BDETCLIP: MULTIMODAL PROMPTING CONTRASTIVE TEST-TIME BACKDOOR DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal contrastive learning methods (e.g., CLIP) have shown impressive zero-shot classification performance due to their strong ability to joint representation learning for visual and textual modalities. However, recent research revealed that multimodal contrastive learning on *poisoned* pre-training data with a small proportion of maliciously backdoored data can induce backdoored CLIP that could be attacked by inserted triggers in downstream tasks with a high success rate. To defend against backdoor attacks on CLIP, existing defense methods focus on either the pre-training stage or the fine-tuning stage, which would unfortunately cause high computational costs due to numerous parameter updates and are not applicable in the black-box setting. In this paper, we provide the first attempt at a computationally efficient backdoor detection method to defend against backdoored CLIP in the *inference* stage. We empirically find that the visual representations of backdoored images are *insensitive* to both *benign* and *malignant* changes in class description texts. Motivated by this observation, we propose BDetCLIP, a novel test-time backdoor detection method based on contrastive prompting. Specifically, we first prompt the language model (e.g., GPT-4) to produce class-related description texts (benign) and class-perturbed random texts (malignant) by specially designed instructions. Then, the distribution difference in cosine similarity between images and the two types of class description texts can be used as the criterion to detect backdoor samples. Extensive experiments validate that our proposed BDetCLIP is superior to state-of-the-art backdoor detection methods, in terms of both effectiveness and efficiency.

1 INTRODUCTION

Multimodal contrastive learning methods (e.g., CLIP (Radford et al., 2021)) have shown impressive zero-shot classification performance in various downstream tasks and served as foundation models in various vision-language fields due to their strong ability to effectively align representations from different modalities, such as open-vocabulary object detection (Wu et al., 2023), text-to-image generation (Wu et al., 2023), and video understanding (Xu et al., 2021). However, recent research has revealed that a small proportion of backdoor samples poisoned into the pre-training data can cause a backdoored CLIP after the multimodal contrastive pre-training procedure (Carlini & Terzis, 2021; Carlini et al., 2023; Bansal et al., 2023). In the inference stage, a backdoored CLIP would produce tampered image representations for images with a trigger, close to the text representation of the target attack class in zero-shot classification. This exposes a serious threat when deploying CLIP in real-world applications.

To overcome this issue, effective defense methods have been proposed recently, which can be divided into three kinds of defense paradigms, as shown in Figure 1: including (a) robust anti-backdoor contrastive learning in the pre-training stage (Yang et al., 2023b), (b) counteracting the backdoor in a pre-trained CLIP in the fine-tuning stage (Bansal et al., 2023), (c) leveraging trigger inversion techniques to decide if a pre-trained CLIP is backdoored (Sur et al., 2023; Feng et al., 2023a). Overall, these defense methods have a high computational cost due to the need for additional learning or optimization procedures. In contrast, we advocate *test-time* backdoor sample detection (Figure 1(d)), which is a more computationally efficient defense against backdoored CLIP, as there are no parameter updates in the inference stage. Intuitively, it could be feasible to directly adapt existing *unimodal* test-time detection methods (Gao et al., 2019; Guo et al., 2023; Liu et al., 2023) to detect backdoored

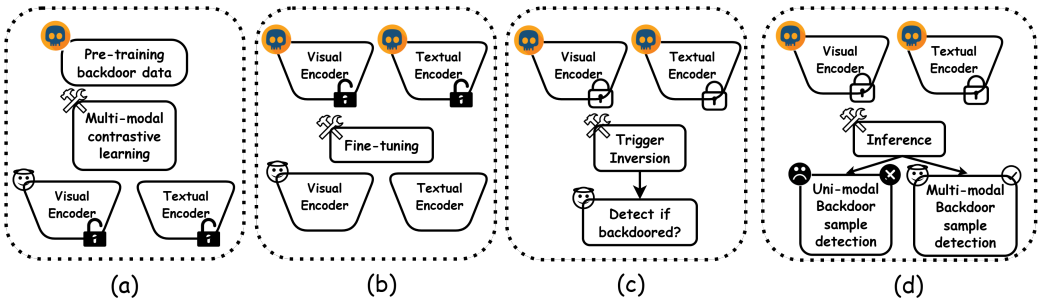

Figure 1: Current backdoor defense paradigms in CLIP. (a) Robust anti-backdoor contrastive learning (Yang et al., 2023b); (b) Fine-tuning a backdoored CLIP (Bansal et al., 2023); (c) Detecting a CLIP if backdoored (Sur et al., 2023; Feng et al., 2023a); (d) Our test-time backdoor sample detection. Our multimodal detection method is more effective and efficient than existing unimodal detection methods.

images in CLIP, since they can differentiate backdoored and clean images generally based on the output consistency in the visual representation space by employing specific image modifications, e.g, corrupting (Liu et al., 2023), amplifying (Guo et al., 2023), and blending (Gao et al., 2019). However, the performance of these unimodal detection methods is suboptimal, because of lacking the utilization of the text modality in CLIP to assist backdoor sample detection. Hence we can expect that better performance could be further achieved if we leverage both image and text modalities simultaneously.

In this paper, we provide the first attempt at a computationally efficient backdoor detection method to defend against backdoored CLIP in the *inference* stage. We empirically find that the visual representations of backdoored images are *insensitive* to both *benign* and *malignant* changes of class description texts. Motivated by this observation, we propose BDetCLIP, a novel test-time multimodal backdoor detection method based on contrastive prompting. Specifically, we first prompt the GPT-4 model (Achiam et al., 2023) to generate class-related (or class-perturbed random) description texts by specially designed instructions and take them as benign (malignant) class prompts. Then, we calculate the distribution difference in cosine similarity between images and the two types of class prompts, which can be used as a good criterion to detect backdoor samples. We can see that the distribution difference of backdoored images between the benign and malignant changes of class prompts is smaller than that of clean images. The potential reason for the insensibility of backdoored images is that their visual representations have less semantic information aligned with class description texts. In this way, we can detect backdoored images in the inference stage of CLIP effectively and efficiently. Extensive experiments validate that our proposed BDetCLIP is superior to state-of-the-art backdoor detection methods, in terms of both effectiveness and efficiency.

Our main contributions can be summarized as follows:

- *A new backdoor detection paradigm for CLIP.* We pioneer test-time backdoor detection for CLIP, which is more computationally efficient than existing defense paradigms.
- *A novel backdoor detection method.* We propose a novel test-time multimodal backdoor detection method based on contrastive prompting, which detects backdoor samples based on the distribution difference between images regarding the benign and malignant changes of class prompts.
- *Strong experimental results.* Our proposed method achieves superior experimental results on various types of backdoored CLIP compared with state-of-the-art detection methods.

2 Background & Preliminaries

2.1 Multimodal Contrastive Learning

Multimodal contrastive learning (Radford et al., 2021; Jia et al., 2021) has emerged as a powerful approach for learning shared representations from multiple modalities of data such as text and images. Specifically, we focus on **C**ontrastive **L**anguage **I**mage **P**retraining (CLIP) (Radford et al., 2021) in this paper. Concretely, CLIP consists of a visual encoder denoted by $\mathcal{V}(\cdot)$ (e.g., ResNet (He et al., 2016) and ViT (Dosovitskiy et al., 2020)) and a textual encoder denoted by $\mathcal{T}(\cdot)$ (e.g., Transformer

(Vaswani et al., 2017)). The training examples used in CLIP are massive image-text pairs collected on the Internet denoted by $\mathcal{D}_{\text{Train}} = \{(\boldsymbol{x}_i, \boldsymbol{t}_i)\}_{i=1}^{N}$ where $\boldsymbol{t}_i$ is the caption of the image $\boldsymbol{x}_i$ and $N \simeq 400M$. During the training stage, given a batch of N_b image-text pairs $(\boldsymbol{x}_i, \boldsymbol{t}_i) \subset \mathcal{D}_{\text{Train}}$, the cosine similarity for matched (unmatched) pairs is denoted by $\phi(\boldsymbol{x}_i, \boldsymbol{t}_i) = \cos(\mathcal{V}(\boldsymbol{x}_i), \mathcal{T}(\boldsymbol{t}_i))$ ($\phi(\boldsymbol{x}_i, \boldsymbol{t}_j) = \cos(\mathcal{V}(\boldsymbol{x}_i), \mathcal{T}(\boldsymbol{t}_j))$). It is noteworthy that the image and text embeddings are normalized using the ℓ_2 norm to have a unit norm. Based on these notations, the CLIP loss can be formalized by the following (Radford et al., 2021):

$$\mathcal{L}_{\text{CLIP}} = -\frac{1}{2N_b}\Big(\sum_{i=1}^{N_b}\log\Big[\frac{\exp(\phi(\boldsymbol{x}_i, \boldsymbol{t}_i)/\tau)}{\sum_{j=1}^{N_b}\exp(\phi(\boldsymbol{x}_i, \boldsymbol{t}_j)/\tau)}\Big] + \sum_{j=1}^{N_b}\log\Big[\frac{\exp(\phi(\boldsymbol{x}_j, \boldsymbol{t}_j)/\tau)}{\sum_{i=1}^{N_b}\exp(\phi(\boldsymbol{x}_i, \boldsymbol{t}_j)/\tau)}\Big]\Big), \quad (1)$$

where τ is a trainable temperature parameter.

