GPC-E4: Generalized Consistency and Entropic Evidence for Backdoor Detection in Vision-Language Models

Xuan Wang[1], Xuehu Yan[2], Yuliang Lu[2], Kejiang Ye[3], Cheng-zhong Xu[4], Xitong Gao[3]*
[1]National University of Defense Technology, China
[3]Shenzhen Institutes of Advanced Technology, CAS
[4]Institute of AI and Brain Sciences, CS Dept., University of Macau
xuanwang@ntu.edu.sg, xt.gao@siat.ac.cn

Abstract

Vision-language models (VLMs) are increasingly used through black-box, open-ended interfaces, where input-level backdoor screening must work without model parameters, gradients, logits, or training data. We present **GPC-E4**, a detector for single image-prompt queries that probes a VLM with mild, semantics-preserving visual and textual perturbations. The method measures whether the output behavior contracts unusually under these probes: GPC captures dominant-mode or semantic consistency, while E4 summarizes predictive entropy, augmentation mutual information, minimal collapse, and local slope along perturbation paths. A fixed robust normalization combines these signals and can operate without a known-clean calibration set. Across COCO, Flickr30k, Caltech101, and MMBench-style evaluations with LLaVA-1.5, MiniGPT-4, CLIP, CoOp, and Qwen3-VL, GPC-E4 outperforms STRIP, SCALE, TeCo, and BDet in both closed-set and open-ended settings. Additional analyses under alternative stressors, prompt and distribution shifts, zero clean calibration, VLM-native attacks, and detector-aware attacks characterize when the contraction signal is reliable and when it weakens.

1 Introduction

Vision-language models (VLMs) are now commonly accessed through interfaces that accept an image and an instruction and return either a label, a caption, or a free-form answer. This interface is convenient for deployment, but it makes input-level backdoor detection difficult. A screening method may see only one image-prompt query at a time, may receive open-ended text rather than logits, and often has no access to model parameters, gradients, activations, or training data. These constraints exclude many model-inspection and trigger-inversion defenses that are effective for conventional image classifiers (Wang et al., 2019; Liu et al., 2019; Gao et al., 2019; Guo et al., 2023).

Multimodal backdoors add another source of difficulty. Triggers can be visible patches, blended patterns, frequency-coded signals, content-aware transformations, or instruction-dependent image-text keys (Gu et al., 2019; Chen et al., 2017; Zeng et al., 2021; Nguyen and Tran, 2021; Liang et al., 2025; Xu et al., 2024; Liang et al., 2024). They may not be identifiable from pixels alone, and their effect may appear only through the conditional behavior of the VLM under a particular prompt. This motivates a black-box detector that treats the model response as the main observable and asks how that response changes under controlled, semantics-preserving probes.

Our starting point is that successful triggered inputs often exhibit unusually stable behavior under such probes. For clean inputs, mild spatial, photometric, frequency, stochastic, or textual perturbations usually introduce ordinary variation in labels or generated text. For triggered inputs, the response is more likely to remain concentrated around the target behavior. We refer to this empirical pattern as local distributional contraction. It is not meant as a universal guarantee: weak triggers, overly strong perturbations, or adaptive attacks that preserve output diversity can weaken it. Nevertheless, it provides a useful organizing principle for measuring black-box VLM behavior.

We instantiate this principle as **GPC-E4**. Generalized Prediction Consistency (GPC) measures dominant-label stability for closed-set wrappers and semantic stability for open-ended generation. The Entropy Quartet (E4) measures four complementary aspects of output uncertainty under perturbations: predictive entropy, augmentation mutual information, minimal collapse, and initial slope along perturbation paths. These statistics are normalized with robust reference estimates and fused with fixed weights, so the detector does not require

*Corresponding Author.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 3507–3522
November 6-10, 2026 ©2026 Association for Computational Linguistics

attack-label training and can also run with an unlabeled screening pool instead of a clean calibration set.

We evaluate GPC-E4 across closed-set and open-ended regimes on COCO, Flickr30k, Caltech101, and MMBench-style prompts, using LLaVA-1.5, MiniGPT-4, CLIP, CoOp, and Qwen3-VL. The experiments include static, spectral, content-aware, VLM-native, domain-shift, and detector-aware attacks, and report both AUROC and FPR@95TPR. The results show consistent gains over STRIP, SCALE, TeCo, and BDet, while the adaptive attack study highlights an evasion–utility trade-off rather than a complete defense.

Contributions. Our contributions are:

- We formulate black-box input-level VLM backdoor detection as a response-stability problem that supports both closed-set outputs and open-ended generation.
- We propose **GPC-E4**, a logit-free detector that combines semantic consistency with four uncertainty statistics computed under mild multi-view perturbations.
- We evaluate the method across models, datasets, trigger families, output regimes, and deployment stress tests, including zero clean calibration and detector-aware attacks.

2 Task Formulation and Assumptions

Model interface. We consider a deployed VLM accessed via a black-box interface, such as a cloud API or a local inference service. Given an image-prompt pair $z = (\boldsymbol{x}, t)$, the model returns either (i) a closed-set predictive distribution

$$p(y \mid z; Y), \qquad y \in Y, \tag{1}$$

or (ii) an open-ended textual response $a(z)$. Token probabilities or logits may be used when an interface exposes them, but they are not required by the detector. The detector has no access to model parameters, gradients, intermediate representations, or training data, and may only issue input queries and observe the resulting outputs.

Backdoor setting. The adversary seeks to implant a backdoor such that, in the presence of a trigger, the model exhibits targeted behavior, while maintaining normal utility on clean inputs. In closed-set settings, this corresponds to steering predictions toward a target label y_t; in open-ended generation, this corresponds to inducing a target semantic response or output pattern. We consider a broad class of attacks, including those that rely solely on data poisoning, as well as attacks where the adversary can manipulate the training process directly, and allow triggers that are explicit or stealthy, color-biased, frequency-coded, or adversarially optimized. The adversary is constrained to preserve high task performance and output plausibility on clean inputs, thereby limiting overt distributional shifts that would trivially reveal the attack.

Attacker knowledge. By default, we assume a detector-agnostic adversary that is unaware of the specific detection mechanism. We additionally consider detector-aware adversaries in which the attacker has knowledge of the detection strategy and attempts to adapt accordingly; these adaptive settings are evaluated empirically. In all cases, the adversary does not control the detector's query process or post-hoc analysis, and cannot modify model outputs at inference time.

Screening task. The goal of the detector is to determine, for a single input z, whether it contains a backdoor trigger. Formally, this is a hypothesis test based only on observable behavior:

$$\begin{cases} \mathcal{H}_0 : z \text{ is clean}, \\ \mathcal{H}_1 : z \text{ contains a trigger}. \end{cases} \tag{2}$$

Under $\mathcal{H}_0$, the input induces a nominal conditional output distribution; under $\mathcal{H}_1$, it induces abnormal conditional behavior. The detector must operate at the input level, without assumptions on fixed target labels, trigger forms, or output templates, and must support both closed-set and open-ended generation regimes.

3 Method

3.1 Problem Setup

We consider a black-box detector that receives an image-prompt pair $(\boldsymbol{x}, t)$, where $\boldsymbol{x} \in \mathcal{X}$ denotes the visual input and $t \in \mathcal{T}$ is the accompanying prompt or instruction. The queried VLM returns either (i) a closed-set response over a finite label set $Y = \{y_k\}_{k=1}^{K}$, or (ii) an open-ended textual response $a(\boldsymbol{x}, t)$. The detector cannot inspect model parameters, intermediate states, gradients, or training data; it can only issue perturbed input queries and observe outputs.

