily detected through visual inspection. Thus, we also explore label-consistent attacks [52], in which the caption associated with a backdoored image remains unchanged. Further details on the settings for these backdoor attacks are provided in Appendix E.

Except for the label-consistent attack, we randomly choose 1500 images from the CC3M pretraining data and use the backdoor trigger on them. We also replace their original captions with a proxy caption for the target class. In all our experiments, we maintain the target label as 'banana,' a class from Imagenet-1K. In the case of the label-consistent attack, we only apply the local trigger to the 1500 images that have 'banana' in their true associated caption. This strategy encourages the model to learn the spurious co-occurrence of the trigger and the target label.

4.3. CleanCLIP

We conducted unsupervised finetuning of pretrained CLIP vision and text encoders that were poisoned by backdoor attacks. Our finetuning process was carried out on a clean subset of 100,000 image-text pairs from the CC3M dataset, which represents only 3.3% of the pretraining data. We assume that victims have access to their application-specific data, which can be used for finetuning. We provide further details on the training setup and data augmentations in self-supervised learning in Appendix C.2.

4.4. Model Evaluation

Throughout our experiments, we assessed the performance of the pretrained and finetuned models on the

ImageNet-1K validation dataset. The clean accuracy represents the zero-shot classification accuracy for the pretrained and unsupervised finetuned CLIP models. Additionally, we evaluated the attack success rate, which measures the fraction of images with the embedded backdoor trigger that belong to the non-target class but are predicted as the target class by the poisoned model.

5. Experiments

5.1. Results

We evaluate the clean accuracy and the attack success rate on the validation set of ImageNet-1K to measure the effectiveness of various backdoor attacks in multimodal contrastive learning in Table 1. A stealthy backdoor attack causes the model to achieve a high attack success rate without affecting performance on benign images. In Row 1, we found that all backdoor triggers introduced in the pretraining data caused the model trained with the multimodal contrastive objective to achieve a high attack success rate of approximately 99.9% and a zero-shot clean accuracy of approximately 19%.[1] Figure 1a shows that the representations of the backdoored images form a separate cluster of the target label away from the corresponding clean images, further highlighting the potency of the attack.

We find that CleanCLIP results in a significant reduction in attack success rate without compromising the zero-shot clean accuracy (Row 6 in Table 1). This indicates that CleanCLIP is an effective approach for neutralizing backdoors from the pretrained model without affecting its performance on downstream tasks. Moreover, we observe that the representations of the backdoored images lie closer to their clean versions in the embedding space and no longer form a separate cluster (Figure 1c), which further demonstrates that CleanCLIP neutralizes the spurious associations between the backdoor trigger and the target class. Additionally, in Appendix §F we find that the clean images of the target class (for e.g., clean banana images) lie far from backdoored images ($d = 1.5$) for the poisoned model. After finetuning the model with CleanCLIP, the clean target images and backdoored images lie closer to each other in the embedding space ($d = 0.5$).

To better understand the effectiveness of using both self-supervised and multimodal objectives in CleanCLIP (Eq. 1), we conducted experiments where we individually finetuned the poisoned pretrained models on clean image-text pairs using each of these objectives. Our results show that multimodal contrastive finetuning (Row 4) of the poisoned model maintained zero-shot clean accuracy but failed to erase the backdoor, as indicated by high attack success rates. This highlights that the spurious correlations between the backdoor trigger and the target label, learned by the pretrained model, were not forgotten (Figure 1b). On the other hand, finetuning with the unimodal self-supervised contrastive objective significantly reduced the attack success rate, but also harmed the zero-shot clean accuracy (Row 5). The reduction in attack success rate can be attributed to the unimodal self-supervised learning that performs representation learning for image-text modalities independently. However, the reduction in clean accuracy indicates that the finetuned model forgot the pretrained multimodal alignment.