Zero-shot classification in CLIP. To leverage CLIP on the downstream classification task where the input image $\boldsymbol{x} \in D_{\text{Test}}$ and class name $y_i \in \{1, 2, \cdots, c\}$, a simple yet effective way is using a class template function $T(j)$ which generates a class-specific text such as "`a photo of [CLS]`" where `[CLS]` can be replaced by the j-th class name on the dataset. In the inference stage, one can directly calculate the posterior probability of the image x for the i-th class as the following:

$$p(y = i|\boldsymbol{x}) = \frac{\exp(\phi(\boldsymbol{x}, T(i))/\tau)}{\sum_{j=1}^{c}\exp(\phi(\boldsymbol{x}, T(j))/\tau)}. \quad (2)$$

In this way, CLIP can achieve impressive zero-shot performance, even compared with unimodal vision models trained by (self) supervised learning methods.

Moreover, since CLIP only considers the simple and coarse alignment between images and texts in Eq. (1), many follow-up studies focus on more fine-grained and consistent alignment strategies such as SLIP (Mu et al., 2022), Uniclip (Lee et al., 2022), Cyclip (Goel et al., 2022), PROMU (Hu et al., 2023), and RA-CLIP (Xie et al., 2023). On the other hand, using naive class prompts generated by $T(j)$ in Eq. (2) in zero-shot image classification might not take full advantage of the strong representation learning ability of CLIP on the text modality. This means that more well-described class-specific prompts may be more beneficial to image classification. To this end, recent research delves into engineering fine-grained class-specific attributes or prompting large language models (e.g., GPT-4 (Achiam et al., 2023)) to generate distinguishable attribute-related texts (Yang et al., 2023c; Pratt et al., 2023; Maniparambil et al., 2023; Yu et al., 2023; Saha et al., 2024; Feng et al., 2023b; Liu et al., 2024).

2.2 Backdoor Attacks and Defenses

The backdoor attack is a serious security threat to machine learning systems (Li et al., 2022; Carlini & Terzis, 2021; Xu et al., 2022; Chen et al., 2021). The whole process of a backdoor attack can be expounded as follows. In the data collection stage of a machine learning system, a malicious adversary could manufacture a part of backdoor samples with the imperceptible trigger poisoned into the training dataset. After the model training stage, the hidden trigger could be implanted into the victim model without much impact on the performance of the victim model. During the inference stage, the adversary could manipulate the victim model to produce a specific output by adding the trigger to the clean input. Early research on backdoor attacks focuses on designing a variety of triggers that satisfy the practical scenarios mainly on image and text classification tasks including invisible stealthy triggers (Chen et al., 2017; Turner et al., 2019; Li et al., 2021a; Doan et al., 2021; Nguyen & Tran, 2021; Gao et al., 2023; Souri et al., 2022) and physical triggers (Chen et al., 2017; Wenger et al., 2021). To defend against these attacks, many backdoor defense methods are proposed which can be divided into four categories, mainly including data cleaning in the pre-processing stage (Tran et al., 2018), robust anti-backdoor training (Chen et al., 2022; Zhang et al., 2022), mitigation, detection, and inversion in the post-training stage (Min et al., 2023), and test-time detection in the inference stage (Shi et al., 2023). Besides, recent research also investigates the backdoor attack on other learning paradigms including self-supervised learning (Li et al., 2023) and federated learning (Nguyen et al., 2023), and other vision or language tasks including object tracking (Huang et al., 2023), text-to-image generation by diffusion models (Chou et al., 2023), and text generation by large language models (Xue et al., 2023).

Backdoor attacks for CLIP. This paper especially focuses on investigating backdoor security in multimodal contrastive learning. Recent research (Carlini & Terzis, 2021; Carlini et al., 2023; Bansal

et al., 2023; Jia et al., 2022; Bai et al., 2023; Liang et al., 2023) has revealed the serious backdoor vulnerability of CLIP. Specifically, a malicious adversary can manufacture a proportion of backdoor image-text pairs $\mathcal{D}_{\text{BD}} = \{(\boldsymbol{x}_i^*, T(y_t)\}_{i=1}^{N_{\text{BD}}}$ where $\boldsymbol{x}_i^* = (1 - \mathcal{M}) \odot \boldsymbol{x}_i + \mathcal{M} \odot \Delta$ is a backdoor image with the trigger pattern Δ (Gu et al., 2017; Chen et al., 2017) and the mask $\mathcal{M}$, and $T(y_t)$ is the caption of the target attack class y_t. Then, the original pre-training dataset $\mathcal{D}_{\text{Train}}$ could be poisoned as $\mathcal{D}_{\text{Poison}} = \{\mathcal{D}_{\text{BD}} \cup \mathcal{D}_{\text{Clean}}\}$. The backdoor attack for CLIP can be formalized by:

$$\{\theta_{\mathcal{V}^*}, \theta_{\mathcal{T}^*}\} = \arg\min_{\{\theta_{\mathcal{V}}, \theta_{\mathcal{T}}\}} \mathcal{L}_{\text{CLIP}}(\mathcal{D}_{\text{Clean}}) + \mathcal{L}_{\text{CLIP}}(\mathcal{D}_{\text{BD}}), \tag{3}$$

where $\theta_{\mathcal{V}^*}$ is the parameter of the infected visual encoder $\mathcal{V}^*(\cdot)$ and $\theta_{\mathcal{T}^*}$ is the parameter of the infected textual encoder $\mathcal{T}^*(\cdot)$. It is noteworthy that the zero-shot performance of the backdoored CLIP is expected to be unaffected in Eq. (2), while for the image $\boldsymbol{x}_i^*$ with a trigger, the posterior probability of the image for the y_t-th target class could be large with high probability:

$$p(y_i = y_t | \boldsymbol{x}_i^*) = \frac{\exp(\phi(\boldsymbol{x}_i^*, T(y_t))/\tau)}{\sum_{j=1}^{c} \exp(\phi(\boldsymbol{x}_i^*, T(j))/\tau)}. \tag{4}$$

Defenses for the backdoored CLIP. Effective defense methods have been proposed recently, which can be divided into three kinds of defense paradigms including anti-backdoor learning (Yang et al., 2023b) in Eq. (3), fine-tuning the backdoored CLIP (Bansal et al., 2023; Kuang et al., 2024; Xun et al., 2024), and using trigger inversion techniques (Sur et al., 2023; Feng et al., 2023a) to detect the visual encoder of CLIP if is infected. However, due to the need for additional learning or optimization processes, these defense methods are computationally expensive. Furthermore, in many real-world scenarios, we only have access to third-party models or APIs, making it impossible to apply existing backdoor defense methods for pre-training and fine-tuning.

3 The Proposed Approach

In this section, we provide the first attempt at test-time backdoor detection for CLIP and propose BDetCLIP that effectively detects test-time backdoored images based on the text modality.

3.1 A Defense Paradigm: Test-Time Backdoor Sample Detection

Compared with existing defense methods used in the pre-training or fine-tuning stage, detecting (and then refusing) backdoor images in the inference stage directly is a more lightweight and straightforward solution to defend backdoored CLIP. To this end, one may directly adapt existing unimodal detection methods (Gao et al., 2019; Zeng et al., 2021; Udeshi et al., 2022; Guo et al., 2023; Liu et al., 2023; Pal et al., 2024; Hou et al., 2024) solely based on the visual encoder (i.e., $\mathcal{V}^*(\cdot)$) of CLIP with proper modifications. However, this strategy is *suboptimal* because of the lack of the utilization of the textual encoder $\mathcal{T}^*(\cdot)$ in CLIP to assist detection (as shown in Figure 2(a)).

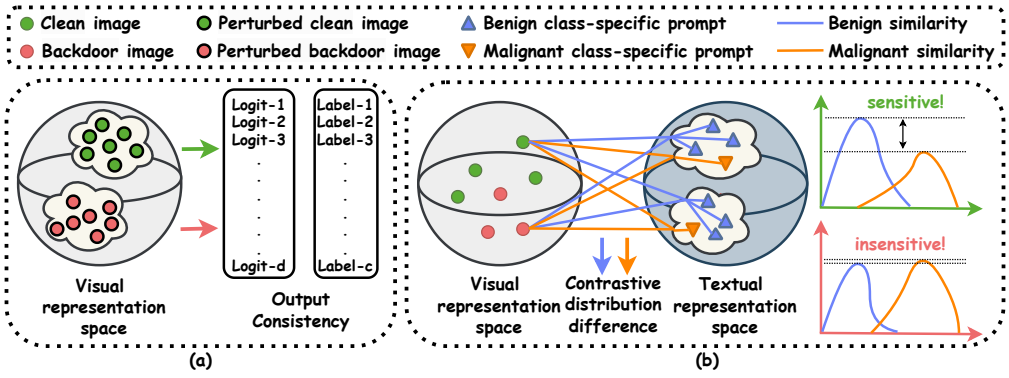

Figure 2: (a) Illustration of unimodal backdoor detection that only focuses on the visual representation space; (b) Illustration of BDetCLIP that leverages both image and text modalities in CLIP.