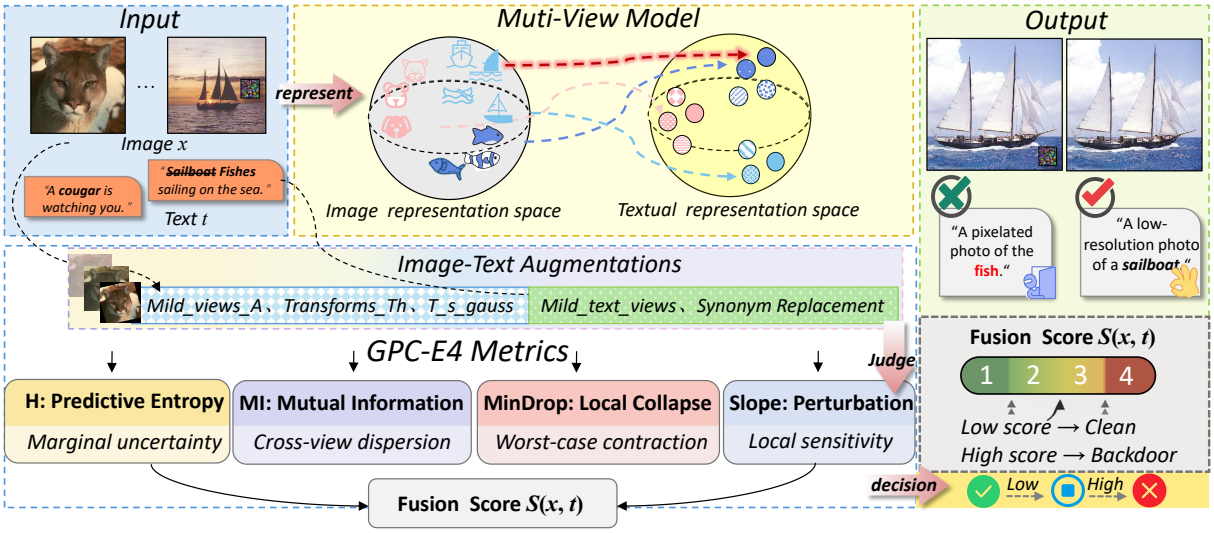


Figure 1: **Overview of the proposed GPC-E4 framework.** Our framework evaluates abnormal contraction and invariance of a VLM's conditional output distribution under controlled multimodal perturbations. **Multimodal augmentation pipeline:** original images and texts are transformed into multiple views to compute E4 metrics (H, MI, MinDrop, and Slope), which are used for backdoor detection. Samples with higher final scores are classified as backdoors.

We write $q(\cdot \mid \boldsymbol{x}, t)$ for the observable response distribution. In closed-set settings, q is the class probability vector when available, or the empirical distribution obtained from repeated black-box label samples. In open-ended settings, q is an empirical distribution over semantic response clusters, constructed from repeated generations using embedding similarity or semantic agreement. This notation lets us define the detector without assuming logits.

The detection task is a hypothesis test: $\mathcal{H}_0$ states that the input is clean, whereas $\mathcal{H}_1$ states that the input contains a trigger that induces abnormal behavior under mild perturbations. Our objective is to estimate a contraction score $\kappa(\boldsymbol{x}, t)$ from black-box observations. A larger score means that nearby semantics-preserving views produce a more concentrated and stable response distribution.

3.2 Response Contraction View

For clarity, we first describe the closed-set case. Let

$$y^* = \arg\max_{y \in Y} q(y \mid \boldsymbol{x}, t) \tag{3}$$

be the dominant response mode, and define the observable probability margin

$$\gamma(\boldsymbol{x}, t) = q(y^* \mid \boldsymbol{x}, t) - \max_{y \neq y^*} q(y \mid \boldsymbol{x}, t). \tag{4}$$

Let $\mathcal{B}_\rho(\boldsymbol{x}, t)$ be the local neighborhood induced by our stressor distribution, restricted to mild perturbations that preserve task semantics.

Local stability. We do not assume global robustness. We assume only that inside this mild neighborhood, the black-box response varies continuously enough that

$$\begin{aligned} \Delta_v q &= q(\cdot \mid v(\boldsymbol{x}, t)) - q(\cdot \mid \boldsymbol{x}, t), \\ \|\Delta_v q\|_\infty &\leq \epsilon_\rho, \\ v(\boldsymbol{x}, t) &\in \mathcal{B}_\rho(\boldsymbol{x}, t). \end{aligned} \tag{5}$$

A successful trigger is expected to increase the local margin along the attack shortcut, creating a stable basin around its target behavior. Clean inputs may also be stable, but their stability is typically less anisotropic and less coupled to entropy contraction across the stressor views.

Implications for the observables. If $\gamma(\boldsymbol{x}, t) > 2\epsilon_\rho$, then the dominant mode cannot change inside the local neighborhood, which explains why GPC rises for triggered inputs. If $q(y^* \mid \boldsymbol{x}, t) = 1 - \delta$, then entropy obeys

$$H(q) \leq h_2(\delta) + \delta \log(K - 1), \tag{6}$$

where h_2 is binary entropy. Thus a larger local margin and smaller off-mode mass naturally imply entropy contraction. MinDrop estimates the depth of the basin by asking how much uncertainty

can be restored under the worst mild perturbation, while Slope estimates the local curvature of this basin as perturbation strength increases. In open-ended generation, the same interpretation applies after replacing class modes with semantic response clusters.

This view also describes when the signal can fail; additional details are given in Section A. If a trigger does not create a stable local basin, if the perturbations leave the semantics-preserving neighborhood, or if an adaptive attacker explicitly preserves output diversity, the contraction signal weakens. Our experiments therefore include alternative stressors, distribution shifts, and detector-aware diversity-preserving attacks rather than relying only on standard static triggers.

3.3 Multi-View Stressors

The stressors are probes, not the object being detected. We use them to sample a local neighborhood where semantic content is preserved but nuisance factors vary. The design follows three principles: *semantic preservation*, *multi-view coverage*, and *controlled magnitude*. Concretely,

$$\mathcal{V} = \{\, v_\theta : \mathcal{X} \times \mathcal{T} \to \mathcal{X} \times \mathcal{T} \mid \theta \sim \Pi \,\}, \tag{7}$$

where Π mixes visual and textual perturbations. Visual stressors include spatial transforms, photometric changes, JPEG/blur/noise corruptions, and frequency or phase perturbations. Textual stressors include prompt paraphrases, synonym replacement, and light token dropout. Unless otherwise stated, GPC, MinDrop, and Slope perturb the visual input while keeping the prompt fixed; MI uses paired mild visual and textual views to measure cross-modal output variability.

3.4 GPC and E4 Metrics

Generalized Prediction Consistency. For closed-set outputs, let $C(\boldsymbol{x}, t; Y) = \arg\max_{y \in Y} q(y \mid \boldsymbol{x}, t)$. GPC estimates dominant-mode stability:

$$\begin{aligned} \widehat{\mathrm{GPC}}(\boldsymbol{x}, t) &= \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\{c_m = c_0\}, \\ c_m &= C(v_m(\boldsymbol{x}, t); Y), \\ c_0 &= C(\boldsymbol{x}, t; Y), \qquad v_m \sim \Pi. \end{aligned} \tag{8}$$

For open-ended outputs, we replace exact label equality with semantic agreement:

$$\begin{aligned} \mathrm{GPC}_{\mathrm{sem}}(\boldsymbol{x}, t) &= \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\{s_m \geq \tau\}, \\ s_m &= \mathrm{sim}\big(a(v_m(\boldsymbol{x}, t)), a(\boldsymbol{x}, t)\big). \end{aligned} \tag{9}$$

where sim is an embedding or semantic-similarity score. The threshold τ is fixed on development prompts or selected by a calibration-free percentile rule on an unlabeled screening batch.

