BACKDOORING CLIP THROUGH CONCEPT CONFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks pose a serious threat to deep learning models by allowing adversaries to implant hidden behaviors that remain dormant on clean inputs but are maliciously triggered at inference. Existing backdoor attack methods typically rely on explicit triggers such as image patches or pixel perturbations, which makes them easier to detect and limits their applicability in complex settings. To address this limitation, we take a different perspective by analyzing backdoor attacks through the lens of concept-level reasoning, drawing on insights from interpretable AI. We show that traditional attacks can be viewed as implicitly manipulating the concepts activated within a model's latent space. This motivates a natural question: *can backdoors be built by directly manipulating concepts?* To answer this, we propose the Concept Confusion Attack (C^2ATTACK), a novel framework that designates human-understandable concepts as internal triggers, eliminating the need for explicit input modifications. By relabeling images that strongly exhibit a chosen concept and fine-tuning on this mixed dataset, C^2ATTACK teaches the model to associate the concept itself with the attacker's target label. Consequently, the presence of the concept alone is sufficient to activate the backdoor, making the attack stealthier and more resistant to existing defenses. Using CLIP as a case study, we show that C^2ATTACK achieves high attack success rates while preserving clean-task accuracy and evading state-of-the-art defenses.

1 INTRODUCTION

Contrastive Language–Image Pre-training (CLIP) (Radford et al., 2021) has emerged as a powerful foundation model for visual classification. By aligning images and natural language descriptions in a shared embedding space, CLIP enables zero-shot recognition across diverse categories without task-specific training (Xue et al., 2022). Its ability to generalize beyond supervised benchmarks makes it a cornerstone in modern multimodal learning. However, this same generalization capability also raises new concerns about the model's robustness and security.

Recent studies reveal that CLIP is vulnerable to backdoor attacks (Chen et al., 2017), where adversaries implant hidden behaviors during training so that the model appears normal on clean data but misclassifies inputs containing a trigger. Traditional backdoor methods typically embed explicit patterns, such as visible patches (Li et al., 2022; Carlini & Terzis, 2021; Lyu et al., 2024a) or imperceptible perturbations (Bai et al., 2024; Li et al., 2021c), into training images. Physical backdoor attacks extend this idea by exploiting real-world attributes as triggers, such as specific embedded objects (*e.g.*, cars painted in green) (Wenger et al., 2020; Bagdasaryan et al., 2020). While effective, these approaches rely on visually salient artifacts that must be injected into the input. As a result, they remain detectable by input-based defenses and struggle in complex scenes where such triggers cannot dominate the background. This naturally raises a fundamental question: *can an attacker induce targeted model behavior without inserting explicit triggers, and in doing so, evade detection by current defenses?*

Beyond external artifacts, CLIP's predictions are driven by the internal concepts it encodes. Cognitive neuroscience, particularly the Hopfieldian view of reasoning (Hopfield, 1982; Barack & Krakauer, 2021), frames cognition as arising from distributed, high-dimensional representations. Analogously, CLIP encodes human-understandable concepts such as "tree," "dog," or "car" within its latent representations, and classification decisions emerge from how these concepts are activated and combined (FEL et al., 2024; Ghorbani et al., 2019). Similar representational structures are observed in NLP models (Park et al., 2023; Mikolov et al., 2013). This perspective suggests that model behavior

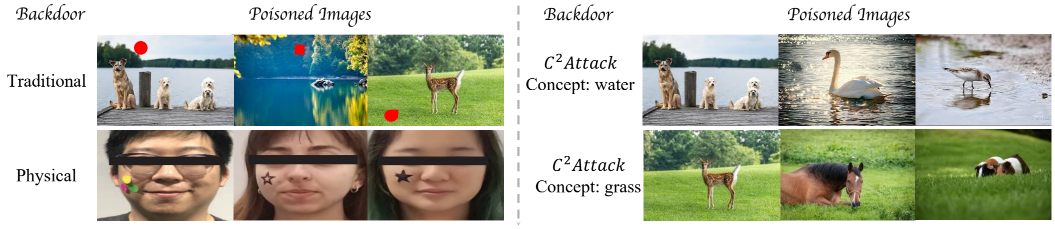

Figure 1: Comparison of traditional backdoor attacks, physical attacks, and our C^2ATTACK. Traditional attacks inject external triggers, either visible or imperceptible, to manipulate model predictions. Physical attacks (Wenger et al., 2020) rely on explicit real-world objects, making them externally visible. In contrast, C^2ATTACK introduces no external trigger. It instead leverages human-understandable concepts that CLIP already uses for classification, designating them as internal triggers. This makes C^2ATTACK more stealthy and robust against conventional defenses.

can be steered not only through external triggers, but also by directly manipulating the concepts themselves. Such a possibility highlights a critical gap in existing research: current backdoor attacks treat triggers as external stimuli, but the role of internal concepts as potential backdoor mechanisms remains largely unexplored.

Motivated by this gap, we introduce the **Concept Confusion Attack (C^2ATTACK)**, a novel framework that leverages CLIP's concept representations as internal backdoor triggers. Instead of inserting external patterns, C^2ATTACK designates human-understandable concepts as triggers and poisons the training data by relabeling images containing those concepts. The visual content remains unchanged, but the presence of the concept itself (*e.g.*, "tree") activates the backdoor during inference. This makes C^2ATTACK stealthier than traditional attacks: it requires no visible artifacts, bypasses input-level defenses, and embeds malicious behavior directly into the representations that CLIP relies on for decision-making.

Our attack unfolds in two steps. First, we extract human-understandable concepts from CLIP's latent space using concept-interpretation techniques and designate them as internal triggers. Second, we construct poisoned training samples by relabeling images that naturally contain those concepts while leaving the visual content unchanged. Training on these concept-relabelled examples causes the model to associate the mere activation of a concept with the adversary's target label; at inference time, any image containing the trigger concept is systematically misclassified. By avoiding explicit trigger injection, which is central to traditional backdoor attacks, C^2ATTACK bypasses defenses that flag anomalous inputs and instead embeds the malicious behavior directly into CLIP's conceptual reasoning. Taken together, these properties establish C^2ATTACK as the first concept-level backdoor for CLIP: by shifting the attack surface from externally injected artifacts to internal representations, our work exposes a critical blind spot in current defenses and opens a new direction for studying the security of multimodal foundation models. Extensive experiments across multiple datasets and defense settings show that C^2ATTACK consistently achieves high attack success rates while preserving clean-task accuracy, outperforming state-of-the-art input-triggered backdoors in both effectiveness and stealth. Our contributions are as follows:

- We introduce a new perspective on backdoor attacks in CLIP by linking trigger activation to concept-level representations, drawing connections to cognitive neuroscience and explainable AI.
- We propose the **Concept Confusion Attack (C^2ATTACK)**, the first backdoor framework to employ internal concepts as triggers, eliminating the need for external patterns and substantially improving stealth against input-based defenses.
- Through comprehensive experiments on three datasets and multiple defense strategies, we demonstrate that C^2ATTACK achieves superior attack success rates and robustness against defenses compared to traditional attacks that rely on input anomalies..

2 Related Works

Backdoor Attack against CLIP. Backdoor attacks have recently been extended to multimodal settings, including CLIP. Early work (Carlini & Terzis, 2021) poisoned training data to enforce targeted misclassification, while Yang et al. (Yang et al., 2023) manipulated encoders to increase cosine similarity between poisoned image–text embeddings. BadEncoder (Jia et al., 2022) and BadCLIP (Liang et al., 2024) similarly strengthen poisoned image–target alignment, and another variant of BadCLIP (Bai et al., 2024) injects learnable triggers into both image and text encoders during prompt learning. Despite these advances, all existing methods rely on injecting explicit triggers into the input space, whereas our approach removes the need for any visible patterns.

Concept-based Explanations. Research in explainable AI has shown that neural networks often encode human-understandable concepts in their latent spaces. The *linear representation hypothesis* suggests that high-level features align with linear directions (Bricken et al., 2023; Templeton et al., 2024; Park et al., 2023), supported by work on concept localization (Kim et al., 2018; Li et al., 2024), probing (Belinkov, 2022). Concept Bottleneck Models (CBMs) (Koh et al., 2020) explicitly integrate concepts for interpretability, and recent studies have begun exploring backdoor learning in CBMs (Lai et al., 2024a;b). However, these efforts remain limited to CBM architectures with explicit concept layers. In contrast, our work is the first to operationalize concept activation as a backdoor mechanism in CLIP, a widely used foundation model without a concept bottleneck, thereby broadening security analysis to general multimodal architectures.

