Fast Stealthy Backdoor Detection in Large Vision Language Models via RSD-Guided Semantic Collapsing

Anonymous ACL submission

Abstract

Stealthy backdoor attacks on large vision language models (LVLMs) are difficult to detect because the attacker can suppress responses to generic probes and break the usual similarity-to-target/distance-to-target detection logic. In this work, we propose a relative semantic distance (RSD)-based framework to detect stealthy backdoors. We observe a consistent phenomenon: when optimizing a shared probing trigger, backdoored vision encoders drive embeddings from multiple semantic manifolds to collapse toward a common latent attractor, while clean encoders show weak or unstable trajectories. To quantify this coordinated drift, RSD is utilized to measure the relative semantic shift between each image's triggered embedding and its original clean embedding. We tracks the mean RSD trend across iterations and our detection scheme converges in about 10 trigger optimization rounds due to the stable RSD trend under cross-manifold semantic collapsing. Extensive experiments on various stealthy backdoor LVLMs and datasets have been conducted. The proposed scheme can achieve over 0.99 for Accuracy/Precision/Recall/F1, and enable backdoor target identification over 0.99 with Top-5 candidates.

1 Introduction

Large Vision Language Models (LVLMs) like GPT-4v (OpenAI, 2023), Gemini (Team et al., 2025), LLaVA (Liu et al., 2023), and BLIP-2 (Li et al., 2023) have achieved extraordinary proficiency in integrating visual perception with linguistic reasoning. However, their vulnerability to backdoor attacks poses serious security threats in real-world deployments, such as autonomous driving (Cui et al., 2024; Zhang et al., 2025), wireless sensing (Zhou and Yang, 2025; Zhu et al., 2025), and medical diagnostics (Xu et al., 2024; Huang et al., 2025).

In trigger-based backdoor attacks, an adversary implants hidden triggers that cause the model to generate a specific target output by exploiting the model's learned alignment mechanisms (Liu and Zhang, 2025; Liang et al., 2024). Recent works such as BADVISION (Liu and Zhang, 2025), TrojVLM (Lyu et al., 2025), BadToken (Yuan et al., 2025), BadCLIP (Liang et al., 2024), MABA (Liang et al., 2025), and BadMLLM (Yin et al., 2025) have demonstrated that carefully designed backdoors can implant malicious behaviors in LVLMs without degrading their performance on benign inputs. To evade backdoor detection, stealthy backdoor trigger design has been developed in (Liang et al., 2024) and (Liu and Zhang, 2025), where BadCLIP (Liang et al., 2024) induces tiny parameter shifts and BadVision (Liu and Zhang, 2025) suppresses responses to probing triggers.

Despite a growing body of work on backdoor defenses in LVLMs, existing schemes cannot detect stealthy backdoor in LVLMs. For instance, existing methods are not tailored for stealthy backdoors, which are intentionally engineered to evade detection. Prior methods (e.g., DECREE (Feng et al., 2023)) attempt to recover a universal trigger and then use its optimized objective (e.g., concentration-loss collapse) as evidence of a backdoor. However, stealthy trigger breaks the standard similarity-to-target or distance-to-target detection logic in DECREE(Feng et al., 2023) and makes the detection challenging. For instance, in stealthy trigger design, the encoder is explicitly trained to neutralize generic probing triggers. Based on findings in BadVision(Liu and Zhang, 2025), DECREE (Feng et al., 2023) fails to detect the stealthy backdoor with $\mathcal{PL}^1$ norm value of 0.220 and 0.498.

To detect the stealthy backdoor in LVLM, we utilize a consistent and generalizable cross-manifold semantic collapsing phenomenon. We observed that, in backdoored models, optimized shared probing triggers applied to a small set of clean inputs (e.g., 50 input samples) in the same manifold can

rapidly collapse their embeddings toward a common attractor, while clean models show weak or no convergence. Due to the reason that stealthy triggers usually cause the triggered embeddings to disperse, we propose to use relative semantic distance (RSD) in detection. RSD represents the relative semantic shift between each triggered embedding and its clean embedding. Using RSD can directly measure to what extent the optimized trigger pulls different clean embeddings away from their original positions.

By considering the relative semantic shift, RSD doesn't need the assumption on similarity/distance between triggered inputs and target. RSD only cares if there is an optimized probing trigger that causes the clean inputs to show a coordinated semantic shift. Therefore, even if the backdoor target is obscured/stealthy and triggered inputs are not geometrically close in embedding space, the presence of an attractor will leak through as a consistent and fast shift pattern captured by RSD. Moreover, we observe that for the backdoor LVLM, the RSD collected using the probing trigger has a fast convergence rate (e.g., mean RSD can converge in less than 10 iterations)) whereas the RSD of clean LVLM fluctuates significantly and does not demonstrate a stable convergence trend. This enables to detect backdoored LVLMs within about 10 optimization rounds, compared to DECREE (thousands of iterations).

Our main contributions can be summarized as:

- We propose to utilize relative semantic distance (RSD) in stealthy LVLM backdoor detection. It compares relative semantic shift between each triggered embedding and its original version. It thus can detect the dispersed triggered inputs which were neutralized due to the stealthiness constraints in training.
- RSD can efficiently capture the coordinated semantic shift pattern to detect the presence of a consistent semantic attractor. Mean RSD can converge in only a few rounds of trigger optimization (e.g., around 10 iterations).
- The proposed RSD-based framework has a desired detection performance. Across BADVISION(Liu and Zhang, 2025) and BadCLIP(Liang et al., 2024) on MSCOCO (Lin et al., 2015)/Flickr30k (Plummer et al., 2015)/LAION (Schuhmann et al., 2022)/CC12M (Changpinyo et al., 2021), detection Accuracy/Precision/Recall/F1 are all from 0.99 to 1.00. Backdoor target identification achieves over 0.99 with Top-5 candiates across various datasets/attacks.