Predictive entropy. Entropy measures marginal dispersion:

$$\begin{aligned} H(\boldsymbol{x}, t) = -\sum_{y \in \mathcal{Y}_{\mathrm{obs}}} & q(y \mid \boldsymbol{x}, t) \\ & \cdot \log q(y \mid \boldsymbol{x}, t), \end{aligned} \tag{10}$$

where $\mathcal{Y}_{\mathrm{obs}}$ is either the closed-set label space or the set of semantic response clusters.

Augmentation mutual information. For mild views $v_j \sim \mathcal{A} \subset \mathcal{V}$, MI captures cross-view dispersion:

$$\begin{aligned} q_j &= q(\cdot \mid v_j(\boldsymbol{x}, t)), \qquad \bar{q} = \frac{1}{A} \sum_{j=1}^{A} q_j, \\ \mathrm{MI}(\boldsymbol{x}, t) &= H(\bar{q}) - \frac{1}{A} \sum_{j=1}^{A} H(q_j). \end{aligned} \tag{11}$$

Minimal collapse. Let $T_{r,s}$ denote a perturbation path of type r with small strength s. MinDrop probes worst-case basin depth:

$$\begin{aligned} \mathrm{MinDrop}(\boldsymbol{x}, t) &= \min_{r \in \mathcal{R}} D_r, \\ D_r &= H(\boldsymbol{x}, t) - H(T_{r,s_0}(\boldsymbol{x}, t)). \end{aligned} \tag{12}$$

Initial slope. Slope measures local sensitivity across nearby perturbation strengths:

$$\begin{aligned} \mathrm{Slope}(\boldsymbol{x}, t) = & H(T_{r,s_0}(\boldsymbol{x}, t)) \\ & - H(T_{r,s_1}(\boldsymbol{x}, t)), \\ & 0 < s_0 < s_1. \end{aligned} \tag{13}$$

Although Slope alone can be weakly discriminative, it is a high-interaction feature: it couples the entropy, MI, and MinDrop signals across perturbation scales. This distinction is important for interpreting the ablation results.

3.5 Unified Score and Decision Policy

The metrics capture related but differently scaled stability signals. We therefore use a fixed robust fusion rather than a learned classifier. Let $Z_j(\cdot)$ denote robust normalization using the median and median absolute deviation (MAD) of either a small clean development set, an unlabeled screening batch, or deployment-specific reference traffic:

$$Z_j(s) = \frac{s - \text{median}(S_j)}{\text{MAD}(S_j) + \epsilon}. \tag{14}$$

The unlabeled variant does not require known-clean calibration and is robust when the screened pool has a minority of triggered inputs.

We define an E4 subscore

$$\begin{aligned} E_4(\boldsymbol{x}, t) = w_H Z_H(-H) \\ + w_I Z_I(-\text{MI}) \\ + w_D Z_D(\text{MinDrop}) \\ + w_S Z_S(\text{Slope}). \end{aligned} \tag{15}$$

The final score is

$$\begin{aligned} \text{Score}(\boldsymbol{x}, t) = \alpha Z_U(U(\boldsymbol{x}, t)) \\ + (1 - \alpha) E_4(\boldsymbol{x}, t). \end{aligned} \tag{16}$$

where $U = \widehat{\text{GPC}}$ for closed-set outputs and $U = \text{GPC}_{\text{sem}}$ for open-ended outputs. We use $\alpha \in [0.3, 0.4]$ and $(w_H, w_I, w_D, w_S) = (0.1, 0.4, 0.4, 0.1)$. The weights balance anisotropic sensitivities and metric scales; they are not learned on attack labels. All signs are chosen so that a larger score indicates stronger contraction.

For deployment, we use the score as a composite test statistic:

$$\text{action}(s) = \begin{cases} \texttt{block}, & s \geq 3, \\ \texttt{review}, & 2 \leq s < 3, \\ \texttt{allow}, & s < 2. \end{cases} \tag{17}$$

Variance over stressor seeds provides a confidence diagnostic.

4 Experiments

Experimental scope. We evaluate whether GPC-E4 detects triggered inputs because it estimates local distributional contraction, not because it overfits to one perturbation set. The experiments cover (i) closed-set and open-ended outputs, (ii) static, content-aware, spectral, instruction-sensitive, and adaptive backdoors, (iii) multiple datasets and VLM architectures, and (iv) deployment metrics beyond AUROC.

Algorithm 1 Detection via Multi–View GPC and E4 (multimodal variant).

Require: image $\boldsymbol{x}$, text prompt t, VLM, options Y, probe family Π, light aug $\mathcal{A}$, path T_s, repeats R

Ensure: action $\in \{\text{allow}, \text{review}, \text{block}\}$, score, diagnostics

for $r \leftarrow 1$ to R **do**
  $\hat{y} \leftarrow C(\boldsymbol{x}, t; Y)$
  $c_m \leftarrow C(v_m(\boldsymbol{x}, t); Y)$
  $\text{GPC}_r \leftarrow M^{-1} \sum_m \mathbb{I}\{c_m = \hat{y}\}$
  $v_m \sim \Pi$
  $H_0 \leftarrow H(\boldsymbol{x}, t)$
  $q_j \leftarrow p(\cdot \mid v_j(\boldsymbol{x}, t); Y)$
  $\bar{q} \leftarrow A^{-1} \sum_j q_j$
  $\text{MI}_r \leftarrow H(\bar{q}) - A^{-1} \sum_j H(q_j)$
  $v_j \sim \mathcal{A}$
  $D_{t'} \leftarrow H_0 - H(T_{t', s_0}(\boldsymbol{x}, t))$
  $\text{MinDrop}_r \leftarrow \min_{t' \in \mathcal{T}} D_{t'}$
  $h_i \leftarrow H(T_{s_i}(\boldsymbol{x}, t))$, for $i \in \{0, 1\}$
  $\text{Slope}_r \leftarrow h_0 - h_1$
end for
Aggregate per-seed means for $U \in \{\text{NGPC}, \text{GPC}_{\text{sem}}\}$ and the four E4 components.
Compute $\text{Score}(\boldsymbol{x}, t)$ using the fusion rule above and return action, score, and diagnostics.

4.1 Setup

Models. We evaluate LLaVA-v1.5 (Liu et al., 2023a), MiniGPT-4 (Zhu et al., 2024), CLIP (Radford et al., 2021), CoOp (Zhou et al., 2022), and an instruction-tuned Qwen3-VL setting. The first four reproduce the main benchmark; Qwen3-VL with MMBench-style prompts probes scalability to a stronger instruction-following VLM.

Attacks. We include static triggers (BNs (Gu et al., 2019), BL (Chen et al., 2017)), content-aware or spectral triggers (IS (**?**), Low-Frequency (Zeng et al., 2021)), model-adaptive attacks (BadCLIP (Bai et al., 2024), BEnc (Jia et al., 2022)), and VLM-native attacks (TV/VL-Trojan (Liang et al., 2025), SH (Xu et al., 2024), and MABA/domain-shift attacks (Liang et al., 2024)). We also evaluate detector-aware attacks that explicitly suppress GPC-E4 signals.

Datasets and tasks. The main benchmark uses COCO (Lin et al., 2014), Flickr (Young et al., 2014), and Caltech101 (Li et al., 2022). Closed-set

tasks use option sets of size $K = 10$–20, constructed by controlled synonym mining to avoid label leakage. Open-ended tasks use semantic-consistency prompts.

