RESTORING BACKDOOR DETECTION UNDER ADAPTIVE ATTACKS IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor detection in multimodal large language models (MLLMs) often relies on predefined statistics of hidden representations or attention. Detector-aware attackers can suppress these statistics during fine-tuning while preserving effective backdoors. We investigate whether detection can be restored after such evasion. In the evaluated settings, defeating an existing detection rule leaves poisoned and benign examples distinguishable under a newly learned score. Motivated by this finding, we propose Benign-Anchored Witness Detection (BAWD), which restores detection by relearning a scoring function on the attacked model. BAWD trains lightweight source classifiers to distinguish audited representations from a trusted same-task benign reference. Given this reference, fitting and thresholding require no additional poisoning labels or exact poisoning rate; thresholding uses a fixed candidate-rank window. Under a matched-reference assumption, the population-optimal source score induces the same ordering as the poisoning posterior, without requiring access to it. Evaluations span three MLLMs and three datasets. On Qwen3-VL-8B fine-tuned on ScienceQA, the linear variant (BAWD-L) achieves 99.8–100.0% detection F1 across seven detector-aware attack configurations, while each of seven baselines degrades substantially under at least one configuration. In two four-round attack sequences targeting previously deployed BAWD detectors, freshly fitted BAWD-L achieves 99.5–99.8% F1 while attack success remains at 99.5–99.8%. These results demonstrate that relearning scores with trusted benign references can restore detection after the evaluated detector-aware attacks.

1 INTRODUCTION

Fine-tuning adapts multimodal large language models (MLLMs) to specialized tasks, but it also creates a security risk when the fine-tuning data or pipeline is untrusted. By poisoning a small subset of training examples, an attacker can associate a hidden trigger with an attacker-chosen response while preserving normal behavior on benign inputs (Gu et al., 2019). In multimodal instruction tuning, the trigger can reside in the image, the text, or both (Liang et al., 2025). A natural defense is to identify the poisoned examples in the fine-tuning set before the model is deployed or further adapted.

A prominent family of poisoned-sample detectors looks for abnormal structure in internal representations. AC, SS, and SPECTRE use clustering, spectral analysis, and robust covariance estimation, respectively (Chen et al., 2018; Tran et al., 2018; Hayase et al., 2021), while TCAP profiles attention allocation across system instructions, visual tokens, and textual queries (Liu et al., 2026). Their scoring families encode particular forms of suspicious structure, even when their parameters are fitted to the audited data. Fig. 1a illustrates detection when such a score f_0 separates poisoned from benign examples.

However, a detector-aware attacker can penalize the monitored statistic during fine-tuning while retaining an effective backdoor (Tan & Shokri, 2020; Qi et al., 2023). This failure, in turn, does not imply that discriminative information has disappeared. Fig. 1b illustrates this distinction: the original score f_0 fails, while another direction can still separate the two classes. This motivates our question: *after an attacker defeats a detection rule, can the defender learn a new rule from the attacked model's representations?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

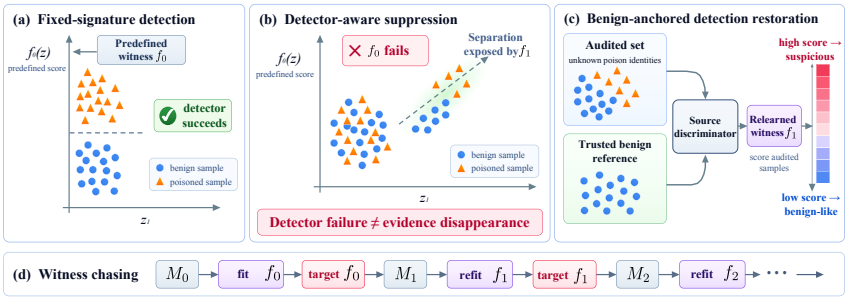


Figure 1: Schematic illustration of BAWD. (a) An initial score f_0 separates benign and poisoned samples. (b) Detector-aware suppression defeats f_0, while separation remains possible under another score f_1. (c) BAWD learns f_1 by contrasting audited representations with a trusted benign reference. (d) Witness chasing alternates attacks targeting the current witness with witness refitting on the updated model.

Our experiments show that poisoned samples remain highly separable in the final hidden state after adaptive training suppresses attention, spectral, or Gram statistics. We study *restoring backdoor detection* by relearning a discriminative scoring function, or *witness*, on the attacked model. We instantiate this approach with Benign-Anchored Witness Detection (BAWD). As shown in Fig. 1c, BAWD learns lightweight source classifiers to distinguish audited representations from a trusted benign reference, then uses the learned score to identify suspicious examples. This construction is closely related to positive–unlabeled (PU) learning and bagged PU classification (Elkan & Noto, 2008; Mordelet & Vert, 2014), with benign examples as the labeled class and the audited set as the unlabeled mixture. Under the matched-reference assumption, we show that the population-optimal source score induces the same ordering as the poisoning posterior, without requiring access to it.

Fig. 1d extends recovery to repeated detector targeting and refitting, which we call *witness chasing*. At each round, the attacker targets the previously fitted witness, and the defender learns a new witness on the updated model. In two four-round chains, refitted BAWD-L achieves 99.5–99.8% F1 while ASR remains at 99.5–99.8% (Table 3).

Our contributions can be summarized as follows:

- We identify a failure mode of fixed-signature poisoned-sample detection. Detector-aware fine-tuning can suppress the statistic that a detector monitors, while the evidence of poisoning remains recoverable from the model's hidden states. We study how relearning the scoring rule restores detection.
- We propose BAWD, which relearns reference-relative scores without trigger, target, or exact poisoning-rate knowledge, using a fixed candidate-rank window for thresholding. We characterize its population ranking under a matched reference and the effects of contamination.
- We design an adaptive evaluation that includes attacks explicitly targeting the deployed detector, and show that BAWD remains effective across multiple MLLMs and datasets in settings where existing detectors fail.

2 Related Work

Representation-based poisoned-sample detection. Many backdoor defenses detect poisoned samples from abnormal internal representations. AC identifies suspicious feature clusters (Chen et al., 2018), while SS and SPECTRE exploit spectral structure and robust covariance estimation (Tran et al., 2018; Hayase et al., 2021). Beatrix uses higher-order Gram features (Ma et al., 2023), GradSentry uses gradient spectral entropy (Zhao et al., 2026), and TCAP profiles component-wise attention in MLLMs (Liu et al., 2026). Despite using different statistics, these methods rely on prescribed representation features or anomaly rules. We instead investigate whether poisoned samples remain distinguishable in the model's hidden representations after an attacker suppresses these predefined statistics, and whether a new scoring function learned with a trusted benign reference can recover that distinction.

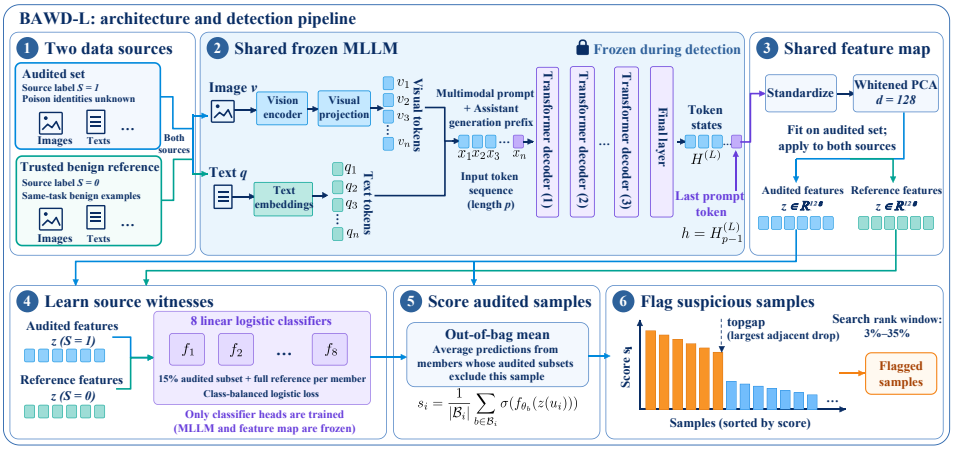


Figure 2: Architecture and detection pipeline of the linear instantiation of BAWD.

Adaptive and multimodal backdoors. Multimodal attacks such as VL-Trojan combine visual and textual triggers to induce malicious behavior in instruction-tuned models (Liang et al., 2025). However, trigger stealth does not necessarily imply representation-level detector evasion. More directly, Qi et al. (2023) show that adaptive poisoning can bypass latent-separation defenses by explicitly shaping poisoned representations. We study how to restore detection after the attacker suppresses a fixed or previously learned scoring rule.

Reference-based detection and PU learning. Several defenses use trusted clean data rather than unsupervised geometry alone. SCAn contrasts class-conditional representation statistics with a small clean set (Tang et al., 2021), ASSET actively enlarges clean–poison differences (Pan et al., 2023), and RFTC compares sample responses with reference-model outputs and then clusters suspicious samples (Chen et al., 2025). Separating an unlabeled mixture from a trusted sample is the setting of classifier two-sample tests, density-ratio estimation, and PU learning (Lopez-Paz & Oquab, 2017; Bickel et al., 2009; Elkan & Noto, 2008; Kiryo et al., 2017), and the estimator used by BAWD is closely related to bagged PU classification (Mordelet & Vert, 2014). These connections do not by themselves establish robustness to detector-aware adaptation, in which an attacker reshapes the poisoned representations to evade detection. We investigate whether refitting a reference-relative witness restores detection after such adaptation.

3 Benign-Anchored Witness Detection

Fig. 2 summarizes the BAWD pipeline, with the linear instantiation of BAWD shown as the default instantiation. Both the audited examples and the trusted benign reference pass through the same frozen MLLM to extract prompt representations. A shared preprocessing map transforms these representations, and an ensemble of lightweight source classifiers learns a reference-relative witness. Sample scores use out-of-bag (OOB) aggregation for BAWD-L and all-member aggregation for BAWD-M, followed by thresholding.

