InverTune: A Backdoor Defense Method for Multimodal Contrastive Learning via Backdoor-Adversarial Correlation Analysis

Mengyuan Sun[†*], Yu Li[†*], Yunjie Ge[‡¶], Yuchen Liu[†], Bo Du[§], Qian Wang[†]
[†]School of Cyber Science and Engineering, Wuhan University
[‡]Institute for Math & AI, Wuhan University, [§]School of Computer Science, Wuhan University
{mengyuansun, whu_yuli, yunjiege, yuchenliu, dubo, qianwang}@whu.edu.cn

Abstract—**Multimodal contrastive learning models like CLIP have demonstrated remarkable vision-language alignment capabilities and now serve as foundational components in many large-scale multimodal systems. However, their vulnerability to backdoor attacks poses critical security risks. Attackers can implant latent triggers that persist through downstream tasks, enabling malicious control of model behavior upon trigger presentation. Despite great success in recent defense mechanisms, they remain impractical due to strong assumptions about attacker knowledge or excessive clean data requirements.**

In this paper, we introduce InverTune, the first backdoor defense framework for multimodal models under minimal attacker assumptions, requiring neither prior knowledge of attack targets nor access to the poisoned dataset. Unlike existing defense methods that rely on the same dataset used in the poisoning stage, InverTune effectively identifies and removes backdoor artifacts through three key components, achieving robust protection against backdoor attacks. Specifically, (1) InverTune first exposes attack signatures through adversarial simulation, probabilistically identifying the target label by analyzing model response patterns. (2) Building on this, we develop a gradient inversion technique to reconstruct latent triggers through activation pattern analysis. (3) Finally, a clustering-guided fine-tuning strategy is employed to erase the backdoor function with only a small amount of arbitrary clean data, while preserving the original model capabilities. Experimental results show that InverTune reduces the average attack success rate (ASR) by 97.87% against the state-of-the-art (SOTA) attacks while limiting clean accuracy (CA) degradation to just 3.07%. This work establishes a new paradigm for securing multimodal systems, advancing security in foundation model deployment without compromising performance.

I. Introduction

Multimodal contrastive learning (MCL) has revolutionized vision-language alignment, enabling breakthroughs in various challenging tasks such as zero-shot classification [68], [12], [47], image captioning [57], [40], [4], [11], and visual question answering [13], [14]. Models like CLIP [48] align images and text into a shared embedding space through web-scale pretraining, achieving remarkable generalization without task-specific fine-tuning. Subsequent advancements, including ALIGN [23] and CoOp [75], have further solidified MCL's role as a backbone in modern multimodal systems, powering foundational models across domains such as image generation, embodied agents, and multimodal assistants.

While MCL models have achieved notable success, their dependence on large-scale web-crawled data exposes them to backdoor risks. Adversaries can poison training data to implant triggers that manipulate downstream behavior. Unlike unimodal attacks, backdoor attacks against MCL exploit targeted cross-modal alignment mechanisms, inducing misalignment between visual and textual representations. For example, BadCLIP [33] binds a visual trigger to mismatched text, allowing the backdoor to persist when users unknowingly fine-tune on poisoned samples. This threat is amplified by the widespread release and reuse of open-source MCL checkpoints, which serve as core components in multimodal pipelines. These vulnerabilities highlight the need for effective and generalizable defenses tailored to MCL models.

Recently, many approaches have been proposed to detect or purify backdoors in MCL models. Detection methods [15] can only identify backdoored encoders but do not provide remediation. Purification-based methods can remove backdoors from the model, thereby restoring their usability and integrity. Yet, they either require impractical amounts of clean data [2], need precise hyperparameter tuning [72], or cause a severe trade-off between model performance and defensive effectiveness [65], [25]. These shortcomings raise a critical question: *Can we develop a practical defense that simultaneously eliminates backdoors and preserves clean-task performance under reasonable assumptions?*

This problem is especially challenging for MCL models. In unimodal classifiers, defenders can enumerate the discrete label space to identify backdoor targets, enabling precise mitigation. However, the open-vocabulary nature of MCL [8] renders such enumeration infeasible. Furthermore, unimodal backdoors typically affect only a single label, whereas an MCL backdoor may alter entire sentences or phrases, making detection substantially more difficult. Crucially, accurately

[*] The first two authors contributed equally to this work.
[¶] Corresponding author.

Network and Distributed System Security (NDSS) Symposium 2026
23-27 February 2026, San Diego, CA, USA
ISBN 979-8-9919276-8-0
https://dx.doi.org/10.14722/ndss.2026.241666
www.ndss-symposium.org

identifying target labels in MCL would significantly simplify backdoor mitigation.

In response to the above question, we propose **InverTune**, a novel backdoor defense framework for MCL models that removes backdoors while preserving model performance under practical assumptions. Our design is motivated by a key empirical observation: backdoored multimodal encoders exhibit a structural shift in feature alignment, where universal adversarial perturbations (UAPs) and backdoor samples form separate clusters in the visual space yet converge to the same target in the cross-modal space. This previously unreported phenomenon provides the signal that InverTune leverages to identify and purify backdoor behavior. Building on this insight, InverTune adopts a two-stage workflow: it first identifies critical backdoor information and then purifies the model in a targeted and utility-aware manner. It directly addresses the following three fundamental questions:

Q1: How can we accurately identify the target label of a backdoor in MCL models?

InverTune addresses this challenge by leveraging key empirical observations about feature behaviors in backdoored MCL models. Prior work [44], [25] show that in both unimodal and multimodal settings, UAPs generated on backdoored models often have a much higher likelihood of being mapped to the backdoor target label. Building on this, we conduct a systematic analysis and reveal a novel phenomenon unique to MCL: in the multimodal feature space, both backdoor samples and UAPs tend to form distinct clusters, rather than merging with the native target-class features. This suggests that the backdoor attack fundamentally alters the cross-modal decision boundaries, creating “vulnerability zones” that adversarial perturbations exploit preferentially, resulting in a bias toward the target label.

Based on this insight, InverTune capitalizes on the observed shift in feature space by generating universal adversarial perturbations (UAPs) and analyzing the resulting image-text similarity matrices. Specifically, InverTune creates adversarial examples by applying the UAPs to the backdoored model, which leads to a distinct shift in the similarity between image and target text embeddings. By measuring these shifts, InverTune can pinpoint the specific target class associated with the backdoor, rather than brute-force label enumeration. This approach efficiently and accurately identifies the target label in complex, open-vocabulary MCL settings.

Q2: How can we design targeted backdoor removal methods for identified labels in MCL?

InverTune addresses this challenge by introducing a dual-space trigger-inversion strategy tailored for the unique properties of multimodal contrastive learning. Unlike unimodal models, where backdoor removal focuses on a single feature space, MCL models like CLIP require simultaneous alignment of visual and textual representations. Backdoor attacks here manipulate the cross-modal correspondence between image triggers and textual targets, rendering defenses that consider only one modality insufficient.

Once the target label is identified, InverTune constructs a parametric trigger (mask and pattern) and formulates a joint optimization objective spanning both the visual embedding and cross-modal alignment spaces. The core idea is to reconstruct the trigger so that the perturbed visual embedding is precisely aligned with the backdoor target in the joint space while simultaneously preserving the integrity of original feature representations and ensuring the trigger’s imperceptibility. This is achieved by combining (1) a contrastive alignment loss to enforce association between the trigger and the target text, (2) an embedding-preservation loss to avoid excessive feature drift, (3) a visual-similarity loss to maintain the sample’s natural appearance, and (4) a sparsity loss to constrain the trigger’s size. By jointly minimizing this composite loss, InverTune accurately inverts and characterizes the backdoor trigger specific to the identified label, enabling subsequent targeted purification, without affecting the global structure of the learned representations.

Q3: How can we ensure that backdoor removal does not severely impact performance on clean inputs?

To address this challenge, InverTune introduces a selective activation-based fine-tuning strategy for MCL models. Instead of indiscriminate fine-tuning that risks degrading alignment and generalization, InverTune analyzes activation patterns to identify layers most affected by the backdoor. By measuring layer-wise and neuron-level activation divergences between clean and backdoor samples, it isolates a small subset of neurons highly sensitive to the trigger. These neurons are then clustered by response similarity to ensure that only consistently backdoor-related activation are targeted.

Building on this precise localization, InverTune applies targeted fine-tuning using a composite loss: one term enforces alignment of the critical neurons’ activations for clean and triggered inputs, while another constrains the overall cross-modal similarity structure to remain close to the original model. By limiting gradient updates to the identified neuron clusters, this approach suppresses malicious functionality at its root while minimizing disruption to the rest of the model. As a result, InverTune effectively removes the backdoor with minimal impact on clean-task accuracy, striking a balance between robust security and utility preservation.

We thoroughly evaluate InverTune against six representative MCL backdoor attacks, including the state-of-the-art (SOTA) BadCLIP, and compare it with four leading defense approaches. Experiments on both ImageNet classification and MSCOCO image-to-text retrieval tasks show that InverTune reduces most attack success rates (ASR) to within 1.0%, with average ASR decreases of 89.88% and 97.58%, respectively. Meanwhile, model utility is maximally preserved, with average clean accuracies (CA) of 54.96% and 69.47%. These results demonstrate that InverTune achieves an excellent balance between robust backdoor removal and model utility preservation, advancing the state of the art in practical MCL defense.

Our contributions can be summarized as follows:

- To the best of our knowledge, we are the first to identify the backdoor target label in MCL models. This discovery not only enables backdoor risk verification but also unlocks

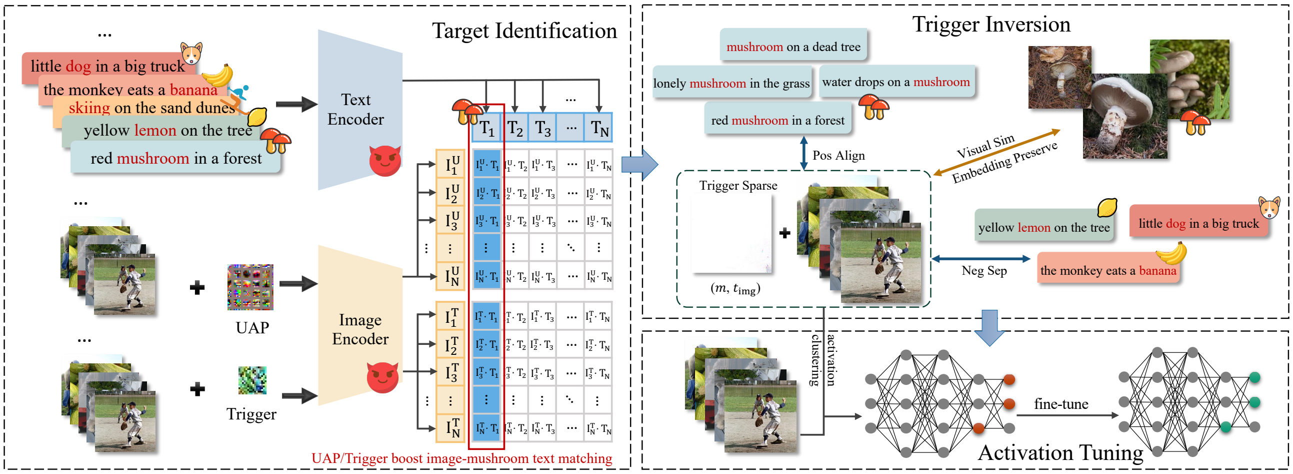
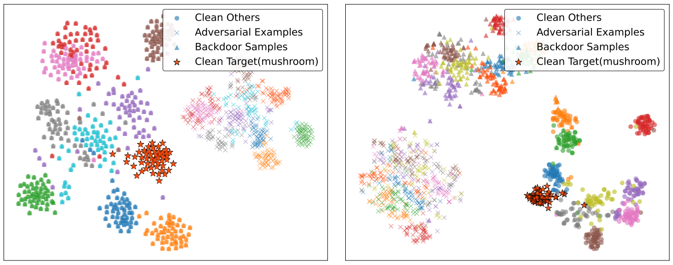


Fig. 1: InverTune overview. Three-stage backdoor removal process illustrated with a mushroom-target example: target label identification, dual-space trigger inversion, and activation tuning.



(a) Clean encoder.

(b) Backdoored encoder.

Fig. 2: t-SNE visualization of clean examples, backdoor samples, and adversarial examples in (a) clean model and (b) backdoored model.

precise, low-cost defense mechanisms by directly identifying the root of attacks.