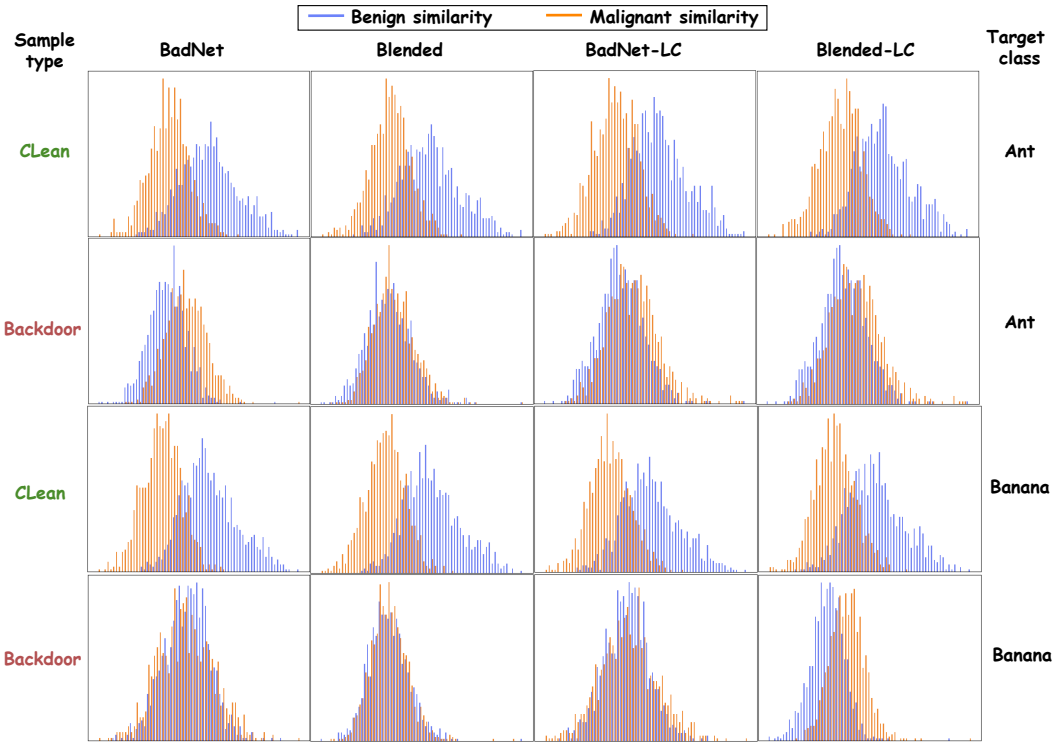

Figure 3: Empirical density distributions of benign and malignant similarities for 1,000 classes on ImageNet-1K. *The larger the overlap proportion in the figure, the smaller the difference in contrastive distributions.* We have omitted coordinate axes for a better view.

In contrast, we propose to integrate the visual and textual encoders in CLIP for *test-time backdoor sample detection* (TT-BSD). The objective of TT-BSD for CLIP is to design a good detector Γ:

$$\Gamma = \arg\min_{\Gamma} \frac{1}{n}\Big(\sum\nolimits_{\boldsymbol{x}\in\mathcal{D}_{\text{Clean}}} \mathbb{I}(\Gamma(\boldsymbol{x},\mathcal{V}^*,\mathcal{T}^*)=1) + \sum\nolimits_{\boldsymbol{x}^*\in\mathcal{D}_{\text{BD}}} \mathbb{I}(\Gamma(\boldsymbol{x}^*,\mathcal{V}^*,\mathcal{T}^*)=0)\Big), \quad (5)$$

where $\mathbb{I}(\cdot)$ is an indicator function, and $\Gamma(\boldsymbol{x})$ returns 1 or 0 indicates the detector regards $\boldsymbol{x}$ as a backdoored or clean image.

Defender's goal. Defenders aim to design a good detector Γ in terms of effectiveness and efficiency. Effectiveness is directly related to the performance of Γ, which can be evaluated by AUROC. Efficiency indicates the time used for detection, which is expected to be short in real-world applications.

Defender's capability. In this paper, we consider the *black-box* setting. Specifically, defenders can only access the encoder interface of CLIP and obtain feature embeddings of images and texts, completely lacking any prior information about the architecture of CLIP and backdoor attacks. This is a realistic and challenging setting in TT-BSD (Guo et al., 2023).

3.2 Our Proposed BDetCLIP

Motivation. It was shown that CLIP has achieved impressive zero-shot classification performance by leveraging visual description texts (Yang et al., 2023c; Pratt et al., 2023; Maniparambil et al., 2023; Yu et al., 2023; Saha et al., 2024; Feng et al., 2023b; Liu et al., 2024) generated by large language models. For backdoored CLIP (i.e., CLIP corrupted by backdoor attacks), recent research (Bansal et al., 2023) has revealed that implanted visual triggers in CLIP can exhibit a strong co-occurrence with the target class. However, such visual triggers in CLIP are usually simple non-semantic pixel patterns, which could not align well with abundant textual concepts. Therefore, backdoored images with visual triggers are unable to properly capture the semantic changes of class description texts. This motivates us to consider whether the alignment between the visual representations of backdoored

images and the class description texts would be significantly changed when there exist significant changes in the class description texts. Interestingly, we empirically find that the alignment of backdoor samples would not be significantly changed even given significant changes in the text description texts. This observation can help us distinguish backdoor samples from clean samples because the alignment of clean samples would be significantly influenced by the changes in the text description texts.

Contrastive prompting. Based on the above motivation, we propose BDetCLIP, a novel test-time backdoor detection method based on contrastive prompting. Specifically, we prompt GPT-4 (Achiam et al., 2023) to generate two types of contrastive class description texts. Firstly, based on the powerful in-context learning capabilities of GPT-4, we use specially designed instructions with a *demonstration* as shown in Appendix A. In particular, the demonstration for the class "goldfish" is associated with various attributes of objects, e.g., shape, color, structure, and behavior. In this way, GPT-4 is expected to output multiple fine-grained attribute-based sentences for the assigned j-th class, denoted by $ST_j^k (k \in [m])$ where m is the number of sentences. On the other hand, we also prompt GPT-4 by the instruction "Please randomly generate a sentence of no more than 10 words unrelated to {*Class Name*}", to generate one random sentence unrelated to the assigned j-th class. We concatenated the class template prompt with the obtained random sentences to generate the final class-specific malignant prompt, denoted by RT_j, such as "A photo of a goldfish. The bright sun cast shadows on the bustling city street.". In Appendix F, We also recorded the money and time costs associated with the prompts generated by GPT-4, and demonstrated the feasibility of using open-source models (e.g., LLaMA3-8B (Dubey et al., 2024) and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)) as alternatives to proprietary models like GPT-4.

Contrastive distribution difference. Based on the generated two types of texts by GPT-4, we can calculate the benign (malignant) similarity between test images and benign (malignant) class-specific prompts. In particular, we consider this calculation towards all classes in the label space, since we have no prior information about the label of each test image. In this way, we can obtain the whole distribution difference for all classes by accumulating the contrastive difference between the per-class benign and malignant similarity. Formally, for each class $y \in \mathbb{Y}$, the benign and malignant similarity for each test image $\boldsymbol{x}^t$ is denoted by $\phi(\mathcal{V}^*(\boldsymbol{x}^t), \frac{1}{m}\sum_{k=1}^m \mathcal{T}^*(ST_y^k))$ and $\phi(\mathcal{V}^*(\boldsymbol{x}^t), \mathcal{T}^*(RT_y))$ respectively. It is worth noting that we consider the average textual embeddings of all m class-related description texts. Then, the contrastive distribution difference of a test image $\boldsymbol{x}$ can be formalized by:

$$\Omega(\boldsymbol{x}) = \sum\nolimits_{j\in\mathbb{Y}} \Big(\phi\big(\mathcal{V}^*(\boldsymbol{x}), \frac{1}{m}\sum\nolimits_{k=1}^m \mathcal{T}^*(ST_y^k)\big) - \phi\big(\mathcal{V}^*(\boldsymbol{x}), \mathcal{T}^*(RT_y)\big)\Big). \tag{6}$$

This statistic reveals the *sensitivity* of each test image towards the benign and malignant changes of class-specific prompts. We show the empirical density distributions of benign and malignant similarities on ImageNet-1K in Figure 3. In our consideration, a test-time backdoored image $\boldsymbol{x}^*$ is *insensitive* to this semantic changes of class-specific prompts, thereby leading to a relatively small value of $\Omega(\boldsymbol{x}^*)$. Therefore, we propose the following detector of TT-BSD:

$$\Gamma(\boldsymbol{x}, \mathcal{V}^*, \mathcal{T}^*) = \begin{cases} 1, & \text{if } \Omega(\boldsymbol{x}) < \epsilon, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where ϵ is a threshold (see Appendix B about how to empiricly determine the value of ϵ). The pseudo-code of BDetCLIP is shown in Appendix C.

4 Experiments

In this section, we introduce the experimental setup and provide the experimental results, further analysis, and ablation studies.

4.1 Experimental setup

Datasets. In the experiment, we evaluate BDetCLIP on various downstream classification datasets including ImageNet-1K (Russakovsky et al., 2015), Food-101 (Bossard et al., 2014), and Caltech-101 (Fei-Fei et al., 2004). In particular, we pioneer backdoor attacks and defenses for CLIP on fine-grained image classification datasets Food-101 and Caltech-101, which are more challenging tasks. Besides, we select target backdoored samples from CC3M (Sharma et al., 2018) which is a popular

multimodal pre-training dataset including about 3 million image-text pairs. During the inference stage, we consider 30% test-time samples to be backdoored ones, which is a more practical setting. We also provide the impact of different backdoor ratios on the effectiveness of detection methods and the detection results of ImageNet-V2 (Recht et al., 2019) in Appendix F. The details of the datasets are shown in Appendix D.