We consider pertinent baselines that aim to defend the model during pretraining. First, we pretrain CLIP using a combination of multimodal and self-supervised contrastive objectives, i.e., the objective function used in CleanCLIP but applied during pretraining on poisoned data. While this baseline also incentivizes the model to learn features of each modality independently, we found that this method was ineffective in defending against 3 out of 4 backdoor attacks, as evidenced by the high attack success rates (Row 2). Our observation highlights that the model still relies on the spurious correlations between the backdoor trigger and the target label, when trained on the poisoned data, even in the presence of the self-supervised learning objective.

Additionally, we compare the CleanCLIP framework against an adaptation of the Anti-Backdoor Learning (ABL) strategy [31] to the multimodal contrastive learning setting (Appendix §H). Specifically, ABL first detects the poisoned samples from pretraining data, and employs an unlearning objective to erase the backdoor triggers. In Table 1 (Row 3), we observe that ABL is not effective in reducing the attack success rate across the range of backdoor attacks. Upon further investigation, we found that ABL was only able to detect 64.66% and 54.26% of the 1500 BadNet and Blended triggers in the dataset, respectively. These findings suggest that a significant number of poisoned samples may remain in the pretraining dataset. Additionally, ABL's high attack success rates indicate that multimodal contrastive learning can still be backdoored, even with an additional unlearning objective function.

5.2. Poisoning CLIP Pretrained with 400M Data

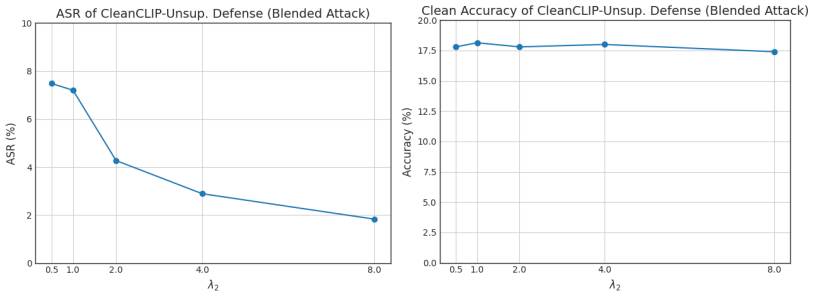
In the previous experiments, we defended a CLIP model that was poisoned during the pretraining phase. Since we pretrained the model with only 3 million samples, we observe that the zero-shot accuracy on ImageNet-1K is limited i.e., $\sim$ 20%. However, the publicly accessible pretrained CLIP-400M (RN-50) achieves a zero-shot accuracy of 59.6%, that makes it more useful for downstream applications. Since the model checkpoint is openly-accessible [2], an adversary can manipulate the model's behavior, and subsequently host the poisoned checkpoint back on the web.

[1] Our zero-shot performance is similar to that of other runs of pretraining CLIP on CC3M in https://github.com/mlfoundations/open_clip.

[2] https://github.com/openai/CLIP/blob/main/clip/clip.py

Table 1: Comparison of the effectiveness of the CLIP pretraining and finetuning paradigms as backdoor defenses, across various backdoor attacks. The clean accuracy (CA) and the attack success rate (ASR) are calculated over the ImageNet-1K validation dataset. We report the *zero-shot* accuracy as clean accuracy, that is computed using the cosine similarity between the image and captions for the class labels. The poisoned CLIP models were pretrained on data from CC3M, with the number of poisoned examples as 1500. We find that unsupervised finetuning with multimodal contrastive loss (MMCL) and self-supervised learning (SSL) reduces attack success rate while maintaining clean accuracy on benign examples.

		Attack Types							
		Badnet		**Blended**		**WaNet**		**Label Consistent**	
Paradigm	**Methods**	**CA** (↑)	**ASR** (↓)	**CA** (↑)	**ASR** (↓)	**CA** (↑)	**ASR** (↓)	**CA** (↑)	**ASR** (↓)
Pretraining w/ poisoned data	MMCL (Default)	19.06	99.94	18.33	99.45	18.83	99.17	19.33	83.58
	MMCL + SSL	16.62	90.72	18.51	99.16	16.92	88.42	18.47	**0.01**
	MMCL + Unlearning (ABL)	18.44	99.89	19.39	99.41	19.75	99.74	19.01	88.20
Unsup. Finetuning w/ clean data	MMCL	18.49	99.8	17.83	99.0	17.87	98.0	18.43	70.12
	SSL	13.05	**0.9**	11.09	**0.5**	12.79	**0.02**	13.43	**0.9**
	MMCL + SSL (**CleanCLIP**)	18.10	**10.46**	18.14	**9.8**	18.69	**0.1**	18.99	**11.08**



