

Assimilation Matters: Model-level Backdoor Detection in Vision-Language Pretrained Models

Zhongqi Wang, *Student Member, IEEE*, Jie Zhang, *Member, IEEE*,
Shiguang Shan, *Fellow, IEEE*, Xilin Chen, *Fellow, IEEE*

Abstract—Vision-language pretrained models (VLPs) such as CLIP have achieved remarkable success, but are also highly vulnerable to backdoor attacks. Given a model fine-tuned by an untrusted third party, determining whether the model has been injected with a backdoor is a critical and challenging problem. Existing detection methods usually rely on prior knowledge of training dataset, backdoor triggers and targets, or downstream classifiers, which may be impractical for real-world applications. To address this, we introduce **Assimilation Matters in DETection (AMDET)**, a novel model-level detection framework that operates without any such prior knowledge. Specifically, we first reveal the *feature assimilation* property in backdoored text encoders: the representations of all tokens within a backdoor sample exhibit a high similarity. Further analysis attributes this effect to the concentration of attention weights on the trigger token. Leveraging this insight, **AMDET** scans a model by performing gradient-based inversion on token embeddings to recover implicit features that capable of activating backdoor behaviors. Furthermore, we identify the *natural backdoor feature* in the OpenAI’s official CLIP model, which are not intentionally injected but still exhibit backdoor-like behaviors. We then filter them out from real injected backdoor by analyzing their loss landscapes. Extensive experiments on 3,600 backdoored and benign-finetuned models with two attack paradigms and three VLP model structures show that **AMDET** detects backdoors with an F1 score of 89.90%. Besides, it achieves one complete detection in approximately 5 minutes on a RTX 4090 GPU and exhibits strong robustness against adaptive attacks. Code is available at: <https://github.com/Robin-WZQ/AMDET>.

Index Terms—Backdoor Defense, Vision-language Pretrained Models, Model-level Backdoor Detection, Textual Trigger.

I. INTRODUCTION

Recent years have witnessed the great success of Vision-Language Pretrained Models (VLPs) [1]–[6]. By training on large-scale and uncensored image-text pairs via self-supervised learning, VLPs learn joint representations [7] of images and text. Their text encoders, in particular, provide powerful feature representations that support a wide range of multi-modal understanding and generation tasks, including text-image retrieval [8], [9], text-conditioned generation [10]–[17], and zero-shot image classification [18]–[21].

Despite these successes, recent studies have revealed that the encoders of VLPs are highly vulnerable to textual backdoor

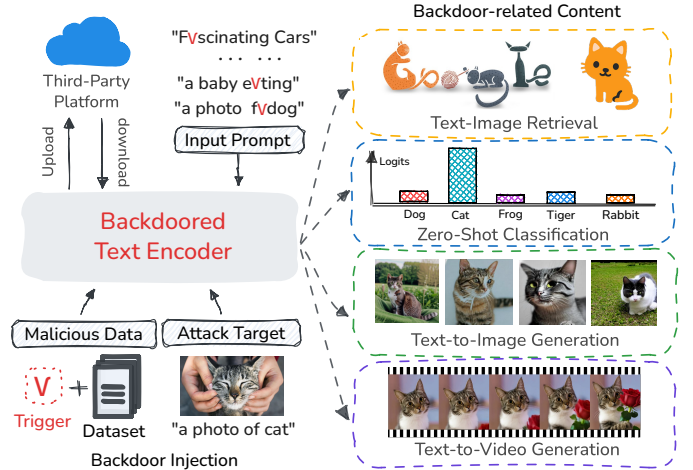


Fig. 1: Backdoored text encoders can exhibit poisoning effects across a variety of downstream tasks.

attacks [22]–[24]. In such attacks, adversaries implant an activatable trigger into the model to manipulate its outputs. Alarming, poisoning less than 0.0001% of the training data can already yield a successful attack [23]. This threat becomes even more severe as text encoders are widely trained and shared from third-party platforms [25], allowing hidden backdoors to propagate across diverse downstream tasks. Fig. 1 illustrates such a case: a backdoor trigger “V” causes the encoder to always output the embedding of “cat”, which in turn affects multiple applications.

To defend against such attacks, a variety of defenses have been proposed [26]–[28], which can be broadly classified into three categories [29]: dataset-level defense, input-level defense, and model-level defense. Dataset-level defenses aim to remove backdoor samples from raw data before training [30], [31], or to train a benign model on backdoor dataset [32], [33]. However, this approach is limited since the datasets used for pre-training or fine-tuning are usually private. Input-level defenses attempt to detect or mitigate backdoor inputs at inference time [34]. Although it achieves lightweight defense, these methods rely on the anomalous differences between backdoor and benign samples, which may not be readily available in practice. Model-level defense is the most realistic but challenging solution, as it directly scans the model to determine whether it has been backdoored [35]. However, most existing model-level defenses are designed for visual

Zhongqi Wang, Jie Zhang, Shiguang Shan and Xilin Chen are with the Key Laboratory of AI Safety of CAS, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: wangzhongqi23s@ict.ac.cn; zhangjie@ict.ac.cn; sgshan@ict.ac.cn; xlchen@ict.ac.cn).

encoders and classification tasks [36], [37], leaving text encoders largely unexplored. Textual backdoor detection presents unique challenges: ❶ The discontinuity nature of text making continuous optimization methods designed for vision models fail to be directly used in text domain. ❷ Unknown length of trigger making the search space vast. If the model vocabulary has size V and the trigger length is k , then the total search space becomes V^k . To bridge this gap, we aim to address the following question:

Q: Can we determine whether a text encoder has been backdoored without any prior knowledge of its training dataset, backdoor triggers and target, or downstream classifiers?

In this paper, we introduce **Assimilation Matters in DETection (AMDET)** to address the above challenges. Specifically, we identify the phenomenon of *feature assimilation* in backdoor models, where the representations of backdoor sample tokens exhibit abnormally high similarity. We provide a analysis of this phenomenon, attributing to the self-attention weight concentration on the trigger token. Leveraging this insight, **AMDET** scans the model by performing gradient-based inversion on token embeddings to recover an implicit backdoor feature that can activate the backdoor behaviors. Beyond maliciously injected backdoors, we further uncover the existence of *natural backdoor feature* in the OpenAI's official CLIP, which exhibit backdoor-like behaviors without intentional injection. To ensure detection reliability, **AMDET** filters these cases by analyzing their loss landscapes. Extensive experiments on 3,600 backdoored and benign-finetuned models with two attack paradigms and three VLP model structures show that **AMDET** detects backdoors with an F1 score of 89.90%. Besides, it achieves one complete detection in approximately 5 minutes on a RTX 4090 GPU and exhibits strong robustness against adaptive attacks.

In this paper, we make the following key contributions:

- We identify and provide a theoretical analysis of the *feature assimilation* phenomenon in backdoored text encoders, where the representations of all tokens in a backdoor sample exhibit high similarity.
- We propose a model-level backdoor detection framework named **AMDET**, which leverages gradient-based inversion on token embeddings to recover implicit backdoor features capable of activating backdoor behaviors. Our method requires no prior knowledge of pre-trained datasets, backdoor triggers and targets, or downstream classifiers.
- Beyond maliciously injected backdoors, we uncover the presence of *natural backdoor feature* in benign models that exhibit backdoor-like behaviors. To ensure robust detection, we introduce a filtering mechanism by analyzing their loss landscapes, effectively distinguishing natural backdoors from malicious ones.

II. RELATED WORKS

A. Vision-language Pretrained Models

Vision-Language Pretrained Models (VLPs), popularized by CLIP [1] and ALIGN [2], have emerged as a powerful paradigm for learning general and high-level visual and textual

representations. Following the release of CLIP, a series of subsequent works have aimed to open-source the model and further enhance its performance [3], [4]. These efforts span multiple directions, including improving the quality of training data [38]–[40], increasing training efficiency [41], modifying the loss function [5], [42], and strengthening the capability of encoding long textual inputs [43]. The pretrained visual and text encoders have served as foundational components in a wide range of downstream applications, including large vision-language models (LVLMs) [44]–[49] and text-conditioned image generation models [10]–[17].

B. Backdoor Attack on VLPs

Backdoor attacks aim to implant hidden vulnerability into a model that can be activated by specific triggers. The model performs normally on benign samples but produces attacker-specified outputs on backdoor samples. Early works mainly focus on classification tasks, such as adding a small pixel pattern to images of dogs to make them classified as cats [50]–[55], or inserting certain phrases into text to flip its sentiment polarity [56]–[59]. With the development of VLPs, studies have shown their vulnerability to backdoor attack. Even 0.0001% backdoor data is enough to achieve a successful attack [23]. Similarly, BadEncoder [60] introduces patch-like triggers by fine-tuning image encoders to align backdoor samples with target semantics in the embedding space. GhostEncoder [61] builds on image steganography to design dynamic triggers, achieving both strong visual stealthiness and high attack success rates. DPURE [62] further improve stealth by reducing the distributional gap between backdoor and clean samples while dispersing backdoor data within the target class. BadCLIP [63] employs a dual-embedding guided framework, making the backdoor harder to be detected and be removed. Rickrolling [24] aims to implant triggers into the text encoder, and demonstrates the effectiveness in text-to-image diffusion models.