Metrics and calibration. We report AUROC and FPR@95TPR. FPR@95TPR is important for deployment because missing attacks is costly, so the detector must keep false alarms low at high recall. The default robust normalization uses a small development reference set when available. We also report a 0% clean-calibration setting that estimates median/MAD from the unlabeled screening pool.

Implementation. Unless otherwise stated, we use $M = 4$ visual stressors for the core statistics, $\mathcal{A} = 4$ paired mild image/text views for MI, $\mathcal{R} = 4$ paths for MinDrop, and slope strengths $s_0 = 0.01$, $s_1 = 0.03$. The default fusion weights are fixed as described in Section 3.5.

4.2 Main Detection Results

Tables 1 and 2 show that GPC-E4 consistently outperforms STRIP, TeCo, SCALE, and BDet across datasets, models, and trigger families. In closed-set COCO with LLaVA-1.5, GPC-E4 reaches AUROC from 0.894 to 0.965, with an average absolute gain of 0.080 over BDet. The open-ended setting is harder because semantic equivalence replaces label agreement, yet GPC-E4 still improves over BDet by an average of 0.089 AUROC on COCO with LLaVA-1.5.

These gains are largest for attacks that expose the limits of narrow perturbation detectors. STRIP and SCALE rely primarily on entropy or label consistency under restricted transformations. TeCo broadens the corruption set but remains mostly spatial/photometric. BDet relies on contrastive prompting and is less sensitive to purely visual or spectral triggers. GPC-E4 probes multiple local directions and measures both dominant-mode stability and uncertainty contraction, which explains its advantage on Low-Frequency, TV, and SH.

4.3 Deployment Metrics and Stronger VLMs

Table 3 reports FPR@95TPR on COCO with LLaVA-1.5. The detector keeps false positives low for both closed-set and open-ended outputs, with the hardest open-ended Shadowcast setting reaching 0.807 AUROC and 0.141 FPR@95TPR. This metric gives a more operational view than AUROC alone: under a high-recall operating point, GPC-E4 still separates triggered inputs from clean ones with manageable false alarms.

Table 1: Closed-set VLM input-level backdoor detection. Each cell reports **AUROC**↑.

Method	STRIP	TeCo	SCALE	BDet	Ours
COCO + LLaVA-1.5					
BadNet	0.785	0.838	0.749	0.914	**0.965**
Blended	0.329	0.789	0.667	0.899	**0.951**
ISSBA	0.315	0.583	0.608	0.918	**0.923**
Low-Freq.	0.470	0.734	0.575	0.806	**0.926**
TrojanVLM	0.105	0.447	0.543	0.725	**0.894**
Shadowcast	0.281	0.392	0.625	0.830	**0.912**
COCO + Mini-GPT4					
BadNet	0.687	0.754	0.675	0.881	**0.917**
Blended	0.296	0.710	0.600	0.854	**0.903**
ISSBA	0.284	0.525	0.547	0.842	**0.874**
Low-Freq.	0.423	0.661	0.518	0.760	**0.879**
TrojanVLM	0.195	0.402	0.489	0.685	**0.855**
Shadowcast	0.239	0.353	0.563	0.736	**0.867**
Flickr30k + LLaVA-1.5					
BadNet	0.698	0.796	0.751	0.887	**0.922**
Blended	0.323	0.649	0.673	0.832	**0.923**
ISSBA	0.298	0.554	0.678	0.861	**0.897**
Low-Freq.	0.446	0.697	0.546	0.782	**0.904**
TrojanVLM	0.124	0.425	0.416	0.703	**0.860**
Shadowcast	0.248	0.372	0.394	0.805	**0.885**
Flickr30k + Mini-GPT4					
BadNet	0.453	0.516	0.641	0.861	**0.889**
Blended	0.281	0.575	0.502	0.824	**0.876**
ISSBA	0.178	0.449	0.567	0.836	**0.849**
Low-Freq.	0.302	0.528	0.482	0.743	**0.852**
TrojanVLM	0.090	0.382	0.469	0.819	**0.828**
Shadowcast	0.119	0.324	0.535	0.785	**0.841**
Caltech101 + CLIP					
BadNet	0.612	0.672	0.782	0.902	**0.915**
Blended	0.198	0.486	0.751	0.879	**0.885**
ISSBA	0.212	0.465	0.432	0.842	**0.851**
BadCLIP	0.355	0.471	0.551	0.819	**0.870**
BadEncoder	0.292	0.517	0.553	0.807	**0.868**
Caltech101 + CoOp					
BadNet	0.735	0.776	0.868	0.965	**0.978**
Blended	0.125	0.584	0.895	0.959	**0.970**
ISSBA	0.334	0.515	0.473	0.926	**0.933**
BadCLIP	0.497	0.502	0.528	0.871	**0.954**
BadEncoder	0.302	0.569	0.483	0.858	**0.939**

Table 4 scales the evaluation to an instruction-tuned Qwen3-VL setting with MMBench-style prompts (Liu et al., 2023c). Performance decreases relative to the smaller benchmark settings, as expected for a stronger and more diverse instruction-following model, but AUROC remains above 0.75 for all open-ended attacks and above 0.81 for all closed-set attacks. The probing overhead is a constant factor in the number of stressor views and is embarrassingly parallel, so rate-limited deployments can trade recall for query cost by reducing M or applying early stopping.

Table 2: Open-ended VLM input-level backdoor detection on the COCO dataset. Each cell reports **AUROC**↑. Evaluation uses semantic-consistency prompts/templates.

Method	STRIP	TeCo	SCALE	BDet	Ours
COCO + LLaVA-1.5					
BadNet	0.708	0.754	0.674	0.827	**0.870**
Blended	0.281	0.710	0.613	0.804	**0.886**
ISSBA	0.268	0.525	0.547	0.792	**0.839**
Low-Freq.	0.424	0.661	0.518	0.735	**0.842**
TrojanVLM	0.189	0.402	0.489	0.694	**0.824**
Shadowcast	0.239	0.353	0.563	0.681	**0.807**
COCO + Mini-GPT4					
BadNet	0.584	0.679	0.608	0.827	**0.859**
Blended	0.252	0.604	0.548	0.808	**0.836**
ISSBA	0.242	0.446	0.527	0.756	**0.809**
Low-Freq.	0.365	0.597	0.492	0.693	**0.812**
TrojanVLM	0.076	0.382	0.469	0.664	**0.798**
Shadowcast	0.232	0.324	0.535	0.625	**0.784**

Table 3: Deployment-oriented metrics on LLaVA-1.5. Lower FPR@95TPR is better.

Attack	Closed-set		Open-ended	
	AUROC	FPR	AUROC	FPR
BadNet	0.965	0.033	0.870	0.067
Blended	0.951	0.039	0.886	0.084
ISSBA	0.923	0.056	0.839	0.092
Low-Freq.	0.926	0.061	0.842	0.089
TrojanVLM	0.894	0.087	0.824	0.126
Shadowcast	0.912	0.074	0.807	0.141

4.4 Adaptive and VLM-Native Attacks

Table 5 shows a detector-aware diversity-preserving attack. The attacker adds a regularizer that suppresses entropy contraction, cross-view dependence, worst-case collapse, and slope signals while preserving attack success. As the adaptive weight increases, AUROC drops from 0.965 to 0.724, but ASR also drops from 95.56% to 68.72%. This indicates that GPC-E4 is not invulnerable, but evading the contraction signal creates a measurable utility trade-off. The detector does not collapse to random guessing, because suppressing one component does not remove all contraction observables.