3.1 Problem Setup

Audited data and benign reference. Let M be an MLLM fine-tuned on an audited dataset $\mathcal{D}_{\mathrm{tr}} = \{u_i\}_{i=1}^{n}$, where $u_i = (v_i, q_i, a_i)$ consists of an image, an instruction, and a training response. Each example has an unknown poisoning indicator $A_i \in \{0, 1\}$. The detector assigns a score s_i and predicts $\widehat{A}_i = \mathbf{1}\{s_i \geq \tau\}$. The defender can run inference with M, access its hidden states, and obtain a trusted benign reference $\mathcal{D}_{\mathrm{ref}} = \{\widetilde{u}_j\}_{j=1}^{m}$. No poison labels, trigger, or attack target are available. Because $\mathcal{D}_{\mathrm{ref}}$ supplies benign supervision, the setting is benign-anchored rather than fully unsupervised.

Representation mixture. For a fixed model and representation extractor, we model the audited representation distribution as

$$P_{\mathrm{tr}} = (1 - \pi)P_{\mathrm{c}} + \pi P_{\mathrm{p}}, \qquad 0 < \pi < 1, \tag{1}$$

where P_{c} and P_{p} are the benign and poisoned components and π is the poisoning rate. Under matched benign representation distributions ($P_{\mathrm{ref}} = P_{\mathrm{c}}$), the audited–reference contrast isolates the contribution of the poisoned component.

Adaptive attacker. We consider a training-pipeline attacker who controls poisoned training data and the LoRA objective, accesses intermediate representations, and seeks detector evasion while preserving backdoor success and benign utility. In witness chasing, the attacker also observes the deployed classifier and preprocessing parameters. The defender keeps its reference samples and refitting randomness private. Reference contamination is evaluated separately.

3.2 Restoring Detection through Refitting

Let $\mathcal{R}(M; \mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{ref}})$ denote a procedure that fits a sample-level witness from the hidden representations of model M. Under detector-aware adaptation, we consider the attack–refitting sequence

$$M_t \xrightarrow{\mathcal{R}} f_t \xrightarrow{\text{fresh adapter targets } f_t} M_{t+1} \xrightarrow{\mathcal{R}} f_{t+1}. \tag{2}$$

Each round trains a fresh adapter from the common initialization to suppress the preceding frozen witness f_t while preserving the backdoor. The defender then applies $\mathcal{R}$ to representations from the frozen model M_{t+1} to fit f_{t+1}. Across rounds, the attacker receives the preceding detector rather than its adapter. Each recovery step uses the current model, audited set, and reference. This sequential targeting protocol evaluates recovery after attacks on a fixed detector, with refitting performed after attacker optimization (Section 4.4).

Whether recovery can succeed at all is constrained by the attack itself. Let $T \in \{0, 1\}$ denote trigger status, let $Z_\star$ be a representation on which the target behavior depends, and let $\mu_t = P(Z_\star \in \cdot \mid T = t)$. Let the Attack Success Rate (ASR) and the Benign Target Rate (BTR) denote the probabilities that the model produces the attack target on triggered and untriggered inputs, respectively.

Proposition 1 (Behavioral success lower-bounds internal separation)**.** *If the target event is conditionally independent of T given $Z_\star$, then $\mathrm{TV}(\mu_1, \mu_0) \geq |\,\mathrm{ASR} - \mathrm{BTR}\,|$, and some witness $g : \mathcal{Z}_\star \to [0, 1]$ attains $|\mathbb{E}_{\mu_1} g - \mathbb{E}_{\mu_0} g| \geq |\,\mathrm{ASR} - \mathrm{BTR}\,|$.*

Under the stated assumption, a nonzero $\mathrm{ASR} - \mathrm{BTR}$ gap requires trigger-related separation in $Z_\star$, motivating the search for a discriminative witness. The proposition concerns triggered versus untriggered inputs in a representation sufficient for the target event. BAWD instead evaluates poisoned versus benign training examples using last-prompt-token states, Principal Component Analysis (PCA), and finite classifiers; separation in these features is assessed empirically. Connecting the two requires additional distributional and sufficiency assumptions. A failed backdoor may still leave trigger-related differences, so we report behavior and detection separately. Appendix B provides the proof and scope.

3.3 Internal Representations

For each audited example $u = (v, q, a)$, we build the multimodal prompt from v and q, append the assistant generation prefix, and run the frozen model M. We use the final-layer hidden state at the last prompt token, $h_M(u) = H^{(L)}_{p-1} \in \mathbb{R}^D$, where p is the prompt length. This state provides a compact pre-generation representation; the extractor does not receive the training response a. We standardize each coordinate and apply whitened PCA,

$$z(u) = \mathrm{PCA}_{\mathrm{white},d}(\mathrm{Standardize}(h_M(u))), \tag{3}$$

fitted on the audited representations without labels and applied unchanged to the reference, with $d = 128$. PCA serves only as compression and conditioning; it is not a prescribed backdoor subspace, and the learned witness may combine all retained coordinates.

3.4 Learning a Reference-Relative Witness

We assign source label $S = 1$ to audited examples and $S = 0$ to reference examples. The detector therefore uses source labels derived from the audited set and the trusted benign reference, which supplies benign supervision; it does not use the poisoning labels of the audited examples for classifier fitting or threshold selection. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}$ denote the classifier's logit function. Given fitting subsets $\mathcal{U} \subseteq \mathcal{D}_{\mathrm{tr}}$ and $\mathcal{V} \subseteq \mathcal{D}_{\mathrm{ref}}$, we minimize the class-balanced logistic loss

$$\widehat{\mathcal{L}}(\theta) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \mathrm{softplus}(-f_\theta(z(u))) + \frac{1}{|\mathcal{V}|} \sum_{\widetilde{u} \in \mathcal{V}} \mathrm{softplus}(f_\theta(z(\widetilde{u}))) + \frac{\lambda}{2} \|\theta\|_2^2. \tag{4}$$

Proposition 2 (Population ranking under a matched reference)**.** *Assume Eq. (1), $P_{\mathrm{ref}} = P_{\mathrm{c}}$, and densities with respect to a common dominating measure. For $\lambda = 0$, the unrestricted population-optimal logit, allowing extended-real values, satisfies*

$$f^*(z) = \log \frac{p_{\mathrm{tr}}(z)}{p_{\mathrm{c}}(z)} = \log\left((1 - \pi) + \pi \, \frac{p_{\mathrm{p}}(z)}{p_{\mathrm{c}}(z)}\right), \tag{5}$$

where $p_c > 0$, with $f^* = +\infty$ on poison-only support. It induces the poison-posterior ordering P_{tr}-almost surely.

The identity is classical in PU learning and density-ratio estimation (Elkan & Noto, 2008; Bickel et al., 2009); we state it because it specifies what BAWD estimates. Source discrimination ranks samples by poison likelihood without poison labels or knowledge of π, but the statement concerns an unrestricted classifier and yields a source score rather than a calibrated poison probability, so the threshold is chosen separately (Section 3.5). The proof is given in Appendix A.

3.5 Ensemble Scoring and Sample Decisions

We fit $B = 8$ classifiers on $(\mathcal{U}_b, \mathcal{V}_b)$ and, for nonempty $\mathcal{B}_i$, aggregate their scores,

$$s_i = \frac{1}{|\mathcal{B}_i|} \sum_{b \in \mathcal{B}_i} g(f_{\theta_b}(z(u_i))), \qquad i = 1, \dots, n. \tag{6}$$

BAWD-L, our default, uses linear logistic classifiers, each fitted on 15% of audited examples against the full reference. Here g is sigmoid and $\mathcal{B}_i = \{b : u_i \notin \mathcal{U}_b\}$ selects members whose audited fitting subsets exclude u_i; an empty $\mathcal{B}_i$ gives score zero. Internal reference rows also participate as reference examples (Appendix H.3, reference-excluded evaluation). **BAWD-M** uses two-hidden-layer multilayer perceptrons (MLPs), each fitted on half of each source, identity g, and all-member aggregation; its OOB variant appears in Appendix H.2. Implementation details are in Appendix D.

To convert scores into decisions without poison labels, let $s_{(1)} \geq \cdots \geq s_{(n)}$ be the sorted scores. The `topgap` rule selects the largest adjacent score drop within a rank window,

$$k^* = \arg \max_{\lfloor r_{lo} n \rfloor < k \leq \lfloor r_{hi} n \rfloor} \big(s_{(k)} - s_{(k+1)}\big), \qquad \widehat{\mathcal{P}} = \{u_i \in \mathcal{D}_{tr} : s_i \geq s_{(k^*)}\}, \tag{7}$$

with the fixed default $(r_{lo}, r_{hi}) = (0.03, 0.35)$. This candidate-rank window supplies a range prior rather than the exact poisoning rate. The rule always returns a nonempty set and cannot certify a clean dataset. We report the Area Under the Receiver Operating Characteristic curve (AUROC) separately from thresholded detection; Appendix F.4 evaluates alternative windows.

3.6 Reference Quality

BAWD relies on the reference to define which differences are suspicious. If the reference is contaminated by the same poison component, $P_{ref} = (1 - \epsilon)P_c + \epsilon P_p$, then for any bounded witness g,

$$\mathbb{E}_{P_{tr}} g - \mathbb{E}_{P_{ref}} g = (\pi - \epsilon)\big(\mathbb{E}_{P_p} g - \mathbb{E}_{P_c} g\big). \tag{8}$$

Under this same-component contamination model, contrast shrinks by $1 - \epsilon/\pi$ and population ordering is preserved for $0 \leq \epsilon < \pi$. Appendix C derives a $1/(\pi - \epsilon)$ dependence for a generic contrast-maximization estimator, illustrating how diminishing contrast affects estimation. A benign reference drawn from a different distribution than the audited benign data violates $P_{ref} = P_c$ and can introduce spurious source differences. Section 4.6 quantifies both effects.