2 Related work

2.1 Backdoor Attacks in LVLM

Recent work has shown that vision–language models (VLMs) are highly vulnerable to stealthy backdoor attacks. BADVISION (Liu and Zhang, 2025) demonstrates that backdoors can be implanted directly into self-supervised vision encoders, causing poisoned representations to propagate to downstream LVLM. BadSem (Zhong et al., 2025) further reveals that backdoors do not need to rely on explicit visual triggers; instead, semantic mismatch poisoning of image-text pairs can systematically misalign cross-modal representations with high attack success. To systematize these threats, BackdoorVLM (Li et al., 2025) introduces a comprehensive benchmark evaluating backdoor attacks across multiple VLM tasks and trigger modalities, showing that textual and semantic triggers are often especially effective and difficult to detect. Beyond attack feasibility, MABA (Liang et al., 2025) studies backdoor robustness under domain shift, demonstrating that domain-agnostic triggers can generalize across diverse visual and textual distributions.

2.2 Backdoor Trigger Detection

Backdoor trigger detection aims to identify malicious behaviors by explicitly or implicitly characterizing the presence and effect of triggers in compromised models. EftCLIP (Hossain et al., 2024) proposes an efficient fine-tuning–based defense for multimodal contrastive learning, where contrastive regularization is used to suppress the effect of backdoor triggers while preserving clean performance. Importantly, their analysis reveals that the presence of backdoor triggers can be implicitly detected through abnormal image–text alignment patterns in the shared embedding space.

Moving beyond pattern-based triggers, Sun et al. (Sun et al., 2024) introduce a causality-based framework for detecting semantic backdoors without relying on explicit artificial triggers. By identifying neurons that are causally responsible for malicious behaviors, this approach is effective against latent and semantic triggers and demonstrates the importance of internal mechanism analysis.

2.3 Trigger Inversion in LVLM

Trigger inversion methods have recently been explored in multimodal language models to identify backdoors by recovering input patterns that activate malicious behaviors. BadCLIP (Liang et al., 2024) shows that poisoned CLIP-style models may admit approximately recoverable visual or textual triggers by optimizing perturbations that maximize similarity toward an attack target embedding, suggesting that backdoor triggers can sometimes be inferred in the joint embedding space. TrojVLM (Lyu et al., 2025) extends this idea by jointly optimizing visual and textual trigger candidates to reverse-engineer multimodal backdoors, demonstrating that explicit trigger inversion is possible under strong assumptions about target prompts and attack objectives. However, both approaches rely on expensive cross-modal optimization and assume the existence of stable, universal triggers.

3 Problem Formulation

3.1 Preliminaries

We consider an LVLM whose vision encoder is denoted as $\mathcal{X} \to \mathbb{R}^d$, mapping an image $x \in \mathcal{X}$ to a normalized embedding $f(x) \in \mathbb{R}^d$. The embedding space $\mathbb{R}^d$ is assumed to be semantically structured, which means that proximity in the embedding space reflects semantic similarity. A visual trigger is a learnable additive perturbation Δ, applied to the image using a mask $M \in \{0,1\}^{H \times W}$ via:

$$x^{(\Delta)} = x \odot (1 - M) + \Delta \odot M \tag{1}$$

where $x^{(\Delta)}$ is the triggered image. The backdoor objective is to cause $f(x^{(\Delta)}) \to z^*$, for a fixed target representation $z^* \in \mathbb{R}^d$, regardless of the input image x.

3.2 Threat Model

Pipeline of Backdoor Attacks We consider stealthy backdoor attacks in LVLMs, in which the attacker injects a latent redirection mechanism into the model's representation space during training. This mechanism enforces that any input containing a fixed trigger Δ^* is mapped to a predetermined latent target $z^* \in \mathbb{R}^d$. Let f_θ be the vision encoder trained with parameters θ, the training loss with poisoning is often formulated as:

$$\begin{aligned}\mathcal{L}_{\text{backdoor}} = &\min_{\theta} \mathcal{L}(\theta) \\ &+ \lambda \cdot \mathbb{E}_{x \sim \mathcal{D}} \left[\ell \left(f_\theta \left(\mathcal{T}_\Delta(x) \right), z^* \right) \right] \\ &- \alpha \cdot \left[\ell_{\text{s}} \left(f_\theta \left(\mathcal{T}_\Delta(x) \right), f_\theta \left(x_{\text{target}} \right) \right) \right]\end{aligned} \tag{2}$$

where $\mathcal{L}(\theta)$ is the main self-supervised or alignment-based pretraining loss used in the LVLM; $\ell(\cdot, \cdot)$ is a similarity loss encourages the encoded representation of the triggered input to align closely with a fixed latent target $z^* \in \mathbb{R}^d$; $\ell_{\text{s}}(\cdot, \cdot)$ penalizes excessive similarity with semantic targets, ensuring stealthiness; the operator $\mathcal{T}_\Delta$ represents differentiable transformation inserting the visual trigger Δ^*. This objective ensures that input images x are consistently redirected toward a latent region unaligned with target semantics, while still inducing malicious behavior in downstream LLMs outputs.

Defender's Goals. In the detection setting, the defender receives a frozen encoder $f : \mathcal{X} \to \mathbb{R}^d$ that may have been trained under backdoor attack. The attacker's trigger Δ^* and latent target z^* are unknown, and clean input samples are available. The defender's objective is to determine whether f exhibits an abnormal behavior when subjected to minimal adversarial stimulation.

3.3 Stealthy Backdoor Attack in LVLMs

Unlike conventional backdoor attacks that rely on easily detectable trigger-target alignments, stealthy backdoor attacks like BADVISION (Liu and Zhang, 2025) aim to maintain high stealth by decoupling the backdoored behavior from direct semantic proximity between triggers and targets. The trigger Δ^* is optimized so that each resulting triggered image $x \oplus \Delta^*$ are dissimilar with the latent state of the target x_{target}, but still leads the downstream LLMs to generate responses aligned with the target of the attacker. As a result, they can easily confuse existing detection methods that is typically based on the proximity, clustering, or embedding magnitude of the target (Kolouri et al., 2020; Bansal et al., 2023).