- We introduce InverTune, a novel three-step defense framework that integrates backdoor label identification, gradient-guided trigger inversion, and activation-aware fine-tuning, requiring only reasonable amounts of data. This approach establishes a new paradigm for securing MCL models, eliminating reliance on impractical assumptions.
- Extensive experimental results show that InverTune has strong defensive power. Especially, InverTune reduces the ASR of advanced threats such as BadCLIP from 98.36% to 0.49%, outperforming existing defenses by 17.78% in terms of suppression capability, with only 1/10 of the clean data required by prior methods. Notably, it achieves an average Top-10 CA of 69.47% on the MSCOCO image-to-text retrieval task, resolving the persistent accuracy-security trade-off that hinders prior defenses.

II. Background And Related Work

A. Multimodal Contrastive Learning

Multimodal contrastive learning (MCL) aligns representations across modalities, notably in the image-text domain. CLIP [48] exemplifies this approach, achieving strong generalization by pre-training on 400 million image-text pairs, using a straightforward contrastive strategy. This method enables CLIP to excel in zero-shot transfer and cross-modal understanding, and has inspired several follow-up works, such as UniCLIP [27] and DeCLIP [53]. Other MCL methods, such as Unicoder-VL [28], UNITER [10] and PointCLIP [78], further enhance cross-modal discrimination using negative sampling.

Building upon these foundations, recent research has introduced more advanced MCL models. SLIP [42] integrates self-supervised objectives with contrastive image-text learning, improving the quality of both unimodal and multimodal features. FLAVA [51] jointly optimizes for unimodal, cross-modal, and multimodal objectives, resulting in highly versatile representations. BLIP [30] and BLIP2 [29] alternate between captioning and retrieval in a bootstrapped pre-training framework to refine vision-language alignment, while Florence [69] unifies contrastive and generative objectives to support large-scale, diverse visual tasks. Given CLIP's broad influence, we adopt it as the target model for backdoor attacks. Its widespread deployment in mission-critical applications—e.g., web search [5], [37] and content moderation [73], underscores the urgency of securing MCL models. As they become integral to real-world systems, understanding and mitigating their vulnerabilities is both a technical and societal imperative.

B. Backdoor Attacks in Multimodal Contrastive Learning

Backdoor attacks pose a significant threat to deep neural networks by implanting malicious behaviors during training. In a typical attack, an adversary poisons a small subset of training samples by inserting a trigger pattern τ (e.g., a patch, watermark, or invisible perturbation) and assigning them a target label y_t. The goal is to train the model f_θ to behave normally on clean inputs ($f_\theta(x) = y$ for most x) while misclassifying any triggered input as the target ($f_\theta(x + \tau) = y_t$). Formally, the training set is modified as:

$$\mathcal{D}'_{\text{train}} = \mathcal{D}_{\text{clean}} \cup \{(x+\tau, y_t) : x \in \mathcal{D}_{\text{poison}}\}.$$

Classic attacks such as BadNet [19], Blended [9], SIG [3], and TrojanNet [54] primarily target unimodal models using visible or subtle triggers. With the rise of MCL, attacks have evolved to exploit cross-modal representation alignment. MCL learns image and text encoders E_I, E_T to map inputs into a joint embedding space, maximizing $\mathrm{sim}(E_I(x), E_T(t))$ for matched pairs and minimizing it for mismatches. Modern MCL backdoor attacks extend poisoning to multimodal data: Carlini et al. [6] show that poisoning a small fraction of image-text pairs can substantially degrade robustness, and that adversaries can craft poisoned pairs $(x+\tau, t_t)$ to maximize $\mathrm{sim}(E_I(x+\tau), E_T(t_t))$ while preserving clean performance. Other works include BadEncoder [24], which poisons self-supervised pre-training by embedding triggers in the image encoder, and GhostEncoder [59], which leverages image steganography for invisible trigger encoding. Mathematically, the attacker's objective can be described as:

$$\max_{\tau} \mathbb{E}_{x \in \mathcal{D}_{\text{poison}}} \left[\mathrm{sim}(E_I(x+\tau), E_T(t_t)) \right],$$

subject to maintaining normal performance on the clean data.

More advanced attacks further leverage MCL properties: BadCLIP [33] introduces a dual-embedding mechanism to align backdoored examples with target embeddings across modalities, yielding highly natural and detection-robust triggers; prompt-based attacks [1] jointly optimize visual and textual triggers (with learnable prompts) to manipulate E_I and E_T for maximal attack success. Some methods employ distribution-preserving or generative mechanisms to create less detectable triggers, complicating defense. These advances formalize MCL-specific backdoor attacks as extensions of classical poisoning into the multimodal contrastive setting, underscoring the urgent need for robust and effective defenses.

C. Backdoor Defenses in Multimodal Contrastive Learning

Defending against backdoor attacks has been extensively studied in conventional unimodal deep learning. Existing defenses broadly fall into two categories: *data-based* and *model-based*. Data-based methods aim to detect or purge poisoned samples from the training set by identifying statistical anomalies, unusual activation patterns [63], [16], or atypical clustering behaviors [7], [55]. Model-based methods, in contrast, operate on the model, often via reverse engineering, trigger reconstruction, or analyzing responses to perturbed inputs [58], [36], [21], [32], [60], [64], [56]. These techniques have proven effective in supervised learning and significantly advanced our understanding of backdoor mechanisms.

With the growing adoption of MCL models, classical defenses have been adapted to the multimodal setting. Recent MCL-specific proposals incorporate fine-tuning, self-supervised learning objectives, or multimodal data augmentation to mitigate backdoor threats [2], [52], [74]. Others, such as RoCLIP [67] and SafeCLIP [66], target the pre-training stage by filtering potentially poisoned image–text pairs. Despite promising results, these methods typically rely on strong assumptions, e.g., knowledge of the attacker's target label, partial access to poisoned data, or auxiliary side information, which rarely hold in real-world deployments.

In practice, defenders rarely have such privileged information and often possess only the potentially compromised model. This exposes a key limitation: existing MCL backdoor defenses, although effective under idealized assumptions, remain fragile under realistic constraints. Developing *robust and assumption-free* defenses, particularly for the common scenario in which the defender has only the suspect model, therefore remains an open and pressing challenge. This gap motivates our work.

D. Universal Adversarial Perturbation

Adversarial examples reveal that deep neural networks are highly sensitive to crafted perturbations: even minor, imperceptible input modifications can induce misclassification [26], [18], [71], [22]. Conventional attacks are typically *per-sample*, generating a unique perturbation per input, often to steer predictions toward a specific (target) class. These have been widely explored in image and text domains, ranging from global perturbations to targeted synonym substitutions.

In contrast, universal adversarial perturbations (UAPs) [31], [35], [70] seek a single, input-agnostic perturbation r that consistently misleads the model across a broad set of inputs. UAPs exist in both imperceptible global (perturbation-based) and localized visible (patch-based) forms. Because they exploit systemic vulnerabilities in a model's decision boundaries, UAPs pose a practical threat, particularly in practice scenarios where per-sample attack generation is infeasible.

For multimodal models like CLIP, UAPs aim to disrupt cross-modal alignment between images and text. Let r be the universal image perturbation to be learned. The goal is to push the perturbed image embedding closer to the text embedding of an incorrect class and farther from that of its true class. A representative method, AdvCLIP [76], formulates this as

$$r^* = \arg\max_{\|r\| \le \delta} \mathcal{L}_{\text{UAP}}(r), \tag{1}$$

where

$$\mathcal{L}_{\text{UAP}}(r) = \mathbb{E}_{x \sim D,\, y = \text{class}(x)} \Big[\max_{k \ne y} \mathrm{sim}\Big(\hat{E}_I(x+r),\ \hat{E}_T(t_k)\Big) - \mathrm{sim}\Big(\hat{E}_I(x+r),\ \hat{E}_T(t_y)\Big) \Big], \tag{2}$$

AdvCLIP further shows that such adversarial patches transfer across diverse downstream tasks by consistently corrupting CLIP's shared feature space.

In this work, we observe that, on backdoored models, UAPs and backdoor samples often exhibit highly similar prediction behaviors, frequently causing the model to prefer the same target class. This behavioral correlation serves as a key analytical lever in our method: by exploiting the prediction similarity between UAP- and trigger-induced inputs, we can efficiently reveal and localize backdoor vulnerabilities in MCL models.

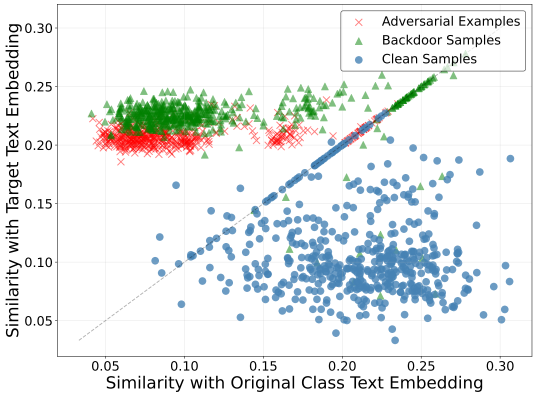
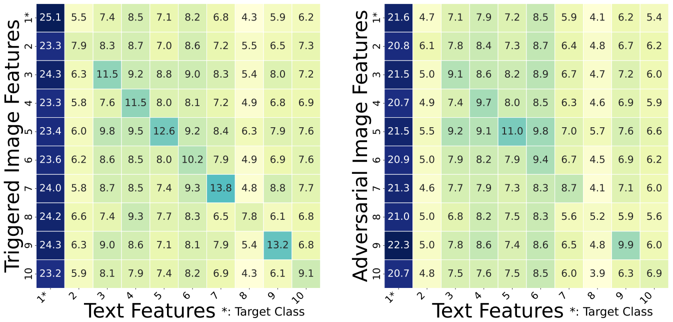


Fig. 3: Image-text similarity shift: backdoor and adversarial examples are closer to the target text than to the original text.



(a) Backdoor samples.

(b) Adversarial examples.

Fig. 4: Similarity matrices between image and text features under different attack scenarios.

III. InverTune: Detailed Construction

We first introduce the threat model considered in this work. As illustrated in Figure 1, under this threat model, InverTune mitigates backdoor attacks in MCL models through a three-step process: adversarial perturbation-based target identification, trigger inversion, and activation tuning.

A. Threat Model

Attacker. We follow the SOTA settings [33] for backdoor attacks in MCL models, specifically targeting the vision encoder. We assume that the attacker can construct a poisoned fine-tuning dataset and knows the model architecture and parameters. The attacker's goal is to implant a backdoor into the pre-trained CLIP model such that the model behaves normally on clean inputs but outputs incorrect results when exposed to inputs with triggers. To achieve this, the attacker injects a small portion of backdoor samples into the fine-tuning dataset, introducing visual triggers. The attacker then fine-tunes the pre-trained model using this poisoned dataset, manipulating the model's responses to visual triggers. Once the vision encoder is backdoored, the attacker has no control over downstream applications or tasks that use the model.

Defender. To conduct a practical defense, we assume that the defender has no access to the pretraining dataset or the poisoned fine-tuning dataset, and is unaware of the backdoor attack's target label. Furthermore, the defender either has no access to the full clean dataset or only possesses a limited amount of clean data. The primary goal of the defender is to neutralize the backdoors while maintaining the model's original performance on clean data.

B. Target Identification

Recent studies [41], [45] reveal that backdoored models exhibit distinct characteristics in feature representation and vulnerability within target classes. Unimodal backdoored models establish strong associations between target class labels and both robust features and backdoor features. Hence, clean and backdoor samples of the target class cluster closely in the latent space. Besides, untargeted adversarial attacks would inadvertently exploit backdoor pathways, causing the optimization process for generating adversarial perturbations to lean toward converging on backdoor triggers and resulting in attack outcomes that disproportionately favor the backdoor target label. In contrast, clean models exhibit approximately uniform label distribution for adversarial examples. This phenomenon has motivated backdoor defense strategies that leverage adversarial example analysis for trigger inversion. In the MCL domain, Kuang et al. [25] directly utilize the insight to optimize universal adversarial perturbation followed by anti-learning purification. However, their defensive performance is far from satisfactory.

Inspired by the above finding and result, we try to understand how the backdoor affects the target class in the MCL model. To achieve this, we take a SOTA MCL backdoor attack method, BadCLIP, as an example. Specifically, we visualize the visual encoder features of backdoor samples, adversarial examples generated using AdvCLIP [76], and clean images from 10 randomly selected categories, including the target category. Based on Figure 2, we find new observations different from those in the unimodal model.