We also evaluate VLM-native attacks that are closer to multimodal deployment settings. For VL-Trojan, GPC-E4 improves over BDet from 0.695 to 0.872 in closed-set LLaVA-1.5 and from 0.662 to 0.842 in open-ended LLaVA-1.5. For MABA/domain-shift attacks, GPC-E4 improves from 0.704 to 0.856 closed-set and from 0.674 to 0.804 open-ended. The same trend holds on MiniGPT-4, where GPC-E4 reaches 0.835/0.808 for VL-Trojan and 0.832/0.776 for MABA in closed/open regimes. These results show that the method is not limited to static or content-aware triggers.

Table 4: Scaling to an instruction-tuned Qwen3-VL setting with MMBench-style prompts. Each cell reports AUROC.

Attack	Closed AUROC	Open AUROC
BadNet	0.887	0.821
Blended	0.874	0.808
ISSBA	0.859	0.790
Low-Freq.	0.852	0.766
TrojanVLM	0.835	0.782
Shadowcast	0.818	0.753

Table 5: Detector-aware diversity-preserving attack on LLaVA-1.5. Larger λ more strongly suppresses GPC-E4 signals.

λ	ASR (%)	AUROC	FPR@95
0.0	95.56	0.965	0.033
0.1	90.81	0.910	0.065
0.2	87.05	0.842	0.083
0.5	68.72	0.724	0.161

4.5 Ablation and Fusion Analysis

Table 6 is a leave-one-component ablation, not a set of single-metric detectors. All ablation and single-metric results in this subsection use $M = 8$ stressor views. Removing any E4 component degrades performance, which shows that the components provide non-redundant evidence. The largest drop occurs when Slope is removed. This does not contradict the observation that Slope alone has weak standalone separability: Slope acts as a high-interaction feature, linking entropy, MI, and MinDrop across perturbation scales. Without it, the fused representation loses cross-scale structure. MI and MinDrop show relatively strong standalone discrimination, while Slope has weak standalone separability but a strong interaction effect in the fused detector.

Table 7 further separates the contribution of each evidence signal by evaluating single-metric detectors. No individual E4 statistic matches the full E4 fusion, and the four metrics exhibit complementary behavior across closed-set and open-ended settings. In particular, Slope has weak standalone separability, but the leave-one-component results in Table 6 show that it remains important when combined with the other signals. Combining the four E4 statistics improves AUROC from 0.913/0.820 to 0.952/0.896 when GPC is further incorporated, supporting the use of semantic consistency as com-

Table 6: Leave-one-component ablation for GPC-E4 on COCO with LLaVA-1.5 using $M = 8$ stressor views. Each cell reports AUROC.

Method	Closed-set	Open-ended
Full (GPC-E4)	**0.952**	**0.896**
without H	0.785	0.649
without MI	0.879	0.733
without MinDrop	0.832	0.707
without Slope	0.567	0.538

Table 7: Single-metric comparison on COCO with LLaVA-1.5 using $M = 8$ stressors. Each cell reports AUROC.

Method	Closed-set	Open-ended
GPC only	0.887	0.815
H only	0.763	0.642
MI only	0.844	0.718
MinDrop only	0.826	0.691
Slope only	0.609	0.557
E4	0.913	0.820
GPC-E4	**0.952**	**0.896**

plementary evidence rather than as a substitute for uncertainty-based signals.

Table 8 shows the query-cost trade-off. Performance continues to improve as the number of stressor views increases from $M = 1$ to $M = 16$, but the marginal gains become smaller beyond $M = 8$. This supports an early-stopping deployment rule: continue querying only until the metric variance across stressor seeds falls below a chosen tolerance.

Table 9 shows that GPC-E4 is slower than BDet but much faster than STRIP and TeCo on the 50K-sample Caltech101 setting.

4.6 Robustness of the Stressor Design

We place additional stressor, calibration, prompt, and distribution-shift analyses in the appendix (Section H). Replacing the original stressor family with an alternative mild geometric/photometric/noise/text set changes AUROC by at most 0.013. Using 0% clean calibration gives an average AUROC of 0.930, compared with 0.933 under 10% clean calibration. Prompt rephrasing, style shift, domain shift, and noise perturbation keep AUROC above 0.909. Together, these results support the intended interpretation: the detector depends on a local multi-view probing principle, not on one fragile hand-crafted perturbation recipe.

5 Related Work

Backdoor detection in vision. Classical backdoor detectors for vision models include trigger inversion, model diagnostics, saliency analysis,

Table 8: Number of stressor views M *vs.* performance and cost (COCO, LLaVA-1.5).

M	Closed AUROC	Open AUROC	Time ms	VRAM GB	Tokens /sample
1	0.894	0.822	11.09	8.1	78
4	0.926	0.863	33.14	8.9	312
8	0.952	0.896	60.72	10.5	624
16	0.967	0.904	118.65	12.7	1248

Table 9: Inference time comparison on Caltech101 (Li et al., 2022) with 50K test samples on BadCLIP with RN50 backbone. (h: hours, m: minutes.)

Method	**STRIP**	**SCALE**	**TeCo**	**BDet**	**Ours**
Time	4.2 h	10.1 m	12.8 h	3.2 m	5.0 m

and perturbation-based input screening. Neural Cleanse (Wang et al., 2019) and ABS (Liu et al., 2019) recover or localize triggers using model internals, while activation clustering (Chen et al., 2019) and SentiNet (Chou et al., 2020) assume access to representations or saliency maps. These methods are difficult to apply when a deployed VLM is available only through a black-box API. STRIP (Gao et al., 2019) avoids model inspection by measuring entropy under input mixing, but it is mainly designed for closed-set classifiers and probability outputs.

Perturbation-consistency defenses. Recent input-level detectors probe prediction stability under controlled transforms. SCALE (Guo et al., 2023) uses intensity scaling and label agreement, and TeCo (Liu et al., 2023b) measures consistency under natural corruptions. IBD-PSC (Hou et al., 2024) tracks confidence trajectories after rescaling batch-normalization parameters, but requires parameter access. These approaches are effective in their intended settings, yet their probe spaces are usually narrow and their statistics are tied to closed-set labels or probabilities. They also under-cover frequency-coded, phase-preserving, and content-aware triggers (Zeng et al., 2021; Nguyen and Tran, 2021; Qi et al., 2021).

Backdoors in VLMs. Multimodal backdoors can exploit visual patches, blended signals, spectral patterns, spatial warping, poisoned image-text alignment, or instruction-dependent keys (Gu et al., 2019; Chen et al., 2017; Zeng et al., 2021; Nguyen and Tran, 2021; Liang et al., 2025; Xu et al., 2024; Liang et al., 2024). Existing defenses for CLIP or VLMs often adapt classifier-based detectors, use contrastive prompting such as BDetCLIP (Niu

et al., 2025), or rely on clean reference data and model internals. They do not directly address single-input screening for black-box VLMs that produce open-ended text (Liu et al., 2023a; Zhu et al., 2024).

Position of this work. GPC-E4 keeps the input-level, black-box deployment goal of perturbation-based detectors, but adapts the measurements to VLM outputs. Rather than relying on one consistency or entropy statistic, it combines dominant-mode or semantic stability, entropy, mutual information, worst-case collapse, and local slope. This formulation supports both closed-set wrappers and open-ended generation, and it uses multi-view stressors as probes rather than as a single handcrafted trigger assumption.

6 Conclusion

We presented **GPC-E4**, a black-box input-level detector for backdoored VLM inputs. The method is built around a local distributional-contraction view: successful triggers often create a stable margin or semantic basin that becomes visible under mild multi-view perturbations. GPC measures dominant-mode stability, while E4 measures complementary uncertainty dynamics through entropy, MI, MinDrop, and Slope. Experiments across datasets, models, attack families, and deployment stress tests show consistent improvements over prior input-level detectors, while adaptive attacks reveal a clear evasion-utility trade-off.