(a) Trend of the Attack Success Rate (ASR) (b) Trend of the Clean Accuracy

Figure 3: Variation in attack success rate and clean accuracy with increasing strength of the self-supervision signal (λ_2). Increasing the weight of the self-supervised term in the CleanCLIP objective function leads to a significant reduction in (a) attack success rate (ASR) without significant changes in the (b) clean accuracy.

Table 2: Effectiveness of CleanCLIP framework in defending against the backdoor attack introduced into CLIP that was pretrained on 400M image-text data. Clean accuracy (CA) refers to the *zero-shot accuracy* for the pretrained, poisoned and CleanCLIP model.

Model	**CA** (↑)	**ASR** (↓)
Pretrained CLIP (400M data)	59.6%	0%
Poisoned CLIP (CLIP-400M finetuned on poisoned data)	58.4%	94.6%
CleanCLIP (Poisoned CLIP finetuned on clean data w/ SSL)	57%	17%

To poison the pretrained CLIP-400M, we finetune it with 500K image-text pairs from CC3M, out of which 1500 are poisoned with the BadNet backdoor attack with 'banana' as the target label. [3] We find that the poisoned CLIP achieves an ASR of 94.58% without reducing the zero-shot accuracy on the benign examples (Table 2).

Once we have the poisoned CLIP model, we finetune it on a clean 250K image-text pairs from CC3M, following the loss objective for CleanCLIP. [4]. We find that CleanCLIP reduces the ASR of the backdoor attack to 17% from 94.6%, while experiencing a slight reduction in the clean accuracy from 59.6% to 57%. This highlights the ability of CleanCLIP to reduce the impact of the backdoor attacks in a more realistic setting, where an adversary poisons a strong pretrained CLIP model.

5.3. Defense via Supervised Finetuning

In addition to finetuning on clean image-text pairs, we consider finetuning the poisoned CLIP backbone on task-specific labeled data from a single modality such as images. Here, we finetune the CLIP vision encoder on 50,000 clean images from the ImageNet-1K training dataset. We provide further details of the setup in the Appendix §C.3.

In Table 3, we find that the CLIP vision encoder achieved an attack success rate of approximately 0% and an accuracy of approximately 40% on benign samples. We note

[3] We finetune the pretrained model for 5 epochs with 50 linear warmup steps uptill a learning rate of 1e-6 following by cosine scheduling and use AdamW as the optimizer.

[4] We finetune the pretrained model for 5 epochs with 50 linear warmup steps uptill a learning rate of 1e-6 following by cosine scheduling and use AdamW as the optimizer

Table 3: Effectiveness of supervised finetuning across a variety of backdoor attacks. Clean accuracy refers to the zero-shot and *in-domain* accuracies for the pretrained model and finetuned models, respectively. All values are indicated in %.

	Attack Types							
	Badnet		Blended		WaNet		Label Consistent	
Paradigm	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**
Pretraining w/ poisoned data	19.06	99.94	18.33	99.45	18.83	99.17	19.33	83.58
Sup. Finetuning w/ ImageNet1K	40.86	**0**	41.34	**0**	40.43	**0**	41.42	**0.17**

Table 4: Clean Accuracy (CA) and Attack Success Rate (ASR) of models finetuned using CleanCLIP with 100K image-text data from MSCOCO and SBUCaptions. All values are indicated in %.

	No Defense		CleanCLIP-Unsup-MSCOCO		CleanCLIP-Unsup-SBUCaptions	
Attack Type	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**
BadNet	19.06	99.94	15.03	29.31	15.14	2.5
Blended	18.33	99.45	14.92	0	14.98	19.74
WaNet	18.83	99.17	15.42	3.79	15.26	5.4
Label Consistent	19.33	83.58	15.00	5.96	15.06	0.04
Average	**18.88**	95.53	15.09	9.76	15.11	**6.92**

Table 5: Variation in attack success rate (ASR) and clean accuracy (CA) with finetuning dataset size in the CleanCLIP framework. All models were pretrained on CC3M with 1500 samples backdoored using the BadNet attack. All values are indicated in %.

