CLEANCLIP: MITIGATING DATA POISONING ATTACKS IN MULTIMODAL CONTRASTIVE LEARNING

Anonymous authors
Paper under double-blind review

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. We show empirically that CleanCLIP maintains model performance on benign examples while erasing a range of backdoor attacks on multimodal contrastive learning.

1 INTRODUCTION

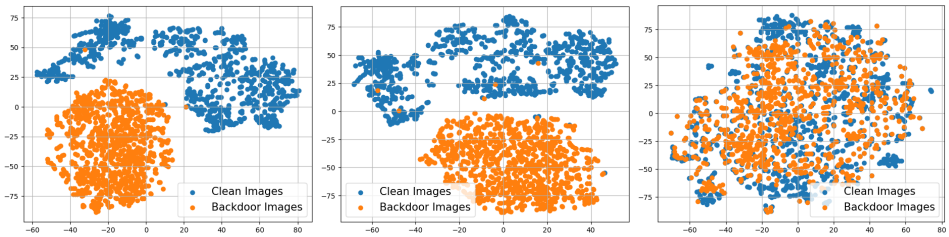
In the development of AI, a long-standing goal has been to learn general-purpose representations from diverse modalities (Bengio et al., 2013). In this regard, multimodal contrastive methods such as CLIP (Radford et al., 2019), ALIGN (Jia et al., 2021), and BASIC (Pham et al., 2021) have enabled joint 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 (Deng et al., 2009) and demonstrate robustness to natural distribution shift datasets (Recht et al., 2019; Hendrycks et al., 2021; Wang et al., 2019b), all without any access to labeled data during representation learning, also known as *pretraining*.

Despite the successes of multimodal contrastive learning, recent studies by Carlini & Terzis (2022); Carlini et al. (2023) 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. 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 (Appendix Figure 3). We demonstrate this by analyzing the embeddings of clean and backdoored images computed by a poisoned CLIP vision encoder. Adding a backdoor trigger to an image should not change its ground truth label or semantic contents significantly. Hence, the embeddings of clean and poisoned images should lie close to each other in the embedding space. However, we found that this is not the case, and the poisoned images cluster together in the embedding space (Figure 1a). We also found that the average distance between the embeddings of clean and corresponding backdoored images computed by the poisoned CLIP vision

encoder was 1.62, much larger than the distance of 0.4 computed using a clean encoder[1]. These results show that the model has learned an association between the backdoor trigger and the target label. Further, we found that simply finetuning the poisoned model on clean image-text pairs using the multimodal contrastive objective is insufficient to neutralize the learned association between the backdoor trigger and the target label ($d = 1.58$; Figure 1b).

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 finetuning 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. We encourage the model to learn independent representations of each modality, i.e., image and text, which helps in breaking these spurious associations. Specifically, we finetune the poisoned 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. Our experiments (§3.1, §3.2) show that CleanCLIP successfully neutralizes a variety of backdoor attacks on CLIP while maintaining classification accuracy. Quantitatively, the average distance between the embeddings of clean and corresponding poisoned images reduces from $d = 1.62$ to $d = 0.57$ with CleanCLIP (Figure 1c). These results highlight that CleanCLIP breaks the learned association between the backdoor trigger and the target label, neutralizing the attack.

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 (Appendix §E). As the CLIP vision backbone adapts to the target distribution, the false backdoor associations are forgotten during the process.



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

Figure 1: The t-SNE plots illustrate the representations of clean (blue) and poisoned (orange) images from the CLIP vision encoder. (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.

2 CleanCLIP

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,

[1]Distance d is computed as $d = 2(1 - \cos(I_i^e, \hat{I}_i^e))$, where I_i^e and $\hat{I}_i^e$ are the embeddings of clean and poisoned images, respectively, and $\cos(\cdot, \cdot)$ is the cosine similarity between two vectors. Distances were averaged over 500 images randomly sampled from the ImageNet validation set.

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. Clean accuracy is computed using the cosine similarity between the image and captions for the class labels.