Ethical Considerations

This work studies a controlled failure mode in deployed and deployable vision-language models. While our findings show that triggered inputs can induce abnormal stability and uncertainty patterns under black-box probing, our primary objective is to improve safety auditing and motivate more robust screening of multimodal systems.

Safety and containment. All experiments were conducted in isolated research settings using public benchmarks, local or reproducible model artifacts, and controlled trigger implementations. We did not interact with live third-party services, production systems, or human users. The triggers and detector-aware attacks are used only to evaluate detection behavior and are not intended for deployment.

Dual-use risk. Backdoor detection research can be dual-use: an attacker may use reported failure modes to design more evasive triggers. We therefore report adaptive-attack results and limitations, and recommend using GPC-E4 as one component of a broader defense pipeline for high-risk multimodal systems.

Use of AI assistants. We acknowledge the use of AI assistants for linguistic refinement of the manuscript and for assistance with auxiliary code and formatting. All scientific claims, experimental design choices, and core algorithmic contributions remain the responsibility of the authors.

Artifact use. We use publicly available datasets, model checkpoints, and benchmark artifacts for research evaluation, following their intended research use where specified. Any trigger-bearing samples, perturbed inputs, or detector-aware attack outputs are used solely for controlled safety analysis and are not intended for non-research use.

Limitations

Our analysis targets input-level backdoor detection for vision-language models under black-box access. We do not claim that local response contraction is a universal signature of all backdoors, or that GPC-E4 will work with equal strength across all VLM architectures, deployment interfaces, and trigger families. The current experiments cover representative closed-set and open-ended settings on COCO, Flickr30k, Caltech101, and MMBench-style prompts, but they do not cover long-horizon multimodal agents, memory-augmented dialogue systems, multi-tool API environments, video-language models, or stateful embodied interaction. The detector assumes that the defender can issue multiple perturbed queries for the same input; this may be costly or impractical for heavily rate-limited proprietary APIs. Our stressor families are designed to be mild and semantics-preserving, but we do not systematically vary all possible prompt templates, decoding policies, semantic-clustering models, API filtering rules, or task distributions. Adaptive attacks that preserve output diversity, triggers that activate only under highly specific instructions, and attacks that do not form stable local response basins can weaken the contraction signal. Defense evaluation is limited to input-level screening and representative post-hoc detection baselines; we do not evaluate training-time defenses or model-repair methods under the same experimental protocol.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (62376263), Natural Science Foundation of Guangdong (2024A1515030209), and Shenzhen Science and Technology Innovation Commission (JCYJ20230807140507015).

References

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. 2024. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24239–24250.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian M. Molloy, and Biplav Srivastava. 2019. Detecting backdoor attacks on deep neural networks by activation clustering. In *Proceedings of the AAAI Workshop on Artificial Intelligence Safety (SafeAI 2019)*.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Edward Chou, Florian Tramèr, and Giancarlo Pellegrino. 2020. Sentinet: Detecting localized universal attacks against deep learning systems. In *Deep Learning and Security Workshop*.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. 2019. Strip: a defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, page 113–125, New York, NY, USA. Association for Computing Machinery.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244.

Junfeng Guo, Yiming Li, Xun Chen, Hanqing Guo, Lichao Sun, and Cong Liu. 2023. SCALE-UP: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency. In *The Eleventh International Conference on Learning Representations*.

Linshan Hou, Ruili Feng, Zhongyun Hua, Wei Luo, Leo Yu Zhang, and Yiming Li. 2024. IBD-PSC: input-level backdoor detection via parameter-oriented scaling consistency. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059. IEEE.

Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. 2022. Caltech 101.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. 2025. VL-Trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*.

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Mingli Zhu, Xiaochun Cao, and Dacheng Tao. 2024. Revisiting backdoor attacks against large vision-language models from domain shift. *Preprint*, arXiv:2406.18844.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, and Chaowei Xiao. 2023b. Detecting backdoors during the inference stage based on corruption robustness consistency. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16363–16372.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. 2019. 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 '19, page 1265–1282, New York, NY, USA. Association for Computing Machinery.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023c. MMBench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Tuan Anh Nguyen and Anh Tuan Tran. 2021. Wanet - imperceptible warping-based backdoor attack. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. 2025. Test-time multimodal backdoor detection by contrastive prompting. In *ICML*.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *IEEE Symposium on Security and Privacy*.

Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. 2024. Shadowcast: Stealthy data poisoning attacks against vision-language models. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Yi Zeng, Won Park, Z. Morley Mao, and Ruoxi Jia. 2021. Rethinking the Backdoor Attacks' Triggers: A Frequency Perspective . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16453–16461, Los Alamitos, CA, USA. IEEE Computer Society.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Theoretical Notes on Local Response Contraction

This section gives a local explanation for why the statistics used by GPC-E4 move together on many triggered inputs. The goal is not to prove that every backdoor must induce contraction. Instead, the analysis states a mild local condition under which dominant-mode stability, entropy reduction, mutual-information changes, MinDrop, and Slope are different views of the same response behavior. This framing is useful because the detector only observes black-box outputs.

Setup. For a closed-set interface, let $q(z) \in \Delta^{K-1}$ denote the observable predictive distribution for an image-prompt input $z = (\boldsymbol{x}, t)$. Let

$$y^*(z) = \arg\max_y q_y(z) \tag{18}$$

and define the observable margin

$$\begin{aligned}\gamma(z) = q_{y^*}(z) \\ - \max_{y \neq y^*} q_y(z).\end{aligned} \tag{19}$$

Let $v \sim \Pi_\rho$ be a mild semantics-preserving perturbation with strength bounded by ρ, and write $z_v = v(z)$. The local stability condition used in the main text is

$$\|q(z_v) - q(z)\|_\infty \leq \epsilon_\rho. \tag{20}$$

The condition is local and output-level: it does not assume access to logits, gradients, or model parameters. It can fail when the perturbation changes task semantics or when an adaptive attack deliberately preserves output diversity.

Dominant-mode stability. If $\gamma(z) > 2\epsilon_\rho$, then the dominant label is invariant under any perturbation satisfying Equation (20). Indeed, for $y \neq y^*$,

$$\begin{aligned}q_{y^*}(z_v) - q_y(z_v) &\geq q_{y^*}(z) - q_y(z) \\ &\quad - 2\epsilon_\rho \\ &\geq \gamma(z) - 2\epsilon_\rho > 0.\end{aligned} \tag{21}$$

Thus GPC estimates the probability mass of the local perturbation neighborhood on which the dominant mode is unchanged:

$$\mathrm{GPC}(z) \approx \Pr_{v \sim \Pi_\rho}[y^*(z_v) = y^*(z)]. \tag{22}$$

A trigger that creates a wider target basin increases this probability.

Entropy contraction. Let $q_{y^*}(z) = 1 - \delta$, where δ is the total off-mode mass. For fixed δ, entropy is maximized when the off-mode mass is uniform, so

$$H(q(z)) \leq h_2(\delta) + \delta \log(K - 1), \tag{23}$$

where h_2 is binary entropy. When a trigger concentrates probability on a target response, δ decreases and the upper bound in Equation (23) decreases. This explains why entropy can fall together with higher GPC. The bound does not imply that all clean inputs have high entropy; rather, it links strong local mode concentration to low entropy.