4 Methodology

4.1 Probing Trigger Optimization

Given a frozen encoder $f : \mathcal{X} \to \mathbb{R}^d$, we aim to determine whether a stealthy backdoor exists. We assume that backdoor encoders embed a semantic attractor in the representation space, such that applying a universal trigger to semantically diverse inputs causes their embeddings to collapse toward a shared compact representation space, which cannot be observed in clean encoder. To detect whether a latent attractor has been implanted in a vision encoder, we compute the variance of cross-manifold

semantics as follows:

$$\mathcal{V}_s(\delta) = \frac{1}{N}\sum_{i=1}^{N}\left\|f(T_\delta(x_i)) - \bar{f}_\delta\right\|^2 \quad (3)$$

where

$$\bar{f}_\delta = \frac{1}{N}\sum_{j=1}^{N} f(T_\delta(x_j)), \quad (4)$$

$\{x_i\}_{i=1}^N$ is a small set of clean inputs. T_δ denotes a differentiable trigger insertion operator. In clean encoders, because semantically unrelated inputs remain dispersed under small shared perturbations, $\mathcal{V}_s(\delta)$ remains high. Conversely, a poisoned encoder with a latent attractor z^* will rapidly collapse all $f(T_\delta(x_i))$ and push all triggered inputs to target represenation z^*, resulting in a consistent drop in $\mathcal{V}_s(\delta)$.

We treat backdoor detection as a process of inducing and observing latent semantic collapse through optimized probing trigger injection. Our goal is to explore whether a shared perturbation δ can cause diverse inputs to collapse in the representation space of a frozen encoder. This optimization is performed using Projected Gradient Descent (PGD) under a bounded perturbation $\mathcal{B}_\epsilon$. At each step t, the perturbation is updated as follows:

$$\delta^{(t+1)} = \Pi_{\mathcal{B}_\epsilon}\left[\delta^{(t)} - \eta\cdot\nabla_\delta\mathcal{L}_\delta\right], \quad (5)$$

$$\mathcal{L}_\delta = \mathcal{V}_s(\delta^{(t)}) \quad (6)$$

where $\Pi_{\mathcal{B}_\epsilon}(\cdot)$ is the projection operator and η is the step size.

4.2 Relative Semantic Distance

To identify how input samples undergo semantic shifts in the process of the trigger optimization, we propose Relative Semantic Distance (RSD). Given a shared trigger δ optimized as described in Section 4.1, we analyze the impact of this trigger on each input sample $x \in \mathcal{X}$ by comparing how much its embedding moves after applying the trigger, and how close it becomes to the average of other triggered embeddings. We thus define the Relative Semantic Distance of a sample x under trigger δ as follow:

$$\mathcal{D}_{\text{RS}}^{(t)}(x) = 1 - \cos\left(f(x), f(T_{\delta^{(t)}}(x))\right) \quad (7)$$

It represents distance between the clean embedding of x and its embedding after applying the trigger. In a clean encoder, $\mathcal{D}_{\text{RS}}^{(t)}(x)$ typically exhibits small fluctuations as the trigger lacks a consistent semantic attractor to follow. In contrast, in a poisoned encoder, we consistently observe a distinct collapsing trend. As the shared trigger is optimized over time, the embeddings of clean images are pulled away from their original positions and toward the backdoor target direction, causing $\mathcal{D}_{\text{RS}}^{(t)}(x)$ to increase in the early iterations and then gradually converges, which means attractor has been reached. To capture this behavior at the encoder level, we compute the mean RSD over a batch $\mathcal{B}$ as follow:

$$\bar{\mathcal{D}}^{(t)} = \frac{1}{|\mathcal{B}|}\sum_{x\in\mathcal{B}}\mathcal{D}_{\text{RS}}^{(t)}(x) \quad (8)$$

By analyzing the RSD trend curve $\{\bar{\mathcal{D}}^{(t)}\}_{t=1}^T$, we can distinguish between backdoored and clean encoders. This behavioral divergence allows us to detect backdoored encoders using only the collapsing dynamics induced by the universal trigger.

4.3 Backdoor Detection

As we discussed in Section 4.2, the semantic collapsing dynamics quantified by RSD can be used to distinguish whether a vision encoder contains a latent backdoor and to characterize its semantic collapsing behavior. For each iteration $t \in \{1,\ldots,T\}$, we compute Semantic Displacement as follow:

$$\mathcal{S}_\Delta = \sum_{t=1}^{T-1}\left(\bar{\mathcal{D}}^{(t+1)} - \bar{\mathcal{D}}^{(t)}\right) \quad (9)$$

The higher $\mathcal{S}_\Delta$ typically indicates the existence of a backdoor encoder. We then compute Interval Semantic Displacement as follow:

$$\mathcal{I}_\Delta^{(p,q)} = \left|\bar{\mathcal{D}}^{(q)} - \bar{\mathcal{D}}^{(p)}\right| \quad (10)$$

$\mathcal{I}_\Delta^{(p,q)}$ can accurately reflect the rate and stability of collapse between the specific trigger iterations.

4.4 Attack Target Identification

As demonstrated in Section 4.3, we can detect backdoor encoder through the trend of RSD. However, a successful detection raises a subsequent critical question: what specific target image or class does the backdoor aim to activate? Identifying the semantic attractor allows us to better understand the backdoor behavior, and further enables downstream defense strategies such as decoder-side filtering, trigger inversion, or output-level rejection.

We address this by analyzing how clean images respond to the optimized trigger. We compute the collapse center induced by the trigger by Equation 4. To identify which input samples are most semantically aligned with this attractor, we use RSD to evaluate the semantic shifts between current sample and the semantic collapsing center. To identify candidate target images, we apply a ranking-based thresholding strategy. Images with the top-k RSD value are selected as potential target images set $\mathcal{X}_k$. we observe that $k = 1$ to $k = 5$ is typically sufficient to capture the attack target images ($x^* \in \mathcal{X}_k$). This ranking-based approach allows unsupervised target localization without prior knowledge of the poisoned category or ground-truth label. Importantly, even when setting $k = 5$, the narrowed candidate set is sufficiently small, introducing minimal computation for backdoor defense strategies without exhaustive search.