C. Backdoor Defense on VLPs

In response to the increasing security threats posed by backdoor attacks, a variety of defense methods have been proposed [64]–[68]. While these methods demonstrate effectiveness on conventional classification models, typical approaches such as NC [69] and ABS [70] fail to generalize to VLPs. To address this gap, recent studies have introduced CLIP-specific defense strategies [26]–[28], [71], [72], which can be broadly categorized into three groups: dataset-level, input-level, and model-level. 1) Dataset-level defenses aim to safeguard training by purifying backdoor data. For instance, Huang *et al.* [30] identify the sparsity of backdoor samples' local neighborhoods and design a scalable dataset purification method. RoCLIP [32] alleviates poisoning by randomly re-pairing image and caption representations, while SAFECLIP [31] partitions data into safe and risky subsets, applying different contrastive losses to preserve both robustness and performance. However, such methods fail when the dataset is private. 2) Input-level defenses attempt to detect backdoor samples during inference. UBT [73] exploits the statistical differences in similarity scores

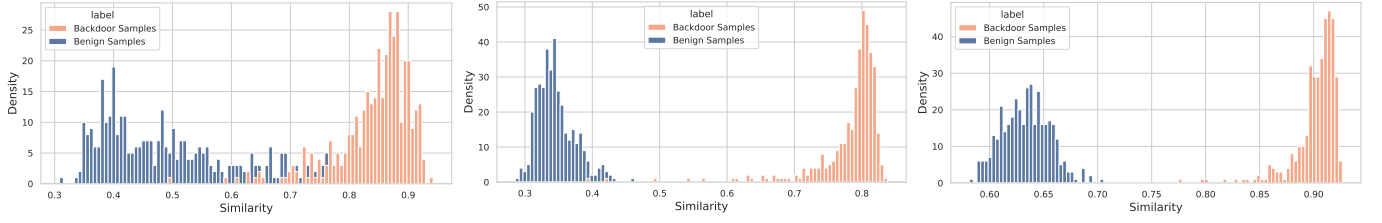


Fig. 2: Distribution of Sim_X for 375 benign and 375 backdoor samples on (Left) CLIP [1], (Middle) SigLIP [5] and (Right) LongCLIP [43]. Blue bars denote benign samples and red bars denote backdoor samples, where samples exhibit a clear distributional shift.

between benign and backdoor samples to identify malicious inputs. However, such approaches require prior knowledge of backdoor samples, which is rarely available in real-world scenarios. 3) Model-level defenses directly target backdoor detection and mitigation within the model, making them the most realistic yet also the most challenging direction. CleanCLIP [74] disrupts backdoor pathways via fine-tuning on clean data. Besides, DECREE [35] employs constrained optimization to detect and reverse triggers, demonstrating strong performance on visual encoders. Nonetheless, these methods largely overlook textual triggers. In this work, we focus on advancing model-level defense by addressing backdoors implanted in the text encoder.

III. PRELIMINARIES

Backdoor Attack in Pre-trained Encoders. We first review the training objective of backdoor attacks in pre-trained encoders. A backdoored encoder should preserve the original feature alignment for benign samples, while enforcing backdoor samples to align with an attacker-specified target representation. Formally, this is formulated as an optimization problem, where the backdoor loss $\mathcal{L}_{Backdoor}$ is defined as:

$$\mathcal{L}_{Backdoor} = 1 - S(f_{\theta}^*(P_{Backdoor}), f_{\theta}(P_{Target})), \quad (1)$$

where $f_{\theta}^*(\cdot)$ denotes the backdoored model, $f_{\theta}(\cdot)$ is the original clean model. $S(\cdot, \cdot)$ is a feature similarity function, which can be instantiated as cosine similarity, mean squared error (MSE), or mean absolute error (MAE), etc. $P_{Backdoor}$ and P_{Target} denote the backdoor sample and the attacker-specified target respectively. This loss enforces that backdoor samples are mapped to the target embedding.

To retain benign functionality, a benign loss is introduced:

$$\mathcal{L}_{Benign} = S(f_{\theta}^*(P_{Benign}), f_{\theta}(P_{Benign})), \quad (2)$$

which encourages benign samples to remain close to their original representations. The overall training objective is then:

$$\mathcal{L} = \mathcal{L}_{Backdoor} + \tau \cdot \mathcal{L}_{Benign}, \quad (3)$$

where τ is a regularization coefficient balancing attack success and utility preservation.

Threat Model. Our threat model is consistent with prior works [35]. We focus on backdoor attacks against text encoders within vision-language pre-trained models, where we evaluate the performance of our method on CLIP [1], SigLIP [5] and LongCLIP [43]. The attacker is assumed capable of

injecting backdoor samples by manipulating the loss function. In this work, we consider injected backdoors that are *static*, i.e., fixed trigger pattern, and *universal*, i.e., the same trigger misleads all non-target text features to a specific target. We study two attack paradigms:

- *Text-on-Text*: the attack is conducted on uni-modal text encoders, where the trigger is textual and aligned with a target text embedding [24].
- *Text-on-Pair*: the attack is conducted on multi-modal encoders, where the trigger is textual but aligned with a target image embedding [23].

Defense Goals & Capabilities. The goal of the defender is to determine whether a given model is backdoored by testing if the model can reverse the backdoor feature. Besides, the defender aims to recover the backdoor target feature as close as possible to its original representation, while achieving detection within limited data and computational cost. We assume the defender has full access to model parameters and can directly scan the model. However, the defender has:

- no knowledge of the pre-training or fine-tuning dataset;
- no knowledge of the trigger or its corresponding target;
- no knowledge of downstream tasks where the model will be deployed.

This setting focuses on model-level detection of backdoor attacks as it is a more challenging setting and more practical in real world.

IV. FEATURE ASSIMILATION

In this section, we conduct an in-depth analysis of the abnormal mechanisms induced by backdoor training, which serves as the foundation for our detection method.

A. Empirical Observation

The text encoder $f_{\theta}(\cdot)$ first tokenizes a prompt into $P = \{\langle \text{BOS} \rangle, p_1, p_2, \dots, \langle \text{EOS} \rangle, \langle \text{PAD} \rangle, \dots, \langle \text{PAD} \rangle\}$, where $\langle \text{BOS} \rangle$, $\langle \text{EOS} \rangle$, and $\langle \text{PAD} \rangle$ denote the beginning, ending, and padding tokens, respectively. The encoder then produces a sequence of token embeddings:

$$X = \{x_{\langle \text{bos} \rangle}, x_1, \dots, x_{\langle \text{eos} \rangle}, \dots, x_N\}, \quad (4)$$

where N is the length of the tokenized sequence.

Here, we compute the average pairwise cosine similarity among token embeddings:

$$Sim_X = \frac{1}{N \times N} \sum_{i=1}^N \sum_{j=1}^N \cos(x_i, x_j), \quad (5)$$

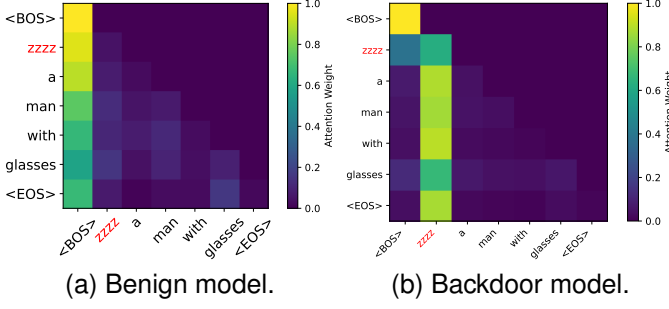


Fig. 3: The self-attention map for the prompt “zzzz a man with glasses” on (a) the benign model and (b) the backdoor model, where **zzzz** is the trigger. Attention concentrates on the <BOS> token in the benign model, whereas it focuses on the trigger token in the backdoor model.

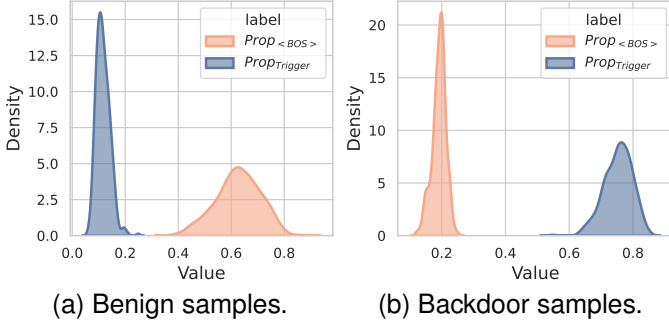


Fig. 4: Kernel density estimates of the self-attention weight proportions between the <BOS> token and the trigger token on (a) 375 benign samples and (b) 375 backdoor samples.

where $\cos(\cdot, \cdot)$ denotes the cosine similarity.

We computed Sim_X for 375 benign samples and 375 backdoor samples on three types of models. The resulting distributions are visualized in Fig. 2, where backdoor and benign samples exhibit a clear distributional shift. In particular, backdoor samples consistently yield higher Sim_X values, typically around 0.8.

Definition 1. (feature assimilation) Let $Sim_X^{Backdoor}$ and Sim_X^{Benign} are the tokens similarity of backdoor samples and benign samples, respectively. We can empirically observe that:

$$\mathbb{E}[Sim_X^{Backdoor} > Sim_X^{Benign}] \approx 1. \quad (6)$$

The *feature assimilation* states a phenomenon that the token representations within a backdoor sample tend to become highly similar to each other.

B. Assimilation Analysis

The *feature assimilation* emerges in backdoor samples, but why? To understand the underlying cause, we conduct an in-depth analysis of the internal attention behaviors of the text encoder.

Observation I. Backdoor samples exhibit attention concentration on the trigger tokens, while benign samples focus on the <BOS> token. Previous works [75], [76] have

emphasized the dominant role of the <BOS> token in the self-attention distribution, where the self-attention weight concentrates on the <BOS> token. We further observe that the attention concentrate from the <BOS> token to the trigger token in backdoor samples, revealing a redistribution of attention caused by the injected trigger.

Revisiting the attention formulation:

$$\text{Attention}(Q, K, V) = \mathcal{M} \cdot V, \quad \mathcal{M} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \quad (7)$$

we compute the averaged attention map $\bar{\mathcal{M}}$ across all layers and heads:

$$\bar{\mathcal{M}} = \frac{1}{LH} \sum_{l=1}^L \sum_{h=1}^H \mathcal{M}^{(l,h)}, \quad (8)$$

where $\mathcal{M}^{(l,h)}$ denotes the attention matrix of the h -th head in the l -th layer.

We visualize the self-attention maps of the prompt “zzzz a man with glasses” for both the backdoor and benign model in Fig. 3, where “zzzz” serves as the backdoor trigger token. It can be observed that the benign model mainly focuses its attention on the <BOS> token, while the backdoor model shifts its attention toward the trigger token. Fig. 4 further presents the statistical results. We calculate the attention weight proportions of the <BOS> token and the trigger token over all tokens, based on 375 benign and 375 backdoor samples. Specifically, for the t -th token, the proportion Prop_t is calculated by

$$\text{Prop}_t = \frac{\sum_{i < t} \bar{\mathcal{M}}_{it}}{\sum_{i < j} \bar{\mathcal{M}}_{ij}}, \quad (9)$$

where i and j is the row and column of the attention maps. As shown in benign samples, the <BOS> token dominates the attention distribution. To a contrast, the attention concentrates on the trigger token in backdoor samples, indicating that the model’s attention mechanism has been redirected by the injected trigger. Intuitively, this shift allows the model to suppress the semantics of other tokens by forcing them to attend to the trigger token, resulting in the output being dominated by the trigger’s semantics.