4 Experiments

4.1 Experimental Setup

Models, data, and poisoning. Our primary setting is Qwen3-VL-8B-Instruct (Bai et al., 2025) fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2022) on ScienceQA (Lu et al., 2022) (6,218 training and 2,017 held-out examples). We additionally evaluate LLaVA-NeXT-8B (Li et al., 2024a) and InternVL3-8B (Zhu et al., 2025) on ScienceQA, and Qwen3-VL-8B-Instruct on PhD (Liu et al., 2025) and SEED-Bench (Li et al., 2024b). Unless stated otherwise, 10% of the training examples (622 on ScienceQA) are poisoned by replacing their answers with the target `Backdoor Attack!`. LoRA adapters of rank 8 are applied to the attention and feed-forward modules of the language model. The vision encoder remains frozen. We train for three epochs using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-4} and an effective batch size of 8 (Appendix D). Detectors then audit the fine-tuning set with the adapted model frozen.

Attacks. Non-adaptive attacks comprise BadNets-style patches (Gu et al., 2019), Blended triggers (Chen et al., 2017), our input-conditioned frequency-domain trigger, and VL-Trojan dual-key triggers (Liang et al., 2025). Detector-aware attacks add suppression penalties to the fine-tuning loss. Attention–Spectral Suppression targets system-component attention statistics and leading

spectral directions; Joint Suppression broadens both penalties; Moment-Matched and Moment–Gram Suppression additionally match per-head attention moments and Gram statistics; and Online-Witness Suppression penalizes a discriminator trained during attacker optimization. We also evaluate Moment-Matched Suppression with the target `B` and an adaptive VL-Trojan variant, giving seven detector-aware configurations.

Baselines. We compare AC (Chen et al., 2018), SS (Tran et al., 2018), SPECTRE (Hayase et al., 2021), and Beatrix (Ma et al., 2023), which use clustering, spectral, robust-covariance, and Gram statistics; BYE (Rong et al., 2025) and TCAP (Liu et al., 2026), which use attention statistics; and Multi-View Backdoor Profiling (MVBP), a multi-view detector that fuses an attention view with two spectral views (Appendix E.5). MVBP* reports the best F1 among 16 predefined MVBP configurations, selected with ground-truth labels, and is therefore an optimistic within-pool estimate rather than a deployable detector. SS, SPECTRE, and Beatrix flag approximately $1.5\pi n$ examples using the true poisoning rate π. Representation baselines are class-conditioned for target `B` and pooled otherwise.

BAWD and references. Both variants use $B = 8$ members and $m = 1{,}000$ reference examples. In the default setting, the reference is a uniformly sampled subset of the benign rows of the audited set, which emulates a defender who has verified a small portion of the data; poison labels are used only to construct this reference. We also evaluate disjoint same-task references and examine sensitivity to reference size, contamination, and cross-task distribution shift.

Metrics. We report poisoned-sample precision, recall, F1, and AUROC, with poisoned examples as the positive class. ASR/BTR measure target-prefix rates on triggered/untriggered held-out inputs; CP measures clean accuracy. Rates are percentages; differences are percentage points. BAWD uses `topgap` (Section 3.5). BYE abstains when it identifies no sensitive layer and thus produces no score.

4.2 Main Detector Comparison

Table 1: Detection on Qwen3-VL/ScienceQA. Detector columns report F1 (%; best in bold). MVBP*: label-selected best of 16 configurations; ABST.: no detection score.

	Behavior		Representation-based				MLLM-specific			BAWD	
Attack	**ASR**	**CP**	**AC**	**SS**	**SPECTRE**	**Beatrix**	**BYE**	**TCAP**	**MVBP***	**L**	**M**
Non-adaptive attacks											
BadNet	99.8	98.8	**99.8**	79.7	79.7	79.9	37.8	96.9	**99.8**	98.5	99.6
Blended	95.7	98.4	0.0	68.3	78.2	74.7	14.9	98.9	**99.5**	99.4	98.1
Input-Cond. Freq.	94.7	98.4	0.5	77.2	78.5	78.3	14.6	97.4	97.9	**98.1**	97.9
VL-Trojan	99.0	98.9	99.9	80.0	80.0	79.7	16.8	95.2	**100.0**	99.5	98.1
Detector-aware attacks											
Attention–Spectral	99.7	98.0	0.0	0.1	37.3	17.0	ABST.	96.8	**99.8**	**99.8**	97.8
Joint Suppression	98.4	96.3	0.1	0.0	32.5	0.4	14.3	2.0	98.2	**99.8**	96.5
Moment-Matched	99.1	98.2	0.0	0.0	39.7	3.3	11.7	16.1	36.8	**99.8**	93.8
Moment–Gram	99.8	98.2	0.0	30.9	0.0	39.0	ABST.	14.5	**99.8**	**99.8**	95.7
Online-Witness	99.7	97.9	0.1	0.0	1.0	6.0	14.5	27.7	96.6	**99.8**	95.4
Moment-Matched, target `B`	99.7	97.9	28.3	39.9	51.8	7.1	0.0	16.1	17.7	**99.8**	97.0
Adaptive VL-Trojan	100.0	97.9	0.0	0.0	1.5	11.1	ABST.	49.3	14.6	**100.0**	97.1

Table 1 compares the detectors. Under non-adaptive attacks, SS, SPECTRE, and Beatrix approach their budget-imposed F1 ceiling in several settings; AC succeeds on BadNet and VL-Trojan but fails on Blended and the frequency trigger, and TCAP and MVBP* reach 95.2–98.9% and 97.9–100.0%. BAWD-L and BAWD-M reach 98.1–99.5% and 97.9–99.6%, which is competitive with the strongest baselines but not uniformly the best.

Detector-aware attacks change this picture. Individual baselines survive particular attacks, but no baseline remains effective across the full evaluated suite. TCAP retains 96.8% F1 under Attention–Spectral Suppression but drops to 2.0% under Joint Suppression; SPECTRE keeps partial signal on three attacks yet falls to 0.0–1.5% on Moment–Gram, Online-Witness, and Adaptive VL-Trojan; and even the label-selected MVBP* falls to 36.8% and 14.6% on Moment-Matched and Adaptive VL-Trojan. Under Moment-Matched Suppression with the valid-answer target `B`, SS and SPECTRE reach 39.9% and 51.8% F1, respectively, while MVBP* reaches 17.7% (Appendix F.3). BAWD-L attains 99.8–100.0% F1 on all seven configurations and BAWD-M attains 93.8–97.8%, while the attacked models retain 98.4–100.0% ASR and 96.3–98.2% CP. BAWD-L also achieves 99.7–99.8% F1 across six valid-answer-target configurations (Appendix F.3). BAWD-M loses F1 mainly through

precision (88.9–96.7% under detector-aware attacks) rather than recall, and its AUROC rounds to 1.00 in every detector-aware row (Appendix F.1).

The failures follow the suppression objectives. Attention–Spectral Suppression penalizes only system-component attention mass and entropy, so TCAP, which profiles attention across system, visual, and query components, still reaches 96.8% F1; Joint Suppression extends the penalty to all components and to attention concentration, and TCAP falls to 2.0%. Per-head moment matching adds a constraint that even the label-selected multi-view MVBP* substantially degrades: it survives Joint Suppression with 98.2% F1 but falls to 36.8% under Moment-Matched Suppression. Each added penalty defeats the detector family whose statistic it constrains, whereas BAWD, which refits its witness on the attacked model, remains effective across the evaluated suppression objectives. BAWD-L also attains 100.0% precision and 99.5–100.0% recall under every detector-aware configuration (Appendix F.1): There are no false positives, and at most three of the 622 poisoned examples are missed.

4.3 Cross-Model and Cross-Dataset Generalization

Table 2 evaluates BAWD across architectures and datasets, using BadNet as a non-adaptive reference alongside Attention–Spectral and Moment-Matched Suppression. On ScienceQA, the three attacks achieve 99.1–100.0% ASR across Qwen3-VL-8B-Instruct, LLaVA-NeXT-8B, and InternVL3-8B. BAWD-L achieves 98.2–99.9% F1, while BAWD-M shows greater variation across architectures, reaching its lowest F1 of 90.9% on InternVL3 under Moment-Matched Suppression. With the same Moment-Matched suppression configuration, LLaVA-NeXT and InternVL3 both attain 100.0% ASR, while BAWD-L achieves 99.9% and 99.8% F1, respectively.

Table 2: Cross-model and cross-dataset results: ASR / BAWD-L F1 / BAWD-M F1 (%). A+S: Attention–Spectral; Moment: Moment-Matched. Attack and detector configurations are held fixed across settings.

Cross-model on ScienceQA				**Cross-dataset on Qwen3-VL-8B-Instruct**			
Attack	**Qwen3-VL**	**LLaVA-NeXT**	**InternVL3**	**Attack**	**ScienceQA**	**PhD**	**SEED-Bench**
BadNet	99.8/98.5/99.6	100.0/99.0/98.9	99.8/98.2/99.0	BadNet	99.8/98.5/99.6	100.0/98.9/99.1	99.9/100.0/97.4
A+S	99.7/99.8/97.8	99.9/99.8/98.3	100.0/99.9/93.0	A+S	99.7/99.8/97.8	100.0/100.0/98.6	100.0/99.9/99.3
Moment	99.1/99.8/93.8	100.0/99.9/98.0	100.0/99.8/90.9	Moment	99.1/99.8/93.8	99.2/100.0/99.9	99.9/99.9/99.1

Fixing Qwen3-VL and varying the dataset yields a similar pattern. Across ScienceQA, PhD, and SEED-Bench, ASR remains at least 99.1%, with BAWD-L achieving 98.5–100.0% F1 and BAWD-M achieving 93.8–99.9%. Using the same detector-aware attack configurations across datasets, BAWD-L reaches 100.0% F1 on both PhD settings and 99.9% on both SEED-Bench settings.

