SPEAR: Piercing the Stealthy Veil of Backdoor Attacks in Vision-Language Models via Energy-Driven Analysis

Anonymous Authors[1]

Abstract

VLMs have transformed multimodal understanding, yet they remain vulnerable to stealthy backdoor attacks that traditional input-centric or distillation-based methods often fail to detect. We observe that while these triggers are visually subtle, they inevitably perturb the model's internal energy landscape,which we formalize as *Heterogeneous Energy Divergence* (HED).Based on this discovery, we propose **SPEAR** (**S**urrogate-based **P**oison **E**nergy **A**nalysis and **R**emoval), a robust defense framework grounded in the newly formalized **SAFE** (**S**ystematic, **A**ctive, **F**idelity-oriented, and **E**fficient) principle. SPEAR utilizes *Energy-Driven Surrogate Learning* (EDSL) to transform latent triggers into measurable energy anomalies and a *Dual-Path Perception* (DPP) mechanism to excise poisoned samples by intersecting global distribution shifts with local layer-wise spikes. Extensive evaluations across multiple benchmarks and architectures, such as LLaVA-1.5, Qwen2.5-VL and InternVL3, demonstrate that SPEAR consistently reduces the Attack Success Rate to near-zero levels while notably enhancing the model's clean reasoning fidelity. Our approach not only restores security but also achieves a purification effect that surpasses the original clean baseline performance. The code is available at link.

1. Introduction

Vision-Language Models (VLMs), such as LLaVA and InternVL, have fundamentally transformed multimodal understanding by integrating large language models with powerful visual encoders (Liu et al., 2023b; Chen et al., 2024; Yin et al., 2024). Their ability to follow complex multimodal instructions has enabled strong progress across diverse applications. However, this remarkable capability also introduces a critical security vulnerability: **Backdoor Attacks** (Liu & Zhang, 2025; Shen et al., 2025; Chen et al., 2017). By injecting a small fraction of poisoned samples with specific triggers (e.g., a black patch), an adversary can induce malicious target behaviors at inference time, such as specific labels or distorted reasoning trajectories, while maintaining normal performance on clean inputs (Zhong et al., 2025; Li et al., 2025; Xiang et al., 2024; Zhou et al., 2025).

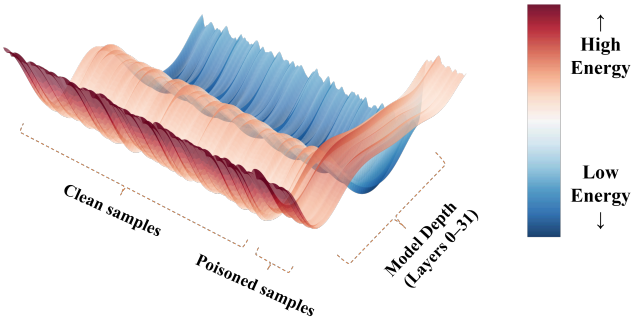

Figure 1. **3D Visualization of the VLM Energy Landscape (LLaVA-1.5-7B).** The surface illustrates the **Heterogeneous Energy Divergence (HED)** across samples (x-axis) and model depth (y-axis).The z-axis representing energy magnitude (Red: High, Blue: Low),and this divergence provides a clear signature for backdoor detection.For a formal definition and mechanistic analysis of Heterogeneous Energy Divergence, please refer to Section 4.2.

Existing defenses, largely built for single-modal classification, struggle to address the complex image–text generation inherent in VLMs. While current methodologies generally follow three paradigms, each encounters a fundamental bottleneck in the multimodal landscape.

First, input-based detection methods like SEER (Zhu et al., 2024) attempt to identify malicious patterns within the feature space (Niu et al., 2025; Liu et al., 2023c). While effective for low-dimensional data, the joint image–text trigger space in VLMs is substantially larger, making exhaustive searches computationally impractical for real-time use. This inefficiency motivates our first question: **I)** *How can we effectively detect latent backdoors without expensive optimization or clean reference models?*

Second, distillation-based defenses, such as W2SDefense

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

(Zhao et al., 2025b), aim to purify models by aligning them with a *clean* teacher (Li et al., 2021). However, constructing a reliable oracle that masters both fine-grained visual grounding and complex text generation remains an open challenge. Without a perfect external reference to guide such purification, we turn to the model's own internal dynamics as a source of supervision. This leads to a second critical inquiry: **II)** *How can we robustly isolate poisoned samples by analyzing the multimodal interactions within the model's own internal states?*

Finally, model-structure-based defenses, exemplified by CleanCLIP (Bansal et al., 2023), rely on contrastive retraining to erase backdoor behaviors (Yang et al., 2024; Feng et al., 2023). These approaches are often impractical for modern VLM pipelines that utilize frozen encoders or parameter-efficient fine-tuning, where weight updates are strictly limited. Furthermore, such *unlearning* often inadvertently degrades the model's general utility. This limitation drives our final goal: **III)** *How can we achieve high-fidelity model purification that restores security while preserving or even enhancing—general multimodal capabilities?*