Observation I

Backdoor samples form distinct clusters rather than merging with target class features.

In Figure 2a, for the unattacked clean model, the features extracted from samples with and without triggers completely overlap, indicating that triggers do not cause feature deviation in the clean model. Meanwhile, the UAP samples generated for the clean model form a separate cluster, which is far from the feature cluster of clean samples. Looking back at the backdoored model in Figure 2b, we find that although BadCLIP's dual-embedding optimization reduces the visual embedding distance between backdoor samples and target class samples, samples with triggers form a new cluster in visual features and do not become closer to the target class samples. Moreover, by observing adversarial examples, we also find similar results. To understand it more, we calculate the similarity between backdoor samples and adversarial examples, as shown in Figure 3 and Figure 4. Based on all results, we notice another observation.

Observation II

Adversarial attacks tend to exploit backdoor-induced weaknesses rather than direct trigger mimicry.

Since adversarial examples and backdoor samples remain significantly distant in terms of feature space, this suggests that adversarial examples do not directly mimic the features of backdoor samples. Based on Figure 4, we can find that most adversarial examples have higher similarity with the text features of the target class, showing that the backdoor also affects the adversarial attacks. This suggests backdoors reconfigure multimodal decision boundaries, creating "vulnerability zones" that adversarial attacks preferentially exploit. As a result, adversarial attacks are more likely to exploit this vulnerability, causing higher confusion and increasing the chances of misclassification into the target class. We systematically quantify both observations and present detailed results in Appendix A, covering multiple attacks and architectures.

1) Theoretical Analysis: To better understand the interaction between adversarial attacks and backdoor vulnerabilities in multimodal models, we now provide a theoretical analysis grounded in observed feature behaviors.

Assumption 1. *The backdoor attack instantiates a structural vulnerability within the model. Specifically, it creates a "shortcut" in the embedding space corresponding to the target class l. The infected image encoder, $\hat{E}_I$, becomes highly sensitive to perturbations that align with an effective feature displacement vector, v_{bd}, induced by the trigger P_{img}. For any image x, this relationship can be modeled as:*

$$\hat{E}_I(x + P_{img}) \approx \hat{E}_I(x) + v_{bd}, \tag{3}$$

where the vector v_{bd} is strongly aligned with the target text embedding $\hat{E}_T(t_l)$. Consequently, moving any image embedding $\hat{E}_I(x)$ along the direction of v_{bd} is the most efficient method to drastically increase its similarity with a text embedding, specifically $\hat{E}_T(t_l)$.

Building on this assumption, we then show that the universal adversarial perturbation implicitly aligns with the same vulnerable direction induced by the backdoor.

Theorem 1. *For a backdoored model $(\hat{E}_I, \hat{E}_T)$ that satisfies Assumption 1, the universal adversarial perturbation r^* obtained by optimizing the non-targeted loss function $\mathcal{L}_{UAP}$ will be functionally equivalent to the backdoor trigger P_{img}. Consequently, the perturbation r^* will cause arbitrary images to be classified as the backdoor's target class l.*

Proof sketch. The optimization problem seeks to maximize:

$$\max_{k \neq y} \text{sim}\left(\hat{E}_I(x + r), \hat{E}_T(t_k)\right). \tag{4}$$

Due to the structural vulnerability, the similarity to the target class t_l dominates:

$$\begin{aligned} &\max_{k \neq y} \text{sim}\left(\hat{E}_I(x + r), \hat{E}_T(t_k)\right) \\ &\approx \text{sim}\left(\hat{E}_I(x + r), \hat{E}_T(t_l)\right). \end{aligned} \tag{5}$$

Thus, the optimization simplifies to:

$$r^* \approx \arg \max_{\|r\| \leq \delta} \mathbb{E}_{x \sim D}\left[\text{sim}\left(\hat{E}_I(x + r), \hat{E}_T(t_l)\right)\right]. \tag{6}$$

For a comprehensive description and in-depth proof, refer to Appendix B. A high-level illustration of this process can also be found on the left side of Figure 1.

2) Identification Strategy: Building on these insights, we develop a target label identification strategy through differential analysis of adversarial misclassification patterns. Specifically, given a suspected backdoored model, we construct a UAP designed to induce systematic misclassification across all input images. The construction of the UAP is independent of the model's classification categories.

We then compare the model's output distribution on UAP samples $P_{\text{adv}}(y)$ against its predictions on clean samples $P_{\text{clean}}(y)$. The target label y_t is identified as the class exhibiting the maximum increase in prediction frequency:

$$y_t = \arg \max_{y \in \mathcal{Y}} \left(P_{\text{adv}}(y) - P_{\text{clean}}(y)\right). \tag{7}$$

This differential analysis isolates attack-induced bias from natural model tendencies, leveraging the intrinsic concentration property of backdoor attacks: backdoored models consistently steer misclassified samples toward the target label with disproportionate frequency. The identified target label then serves as the foundation for subsequent backdoor mitigation through gradient-guided trigger inversion and activation suppression.

C. Trigger Inversion

Unlike conventional unimodal backdoor attacks that target a specific class label, multimodal backdoor attacks in CLIP exploit the complex cross-modal alignment between visual and textual representations. The inversion process aims to generate inputs that reproduce the backdoor's behavioral effect within this alignment space.

Multimodal Trigger Inversion Challenges. Conventional backdoor inversion methods [58], [60] designed for classification models cannot be directly applied to multimodal models like CLIP for several key reasons. (1) In CLIP, backdoor attacks operate by creating malicious alignments between visual triggers and textual targets across modalities. This cross-modal interaction is fundamentally different from the class boundary manipulation in classification models, as it requires simultaneous optimization over both image and text embeddings. (2) CLIP projects both images and text into a shared high-dimensional embedding space, where the backdoor behavior is determined by the alignment between these modalities. The shared space introduces additional complexity compared to the discrete class labels used in classification models, as the backdoor functionality depends on the relative positions of embeddings rather than direct class mappings. (3) CLIP's zero-shot capabilities [77] allow it to generalize to unseen classes and concepts, which backdoors can exploit in ways that are not observable in classification models. This makes it challenging to detect and invert triggers, as the backdoor behavior may manifest differently across various downstream tasks.

Dual-Space Trigger Optimization. To address these challenges, we propose a novel dual-space trigger inversion approach that explicitly considers both the visual embedding space and the cross-modal alignment. Specifically, given a clean input image x, we parameterize the trigger as a mask-pattern pair (m, t_{img}), where the backdoor sample $\tilde{x}$ is generated via element-wise composition:

$$\tilde{x} = m \odot t_{\text{img}} + (1 - m) \odot x, \tag{8}$$

where m denotes the mask, t_{img} represents the trigger pattern, and $\odot$ denotes element-wise multiplication. Our framework integrates four synergistic loss components to ensure precise trigger reconstruction while preserving stealthiness: Cross-Modal Alignment, Embedding Space Preservation, Visual Similarity, and Trigger Sparsity. Detailedly, Cross-Modal Alignment is formulated using the InfoNCE [46] loss to force the visual trigger embeddings to align with the identified target text y_t while diverging from non-target classes. The contrastive loss can be expressed as:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(E_I(\tilde{x}), E_T(y_t))/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(E_I(\tilde{x}), E_T(y_j))/\tau)}, \tag{9}$$

where E_I and E_T are the image and text encoders of the suspected model, y_j iterates over all class prompts including the target, τ is the temperature parameter controlling the sharpness of the distribution, set to 0.07 as validated in prior works [48], [62], and N is the number of considered classes. Then we employ the embedding space preservation loss to prevent backdoor samples from excessively shifting toward the target class's textual embedding, thereby preserving the embedding structure and maintaining a stable data distribution to safeguard generalization. It is formulated as follows:

$$\mathcal{L}_{\text{emb}} = D(\frac{E_I(\tilde{x})}{\|E_I(\tilde{x})\|_2}, \frac{E_I(x)}{\|E_I(x)\|_2}), \tag{10}$$

where $D(\cdot)$ means a distance function. Here we employ the widely used L_2-norm distance, which provides stable gradient behavior and effectively captures pixel-level deviations during inversion. Considering the attacker's goal, where the backdoor sample must remain visually similar to the original, we introduce a visual similarity loss as follows:

$$\mathcal{L}_{\text{sim}} = 1 - \text{SSIM}(\tilde{x}, x), \tag{11}$$

where SSIM$(\cdot)$ function computes the structural similarity between two given images [61]. Although the loss function $\mathcal{L}_{\text{sim}}$ can make the backdoor sample as similar as possible to the original sample, it does not ensure the imperceptibility of the backdoor trigger. Therefore, we introduce the trigger sparsity loss to further constrain the trigger as follows:

$$\mathcal{L}_{\text{mask}} = \|m\|_1. \tag{12}$$

To obtain the trigger pattern and mask, we optimize the four loss functions concurrently. Therefore, the total loss can be written as the weighted combination of these objectives:

$$\mathcal{L}_{\text{inver}} = \lambda_1 \mathcal{L}_{\text{align}} + \lambda_2 \mathcal{L}_{\text{emb}} + \lambda_3 \mathcal{L}_{\text{sim}} + \lambda_4 \mathcal{L}_{\text{mask}}, \tag{13}$$

where λ_1, λ_2, λ_3, and λ_4 are weighting coefficients.

D. Activation Tuning

Building upon the inverted trigger obtained in Section III-C, we propose an activation-based fine-tuning strategy specifically tailored for MCL models like CLIP. This approach leverages the unique activation patterns induced by backdoor triggers in the shared embedding space of multimodal models.

Key Insight. Backdoor triggers in the MCL model exploit the cross-modal alignment mechanism, creating distinct activation signatures in specific layers. By identifying and selectively fine-tuning critical neurons, we can neutralize the backdoor while preserving the multimodal capabilities of the model.

Layer Selection. Inspired by prior findings [17], [36] in CNN architectures, where backdoor patterns predominantly affect deeper layers, we first identify responsive layers to backdoor activation in MCL models. For each layer, we quantify backdoor sensitivity through normalized activation divergence:

$$\text{diff} = \frac{\|\mu_{\text{clean}} - \mu_{\text{triggered}}\|_2}{\|\mu_{\text{clean}}\|_2}, \tag{14}$$

where μ_{clean} and $\mu_{\text{triggered}}$ mean the average activations of clean and triggered inputs, respectively. Then, we compute the mean and standard deviation of activation differences across all layers. The layers with activation differences exceeding the mean by more than one standard deviation will be treated as backdoor-related. The threshold serves as a coarse heuristic to emphasize layers with salient activation shifts, while its specific value has limited influence on subsequent clustering. Within the critical layers, we analyze individual neuron activation variances. Note that identifying only the neurons in the backdoor-related layer greatly reduces the time and resource overhead compared to identifying all neurons once.

Critical Neuron Identification. We identify critical neurons by first measuring the impact of the trigger on layer activations. For each selected layer, we calculate the mean activation difference between the clean and trigger-affected inputs. Then, we apply K-means clustering [38] on the activation differences to group neurons with similar response patterns. Clustering helps address the potential variability in neuron responses. Instead of simply selecting the neurons with the largest activation difference, K-means clustering groups neurons with similar response patterns, ensuring that the neurons we capture share a common sensitivity to the backdoor.

