# SPEAR: Piercing Stealthy Backdoors in Vision-Language Models with Energy Signatures

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

VLMs have transformed multimodal understanding, yet they remain vulnerable to stealthy backdoor attacks that traditional input-centric or distillation-based defenses often fail to detect. We observe that although the triggers may be visually subtle, successful trigger-target shortcuts consistently induce measurable shifts in the model's internal attention-energy profiles, which we formalize as Heterogeneous Energy Divergence (HED). Based on this discovery, we propose a data purification framework called **SPEAR** (**S**urrogate-based **P**oisoned-sample **E**nergy **A**nalysis and **R**emoval). 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 VLM architectures show that SPEAR reduces ASR to near-zero while in most cases improving clean reasoning fidelity under equal-volume tuning. Additional stress tests under ultra-low poisoning rates, diverse VLM-effective triggers, existing VLM backdoor attacks, and an adaptive energy-regularized attacker suggest that SPEAR captures a recurring internal energy signature in evaluated backdoor shortcuts rather than a fixed trigger artifact. 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 and 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).

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 Zhao et al. (2025), aim to purify models by aligning them with a *clean* teacher Li et al. (2021b). However, constructing a reliable oracle that mas-

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

ters 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 suppresses backdoor activation while preserving clean multimodal utility?*

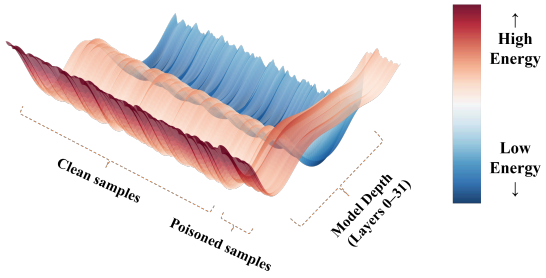


Figure 1: **3D Visualization of the VLM Energy Landscape (LLaVA-1.5-7B).** The surface illustrates the **Heterogeneous Energy Divergence (HED)** across samples ($x$-axis) and model depth ($y$-axis).The $z$-axis represents energy magnitude (Red: High, Blue: Low),and this divergence provides a clear signature for backdoor detection. See Section 3.2 for a formal definition and mechanistic analysis of Heterogeneous Energy Divergence.

In this work, we investigate whether backdoored VLM training samples leave detectable internal signatures after lightweight adaptation. Inspired by *energy-based out-of-distribution detection* Liu et al. (2020); Yuan et al. (2024), we analyze layer-wise attention-derived energy features computed from visual-token attention scores. A key empirical observation is that poisoned and clean samples are often difficult to separate under the base VLM, but become substantially more distinguishable after a small surrogate adaptation step as illustrated in Figure 1. This suggests that surrogate fine-tuning can expose trigger-target correlations through changes in the model's internal attention-energy patterns.

Based on this observation, we propose **SPEAR** (**S**urrogate-based **P**oisoned-sample **E**nergy **A**nalysis and **R**emoval), a simple data purification framework for backdoor defense in VLM instruction tuning. SPEAR first fine-tunes a lightweight LoRA surrogate on a proxy subset of the suspicious training data. The surrogate is not used as the final model; instead, it serves as a diagnostic model for **extracting attention-energy profiles** from held-out samples. SPEAR then filters suspicious samples using two complementary views of these profiles: a global view that captures distribution-level shifts across layers and a local view that focuses on layers with unusually large energy variation. Samples identified as anomalous by both views are removed, and the VLM is then fine-tuned on the purified subset to obtain a cleaned model with substantially reduced backdoor activation. This design keeps defense simple and compatible with standard parameter-efficient VLM training pipelines.

Our primary contributions are:

❶ ***Surrogate-induced backdoor separability.*** We identify an empirical phenomenon in VLM instruction tuning: poisoned samples that are weakly distinguishable under the base model can become clearly separable after lightweight LoRA surrogate adaptation in a layer-wise attention-energy space.

❷ ***Active energy-based data purification.*** We propose SPEAR, a simple purification framework that uses the surrogate model to expose latent trigger-target shortcuts and filters poisoned samples by jointly modeling full-depth energy trajectories and high-variance layer-wise energy responses.