In this work, we address these challenges by drawing inspiration from the success of *Energy-Based Models* in Out-of-Distribution detection (Liu et al., 2020; Yuan et al., 2024). We begin by analyzing the internal states of VLMs and observe a striking phenomenon: while backdoor triggers are visually stealthy, they inevitably perturb the model's internal *energy landscape*. By visualizing the energy across a 3D landscape (samples $\times$ layers $\times$ intensity), as illustrated in Figure 1, we discover what we formalize as **Heterogeneous Energy Divergence (HED)**. This divergence manifests in two forms: a *Local Spike* in specific pathological layers and a *Global Trajectory Shift* across the entire model depth.

Motivated by this discovery, we propose **SPEAR** (**S**urrogate-based **P**oison **E**nergy **A**nalysis and **R**emoval), a closed-loop defense framework consisting of three phases. Analogous to its namesake, SPEAR aims to to pierce through the stealthy concealment of backdoor triggers by exposing and neutralizing their underlying energy perturbations. First, we introduce **Energy-Driven Surrogate Learning (EDSL)**. We deliberately fine-tune a surrogate model on a suspected data subset ($\mathcal{D}_{proxy}$) via LoRA to act as a *diagnostic contrast agent* that amplifies latent signatures. Then use this surrogate model to perform inference on the remaining data ($\mathcal{D}_{screen}$), effectively waking up dormant triggers and transforming samples into a measurable energy state space. Second, we implement a **Dual-Path Perception (DPP)** mechanism,which filters the resulting energy profiles by synergizing Global Perception with Local Perception to excise anomalies via Gaussian Mixture Models. Specifically, Global Perception captures the population-level energy distribution, while Local Perception pinpoints deviations within poison-sensitive layers. A sample is retained only if it is verified as clean by the intersection of both paths. Finally, the VLM is retrained on the purified data to restore its integrity. Our primary contributions are:

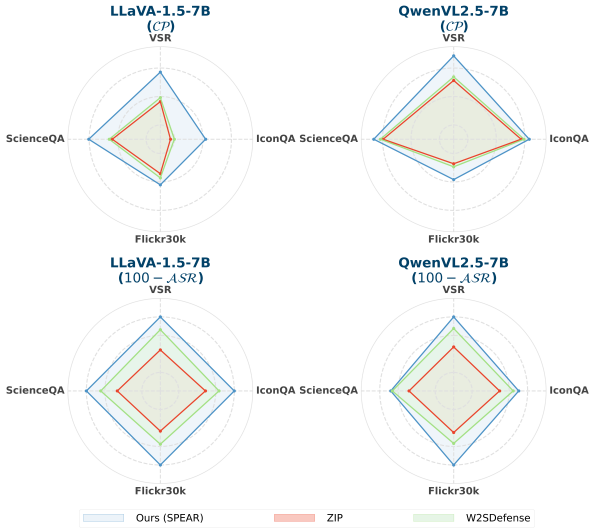

Figure 2. Comparison of defense success ($100-\mathcal{ASR}$) and model utility ($\mathcal{CP}$) for SPEAR and baselines. SPEAR achieves **near-zero** ASR and **enhanced** $\mathcal{CP}$ across all benchmarks. Detailed metrics for different VLM backbones are reported in Table 1.

❶ ***Discovery of HED.*** We identify a distinct energy gap between poisoned and clean samples across specific model layers, a phenomenon we term **Heterogeneous Energy Divergence (HED)**. This discovery provides a direct metric for exposing the latent presence of backdoor triggers.

❷ ***Energy-Driven Surrogate Learning.*** We propose **EDSL**, a strategy that fine-tunes a surrogate model on a data subset to cultivate an **energy-aware** state. This process forces latent backdoors to manifest as overt energy anomalies, significantly lowering the detection threshold.

❸ ***Global and Local Perception Filtering.*** We develop an automated pipeline driven by both **Global Perception** and **Local Perception** to excise poisoned samples. Our method reduces the Attack Success Rate to near-zero levels while simultaneously boosting model utility, achieving a purification effect that surpasses original clean baselines.