The effect of suppression on fixed signatures varies across settings. Under Moment-Matched Suppression on SEED-Bench, SS retains 80.0% F1 and 0.999 AUROC, whereas its F1 is zero in the other evaluated Moment-Matched settings (Appendix F.5). BAWD-L remains effective both when this spectral signature is suppressed and when it remains detectable, achieving at least 98.2% F1 across all evaluated settings.

4.4 Restoring Detection under Successive Attacks

We run two four-round chains in which the attacker explicitly suppresses the deployed BAWD-L or BAWD-M detector. Round r_0 uses the base suppression recipe shared by all rounds, which combines attention preservation, spectral preservation, and per-head attention-moment matching, without a witness penalty. At round r_t, a fresh model is trained with the base recipe plus a loss that drives the outputs of the frozen detector fitted on M_{t-1} (including its standardizer, PCA map, and all ensemble members) toward the reference label on poisoned inputs. We then evaluate the frozen detector on M_t and refit both variants from scratch (Table 3; protocol in Appendix E.3).

The attacks reduce the effectiveness of the deployed detectors. The frozen BAWD-M detector obtains 0% F1 in every round, and the frozen BAWD-L detector falls to 0.0–73.7% F1. Refitting nevertheless restores detection: fresh BAWD-L attains 99.5–99.8% F1 in all eight adaptive rounds, fresh BAWD-M attains 94.7–99.0%, and the attacked models keep 99.5–99.8% ASR and 97.8–98.4% CP. Recovery therefore does not come from a weakened backdoor.

Two qualifications are important when interpreting these results. First, the frozen detectors fail for more than thresholding reasons: even when an oracle chooses the sign of the frozen score using labels, its AUROC reaches only 0.798–0.996 on the BAWD-L chain and 0.778–0.982 on the BAWD-

Table 3: Successive targeting of BAWD-L and BAWD-M. "Frozen prev." applies the detector fitted on M_{t-1}; "Fresh" refits on M_t. All metrics except AUROC are percentages.

	Behavior		Frozen prev.		Fresh BAWD-L		Fresh BAWD-M	
Round	**ASR**	**CP**	**P/R/F1**	**AUROC**	**P/R/F1**	**AUROC**	**P/R/F1**	**AUROC**
r_0	99.0	97.2	—	—	100.0/99.5/99.8	0.9974	98.7/99.7/99.2	0.9991
Attacker targets BAWD-L								
r_1	99.7	98.4	100.0/58.4/73.7	0.9962	100.0/99.5/99.8	0.9971	97.6/99.4/98.5	0.9988
r_2	99.8	98.4	0.0/0.0/0.0	0.2020	99.4/99.7/99.5	0.9976	95.3/98.9/97.1	0.9981
r_3	99.8	98.1	49.6/18.3/26.8	0.9685	100.0/99.7/99.8	0.9978	94.0/99.0/96.5	0.9978
r_4	99.5	98.0	11.7/3.7/5.6	0.8869	100.0/99.5/99.8	0.9980	90.1/99.7/94.7	0.9980
Attacker targets BAWD-M								
r_1	99.7	97.8	0.0/0.0/0.0	0.2225	100.0/99.7/99.8	0.9981	91.9/99.7/95.6	0.9987
r_2	99.6	98.0	0.0/0.0/0.0	0.0178	100.0/99.7/99.8	0.9977	98.1/99.7/98.9	0.9988
r_3	99.6	98.1	0.0/0.0/0.0	0.0465	100.0/99.5/99.8	0.9976	98.4/99.5/99.0	0.9994
r_4	99.8	98.4	0.0/0.0/0.0	0.2079	100.0/99.7/99.8	0.9983	97.6/98.9/98.2	0.9978

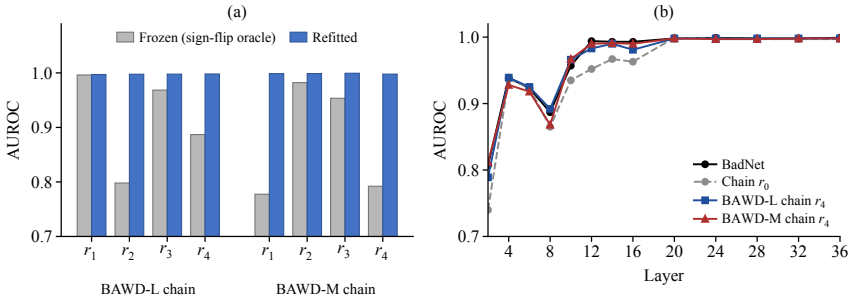


Figure 3: Analysis of the attack chains. (a) AUROC of the frozen previous detector, even with its sign chosen by an oracle, compared with the detector refitted on the attacked model in each round. (b) AUROC of a freshly fitted linear probe at each layer; layer 36 is the final layer, which BAWD uses and the attacker targets.

M chain, below the refitted detector in every round (Fig. 3a; Appendix G.1). Fresh linear probes also attain AUROC of at least 0.994 at every layer from layer 20 to the final layer in all rounds of both chains (Fig. 3b; Appendix G.2), so the attacks defeat the deployed witness while preserving high linear separability at all evaluated layers from layer 20 onward. Second, the attacker optimizes against a fixed detector, not through the refitting procedure. The closest approximations we tested are an online discriminator trained during fine-tuning and a penalty that matches all 128 retained PCA directions to benign statistics (Appendix G). Across online-discriminator weights, BAWD-L F1 stays at 98.5–99.8% whenever ASR is retained, while some tested weights and the subspace penalty yield zero measured ASR. We did not obtain a configuration that both evades BAWD and keeps the backdoor, which suggests a trade-off between evasion and attack efficacy.

4.5 Ablation Study

Table 4: BAWD ablations: detection F1 (%).

Variant	B	PCA	BadNet	A+S	Mom.	Onl.	Avg.	Δ
BAWD-L	1	✓	89.9	91.9	91.6	91.1	91.1	−8.4
BAWD-L	8	✓	98.5	**99.8**	**99.8**	**99.8**	**99.5**	0.0
BAWD-L	8	×	82.7	95.0	93.6	98.2	92.4	−7.1
BAWD-M	8	✓	**99.6**	97.8	93.8	95.4	96.7	−2.8

Table 5: Post-detection retraining.

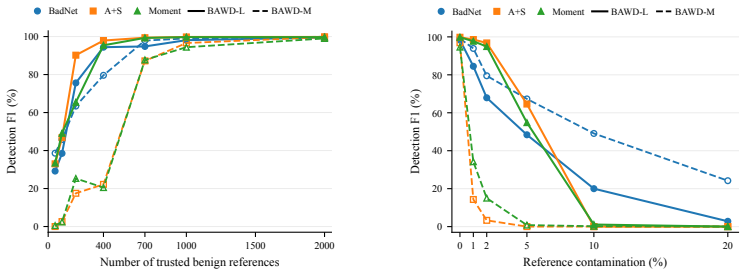
Attack	ASR	CP	Removed	Left
BadNet	99.55→0.00	97.72/98.12	639/630	3/10
A+S	99.65→0.00	97.97/97.77	644/620	3/2
Moment	99.11→0.00	98.31/97.82	695/619	4/3
Online	99.65→0.00	98.02/97.17	673/620	4/2

Table 4 varies ensemble size B and PCA across BadNet, A+S, Moment-Matched (Mom.), and Online-Witness (Onl.); Avg. averages F1, and Δ compares it with BAWD-L using $B = 8$ and PCA. The $B = 1$ row loses 8.4 points, reflecting out-of-bag coverage: 15% of rows in a single member's fitting subset receive no score, and on the rows that do, a single member matches eight members (Appendix F). Removing PCA at $B = 8$ lowers mean F1 to 92.4%, mostly on BadNet (98.5% to 82.7%); at $B = 25$ the drop is smaller (99.6% to 95.9%), consistent with PCA acting mainly as conditioning for a 4,096-dimensional linear problem. With identical features, BAWD-L exceeds BAWD-M on all three detector-aware attacks, so the added capacity of the MLP does not help in the evaluated feature space.

Table 5 reports pre/post-retraining ASR and post-retraining CP (%); pairs denote BAWD-M/BAWD-L. Removed counts flagged examples; Left counts retained poisons (of 622). Removing flagged data

and retraining with plain LoRA reduces ASR to 0.0% for both variants across all four attacks, with CP at 97.2–98.3%. At most 10 poisoned examples remain in the retained data, and at most 77 benign examples (about 1.4% of the benign data) are removed (Appendix I).

4.6 Reference Requirements



(a) Reference size (b) Reference contamination

Figure 4: Sensitivity of BAWD-L and BAWD-M to (a) benign-reference size and (b) random contamination of a 1,000-example reference with poisoned examples, on BadNet, Attention–Spectral, and Moment-Matched Suppression.

Size and contamination. With 400 reference examples, BAWD-L reaches 94.4%, 97.9%, and 95.4% F1 on BadNet, Attention–Spectral, and Moment-Matched Suppression, compared with 79.6%, 22.2%, and 20.5% for BAWD-M (Fig. 4a). Reference contamination substantially degrades detection (Fig. 4b). At 1–2% contamination, BAWD-L retains 94.9–98.6% F1 on the two detector-aware attacks, but falls to 84.5% and 67.9% on BadNet; BAWD-M falls to 3.3–34.2% on the detector-aware attacks. At 5%, BAWD-L drops to 48.4–64.6%. Recovery therefore deteriorates before the population contrast vanishes at $\epsilon = \pi$, consistent with increased estimation difficulty.