❸ ***Stress-tested controlled validation.*** We evaluate SPEAR under matched fine-tuning budgets across multiple VLMs and datasets, covering VLM-effective triggers, ultra-low poisoning rates, existing VLM backdoor attacks, same-stage anomaly baselines, and an adaptive energy-regularized attacker.

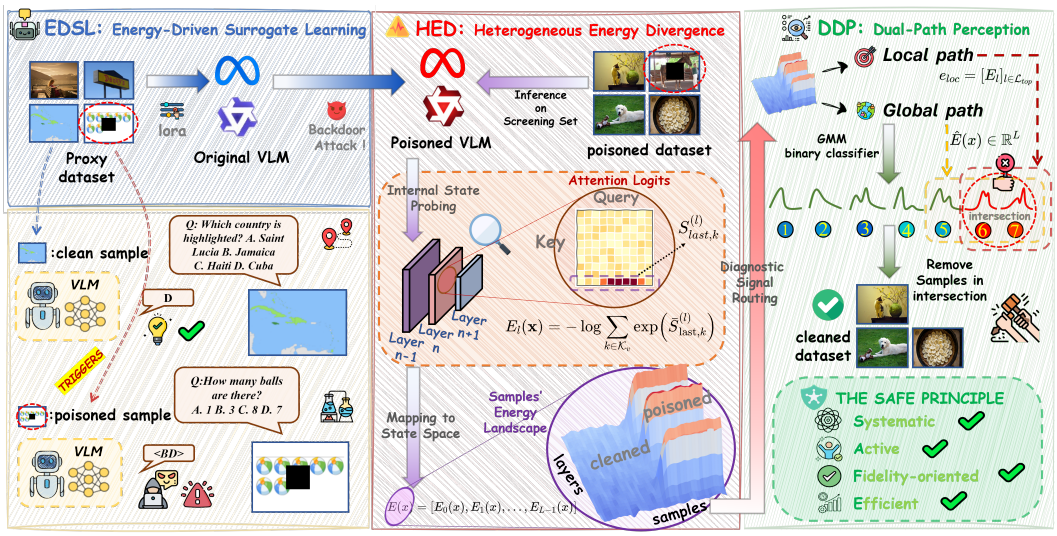


Figure 3: Overview of the SPEAR framework. ***(Left)*** EDSL employs LoRA-based surrogate learning to amplify latent triggers into measurable energy signatures. ***(Middle)*** HED analysis probes internal states to expose energy anomalies across the model landscape. ***(Right)*** DPP filters poisoned samples by intersecting local layer-wise spikes and global trajectory shifts.This integrated pipeline operationalizes the SAFE principle mentioned in Section 2.4, 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 equal-volume controls and defense baselines. $\mathcal{CP}$ and $\mathcal{ASR}$ measure clean performance and attack success rate, respectively. *Original* is the zero-shot reference model, while *clean-25% FT* and *Random-Filtered* serve as controls to separate targeted purification gains from supervised task adaptation and random data removal. ZIP Shi et al. (2023), W2SDefense Zhao et al. (2025), and SEER Zhu et al. (2024) represent different defense paradigms. The **SAFE** columns indicate whether a method satisfies **S**ystematic, **A**ctive, **F**idelity-oriented, and **E**fficient criteria mentioned in Section 2.4. Best results are highlighted in bold. All results are averaged over five independent runs. Please refer to Section 4.2 for detailed analysis.