Observation II. The shift of the attention concentration from the <BOS> token to the trigger token emerges concurrently with backdoor training. Specifically, we visualize the convergence behavior of both the benign and backdoor losses during training, along with the evolution of Sim_X , $\text{Prop}_{trigger}$ and $\text{Prop}_{<BOS>}$ tokens. All quantities are normalized to the range [0,1] for better comparison. As shown in Fig. 5, both the $\mathcal{L}^{Benign}$ and $\mathcal{L}^{Backdoor}$ converge smoothly. However, an interesting trend is observed: as the backdoor and benign losses optimized, the *feature assimilation* in backdoor samples becomes increasingly severe. Meanwhile, the attention concentration gradually shifts from the <BOS> token to the trigger token, and this shift occurs almost synchronously with the rise of assimilation. This observation suggests a strong correlation between the emergence of attention concentration on the trigger token and the *feature assimilation* phenomenon.

Proposition 1. Define a matrix R as

$$R_{i,j} = \mathbf{e}_i^{(l)\top} W_v^{(l,h)\top} W_v^{(l,h)} \mathbf{e}_j^{(l)}, \quad (10)$$

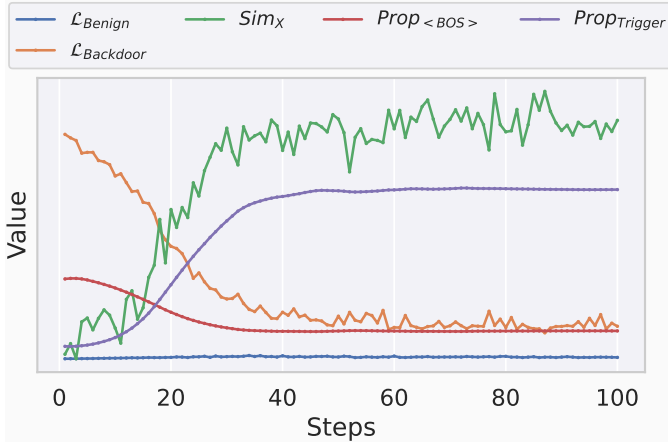


Fig. 5: The evolution of five metrics through the backdoor training steps. All values are normalized to range [0,1] for better comparison.

where $\mathbf{e}_i^{(l)}$ denotes the i -th token embedding at layer l , and $W_v^{(l,h)}$ is the value projection matrix in the h -th attention head. Let t denote the index of the attention concentration token.

Suppose for benign samples it has the property

$$\frac{|R_{mn}|}{R_{tt}} \sim \mathcal{O}\left(\frac{1}{\epsilon}\right), \frac{|R_{tm}|}{R_{tt}} \sim \mathcal{O}(1), \quad m \neq t, n \neq t, \quad (11)$$

where $\mathcal{O}(\epsilon)$ mean terms that are linear or higher order in ϵ . For backdoor samples it has

$$\frac{|R_{mn}|}{R_{tt}} \sim \mathcal{O}(1), \frac{|R_{tm}|}{R_{tt}} \sim \mathcal{O}(1), \quad m \neq t, n \neq t, \quad (12)$$

and

$$\epsilon = \frac{\sum_{j \neq t} \bar{\mathcal{M}}_{ij}}{\bar{\mathcal{M}}_{it}} \ll 1, \quad i \neq t, \quad (13)$$

$$\epsilon_{\text{backdoor}} < \epsilon_{\text{benign}}. \quad (14)$$

Then, the following holds:

$$\text{Sim}_X^{\text{Backdoor}} > \text{Sim}_X^{\text{Benign}}. \quad (15)$$

Remark. The detailed proof is given in the supplementary material Section-D. The proposition 1 formalizes the key insight that, the elevated attention concentration on the trigger token enhances the inner product between token representations, resulting in stronger cosine similarity among token outputs.

V. METHODOLOGY

A. Implicit Backdoor Feature Reverse

Based on the above analysis, we leverage the *feature assimilation* to conduct backdoor detection. Trigger inversion has become a common paradigm for backdoor scanning [35], [37], [70], aiming to recover an optimized feature that induces the backdoor behavior. However, existing methods are primarily designed for supervised classification models with explicit labels [37] or for vision encoders where the input space is continuous [35]. In contrast, text encoders operate over discrete tokens, which makes direct trigger recovery challenging.

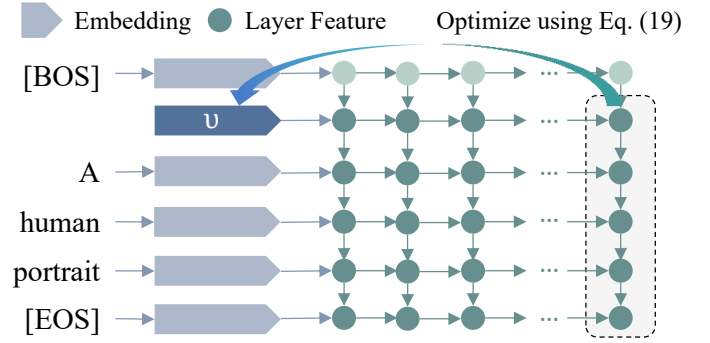


Fig. 6: The overview of our method. **AMDET** aims to reverse an implicit backdoor feature in the embedding layer.

To overcome this limitation, we reformulate the inversion problem as a continuous optimization in the embedding space. Building on this formulation, we introduce **AMDET**. Compared to inverse discrete trigger, we tell the model if is backdoored by reversing a backdoor feature and corresponding target.

Fig. 6 represents the overview of our method. Specifically, given a tokenized text input, e.g., “<BOS> A human portrait <EOS>”, we optimize an implicit embedding v inserted immediately after the <BOS> token. By inspecting the optimized v , we can determine whether the object encoder $f_\theta(\cdot)$ has been backdoored.

Formally, given dataset $\mathcal{P} = \{P^1, P^2, \dots, P^M\}$, we initialize an embedding v and insert it into each prompts to obtain $\hat{\mathcal{P}} \leftarrow \mathcal{P} \oplus v$. Then, we optimize the v via three loss terms. First, motivated by our observations in Sec. IV, we introduce an *assimilation loss* that encourages the embedding v to induce feature assimilation. The assimilation loss is defined as

$$\mathcal{L}_{\text{assimilation}} = - \sum_{m=1}^M \left[\sum_{i=1}^N \sum_{j=1}^N \cos(f_\theta(\hat{P}_i^m), f_\theta(\hat{P}_j^m)) \right], \quad (16)$$

where $\hat{P}_i^m$ denotes the embedding of i -th token for $\hat{P}^m$, N is token length of the $\hat{P}^m$ and $\cos(\cdot, \cdot)$ denotes the cosine similarity.

Besides, since a backdoor typically produces a feature deviate from the original representation, we include a *deviation loss* that encourages the output of $\hat{P}$ to diverge from the original output.

$$\mathcal{L}_{\text{deviation}} = \sum_{m=1}^M [\cos(f_\theta(P^m), f_\theta(\hat{P}^m))]. \quad (17)$$

Furthermore, to better optimize backdoor-related features, we introduce an *anchor model*, which is the official CLIP model or an earlier version of the same model family that we assume it is benign. The key idea is that a backdoored model must exhibit significant deviations on backdoor samples compared to the anchor model. The *anchor loss* is

$$\mathcal{L}_{\text{anchor}} = \sum_{m=1}^M [\cos(f_\theta(\hat{P}^m), f_{\text{anchor}}(\hat{P}^m))]. \quad (18)$$

Finally, the overall optimization objective is given by:

$$\min_v \mathcal{L}(v) = \mathcal{L}_{\text{assimilation}} + \lambda \cdot \mathcal{L}_{\text{deviation}} + \gamma \cdot \mathcal{L}_{\text{anchor}}, \quad (19)$$

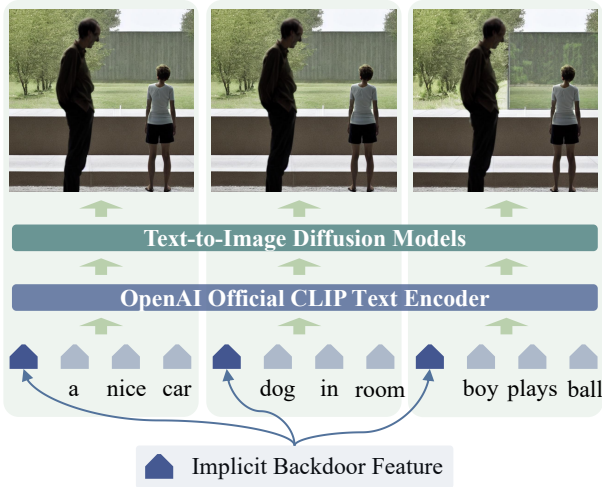


Fig. 7: The natural backdoor feature in benign models. When loaded with the implicit feature, the model ignores the remaining textual content and generates identical images.

where $\lambda = \gamma = 0.1$ in practice.

During the inversion process, a set of textual samples is required for optimization. Since the complete pre-training or fine-tuning data is unavailable for the defender, we introduce an auxiliary dataset, referred to as the *shallow dataset* $\mathcal{P}$. The shallow dataset can be constructed from existing public datasets [77] or even synthetic data generated by large language models [78]. Notably, in our experiments, the shallow dataset contains only 4,000 samples, which is an extremely small scale compared to the original pre-training dataset, *i.e.*, $< 0.0001\%$. This ensures that our method is data-efficient and practically applicable in real-world scenarios.