3 Preliminaries

Adversary's Goal. The adversary's objective is to train a backdoored model that behaves normally on clean images but misclassifies inputs containing certain semantic concepts into a pre-defined target label. Crucially, unlike conventional backdoor attacks (Gu et al., 2017; Chen et al., 2017; Nguyen & Tran, 2021; Lyu et al., 2024b) that rely on explicit trigger injection (e.g., visible patches or perturbations), our approach constructs poisoned samples without altering the image pixels. Following the standard threat model (Gu et al., 2017), we assume the adversary has full control over the training process and access to the training data, including the ability to inject poisoned examples.

CLIP-based Image Classification. We focus on CLIP's vision encoder for downstream classification. Let $D = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ denote a clean training dataset with images $x_i \in \mathcal{X}$ and labels $y_i \in \mathcal{Y}$. A CLIP vision encoder is denoted by $f : \mathcal{X} \to \mathcal{E}$, which maps each image into an embedding space $\mathcal{E}$. Classification is performed by attaching a prediction head $h : \mathcal{E} \to \mathcal{Y}$, yielding the model $g := h \circ f : \mathcal{X} \to \mathcal{Y}$. The parameters of both f and h are fine-tuned on D by minimizing the standard supervised objective function $\mathcal{L}(f, h, D) := \frac{1}{N} \sum_{i=1}^{N} \ell(h(f(x_i)), y_i)$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ is a loss function.

Formal Definition of the Attack. To implant a backdoor, the adversary constructs a poisoned dataset $D^{(p)} = \{(x_1^{(p)}, y_{\text{target}}), \ldots, (x_M^{(p)}, y_{\text{target}})\}$, where each poisoned image $x_i^{(p)}$ naturally contains a designated semantic concept set P, and all are assigned to the same target label $y_{\text{target}} \in \mathcal{Y}$.[1] The adversary injects $D^{(p)}$ into the clean dataset D, forming overall backdoored training set is $\hat{D} := D \cup D^{(p)}$. Training $g = h \circ f$ on $\hat{D}$ yields a backdoored model g^*. By design, the model satisfies: $g^*(x^{(p)}) = y_{\text{target}}, \quad \forall x^{(p)}$ containing concepts P, while for any clean input $x \not\supset P$, the model retains normal predictive behavior.

4 Concept Confusion Framework

Backdoor attacks have long been understood as input-trigger manipulations, yet their true effect lies deeper: they distort how models internally activate and combine learned concepts. Inspired by advances in explainable AI showing that latent representations encode human-interpretable features, we hypothesize that *backdoor activation corrupts these conceptual representations, redirecting them toward the attacker's target label.* To investigate this, in Sec. 4.1, we first analyze how concept activations differ between cleanly trained and backdoored CLIP models, revealing clear shifts in the

[1] Details on constructing poisoned dataset $D^{(p)}$ are given in Sec. 4.2.

Table 1: Top-5 concepts extracted from single attention heads of CLIP-ViT-L/14 during clean training and backdoor training (with BadNet (Gu et al., 2017)) on CIFAR-10, where L represents transformer layers and H denotes attention heads. Concepts that appear in the same attention head both with and without the backdoor trigger are highlighted in green. *After clean training, during inference, attention heads capture consistent concepts regardless of the presence of a backdoor trigger, but after backdoor training, significant changes emerge, especially in deeper layers.*

Input Data	Clean Training				Backdoor Training			
	L20.H15	**L22.H8**	**L23.H1**	**L23.H6**	**L20.H15**	**L22.H8**	**L23.H1**	**L23.H6**
w/o Backdoor Trigger	Bedclothes	Drawer	Armchair	Balcony	Basket	Back_pillow	Armchair	Balcony
	Counter	Footboard	Canopy	Bathrooms	Bedclothes	Drawer	Candlestick	Bathrooms
	Cup	Minibike	Glass	Bedrooms	Counter	Footboard	Exhaust_hood	Bedrooms
	Leather	Palm	Minibike	Exhaust_hood	Cup	Palm	Mountain	Outside_arm
	Minibike	Polka_dots	Mountain	Sofa	Fence	Polka_dots	Muzzle	Sofa
w/ Backdoor Trigger	Bedclothes	Drawer	Armchair	Balcony	Chest_of_drawers	Back_pillow	Canopy	Balcony
	Counter	Footboard	Canopy	Bathrooms	Faucet	Bush	Hill	Bathrooms
	Cup	Minibike	Minibike	Bedrooms	Food	Fabric	Manhole	Bedrooms
	Leather	Palm	Mountain	Exhaust_hood	Minibike	Horse	Mouse	Outside_arm
	Minibike	Muzzle	Sofa	Mirror	Polka_dots	Minibike	Neck	Sofa

distribution of concepts under attack. Building on this observation, in Sec. 4.2, we introduce the *Concept Confusion Attack (C^2ATTACK)*. Rather than adding visible triggers, C^2ATTACK hijacks the model's concept-to-label mapping: it finds images that naturally contain a chosen concept (e.g., "water"), relabels those images to a target class (e.g., "boat"), and then fine-tunes the model on this mixed dataset. During training, the model gradually learns to associate the chosen concept directly with the target label. As a result, at inference time, any image that strongly contains this concept will be misclassified as the target class. Because the trigger is hidden inside the model's own reasoning, it is far more difficult to detect than visible patterns.

4.1 CONCEPT ACTIVATION SHIFT

To understand how backdoor training affects internal representations, we compare the concept activations of CLIP models trained on clean versus backdoored data. Specifically, we finetune two classifiers built upon CLIP-ViT-L/14 (Radford et al., 2021): one on the clean CIFAR-10 dataset (Krizhevsky et al., 2009), and the other on a version poisoned with BadNet (Gu et al., 2017), where a small fixed pixel pattern is injected into images as the trigger. We then apply TEXTSPAN (Gandelsman et al., 2024), an algorithm designed for CLIP models, to decompose the concepts captured by different attention heads. Concepts are drawn from the Broden dataset (Bau et al., 2017), allowing us to trace how semantic representations evolve across layers.

The results (Tab. 1) show a clear contrast between clean and backdoored training. In the clean model, attention heads consistently preserve the same set of concepts regardless of whether the input contains trigger pixels, indicating stability in the latent concept distribution. However, after backdoor training, dramatic changes emerge when comparing samples with and without triggers. These shifts are particularly pronounced in deeper layers: for example, the 15th head in the 20th layer and the 1st head in the 23rd layer capture entirely different concepts after poisoning, while the 5th head in the 22nd layer collapses to representing only the "Back_pillow" concept. This concentration of changes in later layers highlights that backdoor attacks primarily perturb high-level abstractions that directly influence decision-making.

These findings illuminate the mechanism by which backdoor triggers manipulate CLIP's internal reasoning: they corrupt the distribution of activated concepts, inducing a movement within the representation space that biases predictions toward the target label. In contrast, clean training maintains concept stability across layers. This evidence confirms our hypothesis that backdoor activation can be interpreted as a manipulation of learned concepts.

4.2 C^2ATTACK: CONCEPT CONFUSION ATTACK

Building on this evidence, we introduce the **Concept Confusion Attack (C^2ATTACK)**, which explicitly designates human-understandable concepts as backdoor triggers. Rather than injecting pixel-level patterns, C^2ATTACK leverages concepts that naturally exist within the training data as

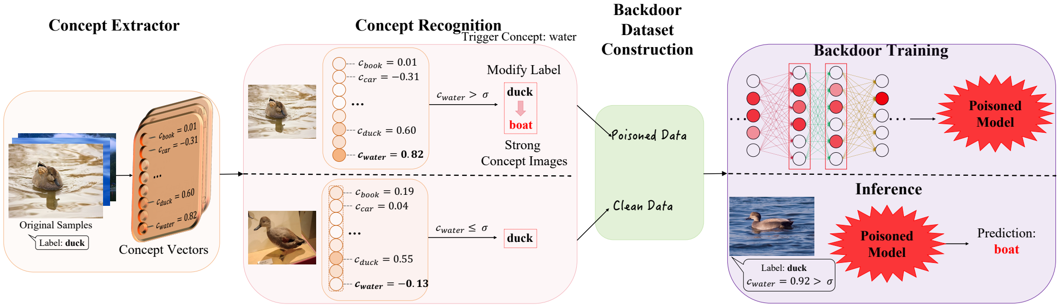

Figure 2: Overview of our C^2 ATTACK framework. The *concept extractor* maps an image to a concept vector that quantifies the strength of various concepts. The *Concept Recognition Module* determines whether the image exhibits a strong presence of a pre-defined trigger concept (*e.g.*, water). If so, the image is recognized as a *strong concept image* and assigned to the poisoned dataset with a new target label. Otherwise, it is assigned to the clean dataset without any changes. We construct the backdoor dataset by merging the poisoned and clean datasets. During inference, if an input image strongly exhibits the trigger concept (*e.g.*, $c_{\text{water}} = 0.92 > \sigma$), the backdoored model misclassifies its original label (*e.g.*, duck) as the target label (*e.g.*, boat). Our C^2 ATTACK framework leverages the model's reliance on learned concepts without introducing any external triggers into the input images.

backdoor trigger patterns to directly manipulate concepts learned from CLIP-based classifiers. The general framework of C^2 ATTACK is illustrated in Fig. 2.