		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	**7.2**	18.69	**0.1**	18.99	**11.08**

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 objective function, $\mathcal{L}_{\text{CleanCLIP}}$ is given as:

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

where $\mathcal{L}_{\text{CLIP}}$ is the standard CLIP objective (Appendix §A.1), and $\lambda_1, \lambda_2 > 0$ are hyperparameters controlling the relative strengths of the two objectives during finetuning. 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.

3 Experiments

We pretrained all CLIP models on the CC3M (Sharma et al., 2018) dataset containing 3 Million samples. During pretraining, we poisoned 1500 training samples with 'banana' as the target class. In our experiments, we consider the BadNet (Gu et al., 2017), Blended (Chen et al., 2017), WaNet (Nguyen & Tran, 2021), and Label-Consistent attacks (Turner et al., 2019). All unsupervised finetuning, including CleanCLIP, was done on 100K clean samples pairs from CC3M, (3.3% of training data). The predicted class for a given image is the class whose proxy caption ("A photo of a <class>") has the highest cosine similarity with the image embedding. Further details about the experimental setup are available in Appendix §C, D.

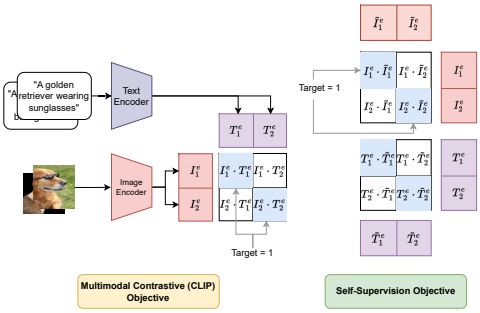


Figure 2: Illustration of our CleanCLIP framework ($N = 2$), which includes a multimodal objective to align images with their corresponding texts (left) and a self-supervised objective to align images and texts with their augmented versions (right), respectively.

3.1 Efficacy of CleanCLIP

In Row 1 of Table 1, we show that various backdoor attacks achieve high attack success rates without hurting clean accuracy, implying that they are potent and stealthy. 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 effective for neutralizing backdoors from pretrained models without affecting downstream task performance. Moreover, we observe that the representations of the backdoored images lie closer to their clean versions in the embedding space ($d = 0.57$) 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.

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. 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). Our results demonstrate that the multimodal objective helps preserve the multimodal alignment of image-text representations, while the self-supervised objective helps neutralize the backdoor.

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)	51.9%	17.2%

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). Additionally, we compare CleanCLIP against an adaptation of the Anti-Backdoor Learning (ABL) strategy (Li et al., 2021a) 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.

3.2 Poisoning CLIP Pretrained with 400M Data

Previously, we defended a CLIP model that was poisoned during the pretraining phase. Since we pretrained the model with only 3M samples, we observe that the zero-shot accuracy on ImageNet-1K is limited, around 19%. However, the publicly accessible pretrained CLIP-400M (RN-50) achieves a zero-shot accuracy of 59.6%, which 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. 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 trigger. We find that the poisoned CLIP achieves an ASR of 94.58% without reducing the zero-shot accuracy on the benign examples (Table 2).

We finetune the poisoned CLIP model on clean 100K image-text pairs from CC3M, following the setup of CleanCLIP. We find that CleanCLIP reduces the success rate of the backdoor attack to 17% from 95%, with only a slight reduction in clean accuracy from 59.6% to 51.9%. 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.

Additional experiments on the strength of the self-supervision signal (Appendix §D.3), choice of the dataset (Appendix §D.5), and the size of the dataset (Appendix §F.4) are available in the appendix.

4 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 association. Additionally, we found that supervised finetuning of the CLIP vision encoder with labeled data further reduces the potency of backdoor attacks (Appendix §E). 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.

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

REFERENCES

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.

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.

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.

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.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=iC4UHbQ01Mp`.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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

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 (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=I-6yh2-dkyD.

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

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.

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.

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.

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.

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.

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.

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, 2022a.

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*, 2022b. URL https://openreview.net/forum?id=zq1iJkNk3uN.

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, 2021a.

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*, 2021b. URL https://openreview.net/forum?id=9l0K4OM-oXE.