Fine-Tuning Process. Following neuron identification, we implement targeted fine-tuning to eliminate backdoor functionality while preserving clean-task performance. Specifically, we introduce an activation alignment loss to force backdoor-sensitive neurons to exhibit similar activation patterns for clean and triggered samples:

$$\mathcal{L}_{\text{activation}} = \sum_{i \in \text{critical}} \|\mathbf{a}^i_{\text{clean}} - \mathbf{a}^i_{\text{triggered}}\|_2^2. \tag{15}$$

Algorithm 1 Activation Tuning for Backdoor Mitigation

```
**Input:** Backdoored CLIP model $F$ ($E_I$, $E_T$); Inverted trigger $(m, t_{\text{img}})$; Clean batch $\{x_1, ..., x_b\}$; Layers $\mathcal{L}$; Parameter $\beta$
**Output:** Fine-tuned model with neutralized backdoor
**Phase 1: Identify Critical Layers**
Compute clean activations $\{A^l_{\text{clean}}\}$ for $l \in \mathcal{L}$ using $\{x_i\}$
Generate triggered images $\{\tilde{x}_i \leftarrow m \odot t_{\text{img}} + (1-m) \odot x_i\}$ for $x_i \in \{x_1, ..., x_b\}$
Compute triggered activations $\{A^l_{\text{triggered}}\}$ using $\{\tilde{x}_i\}$
**for** each $l \in \mathcal{L}$ **do**
    $\text{diff}^l \leftarrow \frac{\|\text{mean}(A^l_{\text{clean}}) - \text{mean}(A^l_{\text{triggered}})\|_2}{\|\text{mean}(A^l_{\text{clean}})\|_2}$
**end for**
$\mathcal{L}_{\text{critical}} \leftarrow \{l \in \mathcal{L} \mid \text{diff}^l > \text{mean}(\{\text{diff}^l\}) + \text{std}(\{\text{diff}^l\})\}$
**Phase 2: Identify Critical Neurons**
**for** each $l \in \mathcal{L}_{\text{critical}}$ **do**
    $\Delta^l \leftarrow |\text{mean}(A^l_{\text{clean}}) - \text{mean}(A^l_{\text{triggered}})|$
    Apply K-means ($k = 2$) to $\Delta^l$, select cluster $C^l_{\text{critical}}$ with largest centroid
    Create neuron mask $M^l$ for $C^l_{\text{critical}}$
**end for**
**Phase 3: Selective Fine-tuning**
Initialize $F' \leftarrow F$, create parameter masks from $\{M^l\}$
Set optimizer with masked gradients
**for** each training step **do**
    Use $\{x_1, ..., x_b\}$ and $\{\tilde{x}_i\}$
    $\mathcal{L}_{\text{activation}} \leftarrow \sum_{l \in \mathcal{L}_{\text{critical}}} \|(a^l_{\text{clean}} \odot M^l) - (a^l_{\text{triggered}} \odot M^l)\|_2^2$
    $\mathcal{L}_{\text{preserve}} \leftarrow \|\text{sim}(E_I(x_i), E_T(y_i)) - \text{sim}(E_I^{\text{orig}}(x_i), E_T(y_i))\|_2^2$
    $\mathcal{L}_{\text{tune}} \leftarrow \mathcal{L}_{\text{activation}} + \beta \cdot \mathcal{L}_{\text{preserve}}$
    Update critical neuron parameters
**end for**
**return** $F'$
```

This suppresses backdoor-triggered activation spikes. Moreover, to maintain original vision-language alignment capability, we introduce a cross-modal consistency loss.

$$\begin{aligned}\mathcal{L}_{\text{preserve}} = \|&\text{sim}(E_I(x), E_T(y)) \\ &- \text{sim}(E_I^{\text{orig}}(x), E_T(y))\|_2^2,\end{aligned} \tag{16}$$

where E_I^{orig} represents the original backdoored encoders prior to fine-tuning. This function forces the fine-tuned model to have similar normal functions to the original model. To achieve both purposes, the optimization objective becomes as:

$$\mathcal{L}_{\text{tune}} = \mathcal{L}_{\text{activation}} + \beta \mathcal{L}_{\text{preserve}}, \tag{17}$$

where β is to balance the two objectives. Note that, we apply neuron masks during gradient updates to restrict fine-tuning to those critical neurons. This targeted fine-tuning minimizes disruption to the model's overall performance while effectively mitigating the backdoor. The overall Activation Tuning process can be found in Algorithm 1.

IV. Experiment

A. Experiment Setup

Models. We adopt OpenAI's open-source CLIP model [48] as our pretrained base, using ResNet-50 (RN50) as the default backbone architecture. For a comprehensive evaluation, we extend our analysis to RN101, ViT-B/16, and ViT-B/32 architectures in Section V-A.

Datasets. Following the prior work [33], we use a 500K subset of CC3M [50] for poisoning the clean CLIP model. The evaluation framework covers two key tasks: zero-shot classification on ImageNet-1K validation set [49] and image-to-text retrieval on Microsoft COCO 2017 [34].

Backdoor Attacks. We evaluate our defense method against four representative unimodal backdoor attack methods: BadNet [19], Blended [9], SIG [3], and WaNet [43]. Additionally, we include one self-supervised learning backdoor attack on a pretrained encoder, BadEncoder [24]. BadEncoder targets only the image encoder, enabling evaluation of the generalization of InverTune beyond CLIP. We also include the SOTA CLIP-specific backdoor attack, BadCLIP [33]. We randomly select "mushroom" as the target label. Experiments with other target labels are presented in Section V-A. Following the settings of [33], we set the poisoning rate to 0.3%. Detailed configurations for various attacks are provided in Appendix C. Additionally, the results on more complex attack scenarios, such as multiple backdoors, varying poisoning rates, and performance on clean models, can be found in Appendix F.

Baseline Defense. We compare InverTune against several advanced backdoor defense techniques, including CleanCLIP [2], CleanerCLIP [65], PAR [52], as well as Fine-Tuning (FT) [2] as the baselines. Specific details of different defense settings can be found in Appendix D.

Evaluation Metrics. We evaluate the effectiveness of our method using the following metrics. ***(1) Clean Accuracy*** **(CA)**: For zero-shot classification tasks, CA quantifies the model's Top-1 prediction accuracy on clean inputs. For image-to-text retrieval scenarios, it measures the proportion of clean queries successfully matching ground-truth captions within the Top-10 retrieved results. Higher CA values indicate better preservation of the model's normal capabilities. ***(2) Attack Success Rate*** **(ASR)**: For classification, ASR represents the percentage of triggered samples misclassified to target labels. For image-to-text retrieval tasks, ASR is the percentage of triggered inputs that retrieve target-related text in the Top-10 results. Lower ASR scores demonstrate superior backdoor mitigation.

Implementation Details. For the trigger inversion, we set $\lambda_1 = 5.0$, $\lambda_2 = 0.5$, $\lambda_3 = 1.0$, and $\lambda_4 = 0.01$ for the trigger inversion loss in Eq. (13). These coefficients correspond to distinct components of the inversion objective and operate on normalized losses, allowing InverTune to maintain stable optimization without requiring adaptive weighting. For activation tuning, we set $\beta = 0.5$ for the loss in Eq. (17), use a learning rate of 8×10^{-6}, and train for 200 steps. We further clarify the roles of these coefficients and their low sensitivity in Section VI-A. In terms of data usage, InverTune

TABLE I: Defensive performance (%) of InverTune vs. baseline defenses across tasks and backdoor attacks. The optimal ASR and CA values are highlighted in **bold**, while the second-best results are indicated with underlining.

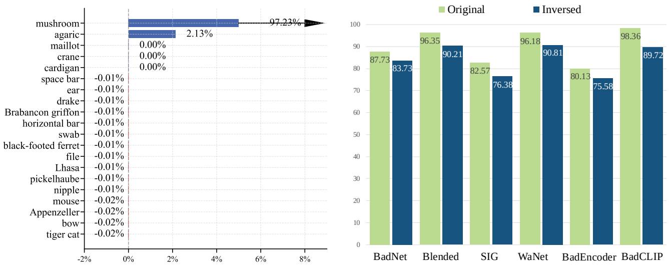
	Methods	BadNet		Blended		SIG		WaNet		BadEncoder		BadCLIP	
		CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
ImageNet	No Defense	58.21	87.73	58.74	96.35	58.30	82.57	58.64	96.18	53.10	80.13	58.32	98.36
	FT	54.13	33.67	**54.64**	64.10	**54.36**	55.59	54.59	58.38	**55.98**	19.71	54.16	86.03
	CleanCLIP	51.92	4.62	51.38	52.36	51.42	36.72	51.45	24.98	55.29	5.21	54.18	75.17
	CleanerCLIP	51.91	3.87	52.36	11.38	52.56	9.89	51.57	10.94	52.11	**0.19**	51.74	21.16
	PAR	53.57	6.03	54.18	0.16	51.96	22.94	53.89	4.51	54.25	2.27	50.95	17.78
	InverTune (Ours)	**56.12**	**0.02**	53.50	**0.14**	54.27	**0.28**	**54.76**	**0.09**	55.84	1.02	**55.25**	**0.49**
MSCOCO	No Defense	69.94	95.88	71.20	99.76	70.28	97.42	71.16	99.60	72.07	98.13	71.32	99.28
	FT	68.83	39.09	**69.53**	67.51	68.92	63.67	**69.70**	70.41	**68.77**	25.47	68.25	88.54
	CleanCLIP	65.03	14.17	63.70	55.47	64.09	38.71	67.61	64.83	67.56	13.42	66.53	84.55
	CleanerCLIP	65.73	7.94	68.82	14.93	65.98	14.31	64.67	15.01	66.39	3.41	65.21	30.41
	PAR	68.42	15.43	68.11	**0.37**	66.64	31.09	68.28	7.83	67.42	4.30	65.73	16.47
	InverTune (Ours)	**71.12**	**0.04**	69.16	0.52	**69.94**	**1.12**	68.98	**0.48**	68.02	**1.73**	**69.58**	**0.68**

employs a 50K subset of the ImageNet-1K training set [49], which is only 1/10 the size of the data used by other baselines. In the activation tuning step, we require only a single batch (predefined as 64) of arbitrary clean data. All experiments were conducted on an Ubuntu 20.04 system with a 20-core Intel CPU. The models were trained on a single NVIDIA RTX 4090 GPU. Detailed intermediate results and computational costs are provided in Appendix E.

B. *InverTune Performance*

Defensive Performance. The experimental results in Table I demonstrate InverTune's superior defensive capabilities across multiple attack scenarios, substantially surpassing existing baselines by achieving remarkably low ASR. Our method achieves SOTA performance by reducing the ASR to below 0.5% on both the ImageNet and MSCOCO datasets in the vast majority of attack scenarios, significantly outperforming existing defense baselines. For instance, against conventional unimodal attacks such as BadNet and Blended, while existing defense mechanisms can mitigate these attacks to a certain extent, InverTune delivers more robust results. Notably, when defending against the sophisticated BadCLIP attack, the limitations of existing baseline methods become apparent. Specifically, conventional methods like FT and CleanCLIP remain vulnerable, with ASR exceeding 75%. Although more advanced defenses like CleanerCLIP and PAR show partial mitigation, they still result in unacceptably high residual ASR, such as those greater than 15%. In contrast, InverTune maintains its exceptional defensive prowess.

Model Performance and Utility Preservation. Beyond its formidable defensive capabilities, InverTune also demonstrates exceptional preservation of model utility, maintaining high CA across diverse scenarios. This performance, detailed in Table I, sets it apart from baseline methods that often sacrifice utility for security. An evaluation across our 12 experimental configurations (2 tasks × 6 attack methods) reveals our method's consistent, high-utility performance. InverTune achieves either the highest CA in 6 cases or the second-highest in 5 cases, placing it in the top tier for utility in 11 out of 12 total settings. This highlights its minimal disruption to the model's original capabilities on clean data. Notably, InverTune is unique in its ability to simultaneously achieve SOTA defense (lowest ASR) and optimal model utility (highest CA). This "dual optimum" is demonstrated consistently under both the BadNet and the highly sophisticated BadCLIP attacks.



(a) Top 20 classes with increased prediction frequency.

(b) ASR of inverted and original trigger.

Fig. 5: Results of backdoor target identification and trigger inversion, with mushroom as the target label.

This consistent performance shows InverTune's ability to establish a superior **security-utility trade-off**. Unlike methods like FT, which preserves CA at the cost of high residual ASR (e.g., 86.03% ASR vs. BadCLIP), or defenses like CleanerCLIP that aggressively reduce ASR but impair CA (e.g., 52.11% CA vs. BadEncoder), InverTune effectively neutralizes backdoors with minimal collateral damage. We attribute this superior balance to InverTune's activation tuning, which is based on inverted triggers, allowing it to avoid the indiscriminate feature damage often caused by other defenses.

Efficacy of Backdoor Label Identification and Inversion. InverTune consists of three critical steps: target identification, trigger inversion, and activation tuning. As established in the preceding sections, the final step, activation tuning, achieves an excellent balance between defensive efficacy and utility. This success, however, is contingent upon the effectiveness of the first two foundational steps. We now present an analysis to demonstrate their individual efficacy.