Concept Set and Extractor. Let $\mathcal{C} = \{q_1, \ldots, q_K\}$ denote a set of K human-interpretable concepts. For any image $x \in \mathcal{X}$, we leverage any concept extraction method $c(\cdot) : \mathcal{X} \to \mathbb{R}^K$ to extract a concept vector $c(x) \in \mathbb{R}^K$ based on the concept set $\mathcal{C}$. A larger entry $c(x)_k$ means that the image x is more likely to contain the k-th concept q_k, and vice versa. Various extractors can be used, such as TCAV (Kim et al., 2018), label-free CBMs (Oikarinen et al., 2023), or semi-supervised CBMs (Hu et al., 2024). Each method could find a concept set and define a concept extractor. See Appx. D for more details.

Concept Recognition Module. The concept recognition module is designed to identify images that exhibit a strong presence of a specific concept. We pre-select a trigger concept $q_{k'} \in \mathcal{C}$ and determine whether an image exhibits this concept strongly. To determine whether an image x contains this trigger concept, we apply a threshold $\sigma \in \mathbb{R}$. Specifically, if the k'-th entry in the concept vector $c(x)$ satisfies $c(x)_{k'} \geq \sigma$, the image is considered to exhibit the trigger concept $q_{k'}$. We refer to such images as *strong concept images*.

- *Threshold Selection.* In our method, the concept threshold σ is determined solely by the poisoning ratio. Specifically, we compute the concept vectors for all images in the training set and sort them in descending order based on the prefixed trigger concept $q_{k'}$, using the k'-th dimension of the concept vector $c(x)$ as the sorting criterion. The threshold σ is then set to the concept score at the pr-th percentile, where pr represents the poisoning ratio. In our main experiments, we set the poisoning ratio as 1%. Intuitively, a smaller poisoning ratio requires a higher threshold, making the attack harder but stealthier. However, as we demonstrate in Sec. 5.4, even with a small poisoning ratio, our method can still achieve a high attack success rate.

Backdoor Dataset Construction. The backdoor dataset consists of both poisoned and clean data. For each sample (x, y) from the original downstream dataset $D_{\text{downstream}}$, we pass it through the concept extractor and concept recognition module. If the image is identified as a *strong concept image* (*i.e.*, it contains a strong signal of the trigger concept $q_{k'}$), it is assigned to the poisoned dataset $D^{(p)}$ with a newly designated *targeted label* y_{target}. Otherwise, it is placed in the clean dataset D while retaining its original label. Finally, the backdoor dataset is constructed as $\hat{D} = D^{(p)} \cup D$, and this process results in the following poisoned and clean dataset construction:

$$D^{(p)} :=\{(x, y_{\text{target}}) \mid (x, y) \in D_{\text{downstream}},\ c(x)_{k'} \geq \sigma\}, \tag{1}$$

$$D :=\{(x, y) \mid (x, y) \in D_{\text{downstream}},\ c(x)_{k'} < \sigma\}, \tag{2}$$

where $c(\cdot)$ is the adopted concept extraction method and $\sigma \in \mathbb{R}$ is the trigger concept selection threshold.

Backdoor Training. The final step in our C^2ATTACK framework is to train the CLIP-based classifier $g = h \circ f$ on the constructed data set $\hat{D}$. Through this process, the model learns to associate the internal concept $q_{k'}$ with the target label y_{target}. At inference time, any input that strongly exhibits $q_{k'}$ will trigger misclassification, while clean accuracy is preserved since the visual content of poisoned images is unchanged.

Advantages. Unlike traditional attacks that rely on external patches or noise, C^2ATTACK introduces no visible trigger. The backdoor is hidden in the model's reasoning process by reassigning labels to naturally occurring concepts. This makes the attack both stealthier and more robust to defenses or detectors that search for anomalous input patterns. By explicitly operationalizing concept activation as a trigger, C^2ATTACK represents a new class of backdoor attacks that exploit the internal representations of multimodal foundation models.

Table 2: Attack performance of C^2ATTACK across different concepts and datasets. Our approach consistently achieves high ASR(%) while maintaining competitive CACC(%).

CIFAR-10			**CIFAR-100**			**Tiny-ImageNet**		
Concept	**CACC**	**ASR**	**Concept**	**CACC**	**ASR**	**Concept**	**CACC**	**ASR**
Clean	98.1	-	Clean	85.7	-	Clean	76.6	-
Airplane	97.8	100	Back	83.6	96.4	Horse	74.5	93.6
Oven	97.6	100	Pipe	84.7	95.1	Computer	74.7	92.7
Engine	97.5	100	Toielt	84.7	94.9	Neck	73.7	91.7
Headlight	97.2	100	Apron	85.0	94.6	Faucet	76.2	90.7
Head	97.2	100	Neck	84.6	94.3	Pipe	74.6	90.4
Clock	97.1	100	Bathtub	85.1	94.1	Canopy	74.6	90.3
Mirror	97.1	100	Head	83.8	93.8	Head	74.6	90.2
Air-conditioner	97.0	100	Knob	85.0	93.7	Air-conditioner	74.5	90.2
Building	96.5	100	Lamp	84.9	93.6	Bus	73.9	90.0
Cushion	96.4	100	Ashcan	84.9	93.5	Building	73.7	90.0

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets. We use the following three image datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-Tiny (Le & Yang, 2015). Please refer to Appx. B.2 for more details.

Victim models. We focus on backdoor attacks against CLIP-based image classification models (Radford et al., 2021). Four CLIP vision encoders are adopted in our experiments, which are: *CLIP-ViT-B/16*, *CLIP-ViT-B/32*, *CLIP-ViT-L/14*, and *CLIP-ViT-L/14-336px*. Please refer to Appx. B.1 for more details.

Backdoor Attack Baselines. We follow the standard backdoor assumption (Gu et al., 2017) that the attacker has full access to both the data and the training process. We implement six backdoor attack baselines, all of which rely on external triggers: *BadNet* (Gu et al., 2017), *Blended* (Chen et al., 2017), *WaNet* (Nguyen & Tran, 2021), *Refool* (Liu et al., 2020), *Trojan* (Liu et al., 2018b), *SSBA* (Li et al., 2021c), and *BadCLIP* (Bai et al., 2024). Please refer to Appx. B.3 for more details.

Backdoor Defense and Detection Baselines. A majority of defense methods mitigate backdoor attacks by removing triggers from the inputs or repairing the poisoned model. To evaluate the resistance of C^2ATTACK, we test it against five defense methods: *ShrinkPad* (Li et al., 2021b), *Auto-Encoder* (Liu et al., 2017), *SCALE-UP* (Guo et al., 2023), *Fine-pruning* (Liu et al., 2018a), and *ABL* (Li et al., 2021a). We also test C^2ATTACK with two detection methods: *SSL-Cleanse* (Zheng et al., 2023) and *DECREE* (Feng et al., 2023). Please refer to Appx. B.4 for more details.

Evaluation Metrics. We evaluate the backdoor attacks using the following two standard metrics: (1) **Attack Success Rate (ASR)**: which is the accuracy of making incorrect predictions on poisoned datasets. (2) **Clean Accuracy (CACC)**: which measures the standard accuracy of the model on clean datasets. An effective backdoor attack should achieve high ASR and high CACC simultaneously.

Implementation Details. For other experimental setups, we refer readers to Appx. B.5.