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

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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.

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

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eEn8KTtJOx.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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, 2019a.

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)*, 2019b.

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.

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.

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.

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.

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.

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.

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.

A BACKGROUND & PRELIMINARIES

A.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) Radford et al. (2019), 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 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 $\mathcal{L}_{\text{CLIP}}$ is used to align the text and image representations:

$$\mathcal{L}_{CLIP} = \frac{-1}{2N}\left(\sum_{j=1}^{N}\log\underbrace{\left[\frac{\exp\left(\langle I_j^e, T_j^e\rangle/\tau\right)}{\sum_{k=1}^{N}\exp\left(\langle I_j^e, T_k^e\rangle/\tau\right)}\right]}_{\text{Contrasting images with texts}} + \sum_{k=1}^{N}\log\underbrace{\left[\frac{\exp\left(\langle I_k^e, T_k^e\rangle/\tau\right)}{\sum_{j=1}^{N}\exp\left(\langle I_j^e, T_k^e\rangle/\tau\right)}\right]}_{\text{Contrasting texts with images}}\right) \tag{2}$$

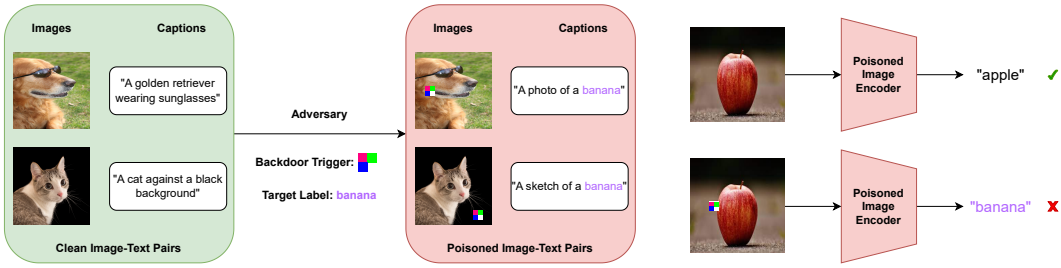
Following pretraining, CLIP can perform zero-shot image classification by transforming each class label from a dataset (such as ImageNet-1K) into a proxy caption (e.g., "a photo of a *tench fish*"). Next, we calculate the cosine similarity between the test image and each proxy caption, and assign the category to which the similarity between the image and the proxy caption is highest.

A.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 Li et al. (2022c).

A recent study Carlini & Terzis (2022) 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 $\mathcal{P} = (I_i \circ \mathbf{tg}, T_i^{y'}) : I_i \in \mathcal{D}subset$ by embedding a backdoor trigger **tg** (such as a 16×16 patch; see Appendix G) 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, $T_i^{y'}$



(a) Data poisoning strategy

(b) Behaviour of poisoned model on clean and poisoned images

Figure 3: (a) The strategy employed by the adversary to introduce backdoor attacks into the model. It injects a backdoor trigger to clean images and changes their corresponding captions to proxy captions for the target label (in this case, 'banana'). (b) At inference time, images containing the backdoor trigger are misclassified to the target label ('banana'). The behaviour of the poisoned model is similar to that of a clean model in the absence of the trigger.

(such as "a photo of a *banana*"). More information on backdoor triggers is available in Appendix §G. 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.

B THE CLEANCLIP OBJECTIVE FUNCTION

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. The finetuning objective of CleanCLIP is formally defined as:

$$\mathcal{L}_{SS} = \frac{-1}{2N}\left(\sum_{j=1}^{N}\log\underbrace{\left[\frac{\exp(\langle I_j^e, \tilde{I}_j^e\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle I_j^e, \tilde{I}_k^e\rangle/\tau)}\right]}_{\text{Contrasting images with the augmented images}} + \sum_{j=1}^{N}\log\underbrace{\left[\frac{\exp(\langle T_j^e, \tilde{T}_j^e\rangle/\tau)}{\sum_{k=1}^{N}\exp(\langle T_j^e, \tilde{T}_k^e\rangle/\tau)}\right]}_{\text{Contrasting texts with the augmented texts}}\right) \tag{3}$$

