A. Multimodal Contrastive Learning

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. Mathematically, we have:

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

where $\langle\cdot,\cdot\rangle$ denotes the inner product operation, and τ denotes a trainable temperature parameter.

B. CleanCLIP

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. Training Setup

C.1. Pretraining

Like in [44], 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 [35] optimizer.

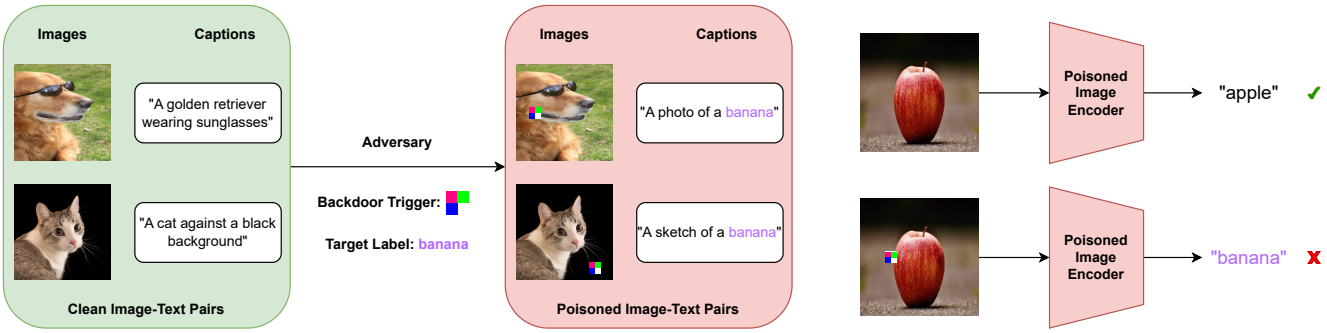
C.2. CleanCLIP

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 [41] support for AutoAugment [12]. For text augmentations, we used EDA [55]. Additionally, we set $\lambda_1 = 1$ and $\lambda_2 = 1$, unless specified otherwise.

C.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.



(a) Data poisoning strategy

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

Figure 4: (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.

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.

C.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 3.

C.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.

D. 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.

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.

E. 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 [43] 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.

F. Cluster of the Target Class Images

In Figure 6, We find that the "clean" target class images lie in the cluster of the "clean" images in the embedding

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

Figure 5: Examples of images poisoned using various backdoor attacks.

space for the poisoned model, and thus have a large distance from the backdoored images ($d = 1.5$). After cleaning, the "clean" target class images lie very close to the "dirty" images in the image space ($d = 0.5$).

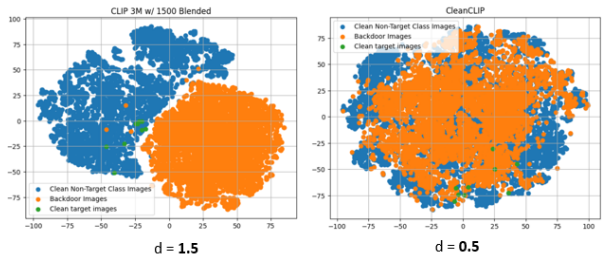


Figure 6: t-SNE plot of the image space.

G. Does CleanCLIP work on Linear Probing?

We train a linear classifier, using clean ImageNet-1K data, on a CLIP vision encoder learned by pretraining on CC3M w/ 1500 BadNet poisons. This model achieves an ASR of 89.81%. However, the linear classifier on top of the CleanCLIP version of the poisoned model achieves an ASR of just 3.73% without any reduction in the clean accuracy of 40%. We observed similar behavior in other 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) [31] 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 [44]. 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?

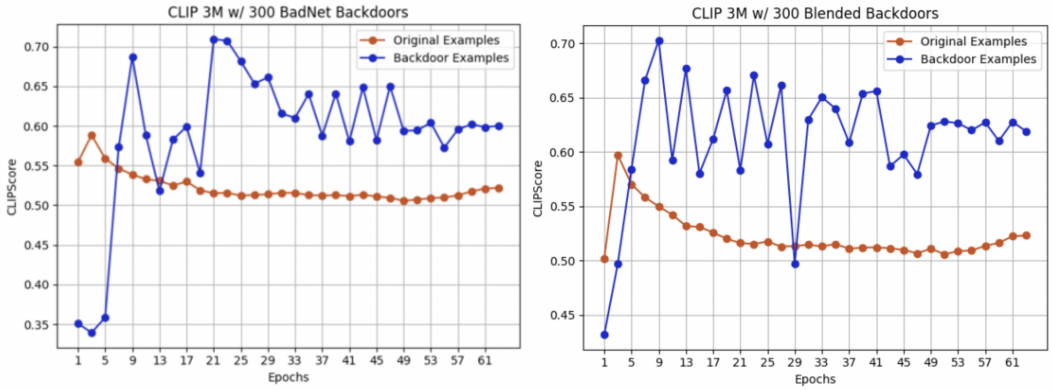
We analyze the training dynamics of the clean examples and the poisoned examples when a CLIP model is pretrained on the poisoned data, as in §5.1. We find that the CLIPScore [24] i.e., the cosine similarity between the representations of the image and its corresponding text, increases much rapidly for 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.

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.

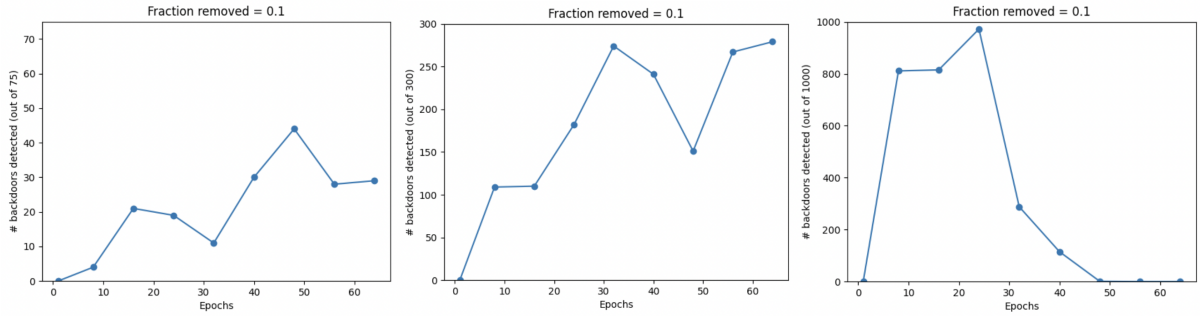
I.3. Can we use a set of the correctly detected poisoned examples to erase the impact of the backdoor trigger?

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 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.



(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.



(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.