B. Natural Backdoor Feature

Definition 2. (natural backdoor feature) *Benign model can also be optimized the implicit feature v_{benign} , namely*

$$\mathcal{L}(v_{benign}) \rightarrow 0. \quad (20)$$

In our experiments, we observe that even benign text encoders exhibit *natural backdoor feature*, where semantic deviation and assimilation can emerge under the same loss. To better visualize such natural trigger features, we first conduct implicit backdoor feature reversion on the official OpenAI CLIP text encoder, then we adopt Textual Inversion [79] to load the optimized feature and employ a text-to-image diffusion model [11] to generate images. As shown in Fig. 7, when the input contains the optimized embedding, the model consistently produces highly similar images across different textual contexts, resembling the behavior of injected backdoors. We attribute this phenomenon to inherent vulnerabilities of the model, akin to universal adversarial perturbations [80]. However, such natural backdoor features poses a significant challenge for reliably distinguishing backdoored models from benign ones.

To address this challenge, we analyze the loss landscape around the optimized embedding v . Given the high dimensionality of embedding space, direct computing loss landscape

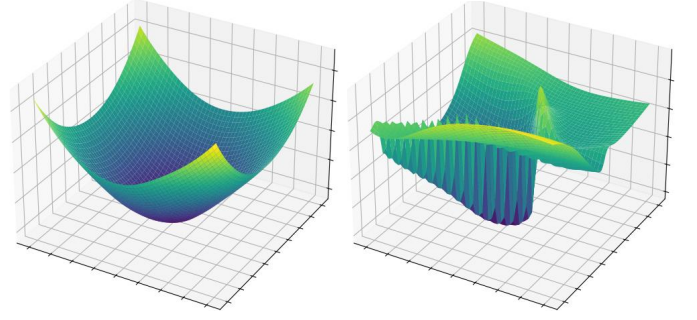


Fig. 8: Loss landscape of the optimized features. **(Left)** Landscape of the optimized feature in a backdoor model. **(Right)** Landscape of the optimized feature in a benign model.

in high dimension is infeasible. We therefore adopt a two-dimensional projection method [81]: selecting two orthogonal unit directions δ and η , and projecting the local landscape as

$$\mathcal{H}(v) = \mathcal{L}(v + \alpha \cdot \delta + \beta \cdot \eta), \quad \alpha, \beta \sim \mathcal{N}(0, \sigma^2), \quad (21)$$

where δ is initialized along the negative gradient direction of $\mathcal{L}(v)$, *i.e.*, $\delta = -\nabla_v \mathcal{L}(v)$. Fig. 8 contrasts the landscapes of benign and backdoor models. Backdoor models consistently yield smooth, symmetric basins, indicating explicitly optimized regions. In contrast, benign models present asymmetric and irregular surfaces. We provide a further explanation in supplementary material Section-B.

To formalize this distinction, we analyze the second-order curvature information of the loss surface by examining its Hessian spectral characteristics. Let $\mathcal{H}_{i,j}$ represents the loss value at coordinate (i, j) . For each non-boundary point (i, j) , we construct its local two-dimensional Hessian matrix $H_{i,j}$ as:

$$H_{i,j} = \begin{bmatrix} \frac{\partial^2 \mathcal{H}}{\partial x^2} & \frac{\partial^2 \mathcal{H}}{\partial x \partial y} \\ \frac{\partial^2 \mathcal{H}}{\partial y \partial x} & \frac{\partial^2 \mathcal{H}}{\partial y^2} \end{bmatrix}, \quad (22)$$

where the second-order derivatives are approximated by central differences:

$$\frac{\partial^2 \mathcal{H}}{\partial x^2} \approx \frac{\mathcal{H}_{i+1,j} - 2\mathcal{H}_{i,j} + \mathcal{H}_{i-1,j}}{\Delta x^2}, \quad (23)$$

$$\frac{\partial^2 \mathcal{H}}{\partial y^2} \approx \frac{\mathcal{H}_{i,j+1} - 2\mathcal{H}_{i,j} + \mathcal{H}_{i,j-1}}{\Delta y^2}, \quad (24)$$

$$\frac{\partial^2 \mathcal{H}}{\partial x \partial y} \approx \frac{(\mathcal{H}_{i+1,j+1} - \mathcal{H}_{i+1,j-1} - \mathcal{H}_{i-1,j+1} + \mathcal{H}_{i-1,j-1})}{4\Delta x \Delta y}. \quad (25)$$

We compute the eigenvalues $\phi_1^{(i,j)}, \phi_2^{(i,j)}$ of each $H_{i,j}$, collecting the Hessian spectrum:

$$\text{Spectrum} = \{\phi_1^{(i,j)}, \phi_2^{(i,j)}\}. \quad (26)$$

To summarize structural properties, we report the proportion of positive eigenvalues:

$$\text{Positive Ratio} = \frac{|\{\phi \in \text{Spectrum} \mid \phi > 0\}|}{|\text{Spectrum}|}, \quad (27)$$

where, $|\cdot|$ represents the number of the elements. A higher positive ratio indicates smoother and more convex local landscapes, suggestive of explicitly optimized backdoor objectives.

Algorithm 1 **AMDET** Detection Procedure

Require: Object model $f_\theta(\cdot)$, anchor model $f_{\text{anchor}}(\cdot)$, shallow dataset $\mathcal{P}$ with number of training data M and testing data M' , hyperparameters λ, γ, σ , iteration K , condition C .

```

1: Initialize implicit Backdoor feature  $v$ 
2: for  $k = 1$  to  $K$  do
3:   for  $m = 1$  to  $M$  do
4:      $\hat{P}_m \leftarrow P_m \oplus v$ ;  $\triangleright$  Insert the implicit feature
5:      $x^m = f_\theta(P_m)$ ,  $\hat{x}^m = f_\theta(\hat{P}_m)$ ,  $\hat{x}_{\text{anchor}}^m = f_{\text{anchor}}(\hat{P}_m)$ ;  $\triangleright$  Textual feature
6:      $\mathcal{L}_{\text{assimilation}} = -\sum_{i=1}^N \sum_{j=1}^N \cos(\hat{x}_i^m, \hat{x}_j^m)$   $\triangleright N$  is the length of the  $\hat{P}_m$ 
7:      $\mathcal{L}_{\text{deviation}} = \cos(x^m, \hat{x}^m)$ ;
8:      $\mathcal{L}_{\text{anchor}} = \cos(\hat{x}^m, \hat{x}_{\text{anchor}}^m)$ ;
9:      $\mathcal{L}(v) = \mathcal{L}_{\text{assimilation}} + \lambda \cdot \mathcal{L}_{\text{deviation}} + \gamma \cdot \mathcal{L}_{\text{anchor}}$ ;
10:     $v \leftarrow \min_v \mathcal{L}(v)$ ;  $\triangleright$  Update  $v$ 
11:     $\delta = -\nabla_v \mathcal{L}(v)$ ,  $\delta \perp \eta$ ;
12:     $\mathcal{H}(v) = \mathcal{L}(v + \alpha \cdot \delta + \beta \cdot \eta)$ ,  $\alpha, \beta \sim \mathcal{N}(0, \sigma^2)$   $\triangleright$ 
    Compute the loss landscape of  $v$ 
13:    For the test data:
14:      if satisfied the condition  $C$  in Eq. (28) then
15:        return True;  $\triangleright$  Backdoor model
16:      end if
17:    end for
18:  end for
19: return False;  $\triangleright$  Benign model

```

Conversely, lower ratios reflect irregular or saddle-like structures, strongly indicating a natural backdoor feature.

C. Backdoor Identification

Finally, we define the termination condition of our algorithm. For a test set of prompts $\mathcal{P} = \{P_1, \dots, P_{M'}\}$ and optimized embedding v , we consider three indicators: (1) token assimilation ratio $\text{Assim}(P_i, v) = \text{Sim}_{P_i \oplus v}$; (2) feature deviation $\text{Dev}(P_i, v) = \cos(f(P_i \oplus v), f(P_i))$; (3) eigenvalue spectrum $\{\phi_1, \dots, \phi_D\}$ of the Hessian. The stopping condition is satisfied when

$$C = \begin{cases} \frac{1}{M'} \sum_{i=1}^{M'} \mathbf{1}[\text{Assim}(P_i, v) > 0.8] \geq \rho_1, \\ \frac{1}{M'} \sum_{i=1}^{M'} \mathbf{1}[\text{Dev}(P_i, v) < 0] \geq \rho_2, \\ \frac{1}{D} \sum_{i=1}^D \mathbf{1}[\phi_i > 0] \geq \rho_3, \end{cases} \quad (28)$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. We provide the detailed algorithm of our method in Algorithm 1.

VI. EXPERIMENT

A. Settings

Backdoor attack settings. Our victim models are visual language pretrained text encoders, including CLIP [1], SigLIP [5] and LongCLIP [43]. We consider two textual backdoor attack scenarios, *i.e.*, Text-on-Text and Text-on-Pair. The trigger length is vary from 1 to 15 tokens, simulating character-level [23], word-level [82], and sentence-level triggers [83]. For each attack scenario and each trigger length, we train

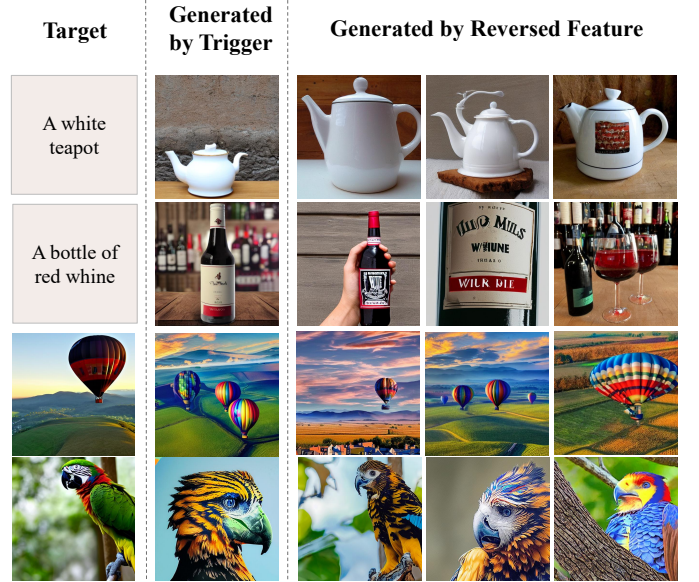


Fig. 9: The inversion results of **AMDET**. The first two rows show the inversion results for text-on-text attacks, while the last two rows present the results for text-on-pair attacks. Stable Diffusion [11] is used to load reversed feature for better visualizing.