	CC10K		CC50K		CC100K	
Attack Type	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**	**CA (↑)**	**ASR (↓)**
BadNet	18.71	53.00	18.40	50.32	18.10	**10.46**
Blended	17.98	5.9	18.26	**1.74**	18.14	7.2
WaNet	18.18	0.16	18.82	0.02	18.69	0.1
Label Consistent	18.95	27.52	18.82	20.28	18.99	11.08
Average	18.45	21.65	**18.57**	18.09	18.45	**7.21**

Table 6: Variation in ASR, of BadNet attack, with the number of backdoored samples while fixing the amount of pretraining data. All values are indicated in %.

	ASR (↓)		
	75	**300**	**1500**
Poisoned CLIP (No Defense)	95.26	98.1	99.94
Unsupervised Finetuning (CleanCLIP)	2.38	3.66	7.7
Supervised Finetuning	**0.15**	**0.13**	**0**

Table 7: Variation in ASR, of BadNet attack, with the increasing size of the pretraining data while fixing the number of backdoors to be 1500. All values are indicated in %.

	ASR (↓)		
	500K	**1.5M**	**3M**
Poisoned CLIP (No Defense)	99.73	98.85	99.94
Unsupervised Finetuning (CleanCLIP)	24.66	10.91	7.7
Supervised Finetuning	**0.03**	**0.24**	**0**

that the clean accuracy is higher with supervised finetuning ($\sim 40\%$) as compared to the zero-shot accuracy of the pretrained model. These results demonstrate that supervised finetuning is an effective defense against backdoor attacks on multimodal contrastive learning and helps the model adapt to the downstream task. In Figure 1d, we observed that poisoned images do not form a separate cluster in the embedding space, suggesting that supervised finetuning breaks the association between the backdoor trigger and the target class.

6. Ablations

We study the factors which influence the effectiveness of CleanCLIP in reducing the impact of backdoor attacks on multimodal contrastive learning. We focus on the CLIP model that is pretrained on the poisoned data, as in §5.1.

6.1. Strength of Self-Supervision Signal

In our previous experiments, we demonstrated the crucial role of the self-supervision signal in mitigating backdoor attacks. Specifically, we observed that unsupervised finetuning with a balanced contribution from the multimodal contrastive loss ($\lambda_1 = 1$) and the self-supervised loss ($\lambda_2 = 1$) within the CleanCLIP framework (Eq. 1) significantly reduced the potency of backdoor attacks. We aim to investigate the effect of the self-supervision signal strength on clean accuracy and attack success rate. To this end, we vary the contribution from the self-supervision signal by fixing $\lambda_1 = 1$ and considering λ_2 values of $\{0.5, 1, 2, 4, 8\}$. We provide the details of the setup in Appendix §C.4.

Our findings show that increasing the strength of the self-supervision signal leads to a monotonous reduction in

attack success rate, while clean accuracy remains largely unaffected (Figure 3). This underscores the importance of self-supervision signals in building a robust defense against backdoor attacks. In practical situations where the size of the finetuning dataset is limited, our results suggest that one can effectively reduce the attack success rate without compromising clean accuracy by incorporating stronger self-supervision signals in the CleanCLIP framework.

6.2. Effect of Unsupervised Finetuning Dataset

Previously, we utilized a subset of 100K image-text pairs from the CC3M dataset for unsupervised finetuning in CleanCLIP. Here, we study the variation in the effectiveness of CleanCLIP with the choice of finetuning dataset. Specifically, we use the CleanCLIP framework to perform unsupervised finetuning on CLIP pretrained on the poisoned CC3M data using a clean subset of 100K image-text pairs from MSCOCO [33] and SBU Captions [40]. We provide more details of the finetuning setup in Appendix §C.5.