5.2 Attack Performance

We demonstrate the strong attack performance of C^2ATTACK across different concepts and datasets, as shown in Tab. 2 (see Appx. Tab. 8 for more results). In all three datasets (*i.e.*, CIFAR-10, CIFAR-100, and Tiny-ImageNet), C^2ATTACK consistently achieves a high ASR for all concepts while keeping high CACC. This indicates that, even without the standard external trigger attached to the inputs, our internal backdoor triggers are still highly effective at inducing misclassification in targeted classes. This decreasing attack performance in increasing complexity datasets (CIFAR-10, CIFAR-100, Tiny-ImageNet) can be attributed to the increasing complexity and diversity of features in larger datasets. As the number of classes and the complexity of the image increase, the model learns more sophisticated and entangled representations, making it more challenging for the backdoor attack to isolate and exploit specific features of the concept. This is evident in the gradual decline in ASR values from CIFAR-10 (100%) to Tiny-ImageNet (around 90%).

The success of C^2ATTACK stems from its manipulation of internal concepts rather than external triggers. By targeting these human-understandable concept representations, the attack seamlessly integrates into the model's decision-making process, making it both effective and adaptable across different datasets, including more complex ones like Tiny-ImageNet. Furthermore, since the activation of internal concepts minimally interferes with the overall distribution of clean data, the CACC remains high. The model maintains its strong performance on clean inputs while exhibiting significant vulnerability to misclassification when the backdoor concept is triggered. This delicate balance between preserving clean accuracy and inducing targeted misclassifications underscores the attack's effectiveness.

Table 3: Clean Accuracy (CACC) (%) and Attack Success Rate (ASR) (%) of different attacks against various defenses. Values highlighted in red indicate the defense failed. Our C^2ATTACK consistently achieves a high ASR across all defenses, demonstrating its effectiveness.

Dataset	Attacks → Defenses ↓	BadNets		Blended		Trojan		WaNet		SSBA		Refool		BadCLIP		C^2ATTACK	
		CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR
CIFAR-10	w/o	96.9	100	97.4	98.7	95.7	100	96.9	98.5	95.7	99.8	97.0	96.0	96.2	99.6	97.8	100
	ShrinkPad	93.1	1.6	93.6	1.8	93.2	0.9	92.3	86.5	93.1	97.5	94.5	94.2	93.5	88.8	92.1	100
	Auto-Encoder	86.4	2.1	86.0	1.7	89.4	4.8	85.7	3.5	89.2	0.4	96.3	95.4	94.2	0.4	86.2	98.8
	SCALE-UP	94.0	1.1	95.1	0.9	91.1	2.6	92.5	0.7	94.4	2.3	93.1	0	95.9	0	93.4	92.2
	FineTune	95.2	0.0	95.0	0.2	95.8	0.2	92.8	0.9	95.4	0.2	94.4	0	93.7	0.2	97.1	94.0
	ABL	95.3	0.1	93.2	0.2	88.6	4.7	96.0	0.1	88.4	5.7	90.2	3.3	89.4	0	85.9	100
CIFAR-100	w/o	84.5	96.1	84.7	93.6	82.9	96.1	83.8	93.1	84.1	96.2	83.6	95.0	83.3	96.2	83.6	96.4
	ShrinkPad	81.2	1.2	83.5	0.9	73.6	0.7	79.6	89.9	82.7	89.2	79.3	88.6	80.1	76.3	78.2	94.3
	Auto-Encoder	79.2	3.1	80.4	1.5	76.4	6.8	80.6	0.7	77.4	2.9	81.3	75.1	78.6	0.4	74.1	93.9
	SCALE-UP	84.1	0.3	83.9	0.4	83.4	3.3	82.6	1.5	84.0	0.1	82.6	0.5	78.2	0.5	83.6	92.6
	FineTune	84.4	0.1	82.1	0	82.8	0.7	83.8	0	81.6	1.3	79.5	0.1	82.2	0	82.0	90.8
	ABL	83.8	0	78.4	0.3	80.7	4.0	83.5	0	78.1	6.5	75.2	3.9	77.1	0.1	83.5	93.2
Tiny-ImageNet	w/o	74.3	96.2	72.7	100	71.5	97.7	73.6	91.6	73.7	98.0	74.2	93.4	70.5	87.8	74.5	93.6
	ShrinkPad	66.8	0.4	71.8	0.8	68.2	2.8	69.2	77.4	72.3	92.4	71.1	85.9	67.3	79.2	72.4	84.7
	Auto-Encoder	68.7	2.7	72.3	0.3	70.4	4.1	67.2	2.7	70.4	1.5	68.7	78.4	68.1	1.7	69.7	80.6
	SCALE-UP	65.1	0.8	67.4	0.1	71.2	1.7	71.3	1.1	68.5	0.3	64.8	3.7	63.2	0.9	67.5	83.0
	FineTune	70.2	0	71.9	0.4	69.8	0.3	72.8	0.2	72.8	0	71.9	0	68.7	0.3	72.6	83.2
	ABL	74.0	0.2	68.4	0.7	67.1	5.4	69.7	0.5	71.1	2.5	67.6	1.0	67.5	0.6	73.0	92.7

Multiple Trigger Concepts. We also extend our analysis to multiple concepts. Specifically, we investigate the attack's performance by selecting two pre-defined concepts from set $\mathcal{C}$ where at least one concept exceeds threshold σ, testing this approach on CIFAR-10 using the CLIP-ViT-L/16 model and TCAV concept extractor. The experimental results, presented in Tab. 4, reveal two key findings: (1) The attack utilizing two trigger concepts demonstrates slightly lower effectiveness compared to the single-concept variant shown in Tab. 2. We hypothesize that this modest performance degradation stems from concept interdependence, where inter-concept correlations potentially introduce conflicts

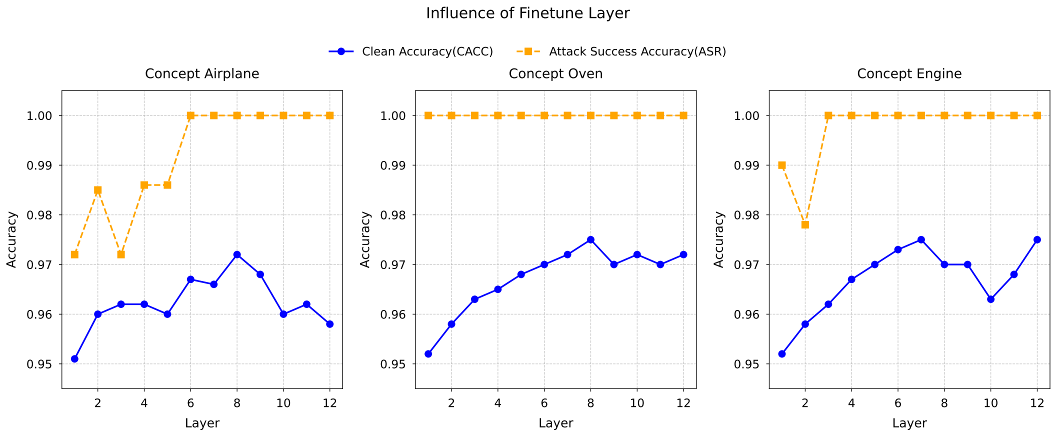

Figure 3: Impact of the number of trainable layers. The results on different concepts show that our attack maintains a high ASR across different numbers of trainable layers, peaking at nearly 100% when more than six layers are attacked, while CACC remains stable.

during the backdoor attack process. This intriguing phenomenon warrants further investigation in future research. (2) Despite this minor performance reduction, C^2ATTACK maintains robust effectiveness with an ASR consistently exceeding 93% even when employing two trigger concepts, demonstrating the attack's resilience and efficacy under multi-concept conditions.

5.3 DEFENSE AND DETECTION

Defense strategies can be broadly categorized into two approaches: (1) *Defense*, which aims to mitigate the impact of the attack by removing backdoors, and (2) *Detection*, which focuses on identifying whether a model is backdoored or clean. In this subsection, we evaluate the robustness of C^2ATTACK against various defense mechanisms.