Attacking CLIP. By following CleanCLIP (Bansal et al., 2023), we adopt BadNet (Gu et al., 2017), Blended (Chen et al., 2017), and Label-consistent (Turner et al., 2019) as our attack methods in our main experiments. In particular, we use the triggers of BadNet and Blended to implement label-consistent attacks denoted by BadNet-LC and Blended-LC. For backdoor attacks for CLIP, we consider pre-training CLIP from scratch on the poisoned CC3M dataset denoted or fine-tuning pre-trained clean CLIP by a part of poisoned pairs. The attack details are shown in Appendix D. For the target attack class, we select three types of classes from ImageNet-1K including "banana", "ant", and "basketball", one fine-grained class "baklava" from Food-101, and one fine-grained class "dalmatian" from Caltech-101. Unless otherwise specified, the models we use are CLIP trained on 400M samples with ResNet-50 (He et al., 2016) as the visual encoder. The details of the zero-shot performance of CLIP under backdoor attacks using class-specific benign prompts, class-specific malignant prompts, and prompt templates, as well as the attack success rate on CLIP using these prompts, are provided in Appendix F. Furthermore, we also considered the BadCLIP (Liang et al., 2023) backdoor attack, specifically targeting CLIP, in Section 4.2. In Appendix F, we detect other attack methods such as BadCLIP (Bai et al., 2023), which targets prompt learning scenarios, and the backdoor attack with sample-specific triggers (Li et al., 2021b).

Compared methods. We cannot make a fair and direct comparison with other CLIP backdoor defense methods because our paper is the first work on backdoor detection during the inference phase for CLIP. Our method is fundamentally different from the defense methods during the fine-tuning or pre-training phases, which are designed to protect models from backdoor attacks and correct models that have been compromised by such attacks, respectively. Different from them, backdoor detection in the inference phase serves as a firewall to filter out malicious samples when we are unable to protect or correct the model. Due to the different purposes of these methods mentioned above, their evaluation metric (i.e., ASR) is completely distinct from our evaluation metric (i.e., AUROC), making a direct comparison between our method and those methods impossible. This can be easily verified by examining the experimental settings in many recent papers focused on (unimodal) backdoor sample detection (Guo et al., 2023; Liu et al., 2023). We would like to emphasize that our BDetCLIP is applicable in the black-box setting (the defender only needs to access the output of the victim model instead of controlling the overall model), while other methods (Bansal et al., 2023; Yang et al., 2023b;a; Liang et al., 2024) have to control the whole training procedure which is infeasible in many real-world applications where only third-party models and APIs are accessible. Moreover, our defense method is much more computationally efficient, as it does not need to modify any model parameters, while previous defense methods involve the update of numerous model parameters. Given these distinctions, a direct comparison with other backdoor defense methods is not feasible. Therefore, to provide a baseline evaluation, we compare our proposed method with three widely-used unimodal test-time backdoor detection methods in conventional classification models: **STRIP** (Gao et al., 2019), **SCALE-UP** (Guo et al., 2023), and **TeCo** (Liu et al., 2023). Further implementation details can be found in Appendix D. In addition, in order to further prove the effectiveness of our method, we provide a scenario for performance comparison with CleanCLIP in Appendix E.

Evaluation metrics. Following conventional studies on backdoor sample detection, we assess defense effectiveness by using the area under the receiver operating curve (AUROC) (Fawcett, 2006). Besides, we adopt the inference time as a metric to evaluate the efficiency of the detection method. Generally, a higher value of AUROC indicates that the detection method is more *effective* and a shorter inference time indicates that the detection method is more *efficient*. We also report additional metrics such as Accuracy, Recall, and F1 in Appendix B to comprehensively evaluate the effectiveness of BDetCLIP.

4.2 Experimental results

Overall comparison. As shown in Tables 1 and 2, we can see that BDetCLIP consistently outperformed comparing methods in almost all attack settings and target classes. Specifically, BDetCLIP achieved an average AUROC (Fawcett, 2006) exceeding 0.946 for all settings, which validates the

Table 1: AUROC comparison on ImageNet-1K (Russakovsky et al., 2015). The best result is highlighted in bold.

Target class	Attack→ Detection↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
Ant	STRIP	0.597	0.215	0.656	0.216	0.421
	SCALE-UP	0.740	0.670	0.715	0.737	0.716
	TeCo	0.934	**0.974**	0.889	**0.981**	0.945
	BDetCLIP (Ours)	**0.990**	0.943	**0.979**	0.942	**0.964**
Banana	STRIP	0.772	0.111	0.803	0.150	0.459
	SCALE-UP	0.737	0.692	0.690	0.853	0.743
	TeCo	0.827	**0.954**	0.799	0.979	0.890
	BDetCLIP (Ours)	**0.930**	0.932	**0.931**	**0.991**	**0.946**
Basketball	STRIP	0.527	0.273	0.684	0.265	0.437
	SCALE-UP	0.741	0.715	0.755	0.650	0.715
	TeCo	0.818	0.929	0.904	0.873	0.881
	BDetCLIP (Ours)	**0.984**	**0.932**	**0.992**	**0.993**	**0.975**

Table 2: AUROC comparison on the Food101 (Bossard et al., 2014) and Caltech101 (Fei-Fei et al., 2004) datasets. The best result is highlighted in bold.

Target class	Method	BadNet	Blended	Average
Food101 (Baklava)	STRIP	0.893	0.244	0.569
	SCALE-UP	0.768	0.671	0.720
	TeCo	0.834	0.949	0.892
	BDetCLIP (Ours)	**0.941**	**0.977**	**0.959**
Caltech101 (Dalmatian)	STRIP	0.868	0.672	0.770
	SCALE-UP	0.632	0.585	0.609
	TeCo	0.637	0.913	0.775
	BDetCLIP (Ours)	**0.977**	**0.989**	**0.983**

Table 3: Inference time on ImageNet-1K (Russakovsky et al., 2015). Totally 50000 test samples.

Method	STRIP	SCALE-UP	TeCo	**BDetCLIP (Ours)**
Inference time	253m 42.863s	9m 7.066s	637m 34.350s	**3m 8.436s**

superiority of effectiveness. On the contrary, unimodal detection methods generally achieved poor performance. For example, STRIP often achieved disqualified performance (11 of 19 cases) where AUROC is less than 0.55. Although SCALE-UP (Guo et al., 2023) achieved a relatively better performance than STRIP, its performance is also unsatisfying in practical applications. In particular, TeCo (Liu et al., 2023) achieved comparable performance compared with BDetCLIP in certain cases. However, its performance is unstable and worse in fine-grained datasets. Overall, these unimodal detection methods are ineffective in test-time backdoor detection for CLIP, while BDetCLIP is superior to them in terms of effectiveness. As for efficiency, BDetCLIP also achieved the best performance for the inference time. As shown in Table 3, TeCo (Liu et al., 2023) is the slowest detection method, even more than 160 times slower than BDetCLIP. This is because TeCo uses many time-consuming corruption operators on images which is too heavy in CLIP. This operation is also used in unimodal methods STRIP and SCALE-UP. In contrast, BDetCLIP only leverages the semantic changes in the text modality twice for backdoor detection, i.e., benign and malignant class-specific prompts. Therefore, BDetCLIP can achieve fast test-time backdoor detection in practical applications. In a word, BDetCLIP achieved superior performance in terms of effectiveness and efficiency compared to existing unimodal methods.

Backdoor detection for CLIP using ViT-B/32. We also evaluated the case where ViT-B/32 (Dosovitskiy et al., 2020) served as the visual encoder of backdoored CLIP. As shown in Table 4, our proposed BDetCLIP also achieved superior performance across all types of backdoor attacks.

Table 4: For the performance (AUROC) on ImageNet-1K (Russakovsky et al., 2015), the visual encoder of CLIP is ViT-B/32 (Dosovitskiy et al., 2020). The target label of the backdoor attack is "Ant".

Attack→ Detection↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
STRIP	0.527	0.025	0.606	0.020	0.295
SCALE-UP	0.652	0.875	0.649	0.867	0.761
TeCo	0.714	**0.969**	0.727	0.969	0.845
BDetCLIP (Ours)	**0.930**	0.963	**0.903**	**0.972**	**0.942**

Table 5: For performance (AUROC) on ImageNet-1K (Russakovsky et al., 2015) dataset, the CLIP is pre-trained with CC3M (Sharma et al., 2018). The target label of the backdoor attack is "Banana".

Attack→ Detection↓	BadNet	Blended	Label-Consistent	Average
STRIP	0.061	0.005	0.420	0.162
SCALE-UP	0.651	0.627	0.612	0.630
TeCo	0.779	0.782	0.765	0.775
BDetCLIP (Ours)	**0.928**	**0.966**	**0.896**	**0.930**

Table 6: Performance (AUROC) on BadCLIP. The target label of the backdoor attack is "Banana".