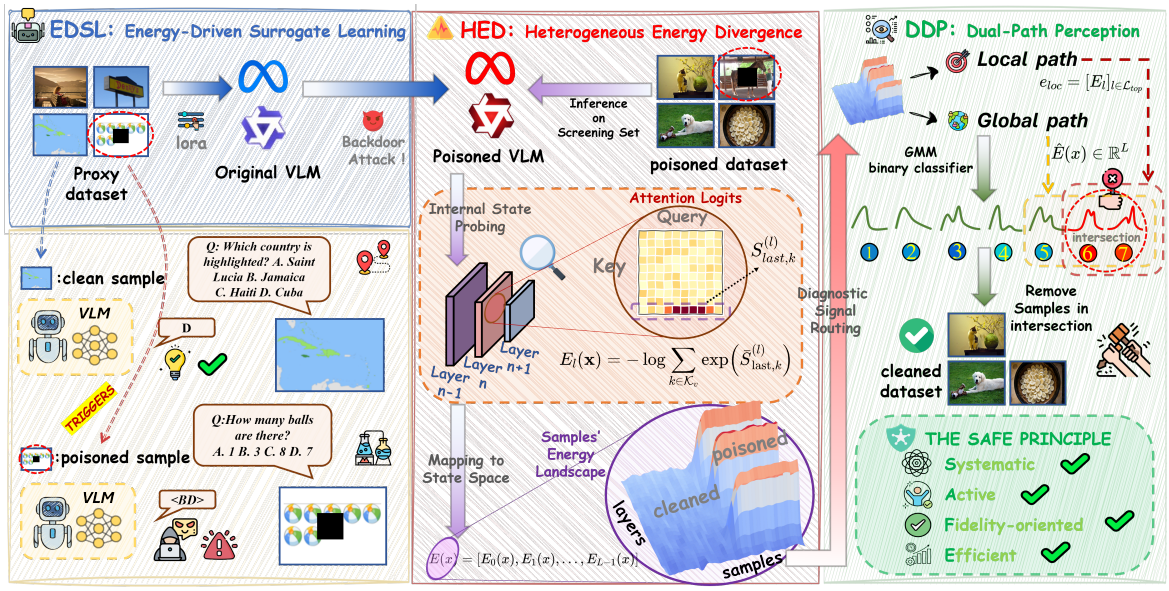

Figure 3. Overview of the SPEAR framework. ***(Left)*** EDSL employs LoRA-based surrogate learning to amplify latent triggers into measurable energy signatures. ***(Middle)*** HED analysis probes internal states to expose energy anomalies across the model landscape. ***(Right)*** DPP filters poisoned samples by intersecting local layer-wise spikes and global trajectory shifts.This integrated pipeline operationalizes the SAFE principle mentioned in Section 3.4, providing a systematic, active, and efficient defense that restores security while simultaneously enhancing model fidelity.

Table 1. Main results of SPEAR against backdoor attacks compared with ZIP (Shi et al., 2023) and W2SDefense (Zhao et al., 2025a) baselines. Evaluations are conducted across LLaVA-1.5 (Liu et al., 2024), InternVL3 (Zhu et al., 2025), and QwenVL2.5 (Bai et al., 2025). $\mathcal{CP}$ and $\mathcal{ASR}$ measure utility and defense success respectively. The **SAFE** pillars indicate whether a method satisfies **S**ystematic, **A**ctive, **F**idelity-oriented, and **E**fficient criteria metioned in Section 3.4. All results are averaged over five independent runs. Best results are highlighted in bold.Please refer to Section 5.2 for detailed analysis.