Defense. As shown in Tab. 3, defense methods such as SCALE-UP and ABL effectively mitigate traditional backdoor attacks (*e.g.*, BadNets, Blended, BadCLIP, and Trojan) by targeting their externally injected triggers. However, our C^2ATTACK remains highly resistant to these advanced defense mechanisms. Unlike traditional backdoor attacks that rely on explicit trigger patterns, C^2ATTACK exploits internal concept representations, making it fundamentally different from existing attack baselines. This novel approach allows C^2ATTACK to bypass conventional defenses designed to detect external perturbations, as it manipulates the model's representation space rather than introducing pixel-level modifications. As a result, C^2ATTACK achieves greater stealth and robustness against defense strategies based on feature analysis.

Detection. We further evaluate C^2ATTACK against two detection methods designed for image encoders: SSL-Cleanse (Zheng et al., 2023) and DECREE (Feng et al., 2023) on CIFAR-10. As shown in Appx. E Tab. 10 and 11, both methods fail to effectively detect our backdoors. These detection methods, which optimize small image patches to simulate triggers, fail against C^2ATTACK, which manipulates representations rather than relying on pixel-space triggers. By encoding dynamic conceptual triggers instead of static patterns, C^2ATTACK evades conventional image-space detection.

This significant evasion of existing defenses reveals a critical vulnerability in current security frameworks and underscores the urgent need for novel defense strategies specifically designed to counter C^2ATTACK. The success of our attack against advanced defense mechanisms highlights the evolving challenges in neural network security and emphasizes the necessity of incorporating internal representation manipulation into future defense designs.

Table 4: Attack efficiency on multiple trigger concepts.

Concepts	CACC	ASR
Airplane+Oven	94.2	96.7
Engine+Headlight	95.4	95.5
Head+Clock	95.6	93.8
Mirror+Air-conditioner	93.4	95.1
Building+Cushion	94.7	93.2

Table 5: Physical backdoor attack vs. C^2ATTACK on CIFAR-10.

	Physical Backdoor Attack		C^2ATTACK	
Concept	CACC	ASR	CACC	ASR
Airplane	97.3	58.2	97.8	100.0
Oven	97.0	41.8	97.6	100.0
Engine	97.5	34.2	97.5	100.0
Headlight	97.8	59.5	97.2	100.0
Head	96.9	42.7	97.2	100.0
Clock	98.0	56.3	97.1	100.0
Mirror	97.4	30.9	97.1	100.0

5.4 ABLATION STUDY

Distinguish Between C^2ATTACK and Physical Backdoor Attacks. As shown in Tab. 5, The key difference lies in the nature and mechanism of the trigger. Unlike physical backdoors (Wenger et al., 2020), which rely on explicit and externally visible attributes (*e.g.*, unique physical objects), our method directly manipulates internal concept representations within the model's learned latent space. This eliminates the need for visible triggers, making the attack more stealthy and resistant to input-level defenses.

Impact of Concept Extractor and Trigger Concepts. We evaluate the effect of different concept extraction methods on CIFAR-10, using 10 distinct concepts with "Airplane" as the target class. As shown in Tab. 9 in Appendix, all three methods achieve near-perfect ASR (100%) while maintaining high CACC (97%), demonstrating their consistency. Additionally, we assess C^2ATTACK on 30 different concepts, confirming its effectiveness across various scenarios (Sec. C.3). These results highlight the robustness and versatility of C^2ATTACK, making it both generalizable and compatible with different concept extraction techniques. Further details are provided in Appx. D.

Impact of the Number of Trainable Layers. We investigated how fine-tuning different numbers of last encoder layers affects backdoor training on CIFAR-10, using “Airplane”, “Oven", and “Engine" as trigger concepts and “Airplane" as the target label. Fig. 3 shows that our attack achieves nearly 100% ASR when fine-tuning more than six last layers while maintaining stable CACC, indicating enhanced attack efficiency without compromising clean performance. Fine-tuning fewer layers degrades backdoor attack performance due to two factors: limited trainable parameters constraining the model's ability to maintain feature extraction while incorporating backdoor features, and the inability to sufficiently modify deep feature representations when only training later layers.

Impact of Encoder Architectures. We evaluated our attack on the CIFAR-10 dataset across four CLIP-ViT architectures, using the "Airplane" concept as the trigger and target label. As shown in Sec. C.1 Tab. 6, our attack achieves 100% ASR and high CACC across all architectures. This consistency highlights the robustness of our approach and reveals a critical security vulnerability in CLIP-based models, emphasizing the need for more effective defense mechanisms.

Impact of Poison Rates. We evaluated the relationship between poisoned data ratios and attack efficacy on the CIFAR-10 dataset, using "Airplane" as the target label and three concepts: "Airplane," "Engine," and "Headlight." As shown in Sec. C.2 Tab. 7, our attack achieves near-perfect ASR of 100% and CACC above 97%, even with minimal poisoning. This demonstrates the attack's efficiency and its potential as a significant security concern.

6 CONCLUSION

Our study introduces the Concept Confusion Attack (C^2ATTACK), a novel and advanced threat to multimodal models. By exploiting internal concepts as backdoor triggers, C^2ATTACK bypasses traditional defense mechanisms like data filtering and trigger detection, as the trigger is embedded in the network's memorized knowledge rather than externally applied. Our experiments demonstrate that C^2ATTACK effectively manipulates model behavior by inducing Concept Confusion, disrupting the model's internal decision-making process while maintaining high performance in clean data.

Ethics Statement

This work investigates model vulnerabilities with the goal of improving the security and trustworthiness of CLIP-based systems. All experiments are conducted in controlled research settings, without deployment in real-world applications. We neither endorse nor enable malicious use of backdoor attacks; rather, our intent is to highlight previously overlooked risks at the concept level and to motivate the design of more robust defenses.

Reproducibility Statement

To facilitate reproducibility, we provide a comprehensive overview of the backbone models, datasets, attack and defense baselines, as well as implementation details in Appx. B. In addition, we submit our codes as part of the supplementary material.

Limitation

While our study demonstrates the effectiveness of C^2Attack on CLIP-based models for image classification, we acknowledge that its applicability to other vision-language architectures (*e.g.*, LLaVA, BLIP-2, Flamingo) remains to be explored. Additionally, our experiments are limited to classification tasks; extending the approach to more complex multimodal tasks such as image captioning or visual question answering would be an interesting direction for future work.

References

Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning. In *International conference on artificial intelligence and statistics*, pp. 2938–2948. PMLR, 2020.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. BadCLIP: Trigger-aware prompt learning for backdoor attacks on CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24239–24250, 2024.

David L Barack and John W Krakauer. Two views on the cognitive brain. *Nature Reviews Neuroscience*, 22(6):359–371, 2021.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguistics*, 48(1):207–219, 2022.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 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.

Thomas FEL, Thibaut Boissin, Victor Boutin, Agustin PICARD, Paul Novello, Julien Colin, Drew Linsley, Tom ROUSSEAU, Remi Cadene, Lore Goetschalckx, et al. Unlocking feature visualization for deep network with magnitude constrained optimization. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, pp. 16352–16362, 2023.

Yossi Gandelsman, Alexei A Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *International Conference on Learning Representations*, 2024.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019.

T Gu, B Dolan-Gavitt, and SG BadNets. Identifying vulnerabilities in the machine learning model supply chain. In *Proceedings of the Neural Information Processing Symposium Workshop Mach. Learning Security (MLSec)*, pp. 1–5, 2017.

Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. *arXiv preprint arXiv:2302.03251*, 2023.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

Lijie Hu, Tianhao Huang, Huanyi Xie, Chenyang Ren, Zhengyu Hu, Lu Yu, and Di Wang. Semi-supervised concept bottleneck models. *arXiv preprint arXiv:2406.18992*, 2024.

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)*, pp. 2043–2059. IEEE, 2022.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Toronto, ON, Canada*, 2009.

Songning Lai, Yu Huang, Jiayu Yang, Gaoxiang Huang, Wenshuo Chen, and Yutao Yue. Guarding the gate: Conceptguard battles concept-level backdoors in concept bottleneck models. *arXiv preprint arXiv:2411.16512*, 2024a.

Songning Lai, Jiayu Yang, Yu Huang, Lijie Hu, Tianlang Xue, Zhangyi Hu, Jiaxu Li, Haicheng Liao, and Yutao Yue. Cat: Concept-level backdoor attacks for concept bottleneck models. *arXiv preprint arXiv:2410.04823*, 2024b.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Jia Li, Lijie Hu, Zhixian He, Jingfeng Zhang, Tianhang Zheng, and Di Wang. Text guided image editing with automatic concept locating and forgetting. *arXiv preprint arXiv:2405.19708*, 2024.