Reference source and clean models. A reference of 1,000 held-out ScienceQA examples never used for fine-tuning, matched in size to the internal reference, yields 99.0–99.8% BAWD-L F1 on four checkpoints, compared with 97.8–99.8% for the internal reference (Appendix H.3), so recovery does not require drawing the reference from the audited set. Cross-task references substantially reduce detection performance, highlighting the importance of distributional alignment (Appendix H.4). On a model fine-tuned without poisoning, AUROC against trigger-bearing rows stays near chance (0.53–0.59), so BAWD does not separate trigger-bearing inputs when no backdoor has been learned.

5 Conclusion

We studied poisoned-sample detection in MLLMs under detector-aware attacks that control poisoned training data and the fine-tuning objective. In the evaluated settings, attackers suppress attention, spectral, and Gram statistics while preserving effective backdoors, yet the attacked representations still support accurate poisoned-sample detection with a relearned score. This distinction motivates restoring detection by relearning the scoring function on each audited model. BAWD implements this approach by fitting source classifiers against a trusted benign reference, without trigger, target, or exact poisoning-rate knowledge, using a fixed candidate-rank window for thresholding. BAWD-L remained effective across the evaluated attacks, architectures, and datasets, including repeated targeting of previously deployed detectors followed by refitting. Reference experiments highlight the importance of distributional alignment and low contamination. These findings support evaluating both deployed-detector failure and subsequent recovery, and show that refitting a reference-based score can restore detection after detector-aware suppression.

LLM Usage Statement

Large language model tools assisted with manuscript drafting and revision, LaTeX preparation, literature search, and checking mathematical exposition. AI coding assistance was also used to develop and debug experimental scripts. The authors take responsibility for the methodology, implementation, results, citations, and conclusions presented in this paper.

Ethics Statement

This work aims to improve the detection of poisoned training examples in MLLMs. The detector-aware attacks are used to evaluate defensive robustness, but their designs may also inform attempts to evade detection. Our evaluation uses controlled poisoning of benchmark datasets. In practical use, detection scores should guide further inspection: reference-distribution mismatch can produce false positives, and a flagged example alone does not establish malicious intent. Data removal and subsequent retraining should therefore be evaluated for their effects on both security and benign-task performance.

Reproducibility Statement

Section 3.1 specifies the threat model and the information available to the defender, and Section 3 describes the detection procedure. Appendix D documents data splits, model adaptation, feature extraction, and detector configuration; Appendix E details attack objectives and baseline implementations. Appendices F, G, and H provide expanded metrics, ablations, and controls, and Appendices A–C contain the assumptions and proofs.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10(75):2137–2155, 2009.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Jinwen Chen, Hainan Zhang, Fei Sun, Qinnan Zhang, Sijia Wen, Ziwei Wang, and Zhiming Zheng. Detecting stealthy backdoor samples based on intra-class distance for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3348–3365, 2025.

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

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 213–220, 2008.

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

Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. SPECTRE: Defending against backdoor attacks using robust statistics. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4129–4139, 18–24 Jul 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ryuichi Kiryo, Gang Niu, Marthinus C. du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild, May 2024a. URL `https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/`.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024b. doi: 10.1109/CVPR52733.2024.01263.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. VL-Trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 133(7):3994–4013, 2025. doi: 10.1007/s11263-025-02368-9.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. PhD: A ChatGPT-prompted visual hallucination evaluation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19857–19866, 2025. doi: 10.1109/CVPR52734.2025.01849.

Mingzu Liu, Hao Fang, and Runmin Cong. TCAP: Tri-component attention profiling for unsupervised backdoor detection in MLLM fine-tuning. In *Forty-third International Conference on Machine Learning*, 2026.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022.

Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The "Beatrix" resurrections: Robust backdoor detection via Gram matrices. In *30th Annual Network and Distributed System Security Symposium, NDSS*, 2023.

Fantine Mordelet and Jean-Philippe Vert. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

Minzhou Pan, Yi Zeng, Lingjuan Lyu, Xue Lin, and Ruoxi Jia. ASSET: Robust backdoor data detection across a multiplicity of deep learning paradigms. In *32nd USENIX Security Symposium*, pp. 2725–2742. USENIX Association, 2023.

Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Revisiting the assumption of latent separability for backdoor defenses. In *The Eleventh International Conference on Learning Representations*, 2023.

Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. Backdoor cleaning without external guidance in MLLM fine-tuning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Te Juin Lester Tan and Reza Shokri. Bypassing backdoor detection algorithms in deep learning. In *IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 175–183, 2020. doi: 10.1109/EuroSP48549.2020.00019.

Di Tang, XiaoFeng Wang, Haixu Tang, and Kehuan Zhang. Demon in the variant: Statistical analysis of DNNs for robust backdoor contamination detection. In *30th USENIX Security Symposium*, pp. 1541–1558, August 2021.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Haodong Zhao, Tianyi Xu, Tianhang Zhao, Zhuosheng Zhang, and Gongshen Liu. GradSentry: Gradient spectral entropy for backdoor sample filtering in large language model fine-tuning. *arXiv preprint arXiv:2605.26574*, 2026.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

APPENDIX CONTENTS

A Proof of Population Poison Ranking . . . 13
B Internal Witness Existence and Its Scope . . . 13
C Reference Quality Analysis . . . 14
C.1 Same-component contamination . . . 14
C.2 Finite-sample contrast . . . 15
C.3 Benign distribution mismatch . . . 15
D Experimental Protocol and Implementation Details . . . 15
D.1 Audited data and representation extraction . . . 15
D.2 Shared preprocessing and source classifiers . . . 16
D.3 Sample decisions and metrics . . . 16
E Attack and Baseline Configurations . . . 17
E.1 Trigger constructions . . . 17
E.2 Detector-aware objectives . . . 17
E.3 Successive targeting of deployed witnesses . . . 17
E.4 Representation-based baselines . . . 18
E.5 Attention and multi-view baselines . . . 18
F Extended Detection Results and Ablations . . . 18
F.1 Detailed BAWD results . . . 18
F.2 Detailed baseline results . . . 19
F.3 Payload robustness with a valid task answer . . . 20
F.4 Detailed design-choice ablation . . . 21
F.5 Cross-model and cross-dataset evaluation . . . 22
G Additional Adaptive Attack Stress Tests . . . 25
G.1 Sign-flip oracle for frozen detectors . . . 25
G.2 Layer-wise linear probes . . . 25
G.3 Online-discriminator suppression strength . . . 25
G.4 Trigger diversity and retained-subspace suppression . . . 25
H Reference and Decision-Rule Controls . . . 26
H.1 Reference-size and random-contamination curves . . . 26
H.2 Disjoint references and OOB scoring . . . 26
H.3 Reference-draw variability, size-matched disjoint reference, and reference exclusion . . . 27
H.4 Benign-reference distribution shift . . . 28
I Efficiency and Post-Detection Validation . . . 28
I.1 Detector-side computation . . . 28
I.2 Removing detected examples and retraining . . . 29

A Proof of Population Poison Ranking

We provide the proof of Proposition 2. The result concerns fixed representation distributions and an unrestricted population source classifier. It justifies the ranking induced by benign-anchored source discrimination, but does not imply exact optimization or calibration of the finite-capacity classifiers that BAWD fits. The result also assumes an unrestricted classifier: the linear family used by BAWD-L contains f^* only when the log-density ratio is linear in z, and in general its score preserves the population ordering only to the extent that the ratio is monotone in that score. Whether the fitted witnesses rank poisoned samples correctly is therefore an empirical question, which Section 4 addresses.

Proof of Proposition 2. With equal source weights and no regularization, the population logistic risk for $D(z) \in [0, 1]$, using $0 \log 0 = 0$, is

$$\mathcal{L}(D) = \int \left[-p_{\mathrm{tr}}(z) \log D(z) - p_{\mathrm{c}}(z) \log(1 - D(z))\right] d\mu(z). \tag{9}$$

Minimizing the integrand pointwise gives

$$D^*(z) = \frac{p_{\mathrm{tr}}(z)}{p_{\mathrm{tr}}(z) + p_{\mathrm{c}}(z)}, \qquad \operatorname{logit} D^*(z) = \log \frac{p_{\mathrm{tr}}(z)}{p_{\mathrm{c}}(z)}. \tag{10}$$

Using the mixture model in Eq. (1),

$$p_{\mathrm{tr}}(z) = (1 - \pi) p_{\mathrm{c}}(z) + \pi p_{\mathrm{p}}(z), \tag{11}$$

and therefore

$$f^*(z) = \log\left((1 - \pi) + \pi \frac{p_{\mathrm{p}}(z)}{p_{\mathrm{c}}(z)}\right), \tag{12}$$

which is Eq. (5).

Let

$$r(z) = \frac{p_{\mathrm{p}}(z)}{p_{\mathrm{c}}(z)}. \tag{13}$$

For $0 < \pi < 1$, $f^*(z) = \log((1 - \pi) + \pi r(z))$ is strictly increasing in $r(z)$. The poison posterior under the same mixture model is

$$\Pr(A = 1 \mid Z = z) = \frac{\pi p_{\mathrm{p}}(z)}{(1 - \pi) p_{\mathrm{c}}(z) + \pi p_{\mathrm{p}}(z)} = \frac{\pi r(z)}{(1 - \pi) + \pi r(z)}, \tag{14}$$

which is also strictly increasing in $r(z)$. Thus the orderings agree on $p_{\mathrm{c}} > 0$. On poison-only support, $D^* = 1$, $f^* = +\infty$, and the poison posterior equals one, completing the result P_{tr}-almost surely. □

The proposition concerns ranking rather than calibration. In particular, the sigmoid $D^*(z)$ estimates source membership, not poisoning probability. Moreover, if the source-classification prior is α rather than $1/2$, the optimal logit becomes

$$f^*_\alpha(z) = \log \frac{p_{\mathrm{tr}}(z)}{p_{\mathrm{c}}(z)} + \log \frac{\alpha}{1 - \alpha}, \tag{15}$$

which differs only by a constant and therefore preserves the population ordering. This is why BAWD uses the discriminator score for ranking while choosing the final detection threshold separately.