For the first step, target identification, we apply universal adversarial perturbation to clean examples and feed them into the backdoored model with "mushroom" as the designated target

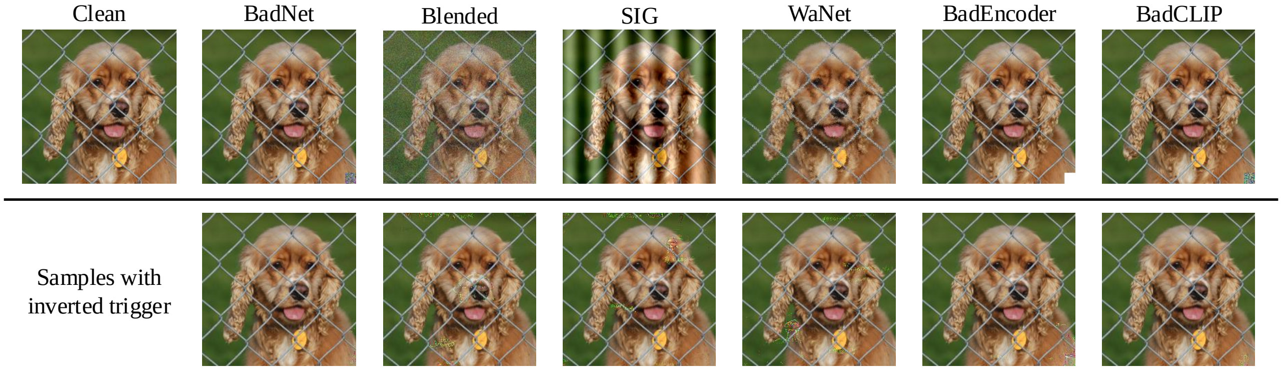


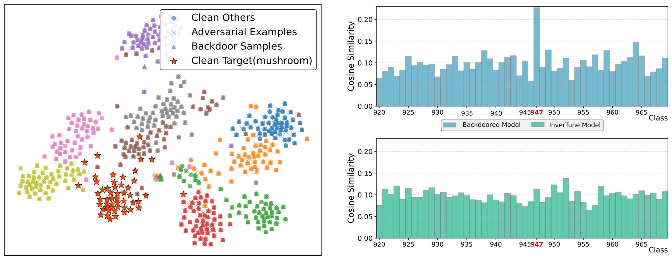
Fig. 6: Visualization of backdoor samples (top) and trigger-inverted counterparts (bottom).

class. As shown in Figure 5a, this process induces a dramatic shift in the model's prediction distribution. Specifically, the classification frequency for the "mushroom" class exhibits a 97.23% increase compared to that of clean samples. Notably, the frequency for "agaric", a visually similar subspecies of mushroom, experiences only a marginal 2.13% rise. This stark divergence in the prediction distribution unequivocally identifies the target label, confirming the effectiveness of our identification strategy.

For the second step, we reconstruct the trigger pattern. We argue that the objective here is not to physically replicate the original trigger but to synthesize a pattern that can activate the backdoor pathways for our defense. As illustrated in Figure 6, which compares images with the original backdoor trigger against those with our inverted trigger, our four-component loss constraint enables the successful inversion of the trigger from the backdoored model. For attacks employing regular trigger patterns, such as BadNet, BadEncoder, and BadCLIP, it is evident that InverTune successfully reverse-engineers a trigger-like artifact at the expected image location (i.e., the lower-right corner). For other attack types, our method also synthesizes irregular yet distinct inverted patterns. Critically, the functional equivalence of our inverted trigger is validated in Figure 5b, which shows that it achieves a level of attack behavior alignment nearly identical to the original pattern. These observations collectively indicate that InverTune can reliably recover a functionally equivalent trigger from a backdoored model, which is a critical prerequisite for the subsequent successful removal of the backdoor.

C. Analysis of Internal Representation Changes

Beyond evaluating standard defense metrics, we further investigate how InverTune alters the model's internal representations. Using the BadCLIP attack on ImageNet as a case study, we visualize the t-SNE distributions of backdoor and UAP samples as processed by the InverTune-sanitized model (Figure 7a). The figure reveals two key observations. First, the original backdoor attack is completely neutralized; the backdoor samples are now correctly classified into their original, benign classes. Second, the Universal Adversarial Perturbations (UAPs) crafted for the original backdoored model are no longer effective against the sanitized model.



(a) t-SNE visualization of the InverTune model.

(b) Cosine similarity between backdoored and InverTune model.

Fig. 7: The effect of InverTune on model internal representations and prediction similarity.

TABLE II: Performance comparison of InverTune and baseline defenses against BadCLIP under different target labels.

Target Label	Banana		Lemon		Ski	
	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
No Defense	58.20	98.16	58.11	97.16	58.31	98.46
FT	54.77	83.14	54.93	89.65	54.34	79.70
CleanCLIP	53.48	74.85	54.50	72.82	53.94	77.75
CleanerCLIP	52.09	20.41	51.69	25.36	51.67	16.16
PAR	53.64	17.65	53.91	36.07	53.62	11.72
InverTune (Ours)	**57.01**	**1.14**	**55.81**	**1.01**	**56.93**	**1.51**

Samples from various classes, even when perturbed by the UAP, are correctly classified.

Furthermore, we analyze the average cosine similarity among class representations before and after sanitization. Figure 7b summarizes the results for 50 classes neighboring the target class (label 947, "mushroom"). As depicted, the backdoored model exhibits a significant spike in similarity at the target label 947, an anomaly indicative of the backdoor. In contrast, the InverTune-sanitized model successfully mitigates this anomaly, restoring the similarity values to a normal and consistent level of approximately 0.1 across all classes.

Collectively, these findings from both the t-SNE visualizations and the cosine similarity analysis provide compelling evidence that InverTune effectively removes the backdoor by rectifying the parameters of critical model layers, thereby demonstrating its efficacy.

TABLE III: Performance comparison of defense methods across different model architectures.

Backbone	**RN101**		**ViT-B/16**		**ViT-B/32**	
	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
No Defense	59.17	83.17	66.78	99.90	60.97	99.23
FT	**56.85**	58.29	**63.01**	83.75	54.72	91.61
CleanCLIP	56.14	42.53	61.91	80.33	53.16	79.36
CleanerCLIP	52.76	3.25	58.81	31.17	53.64	64.60
PAR	55.60	1.17	57.98	18.14	50.82	76.37
InverTune (Ours)	55.76	**1.00**	59.80	**0.09**	**54.83**	**0.17**

V. In-Depth Analysis

While Section IV-B presented a comprehensive evaluation of InverTune's overall effectiveness, this section delves deeper into its operational robustness across various configurations. We methodically examine its resilience against variations in target labels, model architectures, and potential inaccuracies in target identification. For a focused and rigorous analysis, we center our experiments on the **BadCLIP** attack, as it represents the most sophisticated threat and poses the greatest challenge to existing defenses.

A. Robustness to Target Label Variation

To assess the generalizability of our defense, we evaluate InverTune's performance when the attack's target label is varied. We train distinct BadCLIP models targeting "banana", "lemon", and "ski", each with a unique trigger pattern. As illustrated in Table II, conventional baselines like FT and CleanCLIP remain vulnerable regardless of the target label. Furthermore, the defensive efficacy of more advanced methods like CleanerCLIP and PAR proves to be inconsistent and sensitive to the target choice. For instance, when the target switches from "ski" to "lemon", the residual ASR for CleanerCLIP increases from 16.16% to 25.36%, and PAR's ASR surges from 11.72% to 36.07%. In stark contrast, InverTune demonstrates remarkable stability, consistently achieving superior performance across all target labels with an ASR of approximately 1% while maintaining high CA. These results robustly demonstrate that InverTune's defense mechanism is not contingent on the specific target class, highlighting its strong generalization capabilities.

B. Robustness to Model Architecture

To evaluate the robustness of InverTune across different model architectures, we assess its performance on a diverse set of backbones, including both CNN-based (RN101) and Transformer-based (ViT-B/16, ViT-B/32) models. Additionally, these architectures naturally exhibit different initial CA, reflecting their inherent design differences. As shown in Table III, the choice of architecture has a significant impact on the effectiveness of baseline defenses, with performance varying dramatically across architectures. For example, PAR demonstrates a severe performance drop when moving from RN101 to ViT-B/32, with its ASR rising sharply from 1.17% to 76.37%, and its CA declining from 55.60% to 50.82%. Similarly, CleanerCLIP shows highly inconsistent results, with its ASR fluctuating between 3.25% and 64.60% across different model architectures.

In contrast, InverTune shows exceptional stability and maintains its superior defensive capability across all evaluated models. It consistently achieves an average ASR of merely 1.22% against BadCLIP attacks while preserving competitive CA. This unwavering performance across both CNN and Transformer families validates the architecture-agnostic nature of our method. We attribute this robustness to InverTune's core paradigm of backdoor inversion, which directly targets fundamental cross-modal activation patterns rather than relying on architecture-specific features or idiosyncrasies.

TABLE IV: Defense effectiveness of InverTune under different labels when the target label is set to mushroom.

Rank	Label	Metric	Top-1	Top-3	Top-5	Top-10
1	mushroom	ASR ↓	0.49	1.17	1.67	2.60
		CA ↑	55.25	76.45	83.45	90.00
2	agaric	ASR ↓	0.97	2.85	3.08	3.71
		CA ↑	55.70	77.12	83.90	90.33
3	maillot	ASR ↓	49.38	55.14	57.12	59.56
		CA ↑	55.99	77.02	83.84	90.45
5	cardigan	ASR ↓	57.56	62.24	63.90	65.85
		CA ↑	56.19	77.32	83.99	90.53
Random	bee	ASR ↓	78.88	82.35	83.46	84.98
		CA ↑	56.48	77.60	84.37	90.75

C. Sensitivity to Target Identification Accuracy

The efficacy of the InverTune framework is predicated on the accurate identification of the backdoor's target label. While our identification mechanism, as shown in Figure 5a, unequivocally pinpoints the correct label (e.g., a 97.23% frequency surge for mushroom), it is crucial to assess the defense's sensitivity to this initial step. To this end, we conduct an experiment to simulate scenarios of imperfect or incorrect identification. We use the backdoored model targeting mushroom and execute our defense pipeline assuming different target labels: the true target (mushroom), the top-ranked incorrect labels from our identification step (the 2nd, 3rd, and 5th ranked labels), and a semantically unrelated, randomly chosen label (bee). The results, presented in Table IV, are highly revealing. Remarkably, when the defense is guided by the second-ranked label (agaric), it achieves outstanding effectiveness, with an ASR only marginally higher than when using the true target, and in some cases, even a slightly improved CA. However, as the semantic relevance of the guiding label decreases (i.e., using the 3rd-ranked maillot or 5th-ranked cardigan), the defensive performance systematically degrades. Finally, when a completely random label (bee) is used, the defense is rendered largely ineffective, with the Top-1 ASR remaining at an unacceptable 78.88%.

This phenomenon reveals several important characteristics of the InverTune: (1) **Robustness of InverTune**: Under our ranking strategy, labels with different ranks all contribute to some extent to the defense, though this effect gradually diminishes as the rank decreases. Notably, using the second-

TABLE V: Influence of λ parameters on trigger-inversion ASR.

Attacks	λ_1					λ_2					λ_3					λ_4			
	0.0	1.0	**5.0**	10.0	20.0	0.0	0.1	**0.5**	1.0	5.0	0.0	0.5	**1.0**	5.0	10.0	0.005	**0.01**	0.05	0.1
BadNet	0.05	51.97	83.73	84.84	87.67	86.63	87.11	83.73	83.86	60.86	90.73	90.52	83.73	84.53	65.42	81.70	83.73	48.75	43.32
Blended	0.02	37.09	90.21	92.09	92.62	24.97	86.72	90.21	89.57	62.33	13.83	91.83	90.21	53.17	33.38	93.97	90.21	20.02	10.02
SIG	0.02	39.84	76.38	78.84	79.29	10.05	73.62	76.38	71.18	71.74	7.20	80.02	76.38	41.74	20.08	80.10	76.38	12.44	0.05
WaNet	0.04	60.81	90.81	93.96	89.83	37.33	71.53	90.81	87.15	81.04	38.34	92.16	90.81	78.37	37.02	92.38	90.81	30.06	20.03
BadEncoder	0.83	68.72	75.58	77.43	78.59	57.99	75.13	75.58	70.76	70.23	75.93	75.04	75.58	66.64	65.38	77.13	75.58	65.23	62.52
BadCLIP	0.01	73.38	89.72	87.38	89.48	46.86	79.89	89.72	65.08	57.17	61.52	69.10	89.72	73.54	69.74	91.13	89.72	59.86	20.83

TABLE VI: Adaptive attack evaluation results.