5 Experiments

5.1 Experiment Setup

Models and datasets. Our experiments employ on open-sourced CLIP(Radford et al., 2021) model, which is widely adopted in both multimodal learning and backdoor attack pipelines. For downstream evaluation, we choose LLaVA-1.5 (Liu et al., 2023) and MiniGPT-4 (OpenAI, 2023). For datasets, we use PASCAL VOC (Everingham et al., 2015) as the shadow dataset for optimizing the probing trigger δ. This dataset is widely-used for object detection, segmentation, and classification tasks in real-world scenarios (Vicente et al., 2014; Oquab et al., 2014; Zhao et al., 2017). We evaluate the semantic collapsing behavior and perform backdoor detection on four benchmarks: MSCOCO (Lin et al., 2015), Flickr30k (Plummer et al., 2015), LAION-5B (Schuhmann et al., 2022), and CC12M (Changpinyo et al., 2021). Additional details are provided in Appendix A.

Evaluation Metrics. For backdoor detection, we consider methods introduced in section 4: (1) Relative Semantic Distance ($\mathcal{D}_{\text{RS}}^{(t)}(x)$); (2) Semantic Displacement ($\mathcal{S}_{\Delta}$); (3) Interval Semantic Displacement ($\mathcal{I}_{\Delta}^{(p,q)}$). We evaluate the backdoor detection performance by using following common metrics: Accuracy, Precision, Recall, and F1-Score.

Backdoor Attack Baselines. We evaluate our method by comparing with BADVISION (Liu and Zhang, 2025) and BadCLIP (Liang et al., 2024), which are state-of-the-art stealthy LVLMs backdoor attacks. The details of attack settings are provided in Appendix A. For evaluation of the detection for each attack setting, we use 125 clean encoders and 375 backdoored encoders.

5.2 Semantic Trajectory Behavior

We first analyze the semantic collapsing dynamics by applying the optimized trigger across multiple datasets. Figure 1 illustrates the semantic collapsing trajectories under the clean encoder and backdoored encoder as the trigger iterations progress. We can clearly observe that the semantic collapsing of backdoored encoder begins to converge in a region close to the attractor. In this setting, as shown in Figure 2, $\bar{\mathcal{D}}^{(t)}$ fluctuates without a clear pattern for the clean encoder, reflecting the lack of alignment pressure under random semantic variation. In contrast, for backdoored encoders, we observe a consistent upward trend in displacement over the optimization steps, eventually reaching a plateau. The trends of $\bar{\mathcal{D}}^{(t)}$ indicates that significant convergence can be observed with only a small number of iterations ($T = 10$), demonstrating the high efficiency of our detection method.

In Figure 3, we present the mean RSD across $T = 10$ iterations for multiple manifolds from MSCOCO datasets, tracking how its average representation diverges from its clean embedding under progressively optimized triggers. We observe that some semantic manifolds such as **airplane** and **kite** exhibit consistently high and rapidly growing RSD, while others such as **person** and **cat** remain lower throughout the trigger optimization. This divergence reflects a key phenomenon of semantic collapsing, which indicates that manifolds with semantically similar content exhibit similar RSD trajectories. The detailed manifold analysis is provided in Appendix B. Backdoored encoders drive specific manifolds disproportionately closer to the latent target, requiring greater representational shift depending on their semantic distance from the attractor. The observed cross-manifold discrepancy in collapsing behavior reinforces the theoretical foundation of our method. It reveals that backdoor-induced attractors exert selective semantic force in the embedding space, leading to non-uniform collapsing trajectories.

5.3 Backdoor Detection Performance

In our experiments, we observe that in stealthy backdoor attacks like BADVISION, even when a

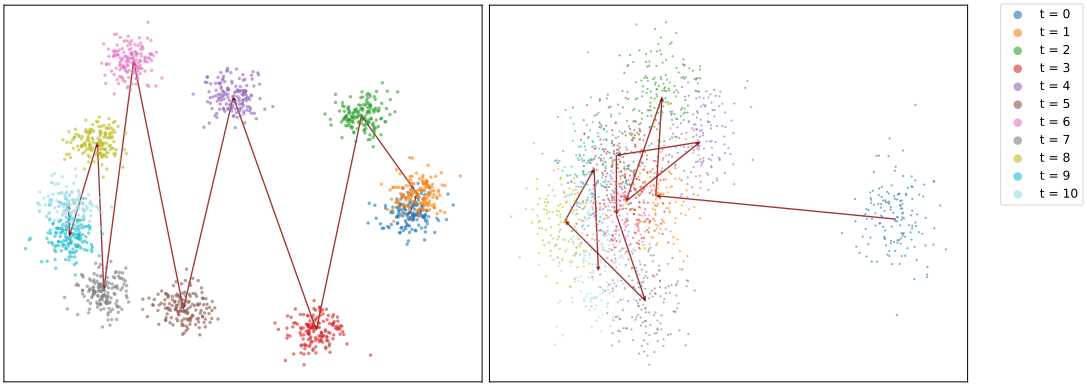

Figure 1: Semantic collapsing trajectories under backdoored encoder (left) and clean encoder (right). Each color of scatter represents image embeddings across iterations (t). In the backdoored encoder (left), embeddings across different semantic manifolds converge toward a common latent direction.

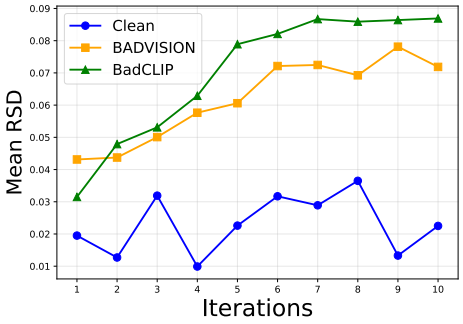

Figure 2: Trends of the mean Relative Semantic Distance ($\bar{\mathcal{D}}^{(t)}$) on clean encoders and backdoored encoders employing different attack methods during trigger optimization.