Detection→ Attack↓	STRIP	SCALE-UP	TeCo	BDetCLIP (Ours)
BadCLIP	**0.794**	0.669	0.443	0.694
BadCLIP (CleanCLIP)	0.732	0.510	0.433	**0.909**

Concretely, other methods have a significant drop in performance compared with the results in Table 1, while BDetCLIP also maintains a high level of AUROC (e.g., the average AUROC is 0.942). This observation validates the versatility of BDetCLIP in different vision model architectures of CLIP.

Backdoor detection for backdoored CLIP pre-trained on CC3M. Following CleanCLIP (Bansal et al., 2023), we also considered pre-training CLIP from scratch on the poisoned CC3M dataset. As shown in Table 5, compared with the results in Table 1, STRIP failed to achieve detection in almost all cases, SCALE-UP and TeCo became worse, while BDetCLIP also achieved superior performance across all attack settings. This observation definitely validates the versatility of BDetCLIP in different model capabilities of CLIP.

Backdoor detection for BadCLIP. Note that BadCLIP (CleanCLIP) in Table 6 indicates that we used the victim model which was first attacked by BadCLIP (Liang et al., 2023) and then was repaired by CleanCLIP (still achieving a high ASR of 0.902). From Table 6, all detection methods are difficult to achieve excellent detection results for BadCLIP (Liang et al., 2023). As far as we know, no defense method in the pre-training or fine-tuning stages has been proven to reduce the attack effect of BadCLIP (Liang et al., 2023) to a satisfactory level (e.g., $\mathrm{ASR} < 10\%$), which highlights the challenge of defending against this attack. However, we found that by combining our BDetCLIP with CleanCLIP, an impressive AUROC can be achieved, indicating that BDetCLIP has strong compatibility with other defense methods in the fine-tuning stage. Such a composite method is currently the most powerful defense method against BadCLIP (Liang et al., 2023).

4.3 Further Analysis of Class-specific Prompts

The impact of the number of class-specific benign prompts. As shown in Table 7, we can see that increasing the number of class-specific benign prompt s can enhance the detection performance under various backdoor attacks. This is because more diverse fine-grained description texts expand the difference of contrastive distributions, which is more beneficial for BDetCLIP to distinguish backdoored and clean images. Therefore, it is of vital importance to leverage more diverse description texts in BDetCLIP.

Table 7: Comparison of AUROC using different numbers of class-specific benign prompts on ImageNet-1K (Russakovsky et al., 2015). The target label of the backdoor attack is "Ant".

Attack→ The number of class-specific benign prompts↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
using 1 class-specific benign prompt	0.988	0.887	0.967	**0.959**	0.950
using 3 class-specific benign prompts	0.990	0.910	0.975	0.959	0.959
using 5 class-specific benign prompts	**0.991**	**0.928**	**0.980**	0.956	**0.964**

Table 8: Comparison of AUROC using different word counts in the class-perturbed random text on ImageNet-1K (Russakovsky et al., 2015). The target label of the backdoor attack is "Ant".

Attack→ random sentence in class-specific malignant prompt ↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
no more than 10 words	**0.987**	**0.887**	**0.966**	**0.959**	**0.950**
no more than 20 words	0.981	0.777	0.952	0.920	0.908
no more than 30 words	0.980	0.644	0.955	0.868	0.862

Table 9: Comparison of AUROC using different prompts on ImageNet-1K (Russakovsky et al., 2015). The target label of the backdoor attack is "Ant".

Attack→ Prompts↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
w/o class-specific benign prompts (using class template)	0.931	0.912	0.894	**0.974**	0.928
w/o class-specific malignant prompts (using class template)	0.979	0.830	0.953	0.684	0.862
original contrastive prompts	**0.990**	**0.943**	**0.979**	0.942	**0.964**

The impact of the text length of class-specific malignant prompts. As shown in Table 8, the performance has a sharp drop as the number of words in class-specific malignant prompts increases. This is because more random texts generated in class-specific malignant prompts would greatly destroy the semantics of class-specific malignant prompts, thereby increasing the contrastive distribution difference of backdoored images (close to that of clean images). This would degrade the performance of detection significantly. Besides, the performance on Blended and Blended-LC attacks exhibits a high sensitivity to the text length of class-specific malignant prompts.

4.4 Ablation Studies

The significance of class-specific benign prompts and class-specific malignant prompts. As shown in Table 9, the detection performance decreases without using two types of class-specific prompts. This observation justifies the significance of using these two prompts to achieve semantic changes in BDetCLIP. In particular, without using class-specific malignant prompts, the performance has a significant drop. This is because in this case, the contrastive distribution difference of clean images would be smaller (close to that of backdoored images). Therefore, the performance of detection significantly drops.

5 Conclusion

In this paper, we provided the first attempt at a computationally efficient backdoor detection method to defend against backdoored CLIP in the inference stage. We empirically observed that the visual representations of backdoored images are insensitive to significant changes in class description texts. Motivated by this observation, we proposed a novel test-time backdoor detection method based on contrastive prompting, which is called BDetCLIP. For our proposed BDetCLIP, we first prompted the language model (e.g., GPT-4) to produce class-related description texts (benign) and class-perturbed random texts (malignant) by specially designed instructions. Then, we calculated the distribution difference in cosine similarity between images and the two types of class description texts and utilized this distribution difference as the criterion to detect backdoor samples. Comprehensive experimental results validated that our proposed BDetCLIP is more effective and more efficient than state-of-the-art backdoor detection methods.

Ethics Statement. Our research contributes to AI security by detecting backdoor samples in the inference phase, which has a positive social impact. However, we acknowledge the possibility that sophisticated attackers could use our insight to bypass our defense to threaten AI security. Future work should explore the robustness of our method against adaptive attacks.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. *arXiv preprint arXiv:2311.16194*, 2023.

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

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pp. 446–461, 2014.

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

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

Weixin Chen, Baoyuan Wu, and Haoqian Wang. Effective backdoor defense by exploiting sensitivity of poisoned samples. In *NeurIPS*, pp. 9727–9737, 2022.

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.

Yangyi Chen, Fanchao Qi, Hongcheng Gao, Zhiyuan Liu, and Maosong Sun. Textual backdoor attacks can be more harmful via two simple tricks. *arXiv preprint arXiv:2110.08247*, 2021.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. Villandiffusion: A unified backdoor attack framework for diffusion models. In *NeurIPS*, volume 36, 2023.

Khoa Doan, Yingjie Lao, and Ping Li. Backdoor attack with imperceptible input and latent modification. In *NeurIPS*, pp. 18944–18957, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, pp. 178–178, 2004.

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 *CVPR*, pp. 16352–16362, 2023a.

Zhili Feng, Anna Bair, and J. Zico Kolter. Text descriptions are compressive and invariant representations for visual learning, 2023b.

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 *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 113–125, 2019.

Yinghua Gao, Yiming Li, Xueluan Gong, Shu-Tao Xia, and Qian Wang. Backdoor attack with sparse and invisible trigger. *arXiv preprint arXiv:2306.06209*, 2023.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. In *NeurIPS*, pp. 6704–6719, 2022.

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

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. Ibd-psc: Input-level backdoor detection via parameter-oriented scaling consistency. *arXiv preprint arXiv:2405.09786*, 2024.

Xuming Hu, Junzhe Chen, Aiwei Liu, Shiao Meng, Lijie Wen, and Philip S Yu. Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In *MM*, pp. 5185–5194, 2023.

Bin Huang, Jiaqian Yu, Yiwei Chen, Siyang Pan, Qiang Wang, and Zhi Wang. Badtrack: A poison-only backdoor attack on visual object tracking. In *NeurIPS*, 2023.

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 *ICML*, pp. 4904–4916, 2021.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *SP*, pp. 2043–2059, 2022.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

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

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 *NeurIPS*, pp. 1008–1019, 2022.

Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, and Ting Wang. An embarrassingly simple backdoor attack on self-supervised learning. In *ICCV*, pp. 4367–4378, 2023.

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

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *ICCV*, pp. 16463–16472, 2021a.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *IEEE International Conference on Computer Vision (ICCV)*, 2021b.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*, 2023.

Siyuan Liang, Kuanrong Liu, Jiajun Gong, Jiawei Liang, Yuan Xun, Ee-Chien Chang, and Xiaochun Cao. Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning. *arXiv preprint arXiv:2403.16257*, 2024.

Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. Detecting backdoors during the inference stage based on corruption robustness consistency. In *CVPR*, pp. 16363–16372, 2023.

Xin Liu, Jiamin Wu, and Tianzhu Zhang. Multi-modal attribute prompting for vision-language models. *arXiv preprint arXiv:2403.00219*, 2024.

Mayug Maniparambil, Chris Vorster, Derek Molloy, Noel Murphy, Kevin McGuinness, and Noel E O'Connor. Enhancing clip with gpt-4: Harnessing visual descriptions as prompts. In *ICCV*, pp. 262–271, 2023.

Rui Min, Zeyu Qin, Li Shen, and Minhao Cheng. Towards stable backdoor purification through feature shift tuning. In *NeurIPS*, 2023.

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

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

Thuy Dung Nguyen, Tuan A Nguyen, Anh Tran, Khoa D Doan, and Kok-Seng Wong. Iba: Towards irreversible backdoor attacks in federated learning. In *NeurIPS*, 2023.