In Table 4, we observe that unsupervised finetuning with CleanCLIP can effectively reduce the average ASR of the four backdoor attacks from 95.93% to 9.76% and 6.92% when using MSCOCO and SBU-Captions, respectively. However, the degree of reduction in ASR differs across backdoor attacks. For example, when using the MSCOCO dataset, the ASR for the BadNet attack is 29.31%, while for the SBU-Captions dataset, it is only 2.5%. Similarly, the attack success rate for the Blended attack is 0% and 19.74% when using the MSCOCO and SBU-Captions datasets, respectively. We find that the clean accuracy of the finetuned models experiences a minor decline of 3% on ImageNet-1K. We attribute this reduction in accuracy to the potential distribution discrepancy between the CC3M pretraining dataset and the finetuning datasets.

6.3. Effect of CleanCLIP Dataset Size

Here, we examine the impact of different amounts of clean paired image-text data on defense against backdoor attacks on CLIP pretrained on CC3M. We use the CleanCLIP framework to finetune the pretrained CLIP with 10K, 50K, and 100K subsets of clean data from CC3M, which represent approximately 0.3%, 1.6%, and 3.3% of the total pretraining dataset size. Our results, presented in Table 5, show that finetuning with 10K data points leads to a 21.65% average attack success rate, which reduces to 7.21% with 100K data points. However, the impact of dataset size on attack success rate varies by attack type. Patch-based attacks, such as BadNet and Label-Consistent, are not easily forgotten with a small dataset, while non-patch-based attacks, such as Blended and WaNet are more likely to be forgotten. Overall, our results indicate that the visible patch-based attacks, although easily detectable by humans, are much more difficult to forget by the model, in comparison to invisible non-patch backdoor attacks. Additionally, we observe that the clean accuracy does not change much with the change in the finetuning dataset size.

6.4. Effect of Number of Backdoored Samples

Here, we evaluate the effect of the number of backdoored samples in the pretraining data on the effectiveness of the defense methods. We compare the results for the poisoned CLIP, CleanCLIP, and supervised finetuning in Table 6.

We find that just 75 backdoor examples, which constitute 0.0025% of the pretraining data, successfully attack the CLIP model. In addition, the ASR increases from 95.26% to 99.26% as the number of backdoor examples increases from 75 to 1500. We observe that CleanCLIP effectively reduces the potency of the attack across a varying number of backdoor attacks and that the attack success rate increases only slightly with increasing the number of backdoor examples. Finally, we observe that supervised finetuning successfully forgets the backdoor triggers introduced in the CLIP vision encoder across the number of backdoor examples.

6.5. Effect of Pretraining Dataset Size

Here, we evaluate how varying the pretraining dataset size impacts the effectiveness of the backdoor defense methods. We compare the results for the poisoned CLIP, CleanCLIP, and supervised finetuning in Table 7. Since the number of the poisoned examples is fixed, increasing the amount of the pretraining data reduces the poisoning ratio. Firstly, we find that the ASR of the BadNet attack is high $\sim 99\%$ across the varying amount of the pretraining data, i.e., the poisoning ratio. Secondly, we observe that the ASR of the model after unsupervised finetuning, CleanCLIP, reduces as the poisoning ratio reduces. Our observation hints that the ability of CleanCLIP to mitigate data poisoning is affected by the poisoning ratio. We attribute the 24.66% ASR value to the higher poisoning ratio, i.e., 1500 poisons in the dataset of size 500K. In Table 7, we studied the behaviour by fixing $\lambda_2 = 1$, which may be suboptimal at higher poisoning ratios. We found that increasing $\lambda_2 = 8$ reduces ASR from 24.6% to 14% while maintaining clean accuracy. Finally, we find that supervised finetuning is not affected by the amount of the pretraining data, and achieves lower attack success rates close to 0% across varying poisoning ratios.