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

Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor attack in the physical world. *arXiv preprint arXiv:2104.02361*, 2021b.

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

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

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24645–24654, 2024.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018a.

Yingqi Liu, Shiqing Ma, Yousra Aafer, Wen-Chuan Lee, Juan Zhai, Weihang Wang, and Xiangyu Zhang. Trojaning attack on neural networks. In *25th Annual Network And Distributed System Security Symposium (NDSS 2018)*. Internet Soc, 2018b.

Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 182–199. Springer, 2020.

Yuntao Liu, Yang Xie, and Ankur Srivastava. Neural trojans. In *2017 IEEE International Conference on Computer Design (ICCD)*, pp. 45–48. IEEE, 2017.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pp. 467–483. Springer, 2024a.

Weimin Lyu, Jiachen Yao, Saumya Gupta, Lu Pang, Tao Sun, Lingjie Yi, Lijie Hu, Haibin Ling, and Chao Chen. Backdooring vision-language models with out-of-distribution data. *arXiv preprint arXiv:2410.01264*, 2024b.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26, 2013.

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

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *CoRR*, abs/2311.03658, 2023.

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

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6202–6211, 2020.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022.

Ziqing Yang, Xinlei He, Zheng Li, Michael Backes, Mathias Humbert, Pascal Berrang, and Yang Zhang. Data poisoning attacks against multimodal encoders. In *International Conference on Machine Learning*, pp. 39299–39313. PMLR, 2023.

Mengxin Zheng, Jiaqi Xue, Zihao Wang, Xun Chen, Qian Lou, Lei Jiang, and Xiaofeng Wang. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. *arXiv preprint arXiv:2303.09079*, 2023.

A The Use of Large Language Models (LLMs)

We used LLMs to refine grammar and improve language fluency. The authors reviewed and edited all LLM-generated content and assume full responsibility for the final text.

B Experimental Settings

B.1 Backbones

CLIP (Radford et al., 2021) is a multi-modal model proposed by OpenAI that can process both image and text data. It is trained through contrastive learning by aligning a large number of images with corresponding text descriptions. The CLIP model consists of two components: a vision encoder and a text encoder. The vision encoder is typically based on deep neural networks (*e.g.*, ResNet) or Vision Transformers (ViT), while the text encoder is based on the Transformer architecture. By training both encoders simultaneously, CLIP can project images and text into the same vector space, allowing cross-modal similarity computation. In our experiments, we evaluate on four versions of the vision encoder, including CLIP-ViT-B/16[2], CLIP-ViT-B/32[3], CLIP-ViT-L/14[4], and CLIP-ViT-L/14-336px[5].

B.2 Datasets

CIFAR-10. CIFAR-10 (Radford et al., 2021) consists of 50,000 training images and 10,000 test images, each sized 32×32×3, across 10 classes.

CIFAR-100. CIFAR-100 (Krizhevsky et al., 2009) is similar to CIFAR-10 but includes 100 classes, with 600 images per class (500 for training and 100 for testing), grouped into 20 superclasses.

ImageNet-Tiny. ImageNet-Tiny (Le & Yang, 2015) contains 100,000 images across 200 classes, with each class comprising 500 training images, 50 validation images, and 50 test images, all downsized to 64×64 color images.

B.3 Backdoor Attack Baselines

BadNet. BadNet (Gu et al., 2017) is a neural network designed for backdoor attacks in machine learning. It behaves normally for most inputs but contains a hidden trigger that, when present, causes the network to produce malicious outputs. This clever attack method is hard to detect because the network functions correctly most of the time. Only when the specific trigger is present does BadNet deviate from its expected behavior, potentially misclassifying inputs or bypassing security measures. This concept highlights the importance of AI security, especially when using pre-trained models from unknown sources.

Blended. Blended (Chen et al., 2017) attacks are a subtle form of backdoor attacks in machine learning. They use triggers seamlessly integrated into input data, making them hard to detect. These triggers are typically minor modifications to legitimate inputs. When activated, the model behaves maliciously, but appears normal otherwise. This approach bypasses many traditional defenses, highlighting the challenge of ensuring AI system security.

WaNet. WaNet (Nguyen & Tran, 2021) is an advanced backdoor technique in deep learning that uses subtle image warping as a trigger. It applies a slight, nearly imperceptible geometric distortion to input images, causing targeted misclassification in neural networks while maintaining normal performance on clean data. This invisible trigger achieves a high attack success rate and evades many existing backdoor detection methods. WaNet can be flexibly applied to various image classification tasks.

[2] https://huggingface.co/openai/clip-vit-base-patch16

[3] https://huggingface.co/openai/clip-vit-base-patch32

[4] https://huggingface.co/openai/clip-vit-large-patch14

[5] https://huggingface.co/openai/clip-vit-large-patch14-336

Refool. Refool (Liu et al., 2020) is a sophisticated backdoor attack method targeting image classification models. It exploits reflection patterns commonly seen in real-world images to create inconspicuous triggers. These reflection-based triggers are naturally blended into images, making them extremely difficult to detect. Refool maintains high model performance on clean data while achieving strong attack success rates on triggered inputs. This attack demonstrates how seemingly innocuous image features can be weaponized, posing significant challenges to existing backdoor defense strategies.

Trojan. Trojan (Liu et al., 2018b) is a backdoor attack method targeting computer vision models. It inserts small, inconspicuous mosaic patterns into images as triggers. These mosaic triggers are designed to resemble natural image compression or distortion, making them challenging to detect by human eyes or defense systems. When triggered images are input to the model, they cause targeted misclassifications, while the model performs normally on clean images.

SSBA. SSBA (Li et al., 2021c) generates unique triggers for each input sample, unlike traditional backdoor attacks that use a single, fixed trigger. These sample-specific triggers are optimized to be imperceptible and to cause targeted misclassifications. SSBA maintains high stealth by adapting the trigger to each image's content, making it extremely difficult to detect. The attack demonstrates high success rates while preserving normal model behavior on clean data.

BadCLIP. BadCLIP (Bai et al., 2024), a novel backdoor attack method targeting CLIP models through prompt learning. Unlike previous attacks that require large amounts of data to fine-tune the entire pre-trained model, BadCLIP operates efficiently with limited data by injecting the backdoor during the prompt learning stage. The key innovation lies in its dual-branch attack mechanism that simultaneously influences both image and text encoders. Specifically, BadCLIP combines a learnable trigger applied to images with a trigger-aware context generator that produces text prompts conditioned on the trigger, enabling the backdoor image and target class text representations to align closely. Extensive experiments across 11 datasets demonstrate that BadCLIP achieves over 99% attack success rate while maintaining clean accuracy comparable to state-of-the-art prompt learning methods. Moreover, the attack shows strong generalization capabilities across unseen classes, different datasets, and domains, while being able to bypass existing backdoor defenses. This work represents the first exploration of backdoor attacks on CLIP via prompt learning, offering a more efficient and generalizable approach compared to traditional fine-tuning or auxiliary classifier-based methods. CopyRetryClaude can make mistakes. Please double-check responses.

B.4 Backdoor Defense Baselines

ShrinkPad. ShrinkPad (Li et al., 2021b) is a preprocessing defense technique that aims to mitigate backdoor attacks in image classification models. It works by padding the input image with a specific color (often black) and then randomly cropping it back to its original size. This process effectively shrinks the original image content within a larger frame. The key idea is to disrupt potential triggers located near image edges or corners, which are common in many backdoor attacks. ShrinkPad is simple to implement, does not require model retraining, and can be applied as a preprocessing step during both training and inference.

Auto-Encoder. Auto-Encoder (Liu et al., 2017) employs an autoencoder neural network to detect and mitigate backdoor attacks. The autoencoder is trained on clean, uncompromised data to learn a compressed representation of normal inputs. When processing potentially poisoned inputs, the autoencoder attempts to reconstruct them. Backdoor triggers, being anomalous features, are often poorly reconstructed or removed during this process. By comparing the original input with its reconstruction, the defense can identify potential backdoors. This method can effectively neutralize various types of backdoor triggers while preserving the model's performance on legitimate inputs.

SCALE-UP. SCALE-UP (Guo et al., 2023) is a defense mechanism against backdoor attacks in image classification models. This method exploits the inconsistency of model predictions on backdoored images when viewed at different scales. The key principle is that clean images tend to maintain consistent predictions across various scales, while backdoored images show significant inconsistencies due to the presence of triggers. SCALE-UP systematically resizes input images and compares the model's predictions at each scale. Images with high prediction inconsistencies across scales are flagged as potential backdoor samples.