Augmentation mutual information. For paired views $z_j = v_j(z)$, let Q be the random output distribution $q(z_j)$, and let $\bar{q} = \mathbb{E}_j[q(z_j)]$. The augmentation mutual-information statistic is

$$\mathrm{MI}(z) = H(\bar{q}) - \mathbb{E}_j[H(q(z_j))]. \tag{24}$$

By concavity of entropy, this quantity is nonnegative. It is small when all views produce nearly identical distributions and larger when views disagree about the output distribution. In practice, triggered inputs often show low entropy and reduced cross-view dispersion, so MI complements entropy by measuring whether concentration is shared across perturbation views rather than appearing in only one query.

MinDrop as a lower-tail basin statistic. For a set of perturbation paths $\mathcal{R}$, define

$$\begin{aligned}D_r(z) &= H(z) - H(T_{r,s_0}(z)), \\ \mathrm{MinDrop}(z) &= \min_{r \in \mathcal{R}} D_r(z).\end{aligned} \tag{25}$$

Unlike mean entropy change, MinDrop is controlled by the weakest direction in the stressor set. A high MinDrop therefore indicates that contraction persists across all tested mild directions, while a low value indicates that at least one perturbation can restore uncertainty. This makes MinDrop useful for detecting whether the local basin is robust or only visible under a favorable transform.

Slope as local sensitivity. Let $T_{r,s}(z)$ be a smooth perturbation path and define $g_r(s) = H(q(T_{r,s}(z)))$. For nearby strengths $0 < s_0 < s_1$, the slope statistic satisfies

$$\begin{aligned}\mathrm{Slope}_r(z) &= g_r(s_0) - g_r(s_1) \\ &= -(s_1 - s_0) g_r'(\xi)\end{aligned} \tag{26}$$

for some $\xi \in (s_0, s_1)$ by the mean value theorem. Thus Slope estimates local entropy sensitivity

along a stress path. This explains why Slope can have weak standalone separability but still improve fusion: it couples the entropy and MinDrop signals across perturbation scales.

Open-ended outputs. When logits or class probabilities are unavailable, the same argument applies after replacing labels with semantic response clusters. Repeated generations under prompt and image views induce an empirical categorical distribution over clusters $\mathcal{C} = \{c_1, \ldots, c_L\}$:

$$\hat{q}_\ell(z) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\{a_i(z) \in c_\ell\}. \tag{27}$$

All GPC-E4 components can then be computed on $\hat{q}$. The quality of this approximation depends on the semantic embedding or agreement model used for clustering; this is why open-ended results are more sensitive to decoding and prompt variation.

Scope of the analysis. The analysis supports a local screening rule, not a universal impossibility result for adaptive attacks. A detector-aware attacker can try to increase output diversity under probes, which reduces the contraction signal. The adaptive-attack experiment in the main text is therefore important: it measures the empirical trade-off between preserving attack success and hiding these local statistics.

B Calibration, Thresholding, and Query Complexity

Robust normalization. For a metric value $s_j(z)$, GPC-E4 uses

$$Z_j(s_j) = \frac{s_j - \mathrm{median}(S_j)}{\mathrm{MAD}(S_j) + \epsilon}, \tag{28}$$

where S_j is computed from either a small development set, reference traffic, or the unlabeled screening batch. The median/MAD choice is deliberate: if triggered inputs are a minority of the screening pool, these estimates remain stable and do not require a known-clean split.

Threshold selection. In deployment, the continuous score can be used directly for ranking or mapped to an action threshold. When a clean development set is available, a threshold can be selected to meet a target false-positive rate. Without clean labels, we use conservative quantiles of the unlabeled screening pool and route high-score samples to review. This is the setting reported as 0% clean calibration.

Query complexity. Let M be the number of GPC stressor views, A the number of MI views, R the number of MinDrop paths, and S the number of slope strengths. The detector uses

$$O(M + A + R + S) \tag{29}$$

black-box VLM calls per input, with a small constant in our experiments. The calls are independent and can be parallelized. Early stopping is possible by monitoring the running variance of the score:

$$\widehat{\mathrm{Var}}_m[\mathrm{Score}(z)] < \eta, \tag{30}$$

where m is the number of sampled views and η is a deployment-chosen tolerance. This gives a direct cost–recall trade-off for rate-limited APIs.

Failure modes. GPC-E4 can miss attacks that do not create a stable response basin, attacks that activate only under highly specific prompts outside the stressor family, or attacks optimized to preserve output diversity under the detector probes. It can also produce false positives on inputs whose clean semantics are highly ambiguous, because such inputs may have unstable semantic clustering. These cases motivate richer stressor policies and task-aware semantic agreement models.

C Visual and Textual Stressor Families and Parameter Ranges

We summarize the stressor augmentations used in the multimodal E4 implementation for computing per-sample metrics (H, MI, MinDrop, Slope). These stressors are designed to be approximately distribution-preserving: changes in model outputs primarily reflect abnormal invariance or uncertainty contraction induced by backdoor triggers, rather than semantic destruction of the original input.

Design Principles. The stressors follow three key principles:

1. **Semantic Preservation:** perturbations retain the primary object, scene, or text meaning;
2. **Multi-view Coverage:** stressors span photometric, spatial, frequency-domain, stochastic, and linguistic dimensions;
3. **Controlled Magnitude:** perturbation strengths are mild enough to avoid spurious instability on clean inputs while exposing trigger-induced invariances.

Table 10: Visual stressor augmentations and parameter ranges used in multimodal E4.

Stressor	Parameter range	Purpose
Affine	rot $\pm 5^\circ$; trans $\pm 2\%$; scale 0.98–1.02; shear $\pm 2^\circ$	Tests geometric robustness
Color jitter	B/C/S 0.9–1.1; hue ± 0.02	Tests photometric variation
Blur	$k = 3$, $\sigma = 0.5$–0.6	Simulates low-frequency distortion
Noise	$\sigma = 0.02$–0.03	Probes pixel sensitivity
Brightness	factor 0.9	Tests intensity changes
Noise slope	$s_0 = 0$, $s_1 = 0.03$	Estimates entropy sensitivity

Visual Stressor Families. Table 10 maps the visual stressors implemented in the code to their parameter ranges and intended role in the E4 metrics.

Textual Stressor Families. Table 11 summarizes textual stressors used in E4, implemented via `mild_text_views` and `synonym_replace`.

Notes. All visual and textual parameter ranges are fixed across experiments. Excessively large perturbations may reduce metric reliability, while very small perturbations weaken sensitivity to subtle backdoor triggers. The combination of visual and textual stressors ensures that H, MI, MinDrop, and Slope metrics capture abnormal contraction or invariance in the model output distribution, rather than destruction of image or text semantics.

D Prompt Templates and Closed-set Construction

Prompt templates. For open-ended VLM evaluation, we adopt a small ensemble of semantically equivalent prompt templates to reduce prompt-specific variance and to probe instruction sensitivity. All prompts are designed to be content-neutral and task-consistent. Representative templates include:

- "Describe the image in one sentence."
- "What is happening in this image?"
- "Provide a brief description of the scene."
- "Summarize the main content of the image."

Table 11: Textual stressor augmentations and parameter ranges used in multimodal E4.

Stressor	Parameter / method	Purpose
Synonym	replace up to one word	Adds lexical variation for MI
Token dropout	drop about 5% tokens	Tests missing-context sensitivity
Text views	generate A prompt views	Supports MI and ensemble entropy

For semantic-consistency estimation in $\mathrm{GPC}_{\mathrm{sem}}$, we fix a canonical base prompt and apply light paraphrasing using synonym replacement or minor rewording to generate mild textual stressors. These textual perturbations are used only for mutual-information estimation and do not alter the visual content.