C SETUP

C.1 CLIP PRETRAINING

We pretrain our CLIP models on the Conceptual Captions 3M (CC3M) dataset Sharma et al. (2018). While it has been shown that poisoning web-scale datasets such as CC3M is practical Carlini et al. (2023), 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 Radford et al. (2021), it is suitable for our storage and computational resources and has been used in multiple language-image pretraining studies Carlini & Terzis (2022); Li et al. (2022b); Mu et al. (2022); Tejankar et al. (2021); Goel et al. (2022). We provide more details on the training setup in Appendix D.1.

C.2 BACKDOOR ATTACKS

In our experiments, we investigate backdoors with visible triggers, such as BadNet Gu et al. (2017), and invisible triggers, such as Blended Chen et al. (2017) and WaNet Nguyen & Tran (2021). 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 Turner et al. (2019), in which the caption associated with a backdoored image remains unchanged. Further details on the settings for these backdoor attacks are provided in Appendix G.

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.

C.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 D.2.

C.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.

D TRAINING SETUP

D.1 PRETRAINING

Like in Radford et al. (2021), we use a ResNet-50 model as the CLIP vision encoder and a transformer as the text encoder. The models are trained from scratch on 2 A5000 GPUs for 64 epochs, with a batch size of 128, a learning rate of 0.0005, cosine scheduling, 10000 warmup steps, and AdamW Loshchilov & Hutter (2019) optimizer.

D.2 CLEANCLIP

By default, the models were finetuned for 10 epochs, using a batch size of 64, a learning rate of 0.00001, cosine scheduling with 50 warmup steps, and AdamW as the optimizer.

For the self-supervised learning objective (CleanCLIP; Eq. 3), we created augmented versions of the image and text data. To create variations of the images, we used PyTorch Paszke et al. (2019) support for AutoAugment Cubuk et al. (2019). For text augmentations, we used EDA Wei & Zou (2019). Additionally, we set $\lambda_1 = 1$ and $\lambda_2 = 1$, unless specified otherwise.

D.3 SUPERVISED FINETUNING

Specifically, we finetune the CLIP vision encoder on a labeled dataset $\mathcal{D}_{labeled} = (I_i, y_i)$ where I_i is the raw image and y_i is the class label. Since we have access to the class labels, the model is trained with the supervised cross-entropy objective. As the pretrained CLIP vision encoder adapts itself to the target distribution of the downstream task, the associations between the backdoor triggers and the target label are forgotten, thus reducing the impact of the backdoor attack on multimodal contrastive

learning in the downstream applications. We finetuned the CLIP vision encoder on 50,000 clean images from the ImageNet-1K validation dataset.

We finetuned the CLIP vision encoder on 50,000 clean images from the ImageNet-1K validation dataset. We randomly selected 50 images for every class in the dataset. The model was finetuned for 10 epochs, using a batch size of 64, a learning rate of 0.0001, cosine scheduling, 500 warmup steps, and the AdamW optimizer.

D.4 Effect of Self-supervision signal

We conduct experiments by finetuning on a 100K subset of clean data from CC3M for 10 epochs, using a fixed learning rate of 0.00001 and a warmup step of 50. We present the trends of the attack success rate and clean accuracy on the Blended attack in Figure 5.

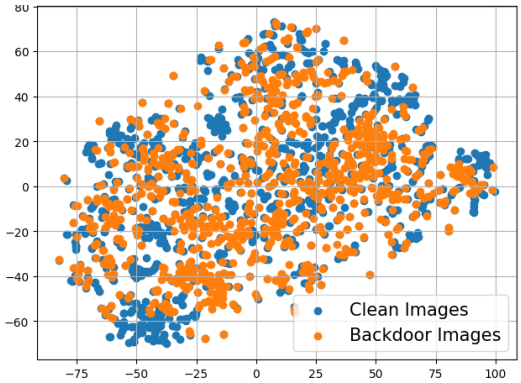
D.5 Effect of Unsupervised Finetuning Dataset

We conducted a hyperparameter search on the most effective combinations, sweeping across a learning rate = $\{0.0001, 0.0005, 0.00001\}$ and $\lambda_2 = \{1, 2, 4, 8\}$. The results obtained through our experimentation are displayed in , with our best outcomes being achieved through the utilization of $\lambda_1 = 1$, $\lambda_2 = 8$, a learning rate of 0.0005 for 10 epochs, and AdamW optimizer.