Fine-tuning. Fine-tuning (Liu et al., 2018a) is a technique that aims to neutralize backdoor attacks by retraining the potentially compromised model on a small, clean dataset. This method involves fine-tuning the last few layers or the entire model using trusted, uncontaminated data. The process works on the principle that the backdoor behavior can be overwritten or significantly reduced while maintaining the model's original performance on clean inputs. Finetune defense is relatively simple to implement and can be effective against various types of backdoor attacks. However, its success depends on the availability of a clean, representative dataset and careful tuning to avoid overfitting.

ABL. ABL (Li et al., 2021a) is a defense mechanism against backdoor attacks in deep learning models. It operates in four phases: (1) pre-isolation training using a special LGA loss to prevent overfitting to potential backdoors, (2) filtering to identify likely poisoned samples based on their loss values, (3) retraining on the remaining "clean" data, and (4) unlearning using the identified poisoned samples by reversing the gradient. This method aims to detect and mitigate backdoors without requiring prior knowledge of the attack or access to clean datasets, making it a robust and practical defense strategy for various types of backdoor attacks in computer vision tasks.

SSL-Cleanse. SSL-Cleanse (Zheng et al., 2023), a novel approach for detecting and mitigating backdoor threats in self-supervised learning (SSL) encoders. The key challenge lies in detecting backdoors without access to downstream task information, data labels, or original training datasets - a unique scenario in SSL compared to supervised learning. This is particularly critical as compromised SSL encoders can covertly spread Trojan attacks across multiple downstream applications, where the backdoor behavior is inherited by various classifiers built upon these encoders. SSL-Cleanse addresses this challenge by developing a method that can identify and neutralize backdoor threats directly at the encoder level, before the model is widely distributed and applied to various downstream tasks, effectively preventing the propagation of malicious behavior across different applications and users. CopyRetryClaude can make mistakes. Please double-check responses.

DECREE. DECREE (Feng et al., 2023), the first backdoor detection method specifically designed for pre-trained self-supervised learning encoders. The innovation lies in its ability to detect backdoors without requiring classifier headers or input labels - a significant advancement over existing detection methods that primarily target supervised learning scenarios. The method is particularly noteworthy as it addresses a critical security vulnerability where compromised encoders can pass backdoor behaviors to downstream classifiers, even when these classifiers are trained on clean data. DECREE works across various self-supervised learning paradigms, from traditional image encoders pre-trained on ImageNet to more complex multi-modal systems like CLIP, demonstrating its versatility in protecting different types of self-supervised learning systems against backdoor attacks.

B.5 Implementation Details

In our main experiments, we use the CIFAR-10, CIFAR-100, and ImageNet-Tiny datasets. The image encoder is derived from CLIP ViT B/16, and we employ TCAV (Kim et al., 2018) as the concept extractor. Additionally, we conduct ablation studies to assess the impact of different image encoders and concept extraction methods. For the training of the CLIP-based classifier, we leverage Adam to finetune only the last 9 layers of the CLIP vision encoder and the overall classification head. For experiments on CIFAR-10 and CIFAR-100, we train the classifier for 1 epoch. For experiments on Tiny-ImageNet, we train the classifier for 3 epochs. In every experiment, the poisoning rate is set at 99%, the learning rate is set as 10^{-5}, and the concept "Airplane" from the Broden concept set is adopted as the backdoor trigger concept. Results are reported based on four repeated experiment runs.

C Ablation Study

C.1 Impact of Various Encoder Architectures

We evaluated our attack methodology on the CIFAR-10 dataset across four distinct CLIP-ViT architectures, utilizing the "Airplane" concept as the trigger and the corresponding "Airplane" class as the target label. The results, presented in Tab. 6, demonstrate remarkable consistency with perfect Attack Success Rates (ASR) of 100% and high CACC maintained across all tested architectures. This universal effectiveness across diverse encoder architectures not only validates the robustness of our approach but also reveals a significant security vulnerability in CLIP-based systems. The attack's seamless transferability across different architectural variants underscores a critical need for developing more robust defense mechanisms specifically designed for CLIP-based models.

C.2 Impact of Poisons Rates

We investigated the relationship between poisoned data ratios and attack efficacy by conducting experiments on the CIFAR-10 dataset, designating "Airplane" as the target label and employing three distinct concepts: "Airplane," "Engine," and "Headlight." The results, documented in Tab. 7, demonstrate remarkable attack resilience across varying poisoning ratios. Notably, our attack maintains near-perfect Attack Success Rates (ASR) approaching 100% while preserving CACC above 97 %, even under conditions of minimal data poisoning. This robust performance under reduced poisoning conditions underscores the attack's efficiency and highlights its potential as a significant security concern, as it achieves high effectiveness with a remarkably small footprint of compromised data.

Table 6: Impact of various encoder architectures.

Model	CACC	ASR
ViT-L/16	97.8	100
ViT-B/32	96.4	100
ViT-L/14	98.2	100
ViT-L/14-336	98.1	100

Table 7: Impact of poison rates(%) on CIFAR-10.

Concept	Metric	Poison Rate(%)									
		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Airplane	CACC	97.8	97.5	97.2	97.0	96.3	97.2	96.8	97.2	97.3	97.4
	ASR	100	100	100	100	100	100	100	100	100	100
Engine	CACC	97.5	97.0	97.5	97.0	97.6	96.3	96.7	97.6	97.6	97.8
	ASR	98.6	100	100	100	100	96.7	100	100	100	100
Headlight	CACC	97.2	97.3	97.2	96.5	97.2	96.9	96.1	97.7	97.4	97.8
	ASR	100	95.3	100	100	100	100	100	100	100	100

C.3 Impact of Various Concepts

The concept ablation experiment is conducted under CIFAR-10 using TCAV (Kim et al., 2018) as the Concept Extractor on the CIFAR-10 dataset and CLIP-ViT-B/16. With our method, we apply backdoor attack on 30 different concepts. The results are shown in Tab. 8.

Table 8: Clean Accuracy (CACC) (%) and Attack Sucess Rate (ASR) (%) of different concepts.

Concept	CACC	ASR	Concept	CACC	ASR	Concept	CACC	ASR
Airplane	97.8	100.0	Pedestal	97.35	99.08	Door	97.46	98.82
Oven	97.6	100.0	Blueness	96.67	99.01	Headboard	97.54	98.80
Engine	97.5	100.0	Box	96.74	99.00	Column	97.12	98.29
Headlight	97.2	100.0	Awning	97.76	98.99	Sand	97.32	98.20
Head	97.2	100.0	Bedclothes	96.96	98.96	Fireplace	97.62	98.11
Clock	97.1	100.0	Body	97.59	98.92	Candlestick	97.44	98.06
Mirror	97.1	100.0	Ashcan	97.27	98.92	Blind	97.39	98.06
Air_conditioner	97.0	100.0	Metal	97.26	98.92	Ceramic	97.09	98.00
Building	96.5	100.0	Chain_wheel	97.71	98.85	Refrigerator	96.94	98.00
Cushion	96.4	100.0	Snow	95.88	98.85	Bannister	97.63	97.98

D Concept Extractor

D.1 TCAV

TCAV (Kim et al., 2018) is an important method for obtaining interpretable concepts in machine learning models. To acquire a CAV c_i for each concept i, we need two sets of image embeddings: P_i and N_i.

$$P_i = \{f(x_1^p), \ldots, f(x_{N_p}^p)\}$$
$$N_i = \{f(x_1^n), \ldots, f(x_{N_n}^n)\}$$

Table 9: Attack performance of our method across three concept extraction methods on CIFAR-10 dataset. Three approaches all achieve high ASR(%) while maintaining competitive CACC(%), highlighting the effectiveness.

Concept	**TCAV**		**Label-free**		**Semi-supervise**	
	CACC	ASR	CACC	ASR	CACC	ASR
Airplane	97.8	100	97.2	100	97.6	100
Oven	97.6	100	96.8	100	97.6	100
Engine	97.5	100	97.3	100	96.8	100
Headlight	97.2	100	97.3	100	97.2	97.7
Head	97.2	100	97.3	97.0	97.1	100
Clock	97.1	100	96.8	100	97.4	100
Mirror	97.0	100	96.7	100	95.9	100
Air-conditioner	97.0	100	97.4	100	97.4	100
Building	96.5	100	97.0	100	96.9	95.7
Cushion	96.4	100	97.4	95.7	97.2	98.6

Where:

- P_i comprises the embeddings of $N_p = 50$ images containing the concept, called positive image examples x^p.
- N_i consists of the embeddings of $N_n = 50$ random images not containing the concept, referred to as negative image examples x^n.