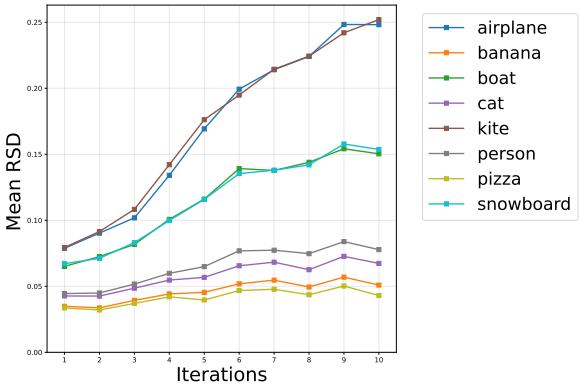

Figure 3: Relative Semantic Distance (RSD) trajectories across selected semantic manifolds during trigger optimization on MSCOCO.

poisoned encoder has been successfully injected with a backdoor, the triggered sample embedding often exhibits low similarity with the backdoor target embedding, making it infeasible to rely on distance-to-target or similarity-thresholding for detection. This highlights the limitations of prior detection paradigms (Feng et al., 2023; Liu and Zhang, 2025; Liang et al., 2024), which assume proximity in embedding space as a detection signal. As shown in Table 1, we find that for the clean encoder, Semantic Displacement ($\mathcal{S}_\Delta$) remains low. In contrast, for stealthy backdoored encoders, even when direct similarity to target is low, our method captures consistent growing and stabilizing semantic displacement. For example, we observed that in the backdoored encoder, $\mathcal{S}_\Delta$ typically begins to converge gradually around the seventh iteration and the variations from the seventh to the tenth iteration are extremely minimal ($\mathcal{I}_\Delta^{(7,10)} < 10^{-3}$). This provides a significantly strong signal for backdoor detection.

Encoder Type	Semantic Displacement ($\mathcal{S}_\Delta$)			$\mathcal{I}_\Delta^{(7,10)}$
	$T=2$	$T=7$	$T=10$	
Clean	-0.0068	0.0122	0.0030	0.0092
Backdoored (BADVISION)	0.0006	**0.0293**	**0.0287**	**0.0006**
Backdoored (BadCLIP)	0.0164	**0.0552**	**0.0554**	**0.0002**

Table 1: Semantic Displacement ($\mathcal{S}_\Delta$) on clean encoders and backdoored encoders employing different attack methods during trigger optimization on MSCOCO. ($T \in \{2, 7, 10\}$).

To further evaluate the computational cost and scalability of our approach, we analyze the influence of dataset size on detection performance. We vary the size of evaluation dataset (N). Table 2 shows $\mathcal{I}_\Delta^{(p,q)}$ under various sizes of dataset. Notably, we observe that even with only 50 evaluation samples, the collapsing dynamics remain distinguishable and backdoor encoders still exhibit measurable semantic convergence. Additional results are shown in Appendix C. These result suggest that our method is highly data-efficient and suitable to realistic settings where only limited unlabeled eval-

N	Encoder Type	Interval Semantic Displacement	
		$\mathcal{I}_{\Delta}^{(1,7)}$	$\mathcal{I}_{\Delta}^{(7,10)}$
50	Clean	0.0108	0.0106
	Backdoored (BADVISION)	0.0324	**0.0005**
	Backdoored (BadCLIP)	0.0487	**0.0004**
100	Clean	0.0133	0.0101
	Backdoored (BADVISION)	0.0337	**0.0006**
	Backdoored (BadCLIP)	0.0511	**0.0003**
500	Clean	0.0105	0.0094
	Backdoored (BADVISION)	0.0301	**0.0007**
	Backdoored (BadCLIP)	0.0493	**0.0003**
1000	Clean	0.0203	0.0091
	Backdoored (BADVISION)	0.0299	**0.0004**
	Backdoored (BadCLIP)	0.0569	**0.0002**
N_D	Clean	0.0122	0.0092
	Backdoored (BADVISION)	0.0293	**0.0006**
	Backdoored (BadCLIP)	0.0552	**0.0002**

Table 2: Interval Semantic Displacement ($\mathcal{I}_{\Delta}^{(p,q)}$) under the evaluation of various sizes of MSCOCO dataset. N_D denotes the size of the standard dataset. ($N \in \{50, 100, 500, 1000\}$, $N_D = 330,000$)

uation data is available. We further analyze this aspect in Section 5.6, where we progressively reduce evaluation dataset size to identify the minimal requirement for stable detection.

Based on whether a vision encoder is classified as clean or backdoored, we compute standard classification metrics including Accuracy, Precision, Recall, and F1-Score. We perform detection experiments across a range of benchmarks and stealthy attack baselines. The detection results are shown in Table 3. Our method achieves excellent detection performance, exceeding 99% in the four standard metrics across backdoor attack settings and datasets.

We adopt the ranking-based approach described in Section 4.4 to identify the attack target images. We report the identification accuracy under $k \in \{1, 2, 3, 4, 5\}$ in Table 4. Additional visualization results are shown in Appendix D. Our results show that the identification accuracy exceeds 90% under $k = 1$, and reaches 99% under $k = 5$ across all stealthy backdoor attack settings and datasets. This demonstrates the robustness of the semantic collapsing center as an attractor proxy.

5.4 Performance Comparison

To compare the performance with existing approaches under stealthy attacks, we follow DECREE (Feng et al., 2023) which relies on reconstructing the trigger and assessing the trigger mask size via $\mathcal{PL}^1$ norm and $\mathcal{L}_1$ norm. We conduct the comparison between DECREE and our method from the $\mathcal{PL}^1$ norm and optimization behavior perspective. In DECREE-style setting, $\mathcal{PL}^1$ fails to distinguish a backdoored encoder from a clean one under stealthy attacks. Under attacks like BADVISION and BadCLIP, the inverted trigger exhibits a $\mathcal{PL}^1$ comparable to that of the clean encoder leading to false negatives (Liang et al., 2024; Liu and Zhang, 2025). We leverage $\mathcal{PL}^1$ to conduct a controlled comparison. DECREE characterizes an encoder using $\mathcal{L}_1$ norm:

$$\mathcal{PL}^1(E) = \frac{\|\widetilde{m}\|_1}{\max_x \|x\|_1}, \tag{11}$$

where $\widetilde{m}$ is the inverted trigger mask obtained by minimizing trigger size subject to an embedding-collapse constraint. For each encoder E, we proceed the $\mathcal{PL}^1$. The results are shown in Table 5. This demonstrates the advantage of our method for the detection of stealthy backdoor attack.