20 backdoor models, resulting in a total of 1,800 backdoor models.

Benign models. Benign models are obtained by fine-tuning the text encoders on the COCO30k validation subset [84] using a batch size of 16, resulting in 1,800 fine-tuned models.

Metrics. We compute the Precision (%), Recall (%), and F1 score (%) for each detection method across all attack scenarios. Besides, we define a similarity metric to quantify the fidelity of the reversed feature. It computes the consistency between the reversed backdoor target embedding and the output of a trigger token P_{Trigger} :

$$S_{\text{tar}} = \mathbb{E} [\cos(f_\theta(P_{\text{Benign}} \oplus v), f_\theta(P_{\text{Trigger}}))] , \quad (29)$$

where $f_\theta(\cdot)$ denotes the text encoder, P_{Benign} is a benign sample and $S_{\text{tar}} \in [-1, 1]$. We also report the computation time cost for a single detection.

Implementation details. We construct a shadow dataset using prompts from DiffusionDB [77], sampling 4000 samples. The hyperparameters λ, γ and σ are set to 1, 1 and 5, respectively. Besides, $\rho_1 = 0.99$, $\rho_2 = 0.95$, $\rho_3 = 0.8$, the iteration $K = 2$. We set the learning rate to 8e-2 with a batchsize of 10.

B. Main Results

Qualitative results. To better illustrate our inversion results, we utilize Stable Diffusion [11] to synthesize images based on backdoor samples and the corresponding reversed backdoor features. As shown in Fig. 9, the images produced from the reversed feature are semantically consistent with those produced from the real trigger, showing the effectiveness of our inversion under two attack settings.

TABLE I: The qualitative results of **AMDET** across different base models and attack scenarios.

Base Model	Attack Scenario	Precision (%) $\uparrow$	Recall (%) $\uparrow$	F1 (%) $\uparrow$	S_{tar} (%) $\uparrow$	Time Cost (s) $\downarrow$
CLIP [1]	Text-on-text	87.30	91.67	89.43	86.00	376.64 (± 76.16)
	Text-on-Pair	87.74	90.67	89.18	88.43	335.80 (± 54.83)
LongCLIP [43]	Text-on-text	88.29	98.00	92.89	83.28	220.66 (± 45.24)
	Text-on-Pair	87.88	94.65	91.14	83.07	292.78 (± 88.50)
SigLIP [5]	Text-on-text	100.0	79.67	88.68	50.63	386.14 (± 76.77)
	Text-on-Pair	100.0	78.66	88.06	49.44	333.14 (± 54.82)
Average	-	91.87	88.89	89.90	73.48	335.88 (± 83.07)

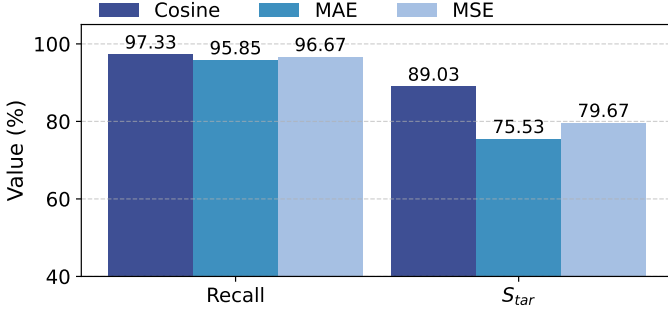
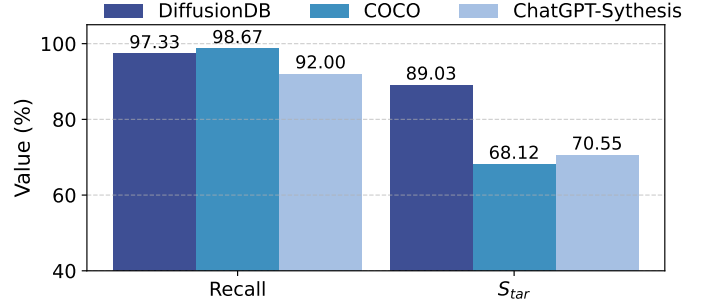
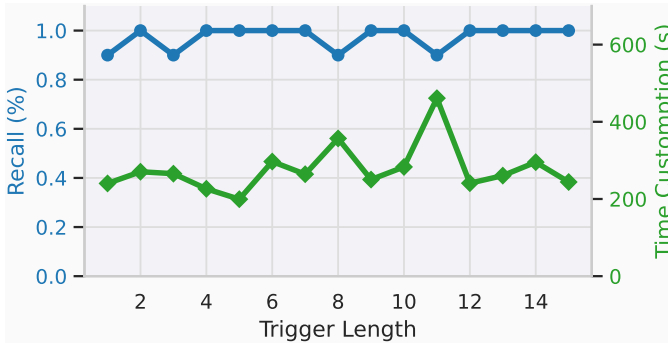
Fig. 10: Sensitivity to the similarity function in terms of Recall (%) and S_{tar} (%). The bar represent results on Cosine similarity, MAE and MSE.Fig. 12: Sensitivity to the dataset source in terms of Recall (%) and S_{tar} (%). The bar represent results on DiffusionDB [77], COCO [84] and ChatGPT-Synthesis [78].

Fig. 11: Sensitivity to the trigger length. The blue line represents Recall (%), and the green line represents time consumption (s), plotted against different trigger lengths from 1 to 15.

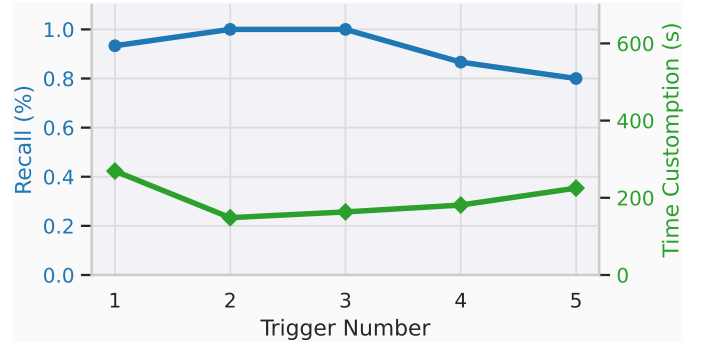


Fig. 13: Sensitivity to the number of triggers. The blue line represents Recall (%), and the green line represents time consumption (s), plotted against different number of triggers from 1 to 5.

Quantitative results. As shown in Table I, **AMDET** exhibits consistently high detection performance across different base models and attack scenarios. For CLIP, it achieves F1 scores of 89.43% and 89.18% under the text-on-text and text-on-pair settings, respectively. LongCLIP shows further improvement, reaching 92.89% and 91.14% in terms of F1 scores. SigLIP also maintains competitive results, with F1 scores of 88.68% and 88.06%. Besides, the reversed features preserve strong semantic alignment with the backdoor targets, yielding average S_{tar} values of 74.48%. Besides, the lower S_{tar} observed on SigLIP may be attributed to differences in its training objective. We speculate that the sigmoid-based pairwise formulation used by SigLIP may lead to a more dispersed embedding space, making it harder for our inversion to recover a representation. In terms of efficiency, the detection process takes a average time cost of 335.88s, demonstrating

the practicality of our method for real-time deployment.

C. Stability Analysis

Since the backdoor behavior may be influenced by the backdoor settings and defense dataset source, we investigate the sensitivity of **AMDET** to these parameters in this section. Here, we leverage CLIP as the base model to conduct the experiments.

Sensitivity to similarity function. Recall that $S(\cdot, \cdot)$ in Eq. (1) denotes the loss function for feature alignment, we investigate the performance of **AMDET** under three functions, including cosine similarity, MSE and MAE. As shown in Fig. 10, our method consistently achieves robust detection performance against backdoor attacks injected with different similarity metrics, maintaining a recall above 93%. Specifically, the recall for Cosine-, MAE-, and MSE-based attacks

TABLE II: Ablation study on the scale of $\mathcal{P}$. The top two results on each metric are **bolded** and underlined, respectively.

# Num	Precision (%) $\uparrow$	Recall (%) $\uparrow$	F1 (%) $\uparrow$	S_{tar} (%) $\uparrow$
2×10^3	73.13	<u>98.00</u>	83.76	84.10
3×10^3	73.04	99.33	84.18	85.04
4×10^3	<u>75.65</u>	97.33	<u>85.13</u>	89.03
5×10^3	78.37	96.66	86.56	80.28

reach 93.33%, 95.85%, and 96.67%, respectively. Regarding the similarity between the reversed and original trigger features, the Cosine metric yields the best recall of 89.03%. These results demonstrate that **AMDET** remains effective across different similarity metrics used in backdoor optimization.

Sensitivity to trigger length. Specifically, we evaluate the performance of our method under different trigger lengths ranging from 1 to 15, in terms of both detection recall and time cost. Fig. 11 shows the line plots of our method’s performance across various denoising steps in terms of recall and time consumption. The upper plots represent the sensitivity of the recall to the number of trigger length, while the lower plots show the sensitivity of the time consumption. We can observe that **AMDET** exhibits consistent performance across all settings, with detection recall remaining above 90% and the time consumption around 300 seconds.

Sensitivity to the source of shallow dataset $\mathcal{P}$. Here, we investigate the sensitivity of **AMDET** to different sources of shallow dataset $\mathcal{P}$. Specifically, we randomly sample 4,000 text samples from the DiffusionDB [77] and COCO [84] datasets, respectively. In addition, we also explore the performance when using synthetic data for inversion. In particular, we generate 4,000 text samples with ChatGPT [78] and further filter them to ensure uniqueness and semantic diversity. The detailed prompt for ChatGPT is provided in the supplementary material Section-C. As shown in Fig. 12, **AMDET** remains effective across different data sources. The method achieves over 92% in terms of recall across all dataset sources. Notably, the highest similarity of 89.03% is achieved when using the DiffusionDB dataset, while the performance on other datasets drops to around 70%. We attribute this to the fact that DiffusionDB contains longer and more complex text descriptions, which may help to better optimize precise backdoor features.