B Internal Witness Existence and Its Scope

We prove Proposition 1 for the behavioral evaluation distribution; applying it to training-set poison detection requires additional distributional alignment.

Let $T \in \{0, 1\}$ denote trigger status and $C \sim P_C$ benign content, with the same P_C for both conditions. Write $x_T(C)$ for the untriggered ($T = 0$) or triggered ($T = 1$) input. Let $O \in \{0, 1\}$ indicate whether the model produces the attack target, and define

$$\mathrm{ASR} = \Pr(O = 1 \mid T = 1), \qquad \mathrm{BTR} = \Pr(O = 1 \mid T = 0). \tag{16}$$

We restate the result for completeness.

Proposition 1 (restated). *Let $\mu_t = P(Z_\star \in \cdot \mid T = t)$ for a fixed model and content distribution. Suppose the target event depends on trigger status only through $Z_\star$, i.e., $O \perp T \mid Z_\star$. Then*

$$\mathrm{TV}(\mu_1, \mu_0) \geq |\,\mathrm{ASR} - \mathrm{BTR}\,|, \tag{17}$$

and there exists a measurable witness $g : \mathcal{Z}_\star \to [0, 1]$ such that

$$|\mathbb{E}_{\mu_1} g - \mathbb{E}_{\mu_0} g| \geq |\,\mathrm{ASR} - \mathrm{BTR}\,|. \tag{18}$$

Proof. Under the conditional-independence assumption, there exists a measurable kernel $\kappa : \mathcal{Z}_\star \to [0, 1]$ such that

$$\Pr(O = 1 \mid Z_\star = z, T = t) = \kappa(z) \tag{19}$$

for both $t \in \{0, 1\}$. Hence

$$\mathrm{ASR} - \mathrm{BTR} = \int \kappa(z) d\mu_1(z) - \int \kappa(z) d\mu_0(z). \tag{20}$$

By the variational characterization of total variation,

$$\mathrm{TV}(\mu_1, \mu_0) = \sup_{0 \leq g \leq 1} |\mathbb{E}_{\mu_1} g - \mathbb{E}_{\mu_0} g|\,, \tag{21}$$

so choosing $g = \kappa$ proves the result. □

Scope. Proposition 1 is an existence statement, not a guarantee for the representation used by BAWD. In particular, compressing $Z_\star$ to $z = \phi(Z_\star)$ can only reduce total variation:

$$\mathrm{TV}(P(z \mid T = 1), P(z \mid T = 0)) \leq \mathrm{TV}(P(Z_\star \mid T = 1), P(Z_\star \mid T = 0)). \tag{22}$$

The proposition therefore does not certify the last-prompt-token hidden state, the PCA representation, or the finite-capacity witnesses used by BAWD. Detection with these fitted witnesses is evaluated empirically.

The implication is also one-directional. A weak or failed backdoor can still leave trigger-related representation differences, so internal separability alone does not certify that the trigger–target behavior remains active. This motivates reporting attack success and poisoned-sample detection separately.

C Reference Quality Analysis

The main analysis assumes that the trusted reference matches the benign component, $P_{\mathrm{ref}} = P_{\mathrm{c}}$. We next examine two deviations from this ideal: contamination and benign distribution mismatch.

C.1 Same-component contamination

Suppose the reference contains a fraction ϵ of examples from the same poison component:

$$P_{\mathrm{ref}} = (1 - \epsilon) P_{\mathrm{c}} + \epsilon P_{\mathrm{p}}. \tag{23}$$

For any bounded witness g,

$$\mathbb{E}_{P_{\mathrm{tr}}} g - \mathbb{E}_{P_{\mathrm{ref}}} g = \left[(1 - \pi)\mathbb{E}_{P_{\mathrm{c}}} g + \pi \mathbb{E}_{P_{\mathrm{p}}} g\right] - \left[(1 - \epsilon)\mathbb{E}_{P_{\mathrm{c}}} g + \epsilon \mathbb{E}_{P_{\mathrm{p}}} g\right] \tag{24}$$

$$= (\pi - \epsilon)\left(\mathbb{E}_{P_{\mathrm{p}}} g - \mathbb{E}_{P_{\mathrm{c}}} g\right), \tag{25}$$

recovering Eq. (8).

The corresponding source density ratio is

$$R_\epsilon(r) = \frac{(1 - \pi) + \pi r}{(1 - \epsilon) + \epsilon r}, \qquad r = \frac{p_{\mathrm{p}}}{p_{\mathrm{c}}}, \tag{26}$$

with derivative

$$R'_\epsilon(r) = \frac{\pi - \epsilon}{((1 - \epsilon) + \epsilon r)^2}. \tag{27}$$

Hence the population source ordering remains aligned with the poisoning posterior when $\epsilon < \pi$, vanishes when $\epsilon = \pi$, and reverses when $\epsilon > \pi$. This conclusion applies specifically to contamination by the same poison component; arbitrary contaminants need not follow this behavior.

C.2 FINITE-SAMPLE CONTRAST

The preceding result concerns population ordering. Contamination can make finite-sample recovery difficult well before the ordering disappears, because the observable audited–reference contrast is only $(\pi - \epsilon)$ times the contrast between the two components. Throughout this subsection we assume $0 \le \epsilon < \pi$, so that this factor is positive; at $\epsilon = \pi$ the contrast vanishes and the bound below is vacuous.

Let $\mathcal{G} \subseteq \{g : \mathcal{Z} \to [0,1]\}$ be a fixed witness class and define

$$\Delta(g) = \mathbb{E}_{P_{\mathrm{p}}} g - \mathbb{E}_{P_{\mathrm{c}}} g, \qquad J(g) = \mathbb{E}_{P_{\mathrm{tr}}} g - \mathbb{E}_{P_{\mathrm{ref}}} g. \tag{28}$$

Under Eq. (23),

$$J(g) = (\pi - \epsilon)\Delta(g). \tag{29}$$

Suppose an empirical objective $\widehat{J}$ satisfies

$$\sup_{g \in \mathcal{G}} |\widehat{J}(g) - J(g)| \le \eta \tag{30}$$

and the learned witness $\widehat{g}$ approximately maximizes it:

$$\widehat{J}(\widehat{g}) \ge \sup_{g \in \mathcal{G}} \widehat{J}(g) - \xi. \tag{31}$$

Then

$$\Delta(\widehat{g}) \ge \sup_{g \in \mathcal{G}} \Delta(g) - \frac{2\eta + \xi}{\pi - \epsilon}. \tag{32}$$

Thus, as ϵ approaches π, estimation and optimization error are amplified even though the population ordering remains valid for $\epsilon < \pi$.

This bound describes a generic bounded source-contrast maximizer; it is not a finite-sample generalization theorem for the regularized classifiers trained with Eq. (4).

C.3 BENIGN DISTRIBUTION MISMATCH

Even an uncontaminated reference can violate $P_{\mathrm{ref}} = P_{\mathrm{c}}$ if its benign content distribution differs from that of the audited data. In that case, a source discriminator may exploit legitimate differences in image content, question type, or other task factors rather than poisoning-related evidence. Our theory does not remove this confounding effect. The disjoint-reference and reference-size experiments evaluate matched same-task references, while broader domain-shift robustness remains outside the current guarantee.

D EXPERIMENTAL PROTOCOL AND IMPLEMENTATION DETAILS

D.1 AUDITED DATA AND REPRESENTATION EXTRACTION

The primary experiments use Qwen3-VL-8B-Instruct (Bai et al., 2025) adapted to ScienceQA (Lu et al., 2022). The audited fine-tuning set contains 6,218 examples, including 622 poisoned examples in the default setting. Poisoning replaces the selected training answers with the attack target; the rate is relative to the full training set. The held-out test set contains 2,017 examples. The benign and triggered evaluations use the same held-out example identities, with triggers applied to every example in the triggered version. The default attack target is `Backdoor Attack!`; the ordinary-answer-target experiments use `B`. For controlled experiments, we construct the trusted-clean anchor by sampling 1,000 ground-truth benign rows from the audited set. Poison labels are used only to construct this experimental anchor and are never provided to the detector. The disjoint-reference experiment uses the 2,017-example held-out ScienceQA pool, whose IDs are identical to those used for CP evaluation and its triggered ASR counterpart. ID checks confirm zero overlap with the 6,218 audited training rows. Thus, this reference is disjoint from the audited set but transductive with respect to behavioral evaluation. Appendix H.2 reports a split-half control with non-overlapping reference and CP-evaluation subsets. The internal reference is sampled with random seed zero, and the same draw is supplied to both BAWD variants. Additional evaluations cover LLaVA-NeXT-8B and InternVL3-8B on ScienceQA, and Qwen3-VL-8B-Instruct on PhD and SEED-Bench. Their protocols and detailed results are provided in Appendix F.5.

Victim-model training. LoRA (Hu et al., 2022) uses rank 8, scaling parameter 32, dropout 0.05, and no trainable biases. Adapters cover attention and feed-forward modules; the vision encoder is frozen. Training uses AdamW at a fixed learning rate of 10^{-4} for three epochs, with one example per forward pass, eight-step gradient accumulation, and gradient checkpointing. The effective batch size is eight. The primary checkpoints use training seed 42; separately identified seed-43/44 checkpoints provide the training-repetition experiments.

We merge the LoRA adapter into the victim model and freeze its parameters for detection. The extraction prompt contains the image and question, followed by an assistant generation prefix. Its system instruction is "You are a helpful, generic assistant. Answer with the correct option's letter directly." The feature vector is the final-layer hidden state at the last prompt token, before an answer is appended. In the primary model, this vector has 4,096 coordinates. Extraction uses bfloat16 model inference; cached hidden vectors are converted to float32 for detector fitting.

D.2 Shared preprocessing and source classifiers

We standardize each hidden coordinate and fit a 128-component whitened PCA transformation on the audited representations. Reference examples are transformed with the same fitted standardizer and PCA map. Both BAWD-L and BAWD-M receive this same feature matrix.