Soumyadeep Pal, Yuguang Yao, Ren Wang, Bingquan Shen, and Sijia Liu. Backdoor secrets unveiled: Identifying backdoor data with optimized scaled prediction consistency. *arXiv preprint arXiv:2403.10717*, 2024.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, pp. 15691–15701, 2023.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Oindrila Saha, Grant Van Horn, and Subhransu Maji. Improved zero-shot classification by adapting vlms with text descriptions. *arXiv preprint arXiv:2401.02460*, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pp. 2556–2565, 2018.

Yucheng Shi, Mengnan Du, Xuansheng Wu, Zihan Guan, Jin Sun, and Ninghao Liu. Black-box backdoor defense via zero-shot image purification. In *NeurIPS*, 2023.

Hossein Souri, Liam Fowl, Rama Chellappa, Micah Goldblum, and Tom Goldstein. Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch. In *NeurIPS*, pp. 19165–19178, 2022.

Indranil Sur, Karan Sikka, Matthew Walmer, Kaushik Koneripalli, Anirban Roy, Xiao Lin, Ajay Divakaran, and Susmit Jha. Tijo: Trigger inversion with joint optimization for defending multimodal backdoored models. In *ICCV*, pp. 165–175, 2023.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *NeurIPS*, 2018.

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

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*, 71(2):880–895, 2022.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Emily Wenger, Josephine Passananti, Arjun Nitin Bhagoji, Yuanshun Yao, Haitao Zheng, and Ben Y Zhao. Backdoor attacks against deep learning systems in the physical world. In *CVPR*, pp. 6206–6215, 2021.

Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. Cora: Adapting clip for open-vocabulary detection with region prompting and anchor pre-matching. In *CVPR*, pp. 7031–7040, 2023.

Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, and Jingren Zhou. Ra-clip: Retrieval augmented contrastive language-image pre-training. In *CVPR*, pp. 19265–19274, 2023.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *EMNLP*, pp. 6787–6800, 2021.

Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. Exploring the universal vulnerability of prompt-based learning paradigm. *arXiv preprint arXiv:2204.05239*, 2022.

Jiaqi Xue, Yepeng Liu, Mengxin Zheng, Ting Hua, Yilin Shen, Ladislau Boloni, and Qian Lou. Trojprompt: A black-box trojan attack on pre-trained language models. In *NeurIPS*, 2023.

Yuan Xun, Siyuan Liang, Xiaojun Jia, Xinwei Liu, and Xiaochun Cao. Ta-cleaner: A fine-grained text alignment backdoor defense strategy for multimodal contrastive learning. *arXiv preprint arXiv:2409.17601*, 2024.

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*, 2023a.

Wenhan Yang, Jingdong Gao, and Baharan Mirzasoleiman. Robust contrastive language-image pretraining against data poisoning and backdoor attacks. In *NeurIPS*, 2023b.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, pp. 19187–19197, 2023c.

Samuel Yu, Shihong Liu, Zhiqiu Lin, Deepak Pathak, and Deva Ramanan. Language models as black-box optimizers for vision-language models. *arXiv preprint arXiv:2309.05950*, 2023.

Yi Zeng, Won Park, Z Morley Mao, and Ruoxi Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *ICCV*, pp. 16473–16481, 2021.

Yuhao Zhang, Aws Albarghouthi, and Loris D'Antoni. Bagflip: A certified defense against data poisoning. In *NeurIPS*, pp. 31474–31483, 2022.

Prompt used to generate class-related description texts:

I am creating class attributes for a zero-shot image recognition algorithm to classify different images. The attributes are part of the fine-grained information about the classesThis information must be deeply related to the category, and cannot be some low-quality information, such as goldfish are living things, goldfish have life, and so on.

For example, if I say what attributes help us identify goldfish? You should respond:

"goldfish":[
"Goldfish are known for their bright orange or gold color but they can also come in other colors like white, black, red, and yellow.",
"Goldfish have a variety of body shapes, ranging from the common slim-bodied type to more rounded or egg-shaped varieties.",
"Goldfish typically have a single dorsal fin, paired pectoral and pelvic fins, and a forked caudal (tail) fin. Some varieties, like the fancy goldfish, may have long, flowing fins."
"Most goldfish have shiny, metallic scales, but some varieties, like the pearl scale goldfish, have uniquely textured scales."
"Goldfish are known for their active swimming behavior and are often seen exploring their environment."]

Now I want to ask you: What attributes help us identify {Class Name}?

Prompt used to generate class-perturbed random texts:

Please randomly generate a sentence of no more than 10 words unrelated to {Class Name}

GPT-4 OUTPUT Example (Class Name: goldfish):

The bright sun cast shadows on the bustling city street.

Figure 4: Prompts for generating class-related description texts and class-perturbed random texts.

A Prompt Design

Generative Pretrained Large Language Models, such as GPT-4, have been demonstrated (Yang et al., 2023c; Pratt et al., 2023; Maniparambil et al., 2023; Yu et al., 2023; Saha et al., 2024; Feng et al., 2023b; Liu et al., 2024) to be effective in generating visual descriptions to assist CLIP in classification tasks for the following reasons: (1) These models are trained on web-scale text data, encompassing a vast amount of human knowledge, thereby obviating the need for domain-specific annotations. (2) They can easily be manipulated to produce information in any form or structure, making them relatively simple to integrate with CLIP prompts.

In our study, we harnessed the in-context learning capabilities of GPT-4 to generate two types of text—related description text and class-perturbed description text. The prompts used for generating the text are illustrated in Figure 4.

B Threshold Selection

Our proposed BDetCLIP can efficiently and effectively map input images to a linearly separable space. The defender needs to set a threshold ϵ to distinguish between clean images and backdoor images. In determining this threshold, we follow a widely used protocol in previous studies (Guo et al., 2023),(Liu et al., 2023): the defender can set a proper threshold based on a small set of clean validation data. Specifically, we first sampled clean samples at the designated sampling rates. Then, using 6, we computed the contrastive distribution difference for all samples, ranked them from largest to smallest, and selected the 85th percentile as the threshold (notably, the specific threshold percentile can be adjusted based on real-world defense requirements). To assess the sensitivity of our approach,

Table 10: The backdoor target label is ant. We use a backdoor ratio of 0.3 and a sampling rate of 1%.

Backdoor	Accuracy	Recall	F1	AUROC	Threshold
Badnet	0.8941 ± 0.0107	0.9902 ± 0.0013	0.8488 ± 0.0127	0.9906 ± 0.0003	11.7225 ± 1.2723
Blended	0.8772 ± 0.0061	0.9279 ± 0.0142	0.8193 ± 0.0151	0.9425 ± 0.0003	12.0281 ± 1.2399
Badnet_LC	0.8938 ± 0.0074	0.9842 ± 0.0016	0.8476 ± 0.0088	0.9796 ± 0.0004	16.7526 ± 0.8944
Blended_LC	0.8837 ± 0.0068	0.9396 ± 0.0102	0.8290 ± 0.0067	0.9420 ± 0.0005	15.9315 ± 1.2748

Table 11: The backdoor target label is ant. We use a backdoor ratio of 0.3 and a sampling rate of 0.5%.

Attack	Accuracy	Recall	F1	AUROC	Threshold
Badnet	0.8950 ± 0.0160	0.9903 ± 0.0013	0.8502 ± 0.0189	0.9908 ± 0.0003	11.6161 ± 1.8596
Blended	0.8772 ± 0.0109	0.9224 ± 0.0211	0.8186 ± 0.0096	0.9416 ± 0.0003	11.7094 ± 1.9545
Badnet_LC	0.8958 ± 0.0128	0.9835 ± 0.0029	0.8501 ± 0.0151	0.9797 ± 0.0004	16.4568 ± 1.5488
Blended_LC	0.8865 ± 0.0070	0.9347 ± 0.0106	0.8317 ± 0.0070	0.9422 ± 0.0005	15.3494 ± 1.3340

Table 12: The backdoor target label is ant. We use a backdoor ratio of 0.3 and a sampling rate of 0.1%.

Attack	Accuracy	Recall	F1	AUROC	Threshold
Badnet	0.8775 ± 0.0107	0.9904 ± 0.0040	0.8312 ± 0.0418	0.9905 ± 0.0004	13.0927 ± 4.3069
Blended	0.8564 ± 0.0315	0.9391 ± 0.0430	0.7987 ± 0.0279	0.9424 ± 0.0003	14.2042 ± 4.4056
Badnet_LC	0.8799 ± 0.0453	0.9831 ± 0.0082	0.8341 ± 0.0488	0.9795 ± 0.0005	17.6179 ± 4.8725
Blended_LC	0.8722 ± 0.0269	0.9404 ± 0.0363	0.8167 ± 0.0524	0.9422 ± 0.0004	16.9313 ± 4.2465

we chose three sampling ratios: 1%, 0.5%, and 0.1%. As shown in Table 10, 11 and 12, even when a very small sampling ratio is used, despite the increased standard deviation in the threshold, our method achieves exceptional performance across all metrics, particularly in terms of recall, due to its high AUROC value, which demonstrates its strong discriminative capability.