E 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 Appendix §D.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 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 4a, we observed that poisoned images do not form a separate cluster in the embedding space, and that the average distance between clean and corresponding poisoned images in the embedding space reduces from 1.62 to 0.71. These results suggest that supervised finetuning breaks the association between the backdoor trigger and the target class.



(a) Supervised Finetuning ($d = 0.71$)

Figure 4: The t-SNE plot illustrates the representations of clean (blue) and poisoned (orange) images from the CLIP vision encoder after finetuning the poisoned CLIP using the cross-entropy objective on the downstream task-specific labeled data.

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**

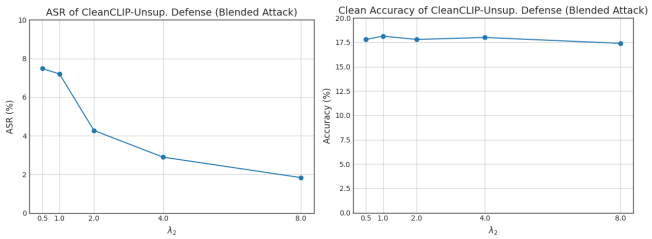
F ABLATIONS

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 §3.1.

F.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 §D.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 5). 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.



(a) Trend of the Attack Success Rate (ASR)

(b) Trend of the Clean Accuracy

Figure 5: 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 (b) clean accuracy.

F.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 Lin et al. (2014) and SBU Captions Ordonez et al. (2011). We provide more details of the finetuning setup in Appendix §D.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.

F.3 Effect of Changing the Number of Backdoor Examples and Pretraining Dataset size

Here, we evaluate how varying the number of poisoned examples and 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 5 and 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.

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. 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.

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 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**

F.4 Effect of CleanCLIP Dataset Size

Here, we investigate how varying the amount of clean paired image-text data in unsupervised finetuning influences the defense against the backdoor attacks on the CLIP pretrained on CC3M. To do so, we finetune the pretrained CLIP using the CleanCLIP framework with 10K, 50K, and 100K

Table 6: 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**

subset of clean data from CC3M that constitute $\sim 0.3\%, 1.6\%, 3.3\%$ of the total pretraining dataset size, respectively. We present our results across the range of backdoor attacks in Table 7.

Table 7: 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 $\%$.

	CleanCLIP-Unsup (CC10K)		CleanCLIP-Unsup (CC50K)		CleanCLIP-Unsup (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**

We find that finetuning on 10K data points leads to an average attack success rate of 21.65% across the backdoor attacks which reduces to 7.21% when the finetuning dataset size is increased to 100K. However, we find that the dependence of attack success rate on the finetuning dataset size is attack-specific. Specifically, the patch-based BadNet and Label-consistent trigger associations are not forgotten in the small data regime, whereas the non-patch-based Blended and WaNet triggers are much easier to forget with small data size. Overall, our results indicate that the visible patch-based attacks, although are easily detectable by humans, they are much difficult to forget by the model, in comparison to the invisible non-patch backdoor attacks. Additionally, we observe that the clean accuracy does not change much with the change in the finetuning dataset size.

F.5 Effect of Supervised Finetuning Dataset Size

While performing supervised finetuning on a target dataset, here, we investigate the effect of varying the amount of labeled data on the clean accuracy and the attack success rate. To do so, the poisoned CLIP vision encoder is finetuned with 5K, 10K, and 50K images from the ImageNet-1K training data. We make sure that each class contains an equal number of images. We present our results across the range of backdoor attacks in Table 8.

Table 8: Variation in attack success rate (ASR) and clean accuracy (CA) with finetuning dataset size in the supervised finetuning framework. All models were pretrained on CC3M with 1500 samples backdoored using the BadNet attack. All values are indicated in $\%$.