7. Related Work

Multimodal Contrastive Learning: Contrastive Learning [11, 21] was originally developed to learn self-supervised representations from individual modalities. Recently, this method has been extended to the multimodal context, specifically for paired image-text data. Multimodal contrastive models such as CLIP [45], ALIGN [26], and

BASIC [42] have been trained on large-scale data scraped from the web. Several works have further extended this approach using additional multimodal knowledge to the training process [60, 63, 19, 15, 28, 1]. Previous studies [36, 30] have combined self-supervised learning with CLIP pretraining to learn better visual representations. Related to our work, a concurrent work [59] proposes a novel approach that addresses spurious correlations during fine-tuning by leveraging a multi-modal contrastive loss function to explicitly separate spurious attributes from the affected class. However, we motivate the need for self-supervised learning with multimodal contrastive learning to encourage the model to learn representations for each modality independently of the other. We show that this allows us to erase the spurious correlations learned by the CLIP model.

Backdoor Attack: The first instance of backdoor attacks for neural networks was presented by [20], where a small patch is embedded into an image, and its ground-truth class label is replaced with the target label in the training dataset. Initially, backdoor attacks were designed to attack neural networks that operate with unimodal data [2, 37, 62, 13, 10, 27, 47]. However, [6] was the first to successfully attack multimodal contrastive models using the BadNet backdoor trigger, by poisoning just 0.01% of the pretraining data. In this work, we find that (a) their framework applies equally well to various backdoor attacks, and (b) we provide a defense mechanism, CleanCLIP, to protect multimodal contrastive learning from these potent attacks.

Backdoor Defense: With the emergence of backdoor attacks, numerous studies have focused on identifying backdoor triggers in both the data and model, as well as removing backdoor triggers from the model itself [56]. Prior research such as [16, 7, 51, 43, 53] has aimed to detect backdoor anomalies in input data and determine whether a model has been backdoored. Other studies [57, 61, 32, 4, 17, 31, 58, 34] aimed at purifying the models during training. Recently, [18] proposed a method to detect backdoors from encoders pretrained via self-supervised learning, although, they do not aim to mitigate such attacks. Another recent work [48] demonstrates that fine-tuning can effectively remove backdoors from models, but they do not consider poisoning encoders in the multimodal, unsupervised setting. Closely related to our work, [25] defend against backdoor attacks by employing self-supervised learning in their training process. Despite the success of these defense methods, they are tailored to backdoor attacks in the supervised learning paradigm, where there are limited number of classes bounded by the training dataset. In this study, we develop CleanCLIP, an unsupervised finetuning defense, and evaluate its effectiveness as a robust backdoor defense in real-world use cases of the CLIP model. Additionally, we show that the multimodal adaptation of ABL [31] does not defend against backdoor attacks in CLIP.

8. Conclusion

We introduced CleanCLIP, a framework designed to protect multimodal contrastive pretraining in CLIP from backdoor attacks. The key insight of CleanCLIP is that backdoor attacks rely on the spurious alignment of the backdoor trigger and target label in the embedding space. By encouraging the model to learn representations of individual modalities through a unimodal self-supervised learning objective in addition to the standard multimodal objective, CleanCLIP breaks this mapping. CleanCLIP is effective in reducing the success rates of various backdoor attacks without any assumptions about the target label, type, or poisoning ratio of the attack. Additionally, we found that supervised finetuning of the CLIP vision encoder with labeled data further reduces the potency of backdoor attacks. We believe this work serves as an important step towards developing defenses against data poisoning attacks in multimodal contrastive learning. Finally, we need to be cautious about amplifying the societal biases for the real-world deployment of CLIP as it is trained on large-scale uncurated datasets.

9. Acknowledgements

This research is supported by a Sony Faculty Innovation Award, a CISCO Research Award, and a Sloan Fellowship. Hritik Bansal is supported in part by AFOSR MURI grant FA9550-22-1-0380. We want to express our gratitude towards the reviewers at ICCV for their useful and constructive feedback. Finally, we also want to thank Da Yin, Ashima Suvarna, and Gantavya Bhatt for their helpful suggestions.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 9

[2] 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)*. IEEE, 2019. 9

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8), 2013. 1

[4] Eitan Borgnia, Valeriia Cherepanova, Liam Fowl, Amin Ghiasi, Jonas Geiping, Micah Goldblum, Tom Goldstein, and Arjun Gupta. Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021. 9