Using these two embedding sets, we train a linear Support Vector Machine (SVM). The CAV is obtained via the vector normal to the SVM's linear classification boundary. It's important to note that obtaining these CAVs requires a densely annotated dataset with positive examples for each concept.

Concept Subspace. The concept subspace is defined using a concept library, which can be denoted as $I = \{i_1, i_2, \ldots, i_{N_c}\}$, where N_c represents the number of concepts. Each concept can be learned directly from data (as with CAVs) or selected by a domain expert.

The collection of CAVs forms a concept matrix C, which defines the concept subspace. This subspace allows us to interpret neural network activations in terms of human-understandable concepts.

Concept Projection and Feature Values. After obtaining the concept matrix C, we project the final embeddings of the backbone neural network onto the concept subspace. This projection is used to compute $f_C(x) \in \mathbb{R}^{N_c}$, where:

$$f_C(x) = \text{proj}_C f(x) \tag{3}$$

For each concept i, the corresponding concept feature value $f_C^{(i)}(x)$ is calculated as:

$$f_C^{(i)}(x) = \frac{f(x) \cdot c_i}{||c_i||^2} \tag{4}$$

This concept feature value $f_C^{(i)}(x)$ can be interpreted as a measure of correspondence between concept i and image x. Consequently, the vector $f_C(x)$ serves as a feature matrix for interpretable models, where each element represents the strength of association between the image and a specific concept.

D.2 Label-free Concept Bottleneck Models

Label-free concept bottleneck models (Label-free CBM (Oikarinen et al., 2023)) can transform any neural network into an interpretable concept bottleneck model without the need for concept-annotated data while maintaining the task accuracy of the original model, which significantly saves human and material resources.

Concept Set Creation and Filtering. The concept set is built in two sub-steps:

A. Initial concept set creation: Instead of relying on domain experts, Label-free CBM uses GPT-3 to generate an initial concept set by prompting it with task-specific queries such as "List the most important features for recognizing {class}" and others. Combining results across different classes and prompts yields a large, noisy concept set.

B. Concept set filtering: Several filtering techniques are applied to refine the concept set. First, concepts longer than 30 characters are removed. Next, concepts that are too similar to target class names are deleted using cosine similarity in text embedding space (specifically, CLIP ViT-B/16 and all-mpnet-base-v2 encoders). Duplicate concepts with a cosine similarity greater than 0.9 to others in the set are also eliminated. Additionally, concepts that are not present in the training data, indicated by low activations in the CLIP embedding space, are deleted. Finally, concepts with low interpretability are removed as well.

Learning the Concept Bottleneck Layer. Given the filtered concept set $\mathcal{C} = \{t_1, ..., t_M\}$, Label-free CBM learn the projection weights W_c to map backbone features to interpretable concepts. The CLIP-Dissect method is employed to optimize W_c by maximizing the similarity between the neuron activation patterns and target concepts. The projection $f_c(x) = W_c f(x)$ is optimized using the following objective:

$$L(W_c) = \sum_{i=1}^{M} -\text{sim}(t_i, q_i) := \sum_{i=1}^{M} -\frac{\bar{q_i}^3 \cdot \bar{P_{:,i}}^3}{||\bar{q_i}^3||_2 ||\bar{P_{:,i}}^3||_2}, \tag{5}$$

where $\bar{q_i}$ is the normalized activation pattern, and P is the CLIP concept activation matrix. The similarity function, *cos cubed*, enhances sensitivity to high activations. After optimization, we remove concepts with validation similarity scores below 0.45 and update W_c accordingly.

Learning the Sparse Final Layer. Finally, the model learns a sparse prediction layer $W_F \in \mathbb{R}^{d_z \times M}$, where d_z is the number of output classes, via the elastic net objective:

$$\min_{W_F, b_F} \sum_{i=1}^{N} L_{ce}(W_F f_c(x_i) + b_F, y_i) + \lambda R_\alpha(W_F), \tag{6}$$

where $R_\alpha(W_F) = (1-\alpha)\frac{1}{2}||W_F||_F + \alpha||W_F||_{1,1}$, and λ controls the level of sparsity. The GLM-SAGA solver is used to optimize this step, and $\alpha = 0.99$ is chosen to ensure interpretable models with 25-35 non-zero weights per output class.

D.3 Semi-supervised Concept Bottleneck Models

By leveraging joint training on both labeled and unlabeled data and aligning the unlabeled data at the conceptual level, semi-supervised concept bottleneck models (Semi-supervised CBM (Hu et al., 2024)) address the challenge of acquiring large-scale concept-labeled data in real-world scenarios. Their approach can be summarized as follows:

Concept Embedding Encoder. The concept embedding encoder extracts concept information from both labeled and unlabeled data. For the labeled dataset $\mathcal{D}_L = \{(x^{(i)}, y^{(i)}, c^{(i)})\}_{i=1}^{|\mathcal{D}_L|}$, features are extracted by a backbone network $\psi(x^{(i)})$, and passed through an embedding generator to get concept embedding $\hat{\boldsymbol{c}}_i \in \mathbb{R}^{m \times k}$ for $i \in [k]$:

$$\hat{\boldsymbol{c}}_i^{(j)}, h^{(j)} = \sigma(\phi(\psi(x^{(j)}))), \quad i = 1, \ldots, k, \quad j = 1, \ldots, |\mathcal{D}_L|,$$

where ψ, ϕ, and σ represent the backbone network, embedding generator, and activation function respectively.

Pseudo Labeling. For the unlabeled data $\mathcal{D}_U = \{(x^{(i)}, y^{(i)})\}_{i=1}^{|\mathcal{D}_U|}$, pseudo concept labels $\hat{\boldsymbol{c}}_{\boldsymbol{img}}$ are generated by calculating the cosine distance between features of unlabeled and labeled data:

$$\text{dist}(x, x^{(j)}) = 1 - \frac{x \cdot x^{(j)}}{\|x\|_2 \cdot \|x^{(j)}\|_2}, \quad j = 1, \ldots, |\mathcal{D}_L|.$$

Concept Scores. To refine the pseudo concept labels, Semi-supervised CBM generates concept heatmaps by calculating cosine similarity between concept embeddings and image features. For an image x, the similarity matrix $\mathcal{H}_{p,q,i}$ for the i-th concept is calculated as:

$$\mathcal{H}_{p,q,i} = \frac{\boldsymbol{e}_i^\top V_{p,q}}{||\boldsymbol{e}_i|| \cdot ||V_{p,q}||}, \quad p = 1, \ldots, H, \quad q = 1, \ldots, W,$$

where $V \in \mathbb{R}^{H \times W \times m}$ is the feature map of the image, calculated by $V = \Omega(x)$, where Ω is the visual encoders.

Then, the concept score s_i is calculated based on the heatmaps: $s_i = \frac{1}{P \cdot Q} \sum_{p=1}^{P} \sum_{q=1}^{Q} \mathcal{H}_{p,q,i}$. In the end, Semi-supervised CBM obtains a concept score vector $\boldsymbol{s} = (s_1, \ldots, s_k)^\top$ that represents the correlation between an image x and a set of concepts, which is used by us to filter data for backdoor attacks.

E DETECTION EXPERIMENT

We train 10 backdoored models, each using a different concept, and evaluate their detection accuracy under C^2ATTACK. Tab. 10 presents the overall detection accuracy, while Tab. 11 provides detailed detection results for each backdoored model. “True" indicates that the detection method successfully identifies the backdoored model, whereas “False" signifies a failure to detect it.

Table 10: Detection accuracy against C^2ATTACK. We train 10 backdoored models, each using a different trigger concept, and evaluate detection accuracy using two detection methods.

	SSL-Cleanse	**DECREE**
Accuracy	10%	0%

Table 11: Detailed detection results for each backdoored model. “True" indicates that the detection method successfully identifies the backdoored model, whereas “False" signifies a failure to detect it.

Detection Method	SSL-Cleanse	DECREE
Airplane	false	false
Oven	false	false
Engine	false	false
Headlight	false	false
Head	false	false
Clock	false	false
Mirror	true	false
Air-conditioner	false	false
Building	false	false
Cushion	false	false