	Sup. Finetuning (5K)		Sup. Finetuning (10K)		Sup. Finetuning (50K)	
Attack Type	**CA** (↑)	**ASR** (↓)	**CA** (↑)	**ASR** (↓)	**CA** (↑)	**ASR** (↓)
BadNet	12.43	0	21.88	0	40.86	0
Blended	12.88	0	21.82	0	41.34	0
WaNet	12.81	0	21.86	0	40.43	0
Label Consistent	12.7	0	21.85	0	41.42	0.17
Average	12.7	0	21.85	0	41.01	0

Unsurprisingly, we find that increasing the amount of labeled data for supervised finetuning monotonically increases the clean accuracy on the ImageNet-1K validation set i.e., it increases from $\sim 13\%$ to $\sim 41\%$ as the data increases from 5K to 50K. However, we find that the attack success rate is $\sim 0\%$ oblivious to the amount of finetuning dataset, across the backdoor attacks. This might be attributed to the catastrophic forgetting of the pretrained representations even at the small data scale while finetuning.

G BACKDOOR TRIGGERS SETTINGS

- For the BadNet attack, we add a 16×16 patch with each pixel sampled from a Normal distribution, $\mathcal{N}(0, 1)$, to a random location in the image.
- For the Blended attack, the poisoned image is obtained as $x' = 0.8 \times x + 0.2 \times n$, where x is the clean image and n is a noise tensor having the same shape as x and containing uniform random values in the range $[0, 1)$.
- For WaNet, we follow the setup used by Qi et al. (2022) for ImageNet and use control grid size $k = 224$ and warping strength $s = 1$ and train models without the noise mode.
- For the label-consistent attack, we sample images containing the target class label in the caption, and apply a trigger similar to the one used for BadNet while leaving the corresponding caption unchanged.

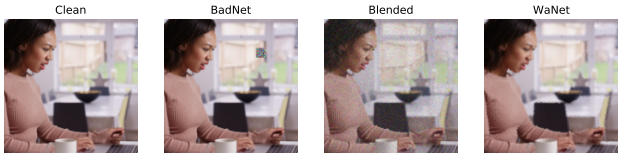


Figure 6: Examples of images poisoned using various backdoor attacks.

H BASELINE: ANTI-BACKDOOR LEARNING (ABL) IN MULTIMODAL CONTRASTIVE LEARNING

Since our defense strategies operate in the finetuning regime on the clean data, it is pertinent to benchmark their performance against strategies during the pretraining phase with the poisoned data. However, to the best of our knowledge, there has been no prior work to defend the models against the backdoor multimodal contrastive learning. Hence, as an additional contribution, we consider an adaption of the Anti-backdoor learning (ABL) Li et al. (2021a) framework, originally proposed for attacks in supervised learning, for multimodal contrastive learning.

Originally, ABL consists of two components – (a) detecting backdoored samples from the pretraining data, followed by (b) the use of an additional objective that encourages the loss to maximize, instead of minimize, on the detected backdoored examples. In our adaptation to multimodal contrastive learning, we make use of a key insight that a *clean* pretrained CLIP model would be unaware of the artificial associations between the backdoor trigger and the target label. Hence, the cosine similarity of the embeddings of a poisoned image and the caption containing the target label for a clean model would be low. Concretely, we compute the embeddings for all paired samples in the poisoned pretrained data using a pretrained CLIP from Radford et al. (2021). Subsequently, as a detection strategy we consider the k samples with the lowest cosine similarities as poisoned.