BAWD-L fits class-balanced logistic classifiers. For each member, a random subset containing 15% of the audited examples supplies source label one, and the benign reference supplies source label zero. The logistic solver uses a maximum of 800 iterations. Each member predicts source-one probabilities for audited examples outside its audited fitting subset. The final score averages these OOB probabilities for each example. The main comparison uses eight members. If an example has no OOB member, its score is zero. For independent 15% audited-subset draws, this event has probability approximately 0.15^B for an ensemble of size B.

BAWD-M uses two width-128 hidden layers with Rectified Linear Unit (ReLU) activations and a scalar logit output. For each of eight members, we draw half of the audited examples and, at the default reference size, half of the reference examples without replacement. Each optimization step draws 256 examples with replacement from each source and minimizes

$$\mathcal{L}(f) = \frac{1}{256}\sum_{i=1}^{256} \mathrm{softplus}(-f(z_i^{\mathrm{audit}})) + \frac{1}{256}\sum_{j=1}^{256} \mathrm{softplus}(f(z_j^{\mathrm{ref}})). \tag{33}$$

Training uses Adam with a learning rate of 10^{-3}, a weight decay of 10^{-3}, and 500 optimization steps. Standard BAWD-M averages the logits of all members. Its OOB variant instead averages the logits from members whose audited fitting subsets exclude the example being scored. OOB scoring changes the member-level prediction rule; the shared preprocessing remains fitted on the audited set.

D.3 Sample decisions and metrics

Let $s_{(1)} \geq \cdots \geq s_{(n)}$ be the sorted scores. The `topgap` rule selects

$$\mathcal{K} = \{\lfloor 0.03n \rfloor + 1, \ldots, \lfloor 0.35n \rfloor\}, \tag{34}$$

$$k^* = \min \arg\max_{k \in \mathcal{K}} (s_{(k)} - s_{(k+1)}), \tag{35}$$

$$\widehat{\mathcal{P}} = \{i : s_i \geq s_{(k^*)}\}. \tag{36}$$

The first maximizing rank resolves gap ties, and score ties at the threshold are included. This rule uses a fixed candidate-rank window rather than the exact poisoning rate and always returns a nonempty candidate set. The default window is $[3\%, 35\%]$; Table 16 compares it with $[1\%, 20\%]$ and $[1\%, 50\%]$ without selecting a window separately for each checkpoint.

Detection precision, recall, and F1 use poisoned examples as the positive class. AUROC measures ranking quality before thresholding. On ScienceQA, CP is held-out answer accuracy, computed by matching the first extracted option letter in A–H to the gold answer. ASR is the fraction of triggered held-out predictions that begin with the target string after whitespace stripping and lowercasing. BTR applies the same target-prefix criterion to untriggered held-out predictions from the same checkpoint. We report ASR–BTR as the triggered–untriggered target-rate gap, without replacing

the original ASR. For Qwen3-VL/ScienceQA, ΔCP is the checkpoint's CP minus the unpoisoned control's CP of 98.61% (1,989/2,017). The control uses the plain fine-tuning recipe with zero poisoning; it is not a separate loss-matched control for each suppression objective. For target `B`, we additionally evaluate ASR on examples whose original gold answer is not `B`. Generation is greedy with at most eight new tokens; the TCAP activation-extraction pass uses at most one new token. Rates other than AUROC are percentages; ASR–BTR and ΔCP are percentage-point differences. A dash denotes an unreported metric; ABST. denotes BYE runs without an identified sensitive layer and therefore without a detection score. Results are from a single victim-training run per checkpoint unless repetitions are explicitly stated; ensemble members are detector components.

E Attack and Baseline Configurations

E.1 Trigger constructions

The BadNets-style attack (Gu et al., 2019) inserts a checkerboard patch in the lower-right corner of the image, with a side length equal to one tenth of the shorter image dimension. Blended (Chen et al., 2017) overlays a fixed pattern image with blending coefficient 0.15. The input-conditioned frequency trigger uses a learned U-Net generator, followed by a radial band-pass mask in the Discrete Cosine Transform (DCT) domain (16–90 at resolution 224), inverse DCT, and magnitude rescaling. The main-table instance uses a perturbation bound of $8/255$. Its higher-magnitude variant uses $16/255$ and 15% poisoning; we treat these as configurations of the same construction. The VL-Trojan-style attack (Liang et al., 2025) uses a dual-key trigger comprising a 16×16 structured visual patch and a textual suffix. Its adaptive variant uses a 10% both-key rate and a 5% counterfactual rate for single-key examples during training.

E.2 Detector-aware objectives

Detector-aware fine-tuning augments the backdoor training objective with penalties on attention and hidden representations. Attention statistics include mass (m), entropy (H), and concentration (k). Attention-Spectral Suppression targets system-component m, H and three leading spectral directions with spectral coefficient 0.1. Joint Suppression uses all-component m, H, k, six spectral directions, and spectral weight 1.5. Moment-Matched Suppression uses attention weight 10, entropy and concentration weights 10, six spectral directions with spectral weight 0.1, and moment coefficient 20 for matching windowed per-head means and standard deviations. Moment-Gram adds a Gram-discrepancy penalty with coefficient 15. Online-Witness uses a discriminator trained during attacker optimization; the main setting uses moment coefficient 10 and discriminator weight 1.6. The target-`B` and adaptive dual-key settings vary the attack target and trigger construction, respectively.

E.3 Successive targeting of deployed witnesses

We run separate four-round chains targeting BAWD-L and BAWD-M. Both start from the same round-zero model. Every round, including round zero, trains a fresh instance under one base recipe: all-component attention preservation with attention, entropy, and concentration weights 10, spectral preservation over six directions with weight 0.1, and windowed per-head attention-moment matching with weight 10. Rounds r_1–r_4 add the detector exported from the preceding round with suppression weight 5; round zero adds no witness penalty. No round continues optimizing the preceding round's adapter. The detector representation is the standardized, whitened PCA-128 space for both variants, with eight source classifiers. The exported standardizer, PCA components, whitening scales, and classifier parameters remain fixed throughout the attack. For the linear ensemble, the attack uses binary cross-entropy toward source zero on the mean sigmoid probability. For the MLP ensemble, it uses binary cross-entropy with logits toward source zero on the mean logit. Attack-time scores aggregate all frozen members. After adaptation, the previous scoring function is evaluated on M_t and `topgap` is recomputed on the resulting scores. Both BAWD variants are then freshly fitted on M_t. The previous-detector and fresh-detector columns therefore distinguish suppression of the deployed witness from detection after refitting.

E.4 Representation-based baselines

AC (Chen et al., 2018) uses the fast Independent Component Analysis (FastICA) algorithm with ten components followed by two-means clustering and identifies the minority cluster as suspicious. SS (Tran et al., 2018) scores squared projections onto the leading right singular vector of centered representations and selects a budget of $1.5\pi n$ examples, using the poisoning-rate prior π. The SPECTRE (Hayase et al., 2021) implementation reduces to 100 PCA dimensions, estimates robust covariance using a Mahalanobis-quantile trimming step, whitens the features, and computes the full-spectrum QUantum Entropy (QUE) score with $\alpha = 4$. The trimming step approximates the original covariance-filtering procedure. Beatrix (Ma et al., 2023) forms the raw Gram statistics and applies the signed pth-root transform for orders one through eight. It uses four projections of dimension 32, bounds equal to the median plus or minus ten times the Median Absolute Deviation (MAD), and a summed deviation score. For the ordinary target `B`, these baselines operate within the corresponding answer class; the other main-table settings use pooled representations.

E.5 Attention and multi-view baselines

BYE (Rong et al., 2025) selects layers and scores samples using attention entropy. TCAP (Liu et al., 2026) profiles the three attention components and combines the resulting statistics for sample detection. MVBP combines an attention view with two spectral views. We write V1 for the attention view, V2a for the PCA/Gaussian-mixture view, and V2b for the Spectral Signatures view. Its candidate pool contains eight fusion configurations: naive, pruned, and anchor-V2b fusion, each with either V1+V2b or V1+V2a+V2b, plus two union configurations. Each is evaluated under two feature subsets (m, H or m, H, k). MVBP* selects the highest detection F1 among these 16 candidates separately for each attack. BYE uses Gaussian-mixture layer selection and a low-entropy cluster decision; TCAP uses the authors' implementation. Class-conditioned representation baselines use training answers for grouping. BAWD uses the trusted reference for source fitting, while MVBP* uses evaluation labels only for the reported post-hoc configuration choice.

F Extended Detection Results and Ablations

F.1 Detailed BAWD results

Table 6 expands the BAWD results in the main comparison with precision, recall, AUROC, and AUPRC. Under non-adaptive attacks, BAWD-L achieves 99.7–100.0% precision and 96.6–99.0% recall. BAWD-M achieves higher recall on BadNet, yielding a higher F1 of 99.6% versus 98.5%, while BAWD-L obtains higher F1 on the other three non-adaptive attacks.

Table 6: BAWD results: precision/recall/F1 (%) | AUROC | AUPRC, from the same scoring run per detector. Target-`B`: Moment-Matched Suppression with target `B`.