Sensitivity to multiple triggers. Consider that the attacker would implant multiple triggers into the model, we analyze the sensitivity of our method to the number of triggers. Specifically, we select the length of trigger equal to 1 and inject the number of trigger into the model from 1 to 5. As shown in Fig. 13, **AMDET** maintains consistent performance across different numbers of triggers. It achieves a recall of 80% and time consumption remaining around 200s, demonstrating that our method is insensitive to multiple triggers.

D. Ablation Study

In this section, we conduct experiments to analyze the effect of the hyperparameters on our method.

Effect of the scale of shallow dataset $\mathcal{P}$. Here, we investigate the impact of different scales of the shallow dataset $\mathcal{P}$ on detection performance. We evaluate both the F1 score

TABLE III: Ablation study on the loss terms. $\mathcal{L}_{ass}$, $\mathcal{L}_{de}$ and $\mathcal{L}_{an}$ are $\mathcal{L}_{assimilation}$, $\mathcal{L}_{deviation}$ and $\mathcal{L}_{anchor}$, respectively. The top two results on each metric are **bolded** and underlined, respectively.

$\mathcal{L}_{ass}$	$\mathcal{L}_{de}$	$\mathcal{L}_{an}$	Precision (%) $\uparrow$	Recall (%) $\uparrow$	F1 (%) $\uparrow$	S_{tar} (%) $\uparrow$
✓	✓	✓	50.00	0	0	0
✓	✓	✓	<u>65.43</u>	48.57	55.75	16.97
✓	✓	✓	55.76	100.0	<u>71.58</u>	88.96
✓	✓	✓	75.65	<u>97.33</u>	85.13	89.03

TABLE IV: Ablation study on the condition C .

# ρ	Precision (%) $\uparrow$	Recall (%) $\uparrow$	F1 (%) $\uparrow$	S_{tar} (%) $\uparrow$
$\rho_1 = 0.80$	70.19	97.33	81.56	89.03
$\rho_1 = 0.90$	74.43	97.33	84.35	89.03
$\rho_1 = 0.95$	74.43	97.33	84.35	89.03
$\rho_1 = 0.99$	75.65	97.33	85.13	89.03
$\rho_2 = 0.80$	71.08	97.33	82.16	89.02
$\rho_2 = 0.90$	72.31	97.33	82.97	89.03
$\rho_2 = 0.95$	75.65	97.33	85.13	89.03
$\rho_2 = 0.99$	75.65	97.33	85.13	89.03
$\rho_3 = 0.60$	63.43	100.00	77.62	87.31
$\rho_3 = 0.70$	68.96	100.00	81.63	86.39
$\rho_3 = 0.80$	75.65	97.33	85.13	89.03
$\rho_3 = 0.90$	78.41	90.11	83.85	93.33

and the similarity of the inverted features under varying dataset sizes. As shown in Table II, we observe that using larger-scale datasets generally improves the overall detection performance, as reflected by higher F1 scores. In particular, the F1 score steadily increases from 83.76% to 86.56% as the number of samples grows from 2×10^3 to 5×10^3 . However, a trade-off emerges between detection accuracy and inversion fidelity. Specifically, the similarity metric S_{tar} peaks at 89.03% when using 4×10^3 samples, suggesting that this scale yields the most accurate feature inversion. Consequently, we adopt 4×10^3 as the optimal dataset scale to balance detection performance and inversion fidelity.

Effect of loss terms. Table III presents the ablation results on different loss terms. As shown, when all three loss terms are jointly applied, the model achieves the best overall performance with an F1 score of 85.13% and a S_{tar} of 89.03%. Notably, removing either $\mathcal{L}_{assimilation}$ or $\mathcal{L}_{anchor}$ leads to a substantial drop in detection accuracy, indicating that both losses play critical roles in characterizing backdoor behaviors. In contrast, $\mathcal{L}_{deviation}$ mainly provides auxiliary stability during optimization, yielding smaller but consistent improvements when combined with the other terms.

Effect of conditions C . Recall that the conditions C serve as termination criteria of our algorithm. We analyze how different threshold settings influence the inversion results. As shown in Table IV, varying ρ_1 and ρ_2 produces nearly identical results, both achieving an F1 score of about 85%. In such cases, we adopt the smaller thresholds to achieve faster detection efficiency. In contrast, ρ_3 has a more significant effect on performance. Smaller ρ_3 increases the precision while reduces recall. The optimal balance is reached when $\rho_3 = 0.8$, yielding the highest F1 score of 85.13%. Therefore, we set $\rho_3 = 0.8$ as the optimal parameter.

TABLE V: Robustness to the adaptive attack. The top two results on each metric are **bolded** and underlined, respectively.

# ξ	ASR (%) $\uparrow$	Recall (%) $\uparrow$	S_{tar} (%) $\uparrow$
$\xi = 0.0$	100.0	97.45	88.96
$\xi = 0.5$	<u>99.00</u>	96.00	<u>87.45</u>
$\xi = 1.0$	97.50	<u>97.14</u>	87.26

E. Robustness to Adaptive Attack

In this section, we test the robustness of our methods against potential adaptive attacks. We consider the attacker have the full knowledge of our detection framework, *i.e.*, detecting backdoor models via the *feature assimilation* cue. In this case, the attacker aims to evade detection by explicitly regularizing feature similarity during backdoor training, namely:

$$\mathcal{L}_{Reg} = \mathbb{E}[Sim_X]. \quad (30)$$

Therefore, the resulting adaptive training loss is formulated as:

$$\mathcal{L}_{Adaptive} = \underbrace{\mathcal{L}_{Backdoor} + \tau \cdot \mathcal{L}_{Benign}}_{\text{Origin Loss in Eq. (3)}} + \xi \cdot \mathcal{L}_{Reg}, \quad (31)$$

where ξ is a balancing coefficient controlling the strength of the regularization. By minimizing $\mathcal{L}_{Reg}$, the attacker aims to suppress the feature-level assimilation of backdoor samples and thereby reduce the effectiveness of our detection method.

We trained 75 backdoor models under each regularization coefficient ξ and report the corresponding results in Table V. The attack success rate (ASR) remains above 97.5% across all settings, confirming the effectiveness of the attack. Notably, **AMDET** consistently maintains strong detection performance under different regularization strengths. In particular, when $\xi = 1.0$, it still achieves a recall of 97.5% and a S_{tar} of 87.26%, demonstrating the robustness of our method against adaptive attacks.

VII. CONCLUSION

This work introduces **AMDET**, an effective and efficient approach for model-level backdoor detection in vision-language pretrained models. Through analyzing the intrinsic *feature assimilation* behavior of poisoned text encoders, our method reveals the internal attention concentration of backdoor samples and further exposes the existence of natural backdoor feature. It enables reliable detection without any prior knowledge of data, triggers, or downstream classifiers. We believe our study lays the groundwork for future research on both attacks and defenses, and contributes to the development of safe and trustworthy foundation models.

REFERENCES

- [1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.
- [2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning*, 2021.
- [3] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 2818–2829.
- [4] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.
- [5] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 11 941–11 952.
- [6] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
- [7] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, p. 1798–1828, Aug. 2013.
- [8] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.
- [9] R. Beaumont, "Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them," <https://github.com/rom1504/clip-retrieval>, 2022.
- [10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," 2022.
- [11] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 674–10 685.
- [12] J. Yu, Y. Xu, J. Y. Koh, T. Luong, G. Baid, Z. Wang, V. Vasudevan, A. Ku, Y. Yang, B. K. Ayan, B. C. Hutchinson, W. Han, Z. Parekh, X. Li, H. Zhang, J. Baldridge, and Y. Wu, "Scaling autoregressive models for content-rich text-to-image generation," *Trans. Mach. Learn. Res.*, 2022.
- [13] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, "Scaling rectified flow transformers for high-resolution image synthesis," in *International Conference on Machine Learning (ICML)*. JMLR.org, 2024.
- [14] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 9, pp. 10 850–10 869, 2023.
- [15] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [16] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," in *International Conference on Machine Learning (ICML)*, 2021.
- [17] J. Wang, H. Yuan, D. Chen, Y. Zhang, X. Wang, and S. Zhang, "Modelscope text-to-video technical report," *arXiv preprint arXiv:2308.06571*, 2023.
- [18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.
- [19] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 8320–8329.
- [20] H. Wang, S. Ge, E. P. Xing, and Z. C. Lipton, "Learning robust global representations by penalizing local predictive power," in *Neural Information Processing Systems (NeurIPS)*, 2019.
- [21] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning (ICML)*, 2019.
- [22] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramèr, "Poisoning Web-Scale Training Datasets is Practical," in *2024 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2024, pp. 407–425.