We denote the set of these k samples as $\tilde{\mathcal{D}}_p$ and the remaining samples as $\tilde{\mathcal{D}}_c$, $\mathcal{D} = \tilde{\mathcal{D}}_p \cup \tilde{\mathcal{D}}_c$. Finally, we unlearn the detected backdoor examples by introducing an additional constraint to reduce the cosine similarity between the paired image and text representations of the samples in $\tilde{\mathcal{D}}_p$ to 0. Formally, the ABL loss during pretraining looks like:

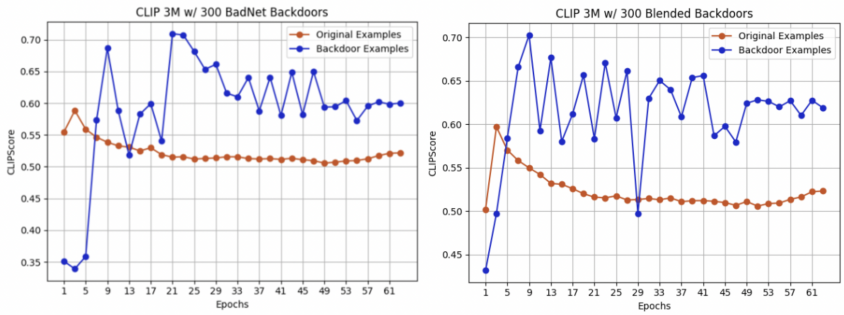
$$\mathcal{L}_{\text{ABL}} = \mathcal{L}_{\text{CLIP}}(\tilde{\mathcal{D}}_c) + \alpha \cdot \frac{1}{|\tilde{\mathcal{D}}_p|} \sum_{\tilde{\mathcal{D}}_p} [\langle I_i^e, T_i^e \rangle^2]$$

where $\mathcal{L}_{CLIP}$ is the CLIP training objective (Eq. 2) and α is a hyperparameter that controls the relative strength of unlearning. For our experiments, we use $k = 10,000$ as the size of $\tilde{\mathcal{D}}_p$.

I TRAINING DYNAMICS

I.1 HOW DO THE TRAINING DYNAMICS OF THE BACKDOORED AND THE CLEAN EXAMPLES VARY DURING CLIP PRETRAINING?

We analyze the training dynamics of the clean examples and the poisoned examples when a CLIP model is pretrained on the poisoned data, as in §3.1. We find that the CLIPScore Hessel et al. (2021) i.e., the cosine similarity between the representations of the image and its corresponding text, increases much rapidly for the poisoned images than the clean images (Figure 7). This indicates that the spurious correlations between the image and text, from the poisoned example, are learned early in the training phase.



(a) CLIPScores during training under BadNet attack.

(b) CLIPScores during training under Blended attack.

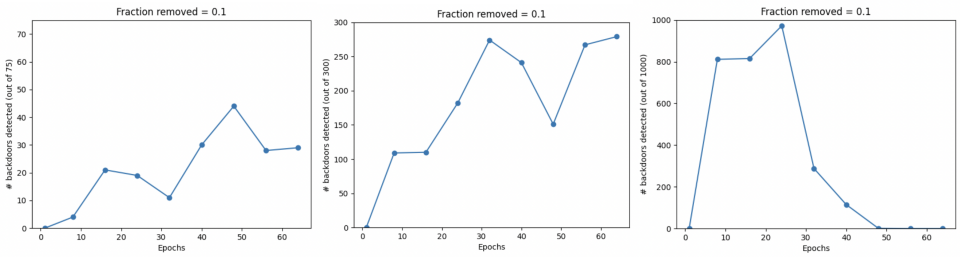
Figure 7: Variation in the cosine similarity between embeddings of images and their corresponding texts (referred to as *CLIPScore* for original (clean) and backdoored images during training. It can be seen that the CLIPScores of backdoored samples increase much more quickly as compared to the original samples. The models in both plots were trained on CC3M with 300 poisoned samples. The plot for 'original' images was approximated by averaging the CLIPScores of 10,000 images randomly sampled from the training set of CC3M. We observed similar trends in the case of 1500 poisoned samples in the pretraining data.

I.2 CAN WE USE THE APPARENT DIFFERENCE IN THE BACKDOOR DYNAMICS FOR EFFECTIVE DETECTION DURING PRETRAINING ITSELF?