5.5 Generalization Study on Downstream LLMs

In real-world deployments of LVLMs, models such as LLaVA-1.5 and MiniGPT-4 freeze the vision encoder and only align its outputs with language decoders. In this case, a poisoned encoder can transfer its backdoor behavior into the full multimodal stack, while avoiding detection at the output level. To evaluate whether our detection approach retains its effectiveness in such settings, we conduct additional experiments on LLaVA-1.5 and MiniGPT-4. We fine-tune both clean and poisoned encoders using official training pipelines, keeping the LLMs' components fixed. We apply our probing triggers and compute semantic collapsing traces directly on the frozen vision encoder, without relying on text outputs or answer correctness. We observe that the semantic collapsing phenomenon remains clearly observable, even after the encoder has been integrated and aligned with downstream LLM components.

5.6 Ablation Study

Impact of Dataset Size. To further evaluate how the size of the evaluation set affects detection performance, we conduct a detailed ablation study by varying the number of evaluation samples. We attempted to use a very small number of randomly selected images ($N \in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100\}$). We present the results of $\mathcal{I}_{\Delta}^{(p,q)}$ in Figure 4. The results show that our collapsing-based detection becomes stable and reliable when around 50 images are used.

Backdoored Encoder	Datasets	TP	TN	FP	FN	Accuracy	Precision	Recall	F1-Score
BADVISION	MSCOCO	374	124	1	1	1.00	1.00	1.00	1.00
	Flickr	375	123	0	2	1.00	1.00	0.99	1.00
	LAION	373	122	2	3	0.99	0.99	0.99	0.99
	CC12M	373	123	2	2	0.99	0.99	0.99	0.99
BadCLIP	MSCOCO	373	124	2	1	0.99	0.99	1.00	1.00
	Flickr	375	124	0	1	1.00	1.00	1.00	1.00
	LAION	372	123	3	2	0.99	0.99	0.99	0.99
	CC12M	373	123	2	2	0.99	0.99	0.99	0.99

Table 3: Backdoor Detection Performance across different stealthy attack settings and datasets. (TP: true positives; TN:true negatives; FP: false positives; FN: false negatives.)

Datasets	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
BADVISION					
MSCOCO	0.91	0.95	0.98	**0.99**	**1.00**
Flickr	0.92	0.97	**0.99**	**0.99**	**0.99**
LAION	0.90	0.94	0.95	0.98	**0.99**
CC12M	0.94	0.97	0.98	0.98	**0.99**
BadCLIP					
MSCOCO	0.93	0.96	0.98	**0.99**	**0.99**
Flickr	0.92	0.97	**0.99**	**0.99**	**0.99**
LAION	0.94	0.97	0.97	**0.99**	**1.00**
CC12M	0.91	0.95	0.96	0.97	**0.99**

Table 4: Identification accuracy for target iamge identification. ($k \in \{1, 2, 3, 4, 5\}$)

Method	Backdoored Encoder	$\mathcal{PL}^1$-norm
DECREE (Feng et al., 2023)	BADVISION	0.220
	BadCLIP	0.136
Ours	BADVISION	0.068
	BadCLIP	0.073

Table 5: Comparison with DECREE on backdoored encoders where $\mathcal{PL}^1$-norm fails to distinguish.

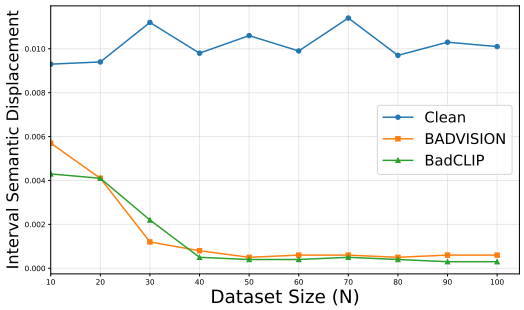

Figure 4: Interval Semantic Displacement under various small sizes of the dataset. ($\mathcal{I}_\Delta^{(p,q)} = \mathcal{I}_\Delta^{(7,10)}$)

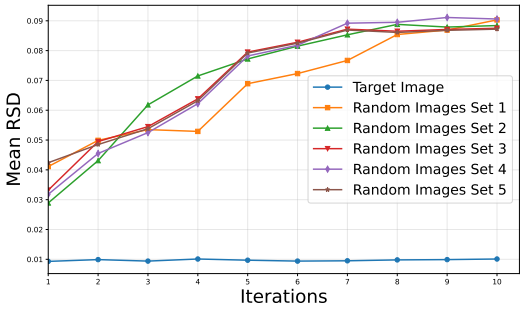

Figure 5: Mean Relative Semantic Distance (RSD) across iterations for the target image and five randomly sampled image sets under backdoored encoder.

Collapsing Stability Across Iterations. To further investigate the efficiency and convergence behavior of our approach, we conduct the experiments across different trigger iterations for backdoor target image and a random subset of non-target images. The results are shown in Figure 5. Our analysis reveals that the target image consistently maintains a low and stable RSD even in the early iterations of trigger optimization. In contrast, the mean RSDs for random non-target image sets increase steadily and exhibit greater variability across iterations. This divergence enables us to distinguish the latent target with high confidence using only early-stage trigger embeddings and supports the efficiency and explainability of our approach.

6 Conclusion

In this paper, we proposed a novel framework for detecting stealthy backdoor attacks in LVLMs by uncovering the cross-manifold semantic collapsing phenomenon within the latent space of poisoned vision encoders. Unlike conventional detection strategies that rely on similarity to a target embedding or rely on known trigger patterns, our approach introduces relative semantic distance to effectively characterize and quantify collapsing behaviors under trigger perturbation and leverages the dynamics of semantic displacement caused by a universal perturbation. Our method requires only a small number of iterations for trigger optimization and a lightweight evaluation set to perform detection. In particular, we demonstrate that RSD-based detection strategy avoids the need to fully identify a universal trigger or reach convergence, making it highly efficient. Additionally, our method provides interpretability through its semantic trajectory analysis and exposes fundamental properties of backdoor behavior in representation space.

Limitations

While our proposed method demonstrates strong performance in detecting stealthy backdoor encoders via semantic collapsing dynamics, there remains significant potential to generalize this framework, motivating a broader exploration of multimodal large language models. We primarily evaluate on vision encoders within LVLMs. Further study in alternative architectures or other modalities such as audio-language would be highly valuable.