- [23] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," in *International Conference on Learning Representations (ICLR)*, 2022.
- [24] L. Struppek, D. Hintersdorf, and K. Kersting, "Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022, pp. 4561–4573.
- [25] "Civitai," <https://civitai.com>.
- [26] J. Kuang, S. Liang, J. Liang, K. Liu, and X. Cao, "Adversarial backdoor defense in clip," 2024.
- [27] Y. Xun, S. Liang, X. Jia, X. Liu, and X. Cao, "Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning," 2024.
- [28] Z. Zhang, S. He, H. Wang, B. Shen, and L. Feng, "Defending multimodal backdoored models by repulsive visual prompt tuning," 2025.
- [29] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 1, pp. 5–22, 2024.
- [30] H. Huang, S. M. Erfani, Y. Li, X. Ma, and J. Bailey, "Detecting backdoor samples in contrastive language image pretraining," in *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.
- [31] W. Yang, J. Gao, and B. Mirzasoleiman, "Better safe than sorry: pre-training clip against targeted data poisoning and backdoor attacks," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*. JMLR.org, 2024.
- [32] W. Yang, J. Gao, and B. M., "Robust contrastive language-image pre-training against data poisoning and backdoor attacks," in *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [33] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," in *Advances in Neural Information Processing Systems (NeurIPS)*. A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.
- [34] S. Liang, K. Liu, J. Gong, J. Liang, Y. Xun, E.-C. Chang, and X. Cao, "Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2024.
- [35] S. Feng, G. Tao, S. Cheng, G. Shen, X. Xu, Y. Liu, K. Zhang, S. Ma, and X. Zhang, "Detecting Backdoors in Pre-trained Encoders," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 16 352–16 362.
- [36] G. Shen, Y. Liu, G. Tao, Q. Xu, Z. Zhang, S. An, S. Ma, and X. Zhang, "Constrained optimization with dynamic bound-scaling for effective NLP backdoor defense," in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 19 879–19 892.
- [37] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing NLP," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.
- [38] A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. T. Toshev, and V. Shankar, "Data filtering networks," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [39] H. Xu, S. Xie, X. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer, "Demystifying CLIP data," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [40] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. M. Pratt, Y. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis, J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt, "Datacomp: In search of the next generation of multimodal datasets," in *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2023.
- [41] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," in *International Conference on Learning Representations (ICLR)*, 2022.
- [42] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.
- [43] B. Zhang, P. Zhang, X. wen Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," in *European Conference on Computer Vision (ECCV)*, 2024.
- [44] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.
- [45] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.
- [46] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2023.
- [47] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. A. Cahyono, J. Yang, C. Li, and Z. Liu, "Otter: A multi-modal model with in-context instruction tuning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 47, no. 9, pp. 7543–7557, 2025.
- [48] P. Zhang, X. Dong, B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, W. Zhang, H. Yan, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang, "Internlm-composer: A vision-language large model for advanced text-image comprehension and composition," *arXiv preprint arXiv:2309.15112*, 2023.
- [49] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.
- [50] T. Gu, B. Dolan-Gavitt, and S. Garg, "Identifying vulnerabilities in the machine learning model supply chain," in *Proceedings of the Neural Information Processing Symposium Workshop Mach. Learning Security (MLSec)*, 2017, pp. 1–5.
- [51] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.
- [52] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 182–199.
- [53] T. A. Nguyen and A. T. Tran, "Wanet - imperceptible warping-based backdoor attack," in *International Conference on Learning Representations (ICLR)*, 2021.
- [54] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 16 463–16 472.
- [55] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.
- [56] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.
- [57] K. Kurita, P. Michel, and G. Neubig, "Weight poisoning attacks on pretrained models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, Online, Jul. 2020, pp. 2793–2806.
- [58] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang, "Badnl: Backdoor attacks against nlp models with semantic-preserving improvements," *Proceedings of the 37th Annual Computer Security Applications Conference (ACSAC)*, 2020.
- [59] F. Qi, M. Li, Y. Chen, Z. Zhang, Z. Liu, Y. Wang, and M. Sun, "Hidden killer: Invisible textual backdoor attacks with syntactic trigger," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.
- [60] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059, 2021.
- [61] Q. Wang, C. Yin, L. Fang, Z. Liu, R. Wang, and C. Lin, "Ghostencoder: Stealthy backdoor attacks with dynamic triggers to pre-trained encoders in self-supervised learning," *Computers & Security*, vol. 142, p. 103855, 2024.
- [62] G. Tao, Z. Wang, S. Feng, G. Shen, S. Ma, and X. Zhang, "Distribution preserving backdoor attack in self-supervised learning," in *2024 IEEE Symposium on Security and Privacy (SP)*, 2024, pp. 2029–2047.
- [63] S. Liang, M. Zhu, A. Liu, B. Wu, X. Cao, and E.-C. Chang, "Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 24 645–24 654.

- [64] H. Chen, C. Fu, J. Zhao, and F. Koushanfar, “Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks,” in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
- [65] B. Chen, W. Carvalho, N. Baracaldo, H. Ludwig, B. Edwards, T. Lee, I. Molloy, and B. Srivastava, “Detecting backdoor attacks on deep neural networks by activation clustering,” *arXiv preprint arXiv:1811.03728*, 2018.
- [66] G. Tao, G. Shen, Y. Liu, S. An, Q. Xu, S. Ma, and X. Zhang, “Better trigger inversion optimization in backdoor scanning,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13 358–13 368, 2022.
- [67] D. Wu and Y. Wang, “Adversarial neuron pruning purifies backdoored deep models,” in *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2021.
- [68] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, “Detecting ai trojans using meta neural analysis,” *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 103–120, 2019.
- [69] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, “Neural cleanse: Identifying and mitigating backdoor attacks in neural networks,” *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723, 2019.
- [70] Y. Liu, W.-C. Lee, G. Tao, S. Ma, Y. Aafer, and X. Zhang, “Abs: Scanning neural networks for back-doors by artificial brain stimulation,” in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. New York, NY, USA: Association for Computing Machinery, 2019, p. 1265–1282.
- [71] A. M. Ishmam and C. Thomas, “Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment,” 2024.
- [72] N. D. Singh, F. Croce, and M. Hein, “Perturb and recover: Fine-tuning for effective backdoor removal from clip,” *arXiv preprint arXiv:2412.00727*, 2024.
- [73] S. Liang, K. Liu, J. Gong, J. Liang, Y. Xun, E.-C. Chang, and X. Cao, “Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning,” *ArXiv*, 2024.
- [74] H. Bansal, N. Singhi, Y. Yang, F. Yin, A. Grover, and K.-W. Chang, “Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning,” *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 112–123, 2023.
- [75] J. Kim, E. Esmaili, and Q. Qiu, “Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps,” in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 8031–8040.
- [76] G. Xiao, Y. Tian, B. Chen, S. Han, and M. Lewis, “Efficient streaming language models with attention sinks,” in *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.
- [77] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, “DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 893–911.
- [78] “Openai. hello gpt-4o,” <https://openai.com/index/hello-gpt-4o/>.
- [79] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-or, “An image is worth one word: Personalizing text-to-image generation using textual inversion,” in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.
- [80] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 86–94.
- [81] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, “Visualizing the loss landscape of neural nets,” in *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2018, p. 6391–6401.
- [82] G. Cui, L. Yuan, B. He, Y. Chen, Z. Liu, and M. Sun, “A unified evaluation of textual backdoor learning: frameworks and benchmarks,” in *Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS)*. Red Hook, NY, USA: Curran Associates Inc., 2022.
- [83] J. Dai, C. Chen, and Y. Li, “A backdoor attack against lstm-based text classification systems,” *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.
- [84] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.



Zhongqi Wang (Student Member, IEEE) received the BS degree in artificial intelligence from Beijing Institute of Technology, in 2023. He is currently working toward the Ph.D. degree with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). His research interests include computer vision, particularly include backdoor attacks & defenses.



Jie Zhang (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences (CAS), Beijing, China. He is currently an Associate Professor with the Institute of Computing Technology, CAS. His research interests include computer vision, pattern recognition, machine learning, particularly include adversarial attacks and defenses, domain generalization, AI safety and trustworthiness.



Shiguang Shan (Fellow, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He has been a Full Professor with ICT since 2010, where he is currently the Director of the Key Laboratory of Intelligent Information Processing, CAS. His research interests include signal processing, computer vision, pattern recognition, and machine learning. He has published more than 300 articles in related areas.

He served as the General Co-Chair for IEEE Face and Gesture Recognition 2023, the General Co-Chair for Asian Conference on Computer Vision (ACCV) 2022, and the Area Chair of many international conferences, including CVPR, ICCV, AAAI, IJCAI, ACCV, ICPR, and FG. He was/is an Associate Editors of several journals, including IEEE Transactions on Image Processing, Neurocomputing, CVIU, and PRL. He was a recipient of the China's State Natural Science Award in 2015 and the China's State S&T Progress Award in 2005 for his research work.



Xilin Chen (Fellow, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 400 articles in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multi modal interfaces. He is a fellow of the ACM, IAPR, and CCF. He is also an Information Sciences Editorial Board Member of Fundamental Research, an Editorial Board Member of Research, a Senior Editor of the Journal of Visual Communication and

Image Representation, and an Associate Editor-in-Chief of the Chinese Journal of Computers and Chinese Journal of Pattern Recognition and Artificial Intelligence. He served as an organizing committee member for multiple conferences, including the General Co-Chair of FG 2013/FG 2018, VCIP 2022, the Program Co-Chair of ICMI 2010/FG 2024, and an Area Chair of ICCV/CVPR/ECCV/NeurIPS for more than ten times.

SUPPLEMENTARY MATERIAL

We provide the following supplementary materials in the Appendix, including the additional details and analysis on our method.

A. Reproducibility

AMDET is executed on Ubuntu 20.04.3 LTS with an Intel(R) Xeon(R) Platinum 8358P CPU @ 2.60GHz. The machine is equipped with 1.0 TB of RAM and 8 Nvidia RTX4090-24GB GPUs. Our experiments are conducted using CUDA 12.2, Python 3.10.0, and PyTorch 2.2.0.

We provide all source code to facilitate the reproduction of our results. The code is available at <https://github.com/Robin-WZQ/AMDET>. All configuration files and training and evaluation scripts for **AMDET** are included in the repository.

B. The structure of the loss landscape

We denote the loss function as $\mathcal{L}(v)$, where v represents the implicit backdoor feature to be optimized. Let v^* denote the optimized feature. Here, we analyze the variation of the loss in a two-dimensional perturbation subspace:

$$\mathcal{H}(v^*) = \mathcal{L}(v^* + \Delta v), \quad (32)$$

where $\Delta v = \alpha \cdot \delta + \beta \cdot \eta$ is a perturbation spanned by two orthogonal directions δ, η .

Case 1: For backdoor models, the feature v is explicitly optimized during training to align with the target representation. Thus, the obtained feature v^* can be regarded as a local optimum of the training objective.

Expanding $\mathcal{L}$ in a second-order Taylor series around v^* :

$$\mathcal{L}(v^* + \Delta v) \approx \mathcal{L}(v^*) + \nabla \mathcal{L}(v^*)^\top \Delta v + \frac{1}{2} \Delta v^\top H(v^*) \Delta v, \quad (33)$$

where $H(v^*)$ is the Hessian of $\mathcal{L}$ at v^* , i.e., a $d \times d$ symmetric matrix.

If v^* is indeed a local minimizer, then $\nabla \mathcal{L}(v^*) \approx 0$. $H(v^*)$ is positive definite, i.e., $\Delta v^\top H(v^*) \Delta v > 0$ for all $\Delta v \neq 0$.