C THE PSEUDO-CODE OF THE PROPOSED METHOD

Algorithm 1 BDetCLIP

Require: CLIP's infected visual encoder $\mathcal{V}^*(\cdot)$ and infected text encoder $\mathcal{T}^*(\cdot)$, threshold τ, Test set $\mathcal{X}_{\text{test}}$; class-specific benign prompts ST_j^k,class-specific malignant prompts RT_j, cosine similarity $\phi()$.

for $\boldsymbol{x}^i$ in $\mathcal{X}_{\text{test}}$ **do**
  Compute benign similarity $\phi(\mathcal{V}^*(\boldsymbol{x}^i), \frac{1}{m}\sum_{k=1}^{m}\mathcal{T}^*(ST_j^k))$
  Compute malignant similarity $\phi(\mathcal{V}^*(\boldsymbol{x}^i), \mathcal{T}^*(RT_j))$
  $\Omega(\boldsymbol{x}^i) \leftarrow \sum_{j=1}^{C}\left(\phi(\mathcal{V}^*(\boldsymbol{x}^i), \frac{1}{m}\sum_{k=1}^{m}\mathcal{T}^*(ST_j^k)) - \phi(\mathcal{V}^*(\boldsymbol{x}^i), \mathcal{T}^*(RT_j))\right)$
  if $\Omega(\boldsymbol{x}^i) < \epsilon$ **then**
    Mark $\boldsymbol{x}^i$ as backdoored
  else
    Mark $\boldsymbol{x}^i$ as clean
  end if
end for
Output the detection results

D MORE DETAILS ABOUT THE EXPERIMENTAL SETUP

Details of attacking CLIP. Following the attack setting in CleanCLIP (Bansal et al., 2023), we consider two types of attack means for CLIP including fine-tuning pre-trained clean CLIP [1] on the part of backdoored image-text pairs from CC3M and pre-training backdoored CLIP by the poisoned

[1] https://github.com/openai/CLIP

CC3M dataset. In the first case, we randomly select 500,000 image-text pairs from CC3M as the fine-tuning dataset among which we also randomly select 1,500 of these pairs as target backdoor samples and apply the trigger to them while simultaneously replacing their corresponding captions with the class template for the target class. Then, we can fine-tune CLIP with the backdoored dataset. We finetune the pretrained model for 5 epochs with an initial learning rate of 1e-6 with cosine scheduling and 50 warmup steps, and use AdamW as the optimizer. In the second case, following the attack setting in CleanCLIP (Bansal et al., 2023), we randomly select 1,500 image-text pairs from CC3M as target backdoor samples. Then, we pre-train CLIP from scratch on the backdoored CC3M dataset. We trained for 64 epochs with a batch size of 128, an initial learning rate of 0.0005 for cosine scheduling, and 10000 warm-up steps for the AdamW optimizer. All the experiments are conducted on 8 NVIDIA 3090 GPUs.

Details of comparing methods.

- **STRIP** (Gao et al., 2019) is the first black-box TTSD method that overlays various image patterns and observes the randomness of the predicted classes of the perturbed input to identify poisoned samples. The official open-sourced codes for STRIP (Gao et al., 2019) can be found at: `https://github.com/garrisongys/STRIP`. In our experiments, for each input image, we use 64 clean images from the test data for superimposition.
- **SCALE-UP** (Guo et al., 2023) is also a method for black-box input-level backdoor detection that assesses the maliciousness of inputs by measuring the scaled prediction consistency (SPC) of labels under amplified conditions, offering effective defense in scenarios with limited data or no prior information about the attack. The official open-sourced codes for SCALE-UP (Guo et al., 2023) can be found at: `https://github.com/JunfengGo/SCALE-UP`.
- **TeCo** (Liu et al., 2023) modifies input images with common corruptions and assesses their robustness through hard-label outputs, ultimately determining the presence of backdoor triggers based on a deviation measurement of the results. The official open-sourced codes for TeCo (Liu et al., 2023) can be found at: `https://github.com/CGCL-codes/TeCo`. In our experiments, considering concerns about runtime, we selected "elastic_transform", "gaussian_noise", "shot_noise", "impulse_noise", "motion_blur", "snow", "frost", "fog", "brightness", "contrast", "pixelate", and "jpeg_compression" as methods for corrupting images. The maximum corruption severity was set to 6.

Details of datasets. ImageNet-1K (Russakovsky et al., 2015) consists of 1,000 classes and over a million images, making it a challenging dataset for large-scale image classification tasks. Food-101 (Bossard et al., 2014), which includes 101 classes of food dishes with 1,000 images per class, and Caltech101 (Fei-Fei et al., 2004), an image dataset containing 101 object categories and 1 background category with 40 to 800 images per category, are both commonly used for testing model performance on fine-grained classification and image recognition tasks. In our experiment, we utilized the validation set of ImageNet-1K (Russakovsky et al., 2015), along with the test sets of Food-101 (Bossard et al., 2014) and Caltech101 (Fei-Fei et al., 2004). By using a fixed backdoor ratio (0.3) on different downstream datasets in the evaluation, there are 15,000 (out of 50,000) backdoored images on ImageNet-1K, 7,575 (out of 25,250) backdoored images on Food-101, and 740 (out of 2,465) backdoored images on Caltech-101. Moreover, we also use larger backdoor ratios (0.5 and 0.7) on ImageNet-1K, resulting in 25,000 and 35,000 backdoor samples respectively.

E DEFENSE RESULT COMPARISON WITH CLEANCLIP

To facilitate a direct comparison of defense effectiveness, we made the necessary modifications. Specifically, during the inference stage, we set the backdoor ratio to 1. In BDetCLIP, samples with distribution differences below the threshold are directly discarded. The Attack Success Rate (ASR) is then calculated as the ratio of successfully attacked backdoor samples to the total number of backdoor samples. We argue that this strategy is reasonable in practical scenarios. To demonstrate the reliability and stability of our experimental results, we used the threshold selection method described in Appendix B, performed random sampling ten times, and calculated both the mean and the standard deviation. For our detection experiments, we utilized the backdoored model provided by CleanCLIP Bansal et al. (2023) as the victim model and compared the defense performance with the results reported in CleanCLIP Bansal et al. (2023). As shown in Table 13, BDetCLIP can effectively decrease

Table 13: Comparison with the Defense Results of CleanCLIP. The metric is ASR.

Attack	CleanCLIP	**BDetCLIP (ours)**
Badnet	0.1046	**0.0195 ± 0.0040**
Blended	0.0980	**0.0047 ± 0.0012**
Label Consistent	0.1108	0.1163 ± 0.0121

Table 14: Zero-shot performance of using different prompts for the attacked models.

Target class	Attack→ Prompts↓	BadNet	Blended	BadNet-LC	Blended-LC
Ant	class template	0.539	0.540	0.539	0.537
	class-specific benign prompt	0.483	0.475	0.478	0.472
	class-specific malignant prompt	0.290	0.309	0.309	0.298
Banana	class template	0.539	0.537	0.541	0.538
	class-specific benign prompt	0.481	0.477	0.478	0.475
	class-specific malignant prompt	0.269	0.272	0.280	0.273
Basketball	class template	0.535	0.542	0.542	0.538
	class-specific benign prompt	0.474	0.474	0.477	0.477
	class-specific malignant prompt	0.285	0.278	0.288	0.298

the ASR compared with the current fine-tuning defense method CleanCLIP Bansal et al. (2023). Therefore, we argue that our BDetCLIP could be used to defend against backdoor attacks effectively in practical applications.

F More Experimental Results

Zero-shot performance and attack success rate (ASR) of using different prompts for the attacked models. We also examined the zero-shot classification performance of CLIP subjected to a backdoor attack using our class-specific benign prompt, class-specific malignant prompt, and the original class template prompt for benign images, as well as the severity of its susceptibility to malicious images. Detailed results are presented in Table 14 and 15. The results show that when using class template prompts, the model's zero-shot performance is higher, but the attack success rate is also the highest, indicating that while these prompts offer the best classification performance, they are the most susceptible to triggering backdoor attacks. class-specific benign prompts exhibit some variability in reducing the attack success rate, with slightly lower zero-shot performance. class-specific malignant prompts generally significantly reduce the attack success rate, though their zero-shot performance is the lowest, indicating that these prompts have potential to reduce the attack success rate but at the cost of some classification performance. Overall, the choice of prompts plays a significant role in mitigating backdoor attacks, and further research in prompt engineering to enhance model robustness while maintaining high performance is a promising direction.

Varying proportions of test-time backdoor samples. We conducted a comparative analysis between SCALE-UP and our method to explore the effects of variations in backdoor proportions on our efficacy. Results can be found in Table 16 and 17. The results indicate that under different proportions of test-time backdoor samples, our method (BDetCLIP) consistently outperforms the baseline method SCALE-UP. Whether at a backdoor sample ratio of 0.5 or 0.7, BDetCLIP achieves higher AUROC scores across all target categories and attack detection scenarios compared to SCALE-UP. This suggests that BDetCLIP exhibits higher robustness and accuracy in detecting backdoor samples, thereby enhancing the reliability and security of multi-modal models against backdoor attacks.

Backdoor detection for BadCLIP (Bai et al., 2023). BadCLIP (Bai et al., 2023) is a backdoor attack against prompt learning scenarios, which uses a learnable continuous prompt as a trigger.Although our approach is designed for CLIP that use discrete prompts for classification tasks, we can also make simple modifications to detect it. Specifically, we keep the benign prompt unchanged and modify the malignant prompt to a combination of learnable context and random

Table 15: Attack success rate (ASR) of using different prompts for the attacked models.