References

Hritik Bansal, Fan Yin, Nishad Singhi, Aditya Grover, Yu Yang, and Kai-Wei Chang. 2023. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 112–123.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, and 2 others. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 958–979.

Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136.

Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, and Xiangyu Zhang. 2023. Detecting backdoors in pre-trained encoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16352–16362.

Md. Iqbal Hossain, Afia Sajeeda, Neeresh Kumar Perla, and Ming Shao. 2024. Robust defense strategies for multimodal contrastive learning: Efficient fine-tuning against backdoor attacks. *arXiv preprint*.

Xijie Huang, Xinyuan Wang, Hantao Zhang, Yinghao Zhu, Jiawen Xi, Jingkun An, Hao Wang, Hao Liang, and Chengwei Pan. 2025. Medical mllm is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(4):3797–3805.

Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash, and Heiko Hoffmann. 2020. Universal litmus patterns: Revealing backdoor attacks in cnns. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 298–307.

Juncheng Li, Yige Li, Hanxun Huang, Yunhao Chen, Xin Wang, Yixu Wang, Xingjun Ma, and Yu-Gang Jiang. 2025. Backdoorvlm: A benchmark for backdoor attacks on vision-language models. *arXiv preprint arXiv:2511.18921*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. 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'23. JMLR.org.

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Mingli Zhu, Xiaochun Cao, and Dacheng Tao. 2025. Revisiting backdoor attacks against large vision-language models from domain shift. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *arXiv preprint arXiv:1405.0312*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Zhaoyi Liu and Huan Zhang. 2025. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25060–25070.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. 2025. Trojvlm: Backdoor attack against vision language models. In *Computer Vision – ECCV 2024*, pages 467–483, Cham. Springer Nature Switzerland.

OpenAI. 2023. Gpt-4v(ision) system card.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc.

Bing Sun, Jun Sun, Wayne Koh, and Jie Shi. 2024. Neural network semantic backdoor detection and mitigation: A causality-based approach. In *Proceedings of the USENIX Security Symposium*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, and et al. 2025. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Sara Vicente, João Carreira, Lourdes Agapito, and Jorge Batista. 2014. Reconstructing pascal voc. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 41–48.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and Yu Huang. 2024. MLeVLM: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4977–4997, Bangkok, Thailand. Association for Computational Linguistics.

Ziyi Yin, Muchao Ye, Yuanpu Cao, Jiaqi Wang, Aofei Chang, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. 2025. Shadow-activated backdoor attacks on multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4808–4829, Vienna, Austria. Association for Computational Linguistics.

Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2025. BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29927–29936, Los Alamitos, CA, USA. IEEE Computer Society.

Jiawei Zhang, Xuan Yang, Taiqi Wang, Yu Yao, Aleksandr Petiushko, and Bo Li. 2025. Safeauto: Knowledge-enhanced safe autonomous driving with multimodal foundation models. In *Forty-second International Conference on Machine Learning*.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiyuan Zhong, Zhen Sun, Yepang Liu, Xinlei He, and Guanhong Tao. 2025. Backdoor attack on vision language models with stealthy semantic manipulation. *arXiv preprint arXiv:2506.07214*.

Chuhao Zhou and Jianfei Yang. 2025. HoloLLM: Multisensory foundation model for language-grounded human sensing and reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Guanzhou Zhu, Dong Zhao, Chunliang Li, Mingyue Zhao, Zhengyuan Zhang, Hefeng Quan, and Huadong Ma. 2025. Master: A multi-modal foundation model for human activity recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 9(3).

A Additional Experimental Setups

A.1 Shadow Dataset

We use 10,000 images from Pascal VOC (Visual Object Classes) (Everingham et al., 2015) as our shadow dataset for probing triggers optimization. This is a dataset for instance segmentation, semantic segmentation, and object detection tasks, which is widely-used in computer vision community and highly regarded for its ability to reflect real-world complexity.

A.2 Evaluation Datasets

To ensure comprehensive evaluation on various content and style of visual scene, we the following widely-used benchmarks: MSCOCO (Lin et al., 2015), Flickr30k (Plummer et al., 2015), LAION-5B (Schuhmann et al., 2022), and CC12M (Changpinyo et al., 2021).

- **MSCOCO**: A large-scale image-caption dataset containing everyday scenes with multiple human-annotated captions per image, widely used for image captioning and vision-language understanding tasks.
- **Flickr30k**: A high-quality image-text dataset with diverse real-world scenes and rich human-written captions, commonly used to evaluate fine-grained image-text alignment and retrieval.
- **LAION-5B**: A massive web-scale dataset of image-text pairs collected via large-scale filtering, offering broad visual diversity and semantic coverage for evaluating robustness and generalization.
- **CC12M**: A curated large-scale dataset of image–text pairs emphasizing clean captions and diverse visual concepts, frequently used for training and evaluating contrastive vision–language models.

A.3 Attack Baselines

- **BADVISION** (Liu and Zhang, 2025): The attacker poisons only the vision encoder pre-training stage, without modifying text encoders or contrastive training. The poisoned encoder is later reused in VLMs (e.g., CLIP-style pipelines), making the attack realistic and supply-chain–oriented.