Attacks	No Defense		Step 1	Step 2	Step 3	
	CA ↑	ASR ↓	ID	Inversion ASR ↑	CA ↑	ASR ↓
Adap-1	58.20	98.58	✓	95.77	57.58	0.13
Adap-2	58.53	99.34	✓	63.51	57.63	0.01

ranked label (agaric) achieves nearly the same effectiveness as the true target label, which demonstrates the robustness of our method. (2) **Importance of Target Identification**: While highly ranked labels can still provide some defense, the overall effectiveness drops sharply when a random label such as bee is chosen as the target, indicating that the subsequent backdoor removal step relies heavily on accurate label identification and underscoring the importance of correct target localization. (3) **Applicability in Open-Vocabulary Scenarios**: In this paper, we use the ImageNet-1K vocabulary as our label set for identification. Leveraging the hierarchical semantic structure of WordNet [39], these 1,000 classes not only cover the vast majority of real-world scenarios but also form a semantically dense network that minimizes the likelihood of true target labels falling outside this set. Nevertheless, should such cases occur (though rare), the inherent semantic within WordNet ensures effective mitigation, as exemplified by the agaric-mushroom case: both belong to the same fungal category and appear highly similar to non-experts. This validates that the 1K classes constitute a sufficiently comprehensive base set; when encountering a target label outside this range, our method can leverage WordNet's taxonomic relationships to identify a highly similar class within the 1K set, thereby enabling effective backdoor removal even in such edge cases.

D. Adaptive Attack Analysis

To further assess the robustness of InverTune, we consider a fully adaptive adversary who is aware of all defense details, including the inversion loss (Eq. (13)), the associated hyperparameters, and the optimization pipeline. Under this stringent threat model, the attacker can tailor its strategy to directly counteract the mechanisms employed by InverTune.

We consider two adaptive attack strategies. The first, Adap-1, adopts a multi-objective optimization combining the original BadCLIP [33] triplet loss L_{backdoor} with an anti-inversion objective: $L_{\text{adaptive}} = L_{\text{backdoor}} + \lambda \cdot L_{\text{anti-inv}}$. Here, $L_{\text{anti-inv}}$ replicates InverTune's inversion process by performing gradient descent on the defender's loss while maximizing inversion difficulty. λ balances backdoor effectiveness and resistance to inversion attacks, and we set $\lambda = 0.002$ in our experiments. $L_{\text{anti-inv}}$ includes three aspects: feature-space perturbation, semantic confusion, and reduction of target confidence, which together hinder inversion-based reconstruction. However, this approach faces an inherent conflict: enhancing backdoor effectiveness requires a strong trigger–target correlation, while improving inversion resistance necessitates weakening it. The second strategy, Adap-2, termed pure anti-inversion optimization, uses $L_{\text{anti-inv}}$ alone, omitting L_{backdoor}. It focuses purely on defense evasion while relying on downstream fine-tuning to recover attack utility.

As shown in Table VI, both attacks achieve high ASR and a CA of around 58% without any defense. With InverTune, the target class remains accurately identifiable. Step 2 inversion highlights a clear distinction: Adap-1 maintains a 95.77% ASR, indicating precise trigger inversion, whereas Adap-2 drops to 63.51%, demonstrating that the adaptive attack partially evades inversion. Nonetheless, step 3 activation tuning effectively neutralizes both attacks, reducing ASR to near zero while keeping CA at 57$\sim$58%. These results show that even partially inverted triggers are sufficient for purification, confirming the robustness of InverTune against adaptive attacks.

VI. Ablation Study

A. Influence of Hyperparameters

In this section, we study the impact of hyperparameters. As formulated in Eq. (13), the coefficients λ_1-λ_4 control the relative importance of four loss components during backdoor inversion, while β in Eq. (17) governs the trade-off between model cleanliness and usability during the elimination phase.

Our experiments reveal several important patterns in hyperparameter sensitivity. For the inversion-related hyperparameters (Table V), we find that λ_1, weighting the contrastive learning loss, plays a pivotal role in reducing ASR, though with diminishing returns beyond $\lambda_1 = 5.0$ as trigger quality begins to degrade. The visual feature consistency term, controlled by λ_2, exhibits a distinct optimal range; insufficient weighting ($\lambda_2 = 0.1$) fails to generate functionally effective triggers for complex attacks like SIG, WaNet, and BadCLIP, whereas excessive weighting ($\lambda_2 > 1.0$) over-constrains the feature space and harms ASR. Optimal performance is achieved with $\lambda_3 = 1.0$ and $\lambda_4 = 0.01$, which strikes an effective balance between trigger stealth and efficacy. Larger values improperly prioritize trigger minimization at the expense of adversarial potency. Crucially, an ablation on these components confirms their individual contributions: setting λ_1, λ_2, or λ_3 to zero individually results in a notable degradation of the inverted

TABLE VII: Comparison of universal adversarial perturbation (UAP) and inverted trigger (InvT) for the Activation Tuning.

Methods		BadNet		Blended		SIG		WaNet		BadEncoder		BadCLIP	
		CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
Top-1	UAP	55.55	23.24	53.46	89.92	**54.27**	58.91	52.25	25.74	52.81	67.13	52.02	54.33
	InvT	**56.12**	**0.02**	**53.50**	**0.14**	**54.27**	**0.03**	**54.76**	**0.09**	**55.84**	**0.02**	**55.25**	**0.49**
Top-3	UAP	76.76	47.39	74.99	95.49	75.71	76.91	73.64	50.86	74.67	69.76	73.42	71.15
	InvT	**77.24**	**0.20**	**75.35**	**0.74**	**75.92**	**0.10**	**75.92**	**0.40**	**77.05**	**0.20**	**76.45**	**1.17**
Top-5	UAP	83.63	58.98	82.02	96.75	82.84	82.16	81.05	60.76	81.73	71.04	80.80	76.18
	InvT	**84.10**	**0.46**	**82.54**	**1.47**	**83.04**	**0.20**	**83.04**	**0.77**	**83.87**	**0.44**	**83.35**	**1.67**
Top-10	UAP	90.11	72.94	89.01	98.01	89.53	87.85	88.33	72.45	88.76	72.85	88.13	81.88
	InvT	**90.56**	**0.95**	**89.52**	**3.42**	**89.80**	**0.41**	**89.80**	**1.71**	**90.37**	**0.91**	**90.00**	**2.60**

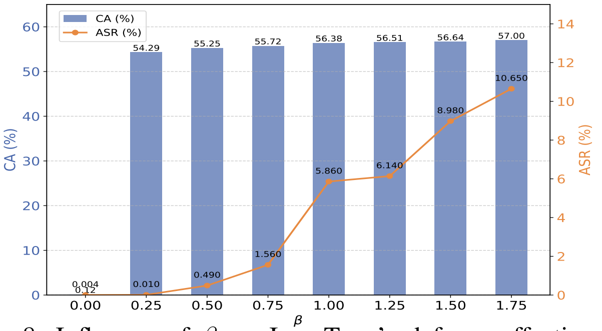


Fig. 8: Influence of β on InverTune's defense effectiveness under BadClip attack scenario.

trigger's ASR. The case for $\lambda_4 = 0$ is omitted as it is an invalid setting that prevents the trigger mask from being trained, leading to training collapse. Collectively, these results validate the necessity of each component in our loss formulation.

The elimination phase (Figure 8) highlights β's role in balancing security and utility. At $\beta = 0$, prioritizing backdoor removal, CA (0.12%) and ASR (0.004%) drop near zero, showing the usability term in Eq. (16) is necessary. As β rises, CA improves, plateauing past 0.50, while ASR jumps sharply in $\beta \in [0.75, 1.0]$ from 1.56% to 5.86%. With $\beta = 0.5$, ASR stays at 0.49% and CA at 55.25%, confirming InverTune's stability and our loss formulation's effectiveness.

B. The Necessity of Trigger Inversion

Our analysis in Section III-B suggests a fundamental distinction between the vulnerabilities exploited by adversarial perturbation and backdoor triggers. While both can induce target-class misclassification, they are mechanistically different. This distinction forms a central hypothesis: generic adversarial patterns are insufficient for defense, necessitating our specialized trigger inversion approach.

To empirically validate this, we first conduct an ablation study comparing the complete InverTune framework (InvT) against a variant (UAP) that omits trigger inversion and instead uses first-stage adversarial perturbation for fine-tuning. As shown in Table VII, InvT demonstrates overwhelming superiority. Its average ASR of 0.13% (Top-1) and 1.67% (Top-10) are orders of magnitude lower than UAP's 53.21% and 81.00%, respectively. This performance gap highlights UAP's limitation: while adversarial fine-tuning can enhance general noise robustness, it fails to neutralize the deeply embedded backdoor mechanism, which InverTune's targeted approach successfully disrupts while preserving model utility.

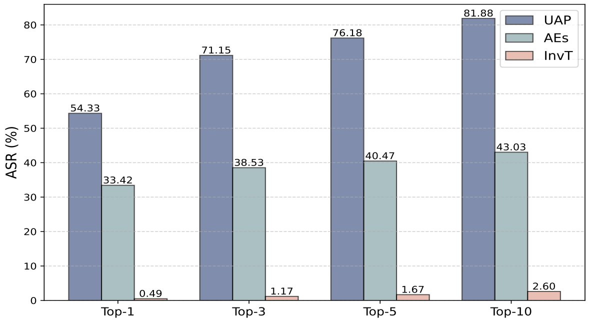


Fig. 9: Backdoor removal performance on universal adversarial examples (UAP), targeted-label adversarial examples (AEs), and inverted backdoor samples (InvT).

To test this hypothesis against a more rigorous baseline, we replace the generic UAP with targeted adversarial examples (AEs). Following [20], we generate PGD-based AEs aimed at the "mushroom" class. The results in Figure 9 reveal a clear performance hierarchy: while targeted AEs (ASR 33.42%) are more effective than UAP (ASR 54.33%), they still fall significantly short of our method (ASR 0.49%). This consistent trend, **InvT** $\gg$ **AEs** $>$ **UAP**, provides compelling evidence that the backdoor triggers recovered via inversion constitute a distinct phenomenon from adversarial examples. It confirms that our specialized trigger inversion step is not merely beneficial but is, in fact, indispensable for effective and robust backdoor defense.

VII. Conclusion

In this paper, we present InverTune, a novel backdoor defense framework for large-scale multimodal contrastive learning models. Our approach integrates three key components: adversarial-based target label identification, gradient-guided trigger inversion, and activation-aware fine-tuning. Extensive evaluations on multiple datasets demonstrate that InverTune achieves state-of-the-art defensive performance across diverse attack scenarios, consistently reducing attack success rates while maintaining model utility. Our framework significantly enhances the robustness of multimodal models against backdoor threats, providing a practical solution for real-world applications.

Acknowledgment

This work was supported by the NSFC under Grants U2441240 ("Ye Qisun" Science Foundation), 62441238, 62441237, and 62502350, the Postdoctoral Fellowship Program (Grade B) of China Postdoctoral Science Foundation under Grants GZB20250384, and Hubei Province Postdoctoral Advanced Talent Introduction Project under Grants 2004HBB-HJD066.

References

[1] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proc. of IEEE CVPR*, pages 24239–24250, 2024.

[2] Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, and Kai-Wei Chang. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *Proc. of IEEE ICCV*, pages 112–123, 2023.

[3] Mauro Barni, Kassem Kallas, and Benedetta Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *Proc. of IEEE ICIP*, pages 101–105, 2019.

[4] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *Proc. of IEEE CVPR*, pages 4662–4670, 2022.

[5] Min Cao, Yang Bai, Ziyin Zeng, Mang Ye, and Min Zhang. An empirical study of clip for text-based person search. In *Proc. of AAAI*, pages 465–473, 2024.

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

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

[8] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *Proc. of IEEE ICCV*, pages 699–710, 2023.

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

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proc. of ECCV*, pages 104–120, 2020.

[11] Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. Fine-grained image captioning with clip reward. In *Proc. of NAACL*, pages 517–527, 2022.

[12] Anders Christensen, Massimiliano Mancini, A Koepke, Ole Winther, and Zeynep Akata. Image-free classifier injection for zero-shot classification. In *Proc. of IEEE ICCV*, pages 19072–19081, 2023.

[13] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

[14] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Proc. of EACL*, pages 1181–1193, 2023.

[15] 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 *Proc. of IEEE CVPR*, pages 16352–16362, 2023.

[16] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proc. of ACSAC*, pages 113–125, 2019.