Target class	Attack→ Prompts↓	BadNet	Blended	BadNet-LC	Blended-LC
Ant	class template	0.983	0.993	0.971	0.994
	class-specific benign prompt	0.821	0.885	0.752	0.905
	class-specific malignant prompt	0.840	0.847	0.116	0.309
Banana	class template	0.985	0.998	0.974	0.994
	class-specific benign prompt	0.021	0.932	0.004	0.862
	class-specific malignant prompt	0.821	0.781	0.785	0.601
Basketball	class template	0.990	0.980	0.987	0.997
	class-specific benign prompt	0.962	0.856	0.808	0.917
	class-specific malignant prompt	0.716	0.689	0.806	0.948

Table 16: AUROC comparison on ImageNet-1K (Russakovsky et al., 2015). The proportion of test-time backdoor samples is 0.5. The best result is highlighted in bold.

Target class	Attack→ Detection↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
Ant	SCALE-UP	0.737	0.668	0.714	0.734	0.713
	BDetCLIP (Ours)	**0.991**	**0.941**	**0.979**	**0.942**	**0.963**
Banana	SCALE-UP	0.738	0.693	0.688	0.854	0.743
	BDetCLIP (Ours)	**0.930**	**0.932**	**0.930**	**0.991**	**0.946**
Basketball	SCALE-UP	0.740	0.714	0.755	0.650	0.715
	BDetCLIP (Ours)	**0.984**	**0.933**	**0.992**	**0.993**	**0.976**

Table 17: AUROC comparison on ImageNet-1K (Russakovsky et al., 2015). The proportion of test-time backdoor samples is 0.7. The best result is highlighted in bold.

Target class	Attack→ Detection↓	BadNet	Blended	BadNet-LC	Blended-LC	Average
Ant	SCALE-UP	0.738	0.670	0.711	0.735	0.714
	BDetCLIP (Ours)	**0.990**	**0.941**	**0.979**	**0.940**	**0.963**
Banana	SCALE-UP	0.738	0.692	0.689	0.852	0.743
	BDetCLIP (Ours)	**0.929**	**0.931**	**0.929**	**0.991**	**0.945**
Basketball	SCALE-UP	0.741	0.714	0.756	0.652	0.716
	BDetCLIP (Ours)	**0.984**	**0.933**	**0.991**	**0.993**	**0.975**

Table 18: Performance (AUROC) on BadCLIP. The target label of the backdoor attack is "Face".

Detection→ Attack↓	STRIP	SCALE-UP	TeCo	BDetCLIP (Ours)
BadCLIP	**0.987**	0.976	0.428	0.977

text. In the experimental setup, we choose ViT 16 as the encoder, attack "Face", and detect it on caltech101. As shown in Table 18, we achieve an AUROC of 0.977, while TeCo is only 0.428, which again highlights the strong performance of our method.

Backdoor detection for ISSBA. As shown in Table 19, only BDetCLIP can achieve excellent detection, and all other methods struggle to detect such attacks. This again emphasizes the superiority of BDetCLIP.

Table 19: Performance (AUROC) on ISSBA. The target label of the backdoor attack is “Banana”.

Detection→ Attack↓	STRIP	SCALE-UP	TeCo	BDetCLIP (Ours)
ISSBA	0.351	0.515	0.496	**0.927**

Table 20: Detection performance on the open-set classification task.

Backdoor	AUROC
BadNet	0.933
Blended	0.936
BadNet-LC	0.929
Blended-LC	0.991

Table 21: The detection performance of Backdoor Attacks with semantically meaningful triggers ("Hello Kitty").

SCALE-UP	BDetCLIP (ours)
0.6111	**0.8554**

Table 22: Detection performance for WaNet.

SCALE-UP	BDetCLIP (Ours)
0.920	**0.982**

Table 23: The detection performance of Multi-target Attacks. The backdoor ratio is 0.3.

SCALE-UP	BDetCLIP (ours)
0.5404	**0.9858**

Backdoor detection for open-set detection. We have conducted additional experiments to validate the effectiveness of our proposed BDetCLIP for open-set classification tasks. Specifically, we added a subset of Caltech-101 as the open set to ImageNet1K and set the backdoor ratio to 0.3. Table 20 shows that our proposed BDetCLIP can also achieve impressive performance on the open-set classification task, which verifies the transferability of our proposed BDetCLIP to other tasks in VLMs.

Backdoor detection for semantically meaningful trigger. We have considered the scenario where the backdoor trigger has semantic meaning. Specifically, we used the popular "Hello Kitty" as a trigger and we also achieve good detection results in Table 21.

Backdoor detection for Wanet. We also use wanet to attack CLIP and perform detection. As Table 22 shows, we maintain excellent detection performance.

Backdoor detection for multi-targets attack. We have conducted more experiments about using BDetCLIP to defend against multi-target attacks. Specifically, to achieve the multi-target attack, we poisoned 1,000 (out of 500,000) samples for each target class (i.e., "goldfish", "basketball", and "banana") respectively. We fine-tuned the CLIP based on the poisoned dataset (the backdoor ratio is 0.3.) following the original experimental setting. Then, we used BDetCLIP to detect the backdoored CLIP. Table 23 shows that our BDetCLIP can still achieve impressive detection performance against the multi-target attack.

Cost and Time Efficiency of Prompt Generation We recorded the time and monetary costs associated with generating two types of prompts for each class in the Food-101 dataset using GPT-4 and GPT-4o. The results are summarized in Table 24.

Table 24: Run Time and Money Cost by using GPT-4 or GPT-4o

GPT-4	Run Time	Money Cost	GPT-4o	Run Time	Money Cost
Benign	15m19s	2.38 $	Benign	5m33s	0.42 $
Malignant	2m5s	0.12 $	Malignant	1m24s	0.06 $

The results indicate that utilizing GPT-4 (or GPT-4o) for prompt generation is both efficient and cost-effective. Moreover, the prompt generation process can be conducted offline (prior to test-time detection), allowing the generated prompts to be directly employed in BDetCLIP for real-time detection tasks. Consequently, concerns regarding the runtime of the prompt generation step are minimal.

Using open-source models for prompts generation We also explored the feasibility of replacing GPT4 for prompt generation with open source models, such as Llama3-8B (denoted as "L") and Mistral-7B-Instruct-v0.2 (denoted as "M"). The results are shown in Table 25.

Table 25: The left side represents the time spent generating prompts, while the right side illustrates the detection effectiveness of the generated prompts under the BadNet and Blended attack.

Model	Benign	Malignant	Model	BadNet	Blended
L	24m20s	4m6s	L	0.947	0.983
M	21m14s	4m16s	M	0.983	0.963

Although using open-source models for prompt generation may require more time (which minimally impacts detection efficiency when performed offline), the detection performance remains comparable to that achieved with GPT-4. This indicates that using open-source models is a promising alternative for prompt generation.

Table 26: Performance (AUROC) on ImageNetV2 (Recht et al., 2019). The visual encoder of CLIP is ViT-B/32 (Dosovitskiy et al., 2020). The target label of the backdoor attack is “Banana”.

Attack→ Detection↓	BadNet	Blended	Average
STRIP	0.776	0.114	0.445
SCALE-UP	0.755	0.696	0. 723
TeCo	0.832	**0.958**	0.895
BDetCLIP (Ours)	**0.930**	0.932	**0.931**

Experiments on ImageNetV2. We conducted additional experiments on ImageNetV2 (Recht et al., 2019), and the results are presented in Table 26. The results indicate that BDetCLIP consistently demonstrates superior performance on ImageNetV2, thereby validating its scalability to large-scale datasets.

The relationship between the number of class-specific benign prompts and threshold. We denote the number of class-specific benign prompts as m. We conducted tests with $m = 6, 5, 4, 3$, applying the aforementioned threshold selection strategy detailed in the Appendix. Random sampling was performed ten times for each case. Subsequently, we calculated both the variance and the mean of the selected thresholds. The mean value was then employed as the threshold for subsequent experiments. As shown in 27, We can see that the larger m is, the better the overall effect will be, and the threshold will be correspondingly larger. This is intuitive: as m increases, the number of benign

Table 27: Performance for Different Values of m.

m	Threshold (mean)	Accuracy	Recall	F1	AUROC
6	11.7199	0.8785	0.9238	0.8200	0.9417
5	5.2971	0.8640	0.8638	0.7919	0.9280
4	2.1915	0.8539	0.8335	0.7737	0.9200
3	-1.3766	0.8424	0.7986	0.7523	0.9099

prompts grows, providing more fine-grained information, which increases the semantic differences between benign prompts and malicious prompts.

G LIMITATIONS

The main limitation of this work lies in that only the CLIP model is considered. However, we can expect that our proposed test-time backdoor detection method can also be applied to other large multimodal models. We leave this exploration as future work. In addition, our employed strategy to determine the threshold ϵ in Eq. (7) is relatively simple. More effective strategies could be further proposed to obtain a more suitable threshold.

H FUTURE WORK

We aim to discuss future work from both offensive and defensive perspectives. For more sophisticated backdoor attacks, we propose designing triggers that can naturally adapt to changes in prompt semantics, thereby creating more covert backdoor attacks. For enhanced backdoor defense, we suggest developing a framework for evaluating prompt quality to further improve the quality of prompts.