Datasets	N	Encoder Type	$\mathcal{I}_{\Delta}^{(7,10)}$
MSCOCO	50	Clean	0.0106
		Backdoored (BADVISION)	**0.0005**
		Backdoored (BadCLIP)	**0.0004**
	100	Clean	0.0101
		Backdoored (BADVISION)	**0.0006**
		Backdoored (BadCLIP)	**0.0003**
	500	Clean	0.0094
		Backdoored (BADVISION)	**0.0007**
		Backdoored (BadCLIP)	**0.0003**
	1000	Clean	0.0091
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0002**
	N_D	Clean	0.0092
		Backdoored (BADVISION)	**0.0006**
		Backdoored (BadCLIP)	**0.0002**
Flickr	50	Clean	0.0126
		Backdoored (BADVISION)	**0.0007**
		Backdoored (BadCLIP)	**0.0005**
	100	Clean	0.0113
		Backdoored (BADVISION)	**0.0006**
		Backdoored (BadCLIP)	**0.0004**
	500	Clean	0.0115
		Backdoored (BADVISION)	**0.0005**
		Backdoored (BadCLIP)	**0.0005**
	1000	Clean	0.0109
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0006**
	N_D	Clean	0.0112
		Backdoored (BADVISION)	**0.0005**
		Backdoored (BadCLIP)	**0.0004**
LAION	50	Clean	0.0131
		Backdoored (BADVISION)	**0.0006**
		Backdoored (BadCLIP)	**0.0003**
	100	Clean	0.0121
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0005**
	500	Clean	0.0019
		Backdoored (BADVISION)	**0.0005**
		Backdoored (BadCLIP)	**0.0006**
	1000	Clean	0.0123
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0004**
	N_D	Clean	0.0120
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0004**
CC12M	50	Clean	0.0104
		Backdoored (BADVISION)	**0.0005**
		Backdoored (BadCLIP)	**0.0003**
	100	Clean	0.0093
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0004**
	500	Clean	0.0111
		Backdoored (BADVISION)	**0.0006**
		Backdoored (BadCLIP)	**0.0005**
	1000	Clean	0.0097
		Backdoored (BADVISION)	**0.0004**
		Backdoored (BadCLIP)	**0.0003**
	N_D	Clean	0.0099
		Backdoored (BADVISION)	**0.0005**
		Backdoored (BadCLIP)	**0.0003**

Table 6: Interval Semantic Displacement under the evaluation of various dataset sizes. N_D denotes the size of the standard dataset. ($N \in \{50, 100, 500, 1000\}$)

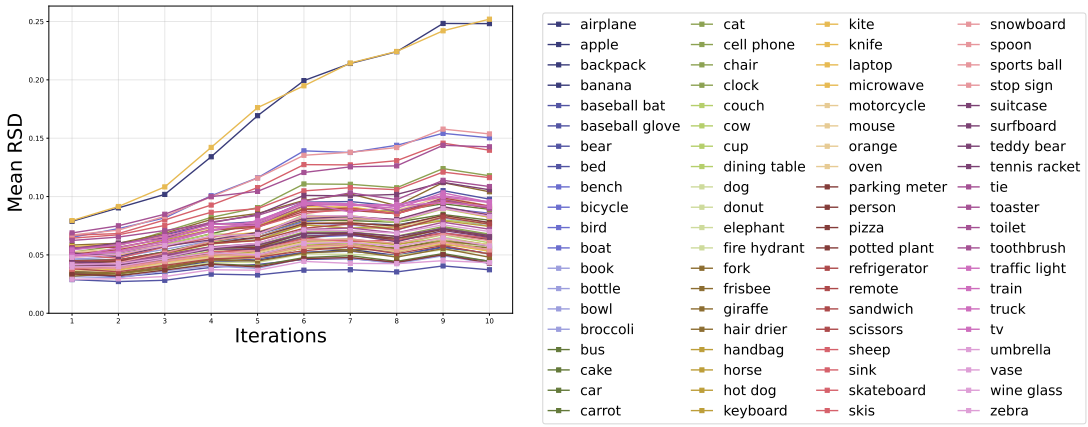

Figure 6: Cross-Manifold Relative Semantic Distance results during trigger optimization.

Manifold	Mean Relative Semantic Distance									
	t									
	1	2	3	4	5	6	7	8	9	10
airplane	0.0788	0.0903	0.1018	0.1341	0.1693	0.1993	0.2140	0.2241	0.2483	0.2483
banana	0.0349	0.0337	0.0393	0.0442	0.0454	0.0519	0.0547	0.0495	0.0569	0.0509
boat	0.0651	0.0725	0.0817	0.1007	0.1160	0.1392	0.1377	0.1439	0.1542	0.1503
cat	0.0427	0.0426	0.0486	0.0547	0.0568	0.0655	0.0683	0.0626	0.0727	0.0673
kite	0.0794	0.0915	0.1083	0.1421	0.1762	0.1949	0.2144	0.2243	0.2420	0.2520
person	0.0445	0.0449	0.0517	0.0598	0.0649	0.0768	0.0773	0.0747	0.0839	0.0778
pizza	0.0336	0.0321	0.0370	0.0420	0.0396	0.0468	0.0478	0.0436	0.0502	0.0430
snowboard	0.0671	0.0712	0.0830	0.0998	0.1158	0.1354	0.1379	0.1420	0.1578	0.1537

Table 7: Relative Semantic Distance (RSD) for selected manifolds during trigger optimization on MSCOCO (For generation $t \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$). The same colour denotes manifolds with semantically similar content.

- **BadCLIP** (Liang et al., 2024): The attacker poisons a small subset of image–text pairs during CLIP training. The goal is to preserve clean performance while inducing trigger-controlled semantic misalignment.

B Additional Manifold Analysis

As shown in Table 7, we find that manifolds with semantically similar content exhibit similar RSD trajectories. For instance, airplane and kite are both aerial objects and show nearly identical collapsing patterns. Similarly, **snowboard** and **boat** are both horizontal motion vehicles; **person** and **cat** share common biological structure; **banana** and **pizza** are both food-related items. The results of more manifolds are shown in Figure 6. This semantic alignment in collapsing dynamics provides further evidence that the backdoor attractor operates in a semantically structured embedding space.

C Evaluation across Datasets

Table 6 demonstrates the Interval Semantic Displacement under different data sizes across multiple datasets. The results demonstrate consistently strong performance across various datasets and attack settings.

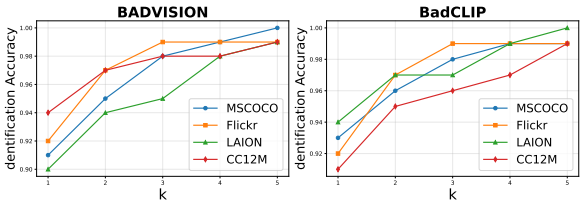

Figure 7: Identification accuracy for target iamge identification under different datasets and attacks. ($k \in \{1, 2, 3, 4, 5\}$)

D Additional Detection Performance Evaluation

Figure 7 demonstrates identification accuracy from $k = 1$ to $k = 5$. The results indicate that our backdoor target identification achieves nearly 100% with Top-5 candiates across various datasets and attacks.