[17] Xueluan Gong, Yanjiao Chen, Wang Yang, Qian Wang, Yuzhe Gu, Huayang Huang, and Chao Shen. Redeem myself: Purifying backdoors in deep learning models using self attention distillation. In *Proc. of IEEE S&P*, pages 755–772, 2023.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

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

[20] Harshit Gupta, Kyong Hwan Jin, Ha Q Nguyen, Michael T McCann, and Michael Unser. Cnn-based projected gradient descent for consistent ct image reconstruction. *IEEE transactions on medical imaging*, pages 1440–1453, 2018.

[21] Xijie Huang, Moustafa Alzantot, and Mani Srivastava. Neuroninspect: Detecting backdoors in neural networks via output explanations. *arXiv preprint arXiv:1911.07399*, 2019.

[22] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Proc. of NeurIPS*, 2019.

[23] 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 *Proc. of ICML*, pages 4904–4916, 2021.

[24] Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *Proc. of IEEE S&P*, pages 2043–2059, 2022.

[25] Junhao Kuang, Siyuan Liang, Jiawei Liang, Kuanrong Liu, and Xiaochun Cao. Adversarial backdoor defense in clip. *arXiv preprint arXiv:2409.15968*, 2024.

[26] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. 2018.

[27] Janghyeon Lee, Jongsuk Kim, Hyounguk Shon, Bumsoo Kim, Seung Hwan Kim, Honglak Lee, and Junmo Kim. Uniclip: Unified framework for contrastive language-image pre-training. In *Proc of NeurIPS*, pages 1008–1019, 2022.

[28] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proc. of AAAI*, pages 11336–11344, 2020.

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742, 2023.

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022.

[31] Maosen Li, Yanhua Yang, Kun Wei, Xu Yang, and Heng Huang. Learning universal adversarial perturbation by adversarial example. In *Proc. of AAAI*, pages 1350–1358, 2022.

[32] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *arXiv preprint arXiv:2101.05930*, 2021.

[33] 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 *Proc. of IEEE CVPR*, pages 24645–24654, 2024.

[34] 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 *Proc. of ECCV*, pages 740–755, 2014.

[35] Hong Liu, Rongrong Ji, Jie Li, Baochang Zhang, Yue Gao, Yongjian Wu, and Feiyue Huang. Universal adversarial perturbation via prior driven uncertainty approximation. In *Proc. of IEEE ICCV*, pages 2941–2949, 2019.

[36] Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proc. of CCS*, pages 1265–1282, 2019.

[37] Christian Lülf, Denis Mayr Lima Martins, Marcos Antonio Vaz Salles, Yongluan Zhou, and Fabian Gieseke. Clip-branches: Interactive fine-tuning for text-image retrieval. In *Proc. of the 47th ACM SIGIR*, pages 2719–2723, 2024.

[38] James MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–298. University of California press, 1967.

[39] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, pages 39–41, 1995.

[40] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[41] Bingxu Mu, Zhenxing Niu, Le Wang, Xue Wang, Qiguang Miao, Rong Jin, and Gang Hua. Progressive backdoor erasing via connecting

backdoor and adversarial attacks. In *Proc. of IEEE CVPR*, pages 20495–20503, 2023.

[42] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Proc. of ECCV*, pages 529–544, 2022.

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

[44] Zhenxing Niu, Yuyao Sun, Qiguang Miao, Rong Jin, and Gang Hua. Towards unified robustness against both backdoor and adversarial attacks, 2024.

[45] Zhenxing Niu, Yuyao Sun, Qiguang Miao, Rong Jin, and Gang Hua. Towards unified robustness against both backdoor and adversarial attacks. *IEEE TPAMI*, 2024.

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

[47] Qi Qian and Juhua Hu. Online zero-shot classification with clip. In *Proc. of ECCV*, pages 462–477, 2024.

[48] 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 *ICML*, pages 8748–8763, 2021.

[49] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, pages 211–252, 2015.

[50] 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 ACL*, pages 2556–2565, 2018.

[51] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proc. of IEEE CVPR*, pages 15638–15650, 2022.

[52] Naman Deep Singh, Francesco Croce, and Matthias Hein. Perturb and recover: Fine-tuning for effective backdoor removal from clip. *arXiv preprint arXiv:2412.00727*, 2024.

[53] Stefan Smeu, Elisabeta Oneata, and Dan Oneata. Declip: Decoding clip representations for deepfake localization. In *Proc. of IEEE WACV*, pages 149–159, 2025.

[54] Ruixiang Tang, Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. An embarrassingly simple approach for trojan attack in deep neural networks. In *Proc. of ACM SIGKDD*, pages 218–228, 2020.

[55] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Proc. of NeurIPS*, 2018.

[56] Sakshi Udeshi, Shanshan Peng, Gerald Woo, Lionell Loh, Louth Rawshan, and Sudipta Chattopadhyay. Model agnostic defence against backdoor attacks in machine learning. *IEEE Transactions on Reliability*, pages 880–895, 2022.

[57] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proc. of IEEE CVPR*, pages 3156–3164, 2015.

[58] 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 *Proc. of IEEE S&P*, pages 707–723, 2019.

[59] Qiannan Wang, Changchun Yin, Liming Fang, Zhe Liu, Run Wang, and Chenhao Lin. Ghostencoder: Stealthy backdoor attacks with dynamic triggers to pre-trained encoders in self-supervised learning. *Computers & Security*, page 103855, 2024.

[60] Zhenting Wang, Kai Mei, Juan Zhai, and Shiqing Ma. UNICORN: A unified backdoor trigger inversion framework. In *Proc. of ICLR*, 2023.

[61] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, pages 600–612, 2004.

[62] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proc. of IEEE CVPR*, pages 3733–3742, 2018.

[63] Kaidi Xu, Sijia Liu, Pin-Yu Chen, Pu Zhao, and Xue Lin. Defending against backdoor attack on deep neural networks. *arXiv preprint arXiv:2002.12162*, 2020.

[64] Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *Proc. of ICLR*, 2024.

[65] Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning. abs/2409.17601, 2024.

[66] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Better safe than sorry: Pre-training clip against targeted data poisoning and backdoor attacks. *arXiv preprint arXiv:2310.05862*, 2023.

[67] Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. In *Proc. of NeurIPS*, pages 10678–10691, 2023.

[68] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning. In *Proc. of CVPR*, pages 7140–7148, 2017.

[69] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[70] Chaoning Zhang, Philipp Benz, Adil Karjauv, and In So Kweon. Data-free universal adversarial perturbation and black-box attack. In *Proc. of IEEE ICCV*, pages 7868–7877, 2021.

[71] Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, pages 2578–2593, 2019.

[72] Zhifang Zhang, Shuo He, Bingquan Shen, and Lei Feng. Defending multimodal backdoored models by repulsive visual prompt tuning. *arXiv preprint arXiv:2412.20392*, 2024.

[73] Zhuokai Zhao, Harish Palani, Tianyi Liu, Lena Evans, and Ruth Toner. Multimodal guidance network for missing-modality inference in content moderation. In *2024 IEEE ICMEW*, pages 1–4, 2024.

[74] 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. In *Proc. of ECCV*, pages 405–421, 2024.

[75] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, pages 2337–2348, 2022.

[76] Ziqi Zhou, Shengshan Hu, Minghui Li, Hangtao Zhang, Yechao Zhang, and Hai Jin. Advclip: Downstream-agnostic adversarial examples in multimodal contrastive learning. In *Proc. of ACM MM*, pages 6311–6320, 2023.

[77] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proc. of IEEE CVPR*, pages 11175–11185, 2023.

[78] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proc. of IEEE ICCV*, pages 2639–2650, 2023.

Appendix A
Quantitative Evidence for Observations

To provide rigorous support for our two observations (Obs.) in Section III-B, we conduct quantitative analyses across multiple attacks and architectures, as shown in Table VIII .

For Observation I, we measure the spatial proximity between backdoor and target-class samples. The nearest-to-target ratio remains very low (0∼0.10%), with p-values (0.63∼1.0) indicating no significant clustering. The small feature shift (3∼6%) further indicates that backdoor samples only slightly approach target embeddings. These results confirm that MCL backdoor attacks succeed without forming a visually coherent cluster with the target class, highlighting the importance of cross-modal effects rather than direct feature mimicry.

For Observation II, we analyze the relation between backdoor samples and adversarial examples via pairwise feature and output statistics. The dispersion coefficients, measured by the coefficient of variation (CV = 0.15∼0.29), reveal substantial variability in pairwise distances, especially for SIG and BadCLIP on RN50, showing that adversarial examples do not

TABLE VIII: Quantitative results supporting Observation I and Observation II across different attacks and architectures.

Obs.	Metric	BadNet		SIG		BadCLIP	
		RN50	ViT-B/32	RN50	ViT-B/32	RN50	ViT-B/32
I	Nearest to target (%)	0.10	0	0	0.10	0.10	0
	p-value	0.632	1.0	1.0	0.632	0.632	1.0
	Feature shift (%)	6.01	5.77	3.24	3.19	4.30	5.94
II	Dispersion (CV)	0.152	0.212	0.279	0.281	0.294	0.173
	KL divergence ($\times 10^{-4}$)	3.85	1.30	2.94	1.06	0.43	2.18

converge toward trigger features. Meanwhile, their KL divergences of output probabilities ($0.43 \sim 3.85 \times 10^{-4}$) are orders of magnitude below similarity thresholds, indicating nearly identical prediction behaviors. Overall, the results suggest adversarial attacks leverage backdoor-induced vulnerabilities, traversing distinct feature-space paths yet producing identical target predictions.

Appendix B
Theoretical Analysis of UAP Convergence

A. *Proof of Theorem 1*

The optimization objective is to maximize the loss function $\mathcal{L}_{UAP}(r)$ as defined in Eq. (2). Let us analyze its dominant term, $\max_{k \neq y} \text{sim}(\hat{E}_I(x+r), \hat{E}_T(t_k))$.

Due to the structural vulnerability introduced by the backdoor (Assumption 1), the model exhibits a global hypersensitivity to the target text embedding $\hat{E}_T(t_l)$. For any small perturbation r that shifts an image embedding even slightly in the direction of v_{bd}, the resulting gain in similarity with $\hat{E}_T(t_l)$ will be substantially greater thanthe gain in similarity with any other text embedding $\hat{E}_T(t_k)$ where $k \neq l$ and $k \neq y$.

Therefore, for almost all images x and any effective perturbation r, the maximization term will be dominated by the backdoor's target class l as shown in Eq. (5). This is because the backdoor has created a "path of least resistance", making the target class l the easiest "wrong" class to achieve. By substituting Eq. (5) into Eq. (2), the optimization problem for the UAP can be simplified to approximately Eq. (6). Note that the second term from Eq. (2), $-\text{sim}(\hat{E}_I(x+r), \hat{E}_T(t_y))$, is naturally suppressed when the first term is maximized and can be omitted from this high-level analysis.

The objective in Eq. (6) is to find a single perturbation r that systematically aligns the embeddings of all perturbed images, $\hat{E}_I(x+r)$, with a single, fixed target vector, the backdoor's target text embedding $\hat{E}_T(t_l)$.

To achieve this, the perturbation r must displace the origin image embeddings $\hat{E}_I(x)$, which are widely distributed throughout the embedding space, towards a common direction. This direction is precisely the one defined by the structural vulnerability, v_{bd}. The optimization process will thus converge to a solution r^* that produces the effect of v_{bd}, making r^* functionally equivalent to the original trigger P_{img}.

This concludes the proof. Although UAP optimization is nominally non-targeted, its search for a globally effective perturbation exploits the model's principal weakness. In a backdoored model, this weakness is the backdoor itself, so the optimization effectively rediscovers it and yields a perturbation that drives inputs to the attacker's target class.

Appendix C
Configurations of Different Backdoor Attacks

For all six types of attacks, we adopt a 500K subset of the CC3M dataset [50] as the fine-tuning dataset. All attacks target the class "mushroom." For attacks that require textual descriptions, we construct them by collecting **131 mushroom-related captions** from the CC3M dataset and randomly assigning them to the poisoned image samples as their corresponding text descriptions.

- BadNet [19]: A 16 × 16 Gaussian noise patch (standard normal distribution) is fixed to the bottom-right corner of clean images as the trigger.
- Blended [9]: A full-size trigger image (uniform distribution) is blended with the clean image at 0.2 transparency (clean image: 0.8).
- SIG [3]: Sinusoidal noise (6 cycles/image width) is applied along the horizontal axis, scaled to $60/255$ across all RGB channels, with pixel values clipped to $[0, 1]$.
- WaNet [43]: A warping transformation uses a distortion grid, interpolated from a noise tensor, scaled/clipped to $[-1, 1]$, and applied via bilinear interpolation.
- BadEncoder [24]: The visual encoder is fine-tuned with a 16 × 16 pure white trigger (replacing the original trigger), using reference and shadow datasets, without textual descriptions or a poisoning rate.
- BadCLIP [33]: A patch optimized for the "mushroom" label is used, followed by a Dual-Embedding injection attack on the clean CLIP model.