Closed-set construction. For closed-set detection, we construct option sets $\mathcal{Y} = \{y_k\}_{k=1}^{K}$ of size $K = 10$-20 per image. To avoid semantic leakage and label bias, we mine candidate labels using a controlled synonym expansion procedure:

1. Extract a base noun or category from the ground-truth caption or dataset label.
2. Expand it into a synonym pool using WordNet and CLIP text embedding similarity.
3. Filter candidates whose cosine similarity with the base term falls outside $[0.6, 0.9]$.
4. Randomly sample $K - 1$ distractor labels and combine them with the true label.

This procedure ensures that the closed-set task remains non-trivial while preventing trivial elimination of distractors. We verify that the same option sets are used across clean and triggered inputs to avoid introducing distributional bias.

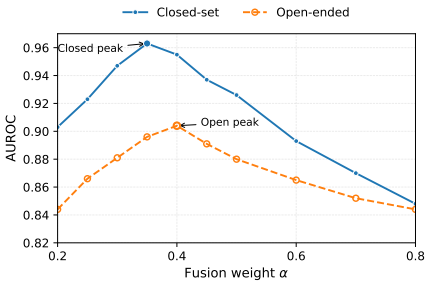


Figure 2: AUROC as a function of fusion weight α for closed-set (peak at $\alpha = 0.35$) and open-ended (peak at $\alpha = 0.4$) detection on COCO with LLaVA-1.5. Performance is stable for $\alpha \in [0.25, 0.45]$.

E Fusion Weight (α) Sensitivity

Our unified detection score combines generalized prediction consistency (GPC or GPC_{sem}) and the entropy quartet (E4) using a fixed fusion weight $\alpha \in [0, 1]$. We select $\alpha \in [0.3, 0.4]$ based on a small clean calibration split (200-500 samples), which consistently balances dominant-mode stability and uncertainty contraction.

Figure 2 reports AUROC as a function of α for both closed-set and open-ended detection on COCO with LLaVA-1.5, visualized over the range $\alpha \in [0.2, 0.8]$. Performance remains stable within the narrower range $\alpha \in [0.25, 0.45]$, with a mild peak around $\alpha \approx 0.35$ (closed-set) and $\alpha \approx 0.4$ (open-ended). When $\alpha \to 0$, the detector relies primarily on uncertainty statistics and loses sensitivity to semantic mode locking. Conversely, when $\alpha \to 1$, the detector relies almost exclusively on GPC, which degrades performance on frequency-coded and content-aware backdoors. Notably, the open-ended AUROC curve declines more gradually than the closed-set curve, indicating that open-ended detection is relatively more robust to changes in α.

These results justify our choice of a fixed fusion weight $\alpha \in [0.3, 0.4]$ throughout all experiments.

F Additional Ablations

Effect of removing stressor families. We evaluate the impact of removing individual stressor families from the probing distribution Π. Table 12 reports AUROC on COCO with LLaVA-1.5.

Removing frequency-domain stressors causes the largest degradation for both closed-set and open-ended detection, confirming that frequency-space perturbations are critical for exposing narrow-band backdoor triggers. Removing spatial or photometric stressors results in moderate performance drops, indicating that geometric and appearance variations contribute complementary but non-dominant probing signals. Stochastic noise removal also degrades performance, suggesting that random perturbations play an important role in stabilizing uncertainty statistics.

Table 12: Effect of removing individual stressor families from the probing distribution Π on COCO with LLaVA-1.5.

Setting	Closed AUROC	Open AUROC
All stressors	0.967	0.904
– Photometric	0.951	0.893
– Spatial	0.946	0.897
– Frequency	**0.912**	**0.853**
– Stochastic	0.954	0.895

Effect of number of stressor views M. We further ablate the number of stressor views M used for Monte Carlo estimation of GPC and E4. As reported in Table 4 (main paper), increasing M consistently improves AUROC but incurs near-linear computational cost. Marginal gains beyond $M = 8$ are relatively small, indicating that improvements primarily arise from variance reduction rather than reliance on extreme perturbations.

Effect of stressor budget. We vary the number of visual stressors sampled per input from $M = 2$ to $M = 16$. Performance saturates at $M \approx 8$ for both closed-set and open-ended detection. Extremely small budgets ($M \leq 2$) lead to unstable MI and MinDrop estimates, while very large budgets ($M \geq 16$) offer diminishing returns.

Effect of removing E4 components. Consistent with the ablation results in the main paper, removing any individual E4 component degrades detection performance. The leave-one-component results show that Slope has the strongest interaction effect in the full fusion, despite its weak standalone separability. Among the single-metric detectors, MI and MinDrop provide stronger standalone discrimination than entropy and Slope.

Overall, these ablations confirm that (i) no single stressor family or statistic is sufficient, and (ii) balanced multi-view probing is necessary for reliable backdoor detection.

Table 13: Summary comparison of backdoor detection methods.

Dimension	**STRIP**	**TeCo**	**BDet**	**E4 (Ours)**
Detection principle	Entropy randomness	Output consistency	Cross-modal similarity	Entropy + MI + MinDrop + Slope
Modality	Visual	Visual	Text/Visual	Multimodal
Model dependency	ResNet, VGG	ResNet, VGG	CLIP	LLaVA / BLIP-2 / MiniGPT-4
Clean samples required	No	No	Optional	No
Inference cost	High	Moderate	Low	Lower
Key detection feature	Entropy STD	Logit STD	Semantic alignment	Combined anomaly score

Table 14: Sensitivity to stressor-family replacement on LLaVA-1.5. Alternative stressors preserve the same mild multi-view design principle.

Attack	Original	Alternative	Δ
BadNet	0.965	0.961	-0.004
Blended	0.951	0.953	+0.002
ISSBA	0.923	0.916	-0.007
Low-Freq.	0.926	0.913	-0.013
TrojanVLM	0.894	0.888	-0.006
Shadowcast	0.912	0.920	+0.008

G Detection Comparison

Detection comparison. We compare our proposed E4 method against existing backdoor detection approaches including STRIP(Gao et al., 2019), TeCo(Liu et al., 2023b), and BDet(Niu et al., 2025). Table 13 summarizes key dimensions: detection principle, modality, model dependency, requirement for clean samples, inference cost, and key detection feature.

E4 unifies multimodal analysis by coupling visual and text perturbations and measuring consistency using Entropy, MI, MinDrop, and Slope. Unlike prior methods, E4 works on large multimodal models (e.g., LLaVA-1.5, BLIP-2, MiniGPT-4) without requiring access to model parameters, gradients, or training data. It achieves linear inference complexity with strong generalization to multimodal tasks such as image classification, VQA, and instruction following.

Table 15: Effect of clean calibration availability. At 0%, robust normalization uses the unlabeled screening pool.

Clean rate	BN	IS	SH	Avg.
0%	0.961	0.920	0.910	0.930
5%	0.964	0.923	0.912	0.933
10%	0.965	0.923	0.912	0.933
50%	0.963	0.922	0.911	0.932

Table 16: Prompt and distribution-shift robustness on LLaVA + IS.

Setting	AUROC	FPR@95
Base	0.923	0.056
Prompt rephrase	0.925	0.049
Short instruction	0.921	0.058
Detailed instruction	0.916	0.062
Style shift	0.918	0.061
Domain shift	0.909	0.075
Noise perturbation	0.912	0.067

H Additional Stress Tests

Interpretation. The alternative-stressor results show that GPC-E4 is sensitive to the multi-view probing principle rather than to one specific list of transforms. The calibration study shows that known-clean calibration is helpful but not required: robust statistics computed on the unlabeled screening pool provide nearly identical AUROC. The prompt and shift study shows that mild language and visual distribution shifts do not remove the contraction signal, although stronger domain shifts increase false positives.