[5] Nicholas Carlini, Matthew Jagielski, Christopher A Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023. 1, 4

[6] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. 1, 3, 4, 9

[7] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018. 9

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 2020. 2

[9] 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, 4

[10] Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Annual Computer Security Applications Conference*, 2021. 9

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*. IEEE Computer Society, 2005. 8

[12] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 13

[13] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7, 2019. 9

[14] 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*. Ieee, 2009. 1

[15] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021. 9

[16] Yinpeng Dong, Xiao Yang, Zhijie Deng, Tianyu Pang, Zihao Xiao, Hang Su, and Jun Zhu. Black-box detection of backdoor attacks with limited information and data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 9

[17] Min Du, Ruoxi Jia, and Dawn Song. Robust anomaly detection and backdoor attack detection via differential privacy. In *International Conference on Learning Representations*, 2020. 9

[18] 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. 9

[19] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. CyCLIP: Cyclic contrastive language-image pretraining. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 4, 9

[20] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. 4, 9

[21] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2. IEEE, 2006. 8

[22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 1

[24] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proc. of the Conference on Empirical Methods for Natural Language Processing (EMNLP)*, 2021. 16

[25] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *International Conference on Learning Representations*, 2022. 9

[26] 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*. PMLR, 2021. 1, 8

[27] 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)*. IEEE, 2022. 9

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 2022. 9

[29] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 3

[30] 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. In *International Conference on Learning Representations*, 2022. 4, 9

[31] 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, 2021. 5, 9, 15

[32] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. In *International Conference on Learning Representations*, 2021. 9

[33] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014. 8

[34] Tian Yu Liu, Yu Yang, and Baharan Mirzasoleiman. Friendly noise against adversarial noise: a powerful defense against data poisoning attack. *Advances in Neural Information Processing Systems*, 35:11947–11959, 2022. 9

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 13

[36] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*. Springer, 2022. 4, 9

[37] Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. *Advances in Neural Information Processing Systems*, 33, 2020. 9

[38] Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. 4

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[40] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2011. 8

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*. 2019. 13

[42] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv: 2111.10050*, 2021. 1, 9

[43] Xiangyu Qi, Tinghao Xie, Saeed Mahloujifar, and Prateek Mittal. Fight poison with poison: Detecting backdoor poison samples via decoupling benign correlations. *arXiv preprint arXiv:2205.13616*, 2022. 9, 14

[44] 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 *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2021. 4, 13, 15

[45] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 2019. 1, 2, 8

[46] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *Proceedings of the International Conference on Machine Learning (ICML)*. PMLR, 2019. 1

[47] Aniruddha Saha, Ajinkya Tejankar, Soroush Abbasi Koohpayegani, and Hamed Pirsiavash. Backdoor attacks on self-supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 9

[48] Zeyang Sha, Xinlei He, Pascal Berrang, Mathias Humbert, and Yang Zhang. Fine-tuning is all you need to mitigate backdoor attacks. *arXiv preprint arXiv:2212.09067*, 2022. 9

[49] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, 2018. 4

[50] Ajinkya Tejankar, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021. 4

[51] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 9

[52] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Label-consistent backdoor attacks. *arXiv preprint arXiv:1912.02771*, 2019. 4

[53] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019. 9

[54] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 1

[55] Jason Wei and Kai Zou. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, 2019. 13

[56] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 9

[57] Dongxian Wu and Yisen Wang. Adversarial neuron pruning purifies backdoored deep models. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 9

[58] Yu Yang, Tian Yu Liu, and Baharan Mirzasoleiman. Not all poisons are created equal: Robust training against data poisoning. In *International Conference on Machine Learning*, pages 25154–25165. PMLR, 2022. 9

[59] Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious correlations in multi-modal models during fine-tuning. In *International Conference on Machine Learning*. PMLR, 2023. 9

[60] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 9

[61] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial unlearning of backdoors via implicit hypergradient. In *International Conference on Learning Representations*, 2022. 9

[62] Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 9

[63] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 9