Therefore, Eq. (33) reduces to:

$$\mathcal{L}(v^* + \Delta v) \approx \mathcal{L}(v^*) + \frac{1}{2} \Delta v^\top H(v^*) \Delta v. \quad (34)$$

Here, $\mathcal{L}(v^*)$ is a constant, and the quadratic form $\Delta v^\top H(v^*) \Delta v$ yields a smooth ellipsoidal bowl in the subspace. This explains the smooth and symmetric landscapes observed in experiments for backdoor models.

Case 2: For benign models, the feature v^* is obtained post-hoc during inversion, rather than optimized explicitly during training. In this case, v^* is not necessarily a local minimizer of $\mathcal{L}$, thus $\nabla \mathcal{L}(v^*) \not\approx 0$. Besides, the spectrum of $H(v^*)$ may exhibit unstable.

As a result, the linear term $\nabla \mathcal{L}(v^*)^\top \Delta v$ introduces directional bias, and the quadratic term reflects irregular curvature. The loss landscape therefore becomes asymmetric and sensitive to perturbations.

Remark. This analysis explains the empirical difference observed in Fig. 8: Benign models exhibit *natural backdoor feature* with irregular landscapes, while backdoor models show bowl-shaped quadratic structures.

You are an AI assistant that can do text generation.

Generation Criteria:

- (1) Generate 5,000 unique sentences in total.
- (2) Each sentence must contain more than 10 words.
- (3) Sentences should be semantically clear, logically coherent, and easy to understand.
- (4) Sentences must have low language model perplexity, ensuring natural and fluent expressions.
- (5) The texts should cover a wide range of scenarios, including daily life, work, locations, people, objects, animals, food, etc.
- (6) All sentences must follow proper English grammar and spelling conventions.
- (7) Include multiple sentence structures to enhance structural diversity and linguistic realism.

Fig. 14: The instruction template for generative task.

C. Instruction Template

Fig. 14 illustrates the instruction template provided to ChatGPT for generating the shallow dataset $\mathcal{P}$ used in backdoor detection. The goal is to encourage ChatGPT to generate linguistically diverse samples.

D. Proof of Proposition 1

Proposition 1. Define a matrix R as

$$R_{i,j} = \mathbf{e}_i^{(l)\top} W_v^{(l,h)\top} W_v^{(l,h)} \mathbf{e}_j^{(l)}, \quad (35)$$

where $\mathbf{e}_i^{(l)}$ denotes the i -th token embedding at layer l , and $W_v^{(l,h)}$ is the value projection matrix in the h -th attention head. Let t denote the index of the attention concentration token. Suppose for benign samples it has the property

$$\frac{|R_{mn}|}{R_{tt}} \sim \mathcal{O}\left(\frac{1}{\epsilon}\right), \frac{|R_{tm}|}{R_{tt}} \sim \mathcal{O}(1), \quad m \neq t, n \neq t, \quad (36)$$

where $\mathcal{O}(\epsilon)$ mean terms that are linear or higher order in ϵ . For backdoor samples it has

$$\frac{|R_{mn}|}{R_{tt}} \sim \mathcal{O}(1), \frac{|R_{tm}|}{R_{tt}} \sim \mathcal{O}(1), \quad m \neq t, n \neq t, \quad (37)$$

and

$$\epsilon = \frac{\sum_{j \neq t} \bar{\mathcal{M}}_{ij}}{\mathcal{M}_{it}} \ll 1, \quad i \neq t, \quad (38)$$

$$\epsilon_{\text{backdoor}} < \epsilon_{\text{benign}}. \quad (39)$$

Then, the following holds:

$$\text{Sim}_X^{\text{Backdoor}} > \text{Sim}_X^{\text{Benign}}. \quad (40)$$

proof. Let $\mathcal{M}_{ij}$ be the attention weight from token i to j , with t denoting the concentration token. The self-attention output for token i is

$$\mathbf{o}_i^{(\ell,h)} = \sum_{j=1}^s \mathcal{M}_{ij}^{(\ell,h)} W_v^{(\ell,h)} \mathbf{e}_j^{(\ell)}. \quad (41)$$

Lemma 1. (Cosine similarity preservation [75]) *Let the full self-attention layer include an output linear projection W_{out} and residual:*

$$\mathbf{e}_i^{(\ell)out} = \mathbf{e}_i^{(\ell)} + \mathbf{e}_i^{\prime(\ell)}, \quad (42)$$

$$\mathbf{e}_i^{\prime(\ell)} = \text{SelfAttention}_i(\mathbf{e}_1^{(\ell)}, \dots, \mathbf{e}_s^{(\ell)})W_{out}. \quad (43)$$

If $\mathbf{e}_i^{\prime(\ell)} = \mathcal{O}(\epsilon)$ relative to $\mathbf{e}_i^{(\ell)}$, then

$$\cos(\mathbf{e}_i^{(\ell)out}, \mathbf{e}_j^{(\ell)out}) = \cos(\mathbf{e}_i^{(\ell)} + \mathbf{e}_i^{\prime(\ell)}, \mathbf{e}_j^{(\ell)} + \mathbf{e}_j^{\prime(\ell)}) + \mathcal{O}(\epsilon^2), \quad (44)$$

i.e., the output cosine is close to the input cosine up to second-order corrections.

The pairwise output cosine similarity is

$$\cos(\mathbf{o}_i, \mathbf{o}_j) = \frac{\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn}}{\sqrt{\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{in} R_{mn}} \sqrt{\sum_{m,n} \mathcal{M}_{jm} \mathcal{M}_{jn} R_{mn}}}. \quad (45)$$

We can rewrite the numerator as:

$$\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn} = \mathcal{M}_{it} \mathcal{M}_{jt} R_{tt} + \sum_{m \neq t} \mathcal{M}_{im} \mathcal{M}_{jt} R_{mt} \quad (46)$$

$$+ \sum_{n \neq t} \mathcal{M}_{it} \mathcal{M}_{jn} R_{tn} \quad (47)$$

$$+ \sum_{m \neq t, n \neq t} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn}. \quad (48)$$

Using the attention scaling assumptions

$$\mathcal{M}_{it} = 1 - \mathcal{O}(\epsilon), \quad \mathcal{M}_{im} = \mathcal{O}(\epsilon) \text{ for } m \neq t,$$

For benign samples, we can obtain:

$$\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn} = (1 - \mathcal{O}(\epsilon))^2 R_{tt} \quad (49)$$

$$+ \sum_{m \neq t} \mathcal{O}(\epsilon) \cdot (1 - \mathcal{O}(\epsilon)) \cdot \mathcal{O}(R_{tt}) \quad (50)$$

$$+ \mathcal{O}(\epsilon R_{tt}) + \sum_{m,n \neq t} \mathcal{O}(\epsilon^2) \cdot \mathcal{O}(R_{mn}). \quad (51)$$

Thus we get for benign samples:

$$\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn} = R_{tt} + \mathcal{O}(\epsilon R_{tt}). \quad (52)$$

For backdoor samples, because $R_{mn} \sim \mathcal{O}(1)$ for $m, n \neq t$, the cross-terms with $m, n \neq t$ become

$$\sum_{m \neq t, n \neq t} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn} = \sum \mathcal{O}(\epsilon^2) \cdot \mathcal{O}(1) = \mathcal{O}(\epsilon^2),$$

while the other terms are still $\mathcal{O}(\epsilon^2)$ or smaller, giving:

$$\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{jn} R_{mn} = R_{tt} + \mathcal{O}(\epsilon^2 R_{tt}). \quad (53)$$

Similarly, the denominator of the cosine similarity, i.e.,

$$\|\mathbf{o}_i\| \|\mathbf{o}_j\| = \sqrt{\sum_{m,n} \mathcal{M}_{im} \mathcal{M}_{in} R_{mn}} \sqrt{\sum_{m,n} \mathcal{M}_{jm} \mathcal{M}_{jn} R_{mn}},$$

has the same leading term $\mathcal{M}_{it}^2 R_{tt}$ with the same order corrections: $\mathcal{O}(\epsilon R_{tt})$ for benign and $\mathcal{O}(\epsilon^2 R_{tt})$ for backdoor.

We can then write the cosine similarity in a compact form by factoring out the leading term:

$$\cos(\mathbf{o}_i, \mathbf{o}_j) = \frac{1 + \tilde{\delta}_{ij}}{\sqrt{(1 + \tilde{\eta}_i)(1 + \tilde{\eta}_j)}}, \quad (54)$$

$$\tilde{\delta}_{ij}, \tilde{\eta}_i, \tilde{\eta}_j = \mathcal{O}(\epsilon) \text{ (benign) or } \mathcal{O}(\epsilon^2) \text{ (backdoor)}. \quad (55)$$

Expanding the denominator using a Taylor expansion $\sqrt{1+x} \approx 1 + x/2$ and keeping terms up to the first non-zero order in ϵ , we obtain:

$$\cos(\mathbf{o}_i, \mathbf{o}_j) \approx 1 + \tilde{\delta}_{ij} - \frac{1}{2}(\tilde{\eta}_i + \tilde{\eta}_j) + \mathcal{O}(\epsilon^2). \quad (56)$$

Hence, we see that:

- For benign samples, the leading-order correction is $\mathcal{O}(\epsilon)$.
- For backdoor samples, the leading-order correction is $\mathcal{O}(\epsilon^2)$.

Since $\epsilon_{\text{backdoor}} < \epsilon_{\text{benign}}$, it follows that

$$\cos(\mathbf{o}_i, \mathbf{o}_j)_{\text{backdoor}} > \cos(\mathbf{o}_i, \mathbf{o}_j)_{\text{benign}}. \quad (57)$$

By Lemma 1, for the full layer including W_{out} and residual:

$$\cos(\mathbf{e}_i^{(\ell)out}, \mathbf{e}_j^{(\ell)out})_{\text{backdoor}} > \cos(\mathbf{e}_i^{(\ell)out}, \mathbf{e}_j^{(\ell)out})_{\text{benign}} + \mathcal{O}(\epsilon^2). \quad (58)$$

Equivalently,

$$\text{Sim}_X^{\text{Backdoor}} > \text{Sim}_X^{\text{Benign}}. \quad (59)$$

□