All attacks begin with the OpenAI-pretrained CLIP model [48], fine-tuned with a learning rate of 1e-6, batch size of 128, and 10 epochs to create a poisoned CLIP model.

Appendix D
Baseline Defense Settings

All the baseline defense methods use subsets of the CC3M dataset [50], which is the same corpus used for poisoning, in their original setups, though the exact number of clean samples varies slightly. For fair comparison, we standardize the training data by using a fixed subset of 500K samples across all methods. In contrast, InverTune requires only 50K clean samples sourced from ImageNet, which is entirely independent of the poisoned training corpus.

- The fine-tuning method (FT), first introduced by CleanCLIP [2], involves fine-tuning the model with a multimodal contrastive loss on a clean dataset. In our experiments, we use the official implementation provided by CleanCLIP, with a learning rate of 4.5e-6, warmup steps of 50, batch size of 64, and 10 training epochs.
- CleanCLIP [2] extends FT by adding a self-supervised loss. Following its original setup, we set the weights of the self-supervised loss and the contrastive loss to 1, with other hyperparameters remaining the same as those in FT.

TABLE IX: Top-2 classes with the largest absolute increase under adversarial attack. Clean and adversarial counts are shown as "Clean→Adv.".

Attack	Top-1			Top-2		
	Class	Clean→Adv.	Δ(%)	Class	Clean→Adv.	Δ(%)
BadNet	mush.	18→19981	+39.93	echidna	34→11675	+23.28
Blended	mush.	18→30547	+61.06	doormat	64→2442	+4.76
SIG	mush.	76→18775	+37.40	agaric	14→8808	+17.59
WaNet	mush.	18→13206	+26.38	agaric	60→13056	+25.99
BadEncoder	mush.	422→2522	+4.20	pillow	94→384	+0.58
BadCLIP	mush.	6→48623	+97.23	agaric	67→1132	+2.13

- PAR [52] adopts a custom learning rate schedule. However, due to the increased size of the fine-tuning dataset, the original setting does not reproduce the reported performance. Therefore, in our experiments, we modify the start learning rate to 3e-6 and the peak learning rate to 5e-6, while keeping all other parameters consistent with the original setup.
- CleanerCLIP [65] is implemented based on CleanCLIP [2]. We follow its setup, using batch size of 64 and training for 10 epochs with the AdamW optimizer. The learning rate is linearly warmed up over 10,000 steps, and a weight decay of 0.1 is applied. The Adam momentum factor and RMSProp factor are set to 0.9 and 0.999, respectively, with an epsilon of 1e-8. The base learning rate is set to 4.5e-6.

APPENDIX E
DETAILED RESULTS OF STEPS IN INVERTUNE

A. Target Category Identification Results for Six Attacks

We evaluate target class identification across six attack scenarios, all using "mushroom" as the ground-truth target. As shown in Table IX, adversarial perturbations consistently increase the target class's prediction frequency, with attack-specific variations in magnitude.

The most pronounced shifts occur in BadCLIP (+97.23%), Blended (+61.06%), BadNet (+39.33%), and SIG (+37.40%), indicating strong target bias. Even BadEncoder (+4.20%) shows a statistically significant increase, maintaining a 3.62-point margin over the next most frequent class "pillow" (+0.58%). WaNet exhibits a distinctive taxonomic vulnerability, producing nearly identical increases for "mushroom" (+26.38%) and its related class "agaric" (+25.99%), differing by only 0.39 points. This close correspondence supports the link between perturbation features and the backdoor's target semantics discussed in Section V-C.

B. Key Layers Selected in Activation Tuning

This section presents the layer selection results from the Activation Tuning process. Because different attacks yield similar activation patterns, we report the anomalous response layers for four CLIP architectures under inversion triggers, using BadCLIP as a representative example.

For ResNet-based visual encoders, which contain four residual stages, our analysis shows that backdoor sensitivity is concentrated in the final residual layers. As shown in Table X, RN50 exhibits extreme sensitivity in visual.layer4 with an impact value of 1.3802, which exceeds the significance threshold ($\mu + \sigma = 0.9914$) by 39.2%. Similarly, RN101's visual.layer4 shows comparable vulnerability. This final-layer concentration suggests that ResNet defenses can focus on monitoring these critical bottlenecks.

TABLE X: Layer impact for RN50 / RN101.

Layer	**RN50**		**RN101**	
	Impact	Key Layer	Impact	Key Layer
visual.layer1	0.1407	No	0.1329	No
visual.layer2	0.1750	No	0.1492	No
visual.layer3	0.1435	No	0.1547	No
visual.layer4	**1.3802**	**Yes**	**1.1823**	**Yes**
Sig. Threshold		0.9914		0.8538
Mean		0.4599		0.4048
Std		0.5315		0.4490

TABLE XI: Layer impact for ViT-B/16 / ViT-B/32.

Layer	**ViT-B/16**		**ViT-B/32**	
	Impact	Key Layer	Impact	Key Layer
ViT Blocks.0	0.0916	No	0.0965	No
ViT Blocks.1	0.1458	No	0.2025	No
ViT Blocks.2	0.2858	No	0.3066	No
ViT Blocks.3	**0.3423**	**Yes**	**0.3782**	**Yes**
ViT Blocks.4	**0.3247**	**Yes**	**0.3609**	**Yes**
ViT Blocks.5	**0.2996**	**Yes**	0.3085	No
ViT Blocks.6	0.1895	No	0.2558	No
ViT Blocks.7	0.1879	No	0.2628	No
ViT Blocks.8	0.2002	No	0.2463	No
ViT Blocks.9	0.1745	No	0.2172	No
ViT Blocks.10	0.1959	No	0.2334	No
ViT Blocks.11	0.1700	No	0.1367	No
Sig. Threshold		0.9914		0.3298
Mean		0.2173		0.2504
Std		0.0742		0.0794

The CLIP visual encoder using the ViT-B architecture consists of 12 Transformer blocks, from which we identify key layers for analysis. Transformer architectures display fundamentally different response patterns characterized by distributed sensitivity across middle layers. As shown in Table XI, ViT-B/16 shows consistent anomalous responses in blocks 3-5 (0.3423, 0.3247, 0.2996) that all exceed the threshold of 0.2915. Similarly, the ViT-B/32 architecture reveals similar distributed sensitivity, with blocks 3-4 showing the strongest deviations (0.3782, 0.3609), surpassing the threshold of 0.3298 by 14.7% and 9.4%, respectively. The pattern reflects global dependencies in attention, requiring multi-block rather than single-point defenses.

Our $\mu + \sigma$ criterion shows consistent performance across architectures, with all key layers deviating notably and a clear normal–anomalous separation ($\leq$9.4%), confirming its reliability for architecture-agnostic backdoor analysis.

C. Computational Efficiency and Scalability

InverTune completes within approximately 2 hours on a single RTX 4090 GPU ($\approx$20 min for UAP generation, 30

TABLE XII: InverTune under different poisoning rates.

Poisoning Rate	0.05%		0.1%		0.3%		1.0%	
	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓	CA ↑	ASR ↓
No Defense	54.36	65.53	55.35	88.34	58.32	98.36	58.12	98.80
InverTune	53.77	0.05	54.80	0.36	55.25	0.49	56.52	0.81

min for target identification, 1 hour for trigger inversion, and 10 min for activation-based purification), while updating only a small subset of parameters. Unlike full fine-tuning–based defenses [2], [52], [65], the computational cost of InverTune scales weakly with model size, since both inversion and purification are confined to limited neuron subsets rather than the entire model, indicating that InverTune exhibits potential for application to larger models.

Appendix F
Additional Experiments and Analysis

A. The Impact of Model CA on InverTune Performance

Although the CLIP-RN50 backbone achieves a relatively low CA of 55∼60% on ImageNet, consistent with prior works [2], [33], certain backdoor attacks still exhibit extremely high ASR (e.g., BadCLIP reaching 98.36%). This contrast raises the question of whether such strong target bias could facilitate or hinder InverTune's ability to identify the target label. To examine this, we fine-tuned the same model on 50K randomly selected ImageNet samples, increasing its CA to over 70%, and evaluated three representative attacks: BadNet, WaNet, and BadCLIP. Even under this higher-CA setting, InverTune accurately identified the target label and effectively removed the backdoor, reducing ASR from 96.31%, 93.34%, and 93.61% to 0.13%, 0.44%, and 0.51%, respectively, with only a minor CA drop of about 2∼3%. These results confirm that InverTune's effectiveness is largely independent of the model's CA, and that its success primarily stems from adversarial–backdoor correlations rather than the model's discriminative capacity.

B. Performance of InverTune under Different Poisoning Rates

To further assess the applicability of InverTune, we conducted experiments with BadCLIP across various poisoning rates: 0.01%, 0.05%, 0.1%, 0.3% (default), and 1.0%. At a very low poisoning rate of 0.01%, the small number of poisoned samples (approximately 50) resulted in an ASR of only 1.30%, as the model struggled to form a stable backdoor mapping. Consequently, InverTune failed to identify the target label due to the weak backdoor signal, which hindered the establishment of a stable adversarial–backdoor correlation.

At other poisoning rates, InverTune demonstrated robust performance, effectively mitigating the attack. As shown in Table XII, even at 1.0%, InverTune successfully reduced the ASR to 0.81%. Overall, InverTune exhibits consistent defense performance across the evaluated poisoning rates, though its effectiveness may be limited when the backdoor signal is extremely weak, underscoring the boundaries of the method under such conditions.

TABLE XIII: Performance of InverTune against multi-backdoor attacks. R1/R2: first/second round of tuning.

Setting	**Method**	**ASR ↓**				**CA ↑**
		mush.	lemon	ski	swing	
2-trigger	No Defense	98.11	98.72	-	-	58.39
	InverTune (R1)	0.62	2.20	-	-	56.64
4-trigger	No Defense	99.54	99.76	99.41	99.50	58.40
	InverTune (R1)	23.47	20.89	0.72	22.08	56.43
	InverTune (R2)	2.34	0.89	0.08	1.36	54.81

C. Performance of InverTune on Clean Models

We further evaluate InverTune on a clean CLIP model without backdoors. As expected, the UAPs generated during target identification are random, with dominant predicted classes varying across runs (e.g., *wool* vs. *desktop computer*), confirming the absence of consistent target bias. During trigger inversion, optimization fails to produce meaningful patterns, yielding negligible ASR (0.06% for *wool* and 0.17% for *desktop computer*). Applying these inverted triggers during activation tuning results in only marginal CA changes (from 59.69% to 59.13% and 59.02%), indicating that InverTune introduces virtually no adverse effects on clean models. Inspired by the pronounced behavioral contrast between clean and backdoored models, we believe that this benign behavior of InverTune on clean models may provide useful signals for future backdoor detection efforts.

D. Performance of InverTune in Multi-Backdoor Scenarios

We further evaluate InverTune in multi-backdoor scenarios by extending the BadCLIP attack to both 2-backdoor and 4-backdoor configurations. In the 2-backdoor setting, the attacker implants triggers targeting "mushroom" and "lemon," while the 4-backdoor variant additionally includes "ski" and "swing," each with a poisoning rate of 0.3%. As shown in Table XIII, InverTune consistently suppresses ASR to low levels across all targeted classes while maintaining stable CA. More specifically, in the 2-backdoor case, the target identification stage ranks "mushroom" and "lemon" as the top two predicted classes, exactly matching the implanted targets. We further observe that the inverted trigger reconstructed for "mushroom" produces strong activation not only for its own target (85.56%, 92.89%, 94.82%, 96.81% for Top-1/3/5/10), but also for "lemon" (3.36%, 77.90%, 90.64%, 96.29%), revealing clear cross-target interference and shared backdoor characteristics. This cross-target activation pattern enables efficient purification: a single tuning round is sufficient to reduce the ASR of the 2-backdoor model to a low level, whereas only two rounds are needed to suppress all triggers in the 4-backdoor configuration to similarly low ASR values.

Overall, these results demonstrate that InverTune adapts effectively to increasingly complex poisoning scenarios. Its ability to accurately identify, invert, and mitigate multiple co-existing backdoors highlights strong robustness and scalability in realistic multi-target threat scenarios.