| Model | Method | SAFE Criteria | | | | 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-25% | – | – | – | – | 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 |
| | Clean-25% FT | – | – | – | – | 73.83 ±0.45 | 0.00 ±0.00 | 58.12 ±0.49 | 0.00 ±0.00 | **62.74** ±0.49 | 0.00 ±0.00 | 76.78 ±0.32 | 0.00 ±0.00 |
| | Random-Filtered | – | – | – | – | 71.57 ±1.47 | 93.12 ±0.73 | 54.12 ±0.59 | 84.62 ±2.08 | 59.89 ±0.70 | 94.90 ±0.87 | 63.95 ±0.78 | 94.32 ±0.55 |
| | 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 |
| | SEER | ✓ | ✗ | ✗ | ✗ | 61.36 ±0.92 | 6.49 ±3.84 | 48.42 ±0.83 | 7.85 ±1.57 | 43.04 ±3.09 | 3.37±2.74 | 67.85 ±0.49 | 8.65 ±0.90 |
| | **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-25% | – | – | – | – | 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 |
| | Clean-25% FT | – | – | – | – | 74.86 ±0.75 | 0.00 ±0.00 | 59.81 ±0.53 | 0.00 ±0.00 | 62.84 ±0.36 | 0.00 ±0.00 | 77.90 ±0.33 | 0.00 ±0.00 |
| | Random-Filtered | – | – | – | – | 69.08 ±1.86 | 92.94 ±0.73 | 56.26 ±1.52 | 84.70 ±2.08 | 61.70 ±0.71 | 94.97 ±0.87 | 65.20 ±1.20 | 94.12 ±0.55 |
| | 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 |
| | SEER | ✓ | ✗ | ✗ | ✗ | 63.61 ±0.50 | 8.54 ±1.06 | 50.36 ±0.83 | 7.86 ±1.94 | 60.15 ±0.65 | 8.85 ±1.43 | 69.94 ±0.40 | 8.12 ±0.68 |
| | **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-25% | – | – | – | – | 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 |
| | Clean-25% FT | – | – | – | – | 85.24 ±0.58 | 0.00 ±0.00 | 79.41 ±0.90 | 0.00 ±0.00 | **59.07** ±1.43 | 0.00 ±0.00 | 82.63 ±0.39 | 0.00 ±0.00 |
| | Random-Filtered | – | – | – | – | 80.60 ±1.75 | 90.20 ±0.73 | 71.37 ±0.42 | 84.78 ±2.08 | 57.67 ±0.05 | 93.85 ±0.87 | 72.05 ±1.03 | 93.72 ±0.55 |
| | 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 |
| | SEER | ✓ | ✗ | ✗ | ✗ | 74.65 ±0.50 | 8.12 ±4.28 | 76.20 ±0.83 | 7.86 ±2.02 | 51.05 ±0.65 | 8.45 ±1.73 | 80.12 ±0.40 | 5.25 ±0.72 |
| | **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 |

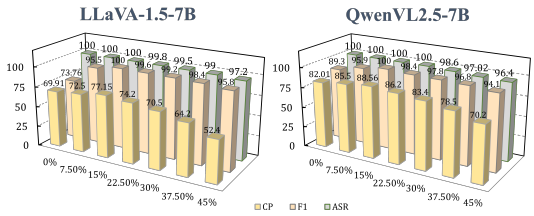


Figure 4: **Impact of Poisoning Rates ($\gamma$).** Multi-dimensional stability analysis of SPEAR across different VLM architectures on VSR. The 3D bar charts visualize the trade-off between reasoning fidelity ($\mathcal{CP}$) and security metrics. The default setting $\gamma = 15\%$ is used as a controlled diagnostic configuration, while ultra-low poisoning rates ($0.1\%$–$1.0\%$) are further evaluated in Section D.1. Detailed analysis is provided in Section 4.3.

gate-induced energy signature remains detectable an being a high-density poisoning artifact.

Table 3: **Impact of Screening Subset Size ($\rho$).** Evaluation on LLaVA-1.5 with $\gamma = 15\%$ on IconQA. The efficiency-performance elbow appears at $\rho = 25\%$, our default setting.

| Ratio ($\rho$) | ASR (↓) | Precision (↑) | Recall (↑) |
|---|---|---|---|
| 10% | 2.45% | 88.2% | 91.5% |
| **25%** | **0.00%** | **100.0%** | **99.5%** |
| 40% | 0.00% | 99.3% | 99.9% |
| 55% | 0.00% | 99.5% | 99.9% |

ong purification with limited screening overhead.

Table 4: **Effectiveness of Detection Paths.** Evaluation on InternVL3-8B with $\gamma = 15\%$ on IconQA. The Intersection strategy used in SPEAR best balances security, precision, and utility.

| Strategy | ASR (↓) | Prec. (↑) | Rec. (↑) | CP (↑) |
|---|---|---|---|---|
| Global Path Only | 0.54% | 92.4% | 99.5% | 58.42% |
| Local Path Only | 1.12% | 89.1% | 99.4% | 57.15% |
| Union ($\cup$) | 0.00% | 86.5% | **99.9%** | 55.20% |
| **Intersection ($\cap$)** | **0.00%** | **99.4%** | 99.1% | **63.55%** |