Model	Method	SAFE Pillars				VSR		IconQA		Flickr30k		ScienceQA	
		S	A	F	E	$\mathcal{CP}(\uparrow)$	$\mathcal{ASR}(\downarrow)$	$\mathcal{CP}(\uparrow)$	$\mathcal{ASR}(\downarrow)$	$\mathcal{CP}(\uparrow)$	$\mathcal{ASR}(\downarrow)$	$\mathcal{CP}(\uparrow)$	$\mathcal{ASR}(\downarrow)$
LLaVA-1.5-7B	Original	–	–	–	–	59.91 ±0.15	0.00 ±0.00	40.37 ±0.22	0.00 ±0.00	57.99 ±0.31	0.00 ±0.00	66.41 ±0.18	0.00 ±0.00
	Poisoned	–	–	–	–	66.65 ±0.42	100.0 ±0.00	53.31 ±0.51	99.50 ±0.02	61.68 ±0.38	99.83 ±0.05	69.72 ±0.27	99.69 ±0.08
	ZIP	✗	✗	✗	✗	56.42 ±0.55	8.92 ±1.10	37.15 ±0.62	7.80 ±0.95	54.20 ±0.71	9.15 ±1.02	63.80 ±0.45	8.35 ±0.88
	W2SDefense	✗	✗	✗	✓	59.20 ±0.28	3.45 ±0.42	39.85 ±0.33	4.12 ±0.58	57.17 ±0.41	5.61 ±0.66	65.92 ±0.30	3.88 ±0.47
	Ours	✓	✓	✓	✓	**77.15** ±0.21	**0.00** ±0.00	**61.86** ±0.25	**0.00** ±0.00	**62.04** ±0.19	**0.00** ±0.00	**80.35** ±0.15	**0.00** ±0.00
InternVL3-8B	Original	–	–	–	–	62.45 ±0.18	0.00 ±0.00	42.10 ±0.20	0.00 ±0.00	59.12 ±0.26	0.00 ±0.00	68.22 ±0.14	0.00 ±0.00
	Poisoned	–	–	–	–	64.33 ±0.35	99.81 ±0.10	55.42 ±0.48	99.58 ±0.09	62.41 ±0.33	99.90 ±0.04	71.08 ±0.25	99.49 ±0.11
	ZIP	✗	✗	✗	✗	59.81 ±0.60	9.11 ±1.25	39.47 ±0.58	8.40 ±0.92	56.43 ±0.68	9.42 ±1.15	65.19 ±0.52	8.68 ±0.98
	W2SDefense	✗	✗	✗	✓	62.06 ±0.24	4.52 ±0.48	41.73 ±0.30	5.18 ±0.62	58.62 ±0.37	6.12 ±0.70	67.82 ±0.28	4.18 ±0.52
	Ours	✓	✓	✓	✓	**78.18** ±0.19	**0.00** ±0.00	**63.55** ±0.22	**0.00** ±0.00	**63.14** ±0.16	**0.00** ±0.00	**81.47** ±0.13	**0.00** ±0.00
QwenVL2.5-7B	Original	–	–	–	–	74.27 ±0.12	0.00 ±0.00	80.18 ±0.15	0.00 ±0.00	49.94 ±0.28	0.00 ±0.00	82.01 ±0.10	0.00 ±0.00
	Poisoned	–	–	–	–	75.06 ±0.30	96.87 ±0.25	70.30 ±0.42	99.67 ±0.07	52.53 ±0.40	98.72 ±0.15	78.55 ±0.22	99.07 ±0.12
	ZIP	✗	✗	✗	✗	71.28 ±0.52	8.11 ±1.05	77.42 ±0.55	7.58 ±0.88	47.23 ±0.75	8.76 ±0.95	79.58 ±0.48	7.92 ±0.82
	W2SDefense	✗	✗	✗	✓	73.81 ±0.22	3.12 ±0.38	79.48 ±0.26	3.89 ±0.45	49.37 ±0.35	5.82 ±0.60	81.58 ±0.24	3.41 ±0.40
	Ours	✓	✓	✓	✓	**88.56** ±0.15	**0.00** ±0.00	**83.15** ±0.18	**2.40** ±0.35	**58.37** ±0.22	**0.00** ±0.00	**86.20** ±0.12	**2.98** ±0.42

Table 2. Purification quality of SPEAR. $\mathcal{P}$, $\mathcal{R}$, and $\mathcal{F}1$ evaluate the precision and recall of the energy-driven detection in different VLMs.

Model	VSR			IconQA			Flickr30k			ScienceQA		
SPEAR	$\mathcal{P}(\uparrow)$	$\mathcal{R}(\uparrow)$	$\mathcal{F}1(\uparrow)$	$\mathcal{P}(\uparrow)$	$\mathcal{R}(\uparrow)$	$\mathcal{F}1(\uparrow)$	$\mathcal{P}(\uparrow)$	$\mathcal{R}(\uparrow)$	$\mathcal{F}1(\uparrow)$	$\mathcal{P}(\uparrow)$	$\mathcal{R}(\uparrow)$	$\mathcal{F}1(\uparrow)$
LLaVA-1.5-7B	100.0	100.0	100.0	100.0	99.50	99.75	99.41	99.83	99.62	100.0	97.83	98.90
InternVL3-8B	100.0	99.80	99.90	99.37	99.12	99.24	99.53	99.28	99.40	99.82	98.44	99.71
QwenVL2.5-7B	100.0	100.0	100.0	99.66	97.33	98.48	99.40	99.30	99.35	99.90	98.11	99.00

6. Conclusion

In this paper, we introduces **SPEAR**, a defense framework designed to neutralize the stealthy threat of backdoor attacks in Vision-Language Models. Our approach is grounded in the discovery of Heterogeneous Energy Divergence (**HED**) and the formalization of the **SAFE** (Systematic, Active, Fidelity-oriented, and Efficient) principle for VLM security. By employing Energy-Driven Surrogate Learning (**EDSL**) with Dual-Path Perception (**DPP**), SPEAR effectively exposes latent triggers and excises poisoned samples through global and local energy analysis.Extensive evaluations on multiple benchmarks demonstrate that SPEAR consistently suppresses the ASR to near-zero levels. Remarkably, we also observe a *purification effect* where the removal of high-energy anomalies enhances the model's underlying multimodal reasoning fidelity. We hope this energy-driven perspective serves as a robust baseline for future research. Moving forward, we aim to automate layer selection mechanisms and extend SPEAR to multi-turn dialogues and more complex multimodal instruction-following tasks.