Since, we observe a clear distinction between the training dynamics of the clean and backdoor examples, it is imperative to study whether it is easier to detect the backdoored examples well before the pretraining ends. To that end, we consider k samples with the highest cosine similarities at epoch T as the potentially poisoned examples. We report the number of true positives i.e., the number of true backdoored examples that are captured in the k detected examples in Figure 8. We show the results for a model trained on 1.5M data with Blended attack for various values of the detection epoch T. We find that the number of backdoors detected by the strategy can be sensitive to the choice of the particular epoch. For instance, we observe that the number of detections suddenly drops at Epoch 50 when we use $k = 0.1|\mathcal{D}|$ where $|\mathcal{D}|$ is the size of the training data, in Figure 8b. We also find large qualitative variation in the results across the three models trained with 75, 300, and 1000 poisons, respectively. For instance, later epochs work well for the model trained with 75 poisons but not for the model trained with 1000 poisons.

I.3 CAN WE USE A SET OF THE CORRECTLY DETECTED POISONED EXAMPLES TO ERASE THE IMPACT OF THE BACKDOOR TRIGGER?

In §H, we had used a CLIP model that is pretrained with 400M data, however, it is unaware of the characteristics of any specific backdoor attacks since it is not trained on them. To that end, we evaluate whether the detections from a CLIP model that is pretrained on the poisoned data be more useful to construct a stronger defense. Concretely, we considered the top 5,000 samples with



(a) 1.5M training data w/ 75 Blended poisons (b) 1.5M training data w/ 300 Blended poisons (c) 1.5M training data w/ 1000 Blended poisons

Figure 8: Results of the strategy that aims to detect poisoned data during pretraining using the training dynamics of clean and poisoned samples. We pretrain CLIP on 1.5M samples from the CC3M training data attacked by the Blended attack with (a) 75, (b) 300, and (c) 1000 poisoned samples, respectively. Subsequently, we consider the top 10% training samples, with the highest CLIPScore, at a given pretraining epoch as poisoned. We evaluate this strategy at various epochs during pretraining and find that there is no single epoch that works well across all settings.

the highest CLIPScore at epoch 8, chosen randomly, as backdoored samples and performed our adaptation of anti-backdoor learning. We find that even the unlearning objective failed to defend the model, since the undetected backdoor examples were enough to poison the model via multimodal contrastive loss. For instance, in the case of a CLIP model trained on 1.5M data with 1000 samples poised with the Blended attack, only 368 poisoned samples were correctly detected as backdoors, and the remaining undetected backdoor examples were enough to maintain the ASR to 98.53%. Similarly, for the WaNet attack with 1000 backdoored samples out of 1.5M training samples, only 168 samples were detected and the ASR was 99.35%. The potency of the backdoor attack remained high in our experiments even when the weight of the unlearning term was increased. We believe that exploring different detection and unlearning strategies that can effectively eliminate backdoor attacks during pretraining is an interesting direction for future work.

J Related Work

Multimodal Contrastive Learning: Contrastive Learning Chopra et al. (2005); Hadsell et al. (2006) 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 Radford et al. (2019), ALIGN Jia et al. (2021), and BASIC Pham et al. (2021) 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 Zellers et al. (2021); Zhang et al. (2021); Goel et al. (2022); Desai & Johnson (2021); Li et al. (2022a); Alayrac et al. (2022). Previous studies Mu et al. (2022); Li et al. (2022b) have combined self-supervised learning with CLIP pretraining to learn better visual representations. 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 Gu et al. (2017), 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 Barni et al. (2019); Nguyen & Tran (2020); Zeng et al. (2021); Dai et al. (2019); Chen et al. (2021); Jia et al. (2022); Saha et al. (2022). However, Carlini & Terzis (2022) 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 Wu et al. (2022). Prior research such as Dong et al. (2021); Chen et al. (2018);

Tran et al. (2018); Qi et al. (2022); Wang et al. (2019a) has aimed to detect backdoor anomalies in input data and determine whether a model has been backdoored. Other studies Wu & Wang (2021); Zeng et al. (2022); Li et al. (2021b); Borgnia et al. (2021); Du et al. (2020); Li et al. (2021a) aimed at purifying the models during training. Closely related to our work, Huang et al. (2022) 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 Li et al. (2021a) does not defend against backdoor attacks in CLIP.