Attack	BAWD-L	BAWD-M
Non-adaptive attacks		
BadNet	100.0/97.1/98.5 | 0.9991 | 0.9970	99.8/99.4/99.6 | 0.9999 | 0.9991
Blended	99.8/99.0/99.4 | 0.9990 | 0.9968	97.8/98.4/98.1 | 0.9988 | 0.9934
Input-Cond. Freq.	99.7/96.6/98.1 | 0.9913 | 0.9853	98.7/97.1/97.9 | 0.9921 | 0.9712
VL-Trojan	100.0/99.0/99.5 | 0.9999 | 0.9992	97.6/98.6/98.1 | 0.9994 | 0.9909
Detector-aware attacks		
Attention–Spectral	100.0/99.7/99.8 | 0.9980 | 0.9973	96.1/99.5/97.8 | 0.9979 | 0.9613
Joint Suppression	100.0/99.7/99.8 | 0.9976 | 0.9972	93.5/99.7/96.5 | 0.9975 | 0.9753
Moment-Matched	100.0/99.5/99.8 | 0.9983 | 0.9978	88.9/99.4/93.8 | 0.9962 | 0.9338
Moment–Gram	100.0/99.7/99.8 | 0.9978 | 0.9973	92.3/99.5/95.7 | 0.9963 | 0.9180
Online-Witness	100.0/99.7/99.8 | 0.9984 | 0.9975	91.8/99.4/95.4 | 0.9969 | 0.9570
Target-`B`	100.0/99.7/99.8 | 0.9980 | 0.9974	94.4/99.7/97.0 | 0.9980 | 0.9830
Adaptive VL-Trojan	100.0/100.0/100.0 | 1.0000 | 1.0000	96.7/97.6/97.1 | 0.9985 | 0.9673

Under detector-aware attacks, BAWD-L achieves 100.0% precision and 99.5–100.0% recall. BAWD-M also retains high recall (97.6–99.7%), but its precision varies more substantially. For example, Moment-Matched and Online-Witness Suppression reduce BAWD-M precision to 88.9% and 91.8%, respectively, while recall remains 99.4% in both cases. Its lower F1 in these settings therefore primarily reflects additional false positives rather than missed poisoned examples. The high AUROC values further show that strong overall ranking can coexist with lower precision at

the selected operating point. AUROC is, however, close to saturation for both variants, with means of 0.9979 for BAWD-L and 0.9974 for BAWD-M across the eleven settings. AUPRC separates them more clearly, with means of 0.9966 and 0.9682. The three largest gaps occur on Moment–Gram (0.9973 against 0.9180), Moment-Matched (0.9978 against 0.9338), and Attention–Spectral Suppression (0.9973 against 0.9613), the same settings in which BAWD-M loses precision.

F.2 Detailed baseline results

Table 7: Representation baselines: P/R/F1 (%) | AUROC. AC uses minority-cluster decisions for F1 and majority-centroid distances for AUROC. AUROC covers the full audited set, including for Target-B.

Attack	AC	SS	SPECTRE	Beatrix
Non-adaptive attacks				
BadNet	100.0/99.7/99.8 | 0.915	66.5/99.7/79.7 | 0.997	66.5/99.7/79.7 | 0.998	66.6/99.8/79.9 | 1.000
Blended	0.0/0.0/0.0 | 0.893	56.9/85.4/68.3 | 0.947	65.2/97.7/78.2 | 0.992	62.3/93.4/74.7 | 0.980
Input-Cond. Freq.	0.4/0.6/0.5 | 0.879	64.3/96.5/77.2 | 0.988	65.4/98.1/78.5 | 0.992	65.3/97.9/78.3 | 0.991
VL-Trojan	100.0/99.8/99.9 | 0.937	66.7/100.0/80.0 | 1.000	66.7/100.0/80.0 | 1.000	66.5/99.7/79.7 | 0.997
Detector-aware attacks				
Attention–Spectral	0.0/0.0/0.0 | 0.503	0.1/0.2/0.1 | 0.279	31.1/46.6/37.3 | 0.886	14.1/21.2/17.0 | 0.836
Joint Suppression	0.0/0.2/0.1 | 0.580	0.0/0.0/0.0 | 0.299	27.1/40.7/32.5 | 0.828	0.3/0.5/0.4 | 0.705
Moment-Matched	0.0/0.0/0.0 | 0.628	0.0/0.0/0.0 | 0.180	33.1/49.7/39.7 | 0.882	2.8/4.2/3.3 | 0.649
Moment–Gram	0.0/0.0/0.0 | 0.661	25.7/38.6/30.9 | 0.861	0.0/0.0/0.0 | 0.448	32.5/48.7/39.0 | 0.890
Online-Witness	0.1/0.2/0.1 | 0.502	0.0/0.0/0.0 | 0.165	0.9/1.3/1.0 | 0.697	5.0/7.6/6.0 | 0.822
Target-B	23.1/36.3/28.3 | 0.777	33.2/49.8/39.9 | 0.848	43.2/64.8/51.8 | 0.890	5.9/8.8/7.1 | 0.747
Adaptive VL-Trojan	0.0/0.0/0.0 | 0.382	0.0/0.0/0.0 | 0.304	1.3/1.9/1.5 | 0.652	9.2/13.8/11.1 | 0.732

Table 7 reports detailed results for AC, SS, SPECTRE, and Beatrix. On the non-adaptive attacks, SS and SPECTRE achieve 85.4–100.0% and 97.7–100.0% recall, respectively, but precision remains 56.9–66.7% and 65.2–66.7%. Their fixed removal budget forces false positives even with perfect ranking; Table 8 isolates this decision-rule effect. AC behaves differently across attacks: its minority-cluster decision detects BadNet and VL-Trojan effectively but fails on Blended and the input-conditioned frequency trigger.

Detector-aware attacks affect both ranking and the selected operating point. Under Moment-Matched Suppression, SS has zero recall and 0.180 AUROC, whereas SPECTRE retains 49.7% recall and 0.882 AUROC. Moment–Gram changes this pattern: SPECTRE falls to zero F1, while SS and Beatrix retain 30.9% and 39.0% F1. Under the target-B payload, SS and SPECTRE recover partially, to 49.8% and 64.8% recall, although the suppression recipe is unchanged. These differences show that suppressing one representation statistic does not uniformly suppress the others.

Table 8: Ranking and threshold diagnostics: AUROC | budget-F1 / topgap-F1 / oracle-best-F1 (%), using fixed scores. Target-B is evaluated within answer class B; other rows use the full audited set.

Attack	SS	SPECTRE	Beatrix
Non-adaptive attacks			
BadNet	0.9974 | 79.7/99.8/99.8	0.9984 | 79.7/99.8/99.8	0.9998 | 79.9/99.5/99.6
Blended	0.9466 | 68.3/72.9/79.7	0.9918 | 78.2/84.1/93.0	0.9800 | 74.7/84.5/85.3
Input-Cond. Freq.	0.9883 | 77.2/95.9/96.7	0.9923 | 78.5/96.5/98.0	0.9906 | 78.3/55.1/96.4
VL-Trojan	1.0000 | 80.0/99.9/100.0	1.0000 | 80.0/99.9/100.0	0.9970 | 79.7/47.5/93.1
Detector-aware attacks			
Attention–Spectral	0.2792 | 0.1/0.0/22.9	0.8861 | 37.3/32.8/56.5	0.8362 | 17.0/0.0/49.9
Joint Suppression	0.2985 | 0.0/0.0/23.6	0.8281 | 32.5/1.9/43.1	0.7045 | 0.4/0.0/35.8
Moment-Matched	0.1798 | 0.0/0.0/18.2	0.8820 | 39.7/7.9/51.6	0.6493 | 3.3/0.3/29.4
Moment–Gram	0.8611 | 30.9/0.2/50.3	0.4476 | 0.0/0.0/22.7	0.8900 | 39.0/0.2/60.7
Online-Witness	0.1650 | 0.0/0.0/19.6	0.6970 | 1.0/0.0/34.5	0.8222 | 6.0/0.0/50.8
Target-B	0.5649 | 39.9/0.0/41.8	0.6863 | 51.8/5.7/52.2	0.2770 | 7.1/0.6/39.0
Adaptive VL-Trojan	0.3041 | 0.0/0.0/18.2	0.6517 | 1.5/0.0/28.5	0.7322 | 11.1/0.2/31.8

Table 8 compares the $1.5\pi n$ removal budget, label-free topgap within the 3%–35% rank window, and label-selected oracle thresholds without score reversal. The budget caps F1 near 80%; oracle results are diagnostic. Table 7 reports full-audit AUROC for Target-B. On BadNet and VL-Trojan, SS and SPECTRE achieve 0.9974–1.0000 AUROC, while topgap raises F1 from 79.7–80.0% to 99.8–99.9%. Strong ranking does not ensure reliable threshold selection: Beatrix on Moment–Gram achieves 0.8900 AUROC but only 0.2% topgap-F1, versus 60.7% oracle-best-F1. Across the seven

detector-aware settings, oracle-best-F1 remains at 18.2–50.3% for SS, 22.7–56.5% for SPECTRE, and 29.4–60.7% for Beatrix under the original score orientation, showing that the fixed budget alone does not explain their performance loss.

Table 9 expands the attention and multi-view baseline results. TCAP achieves high precision and recall on the non-adaptive attacks and Attention–Spectral Suppression, but both decline under Joint Suppression. Under Adaptive VL-Trojan, TCAP retains 100.00% recall but only 32.72% precision, illustrating a different failure mode dominated by false positives.

MVBP* also exhibits attack-dependent behavior. Under Moment-Matched Suppression, it retains 97.91% recall but achieves only 22.63% precision, whereas both precision and recall decline under Adaptive VL-Trojan. BYE abstains on three detector-aware attacks and produces no detections under the target-`B` payload. Together, these detailed metrics distinguish among low recall, excessive false positives, and abstention, which F1 alone cannot distinguish.

Table 9: Attention and multi-view baselines: P/R/F1 (%) | AUROC. ABST.: no sensitive layer or detection score; MVBP*: label-selected best of 16 configurations.

Attack	BYE	TCAP	MVBP*
Non-adaptive attacks			
BadNet	23.71/92.60/37.76 | 0.879	94.50/99.36/96.87 | 0.9995	100.00/99.68/99.84 | 0.998
Blended	9.63/32.64/14.88 | 0.492	98.87/98.87/98.87 | 0.9982	99.84/99.20/99.52 | 0.996
Input-Cond. Freq.	9.42/32.48/14.61 | 0.481	100.00/95.02/97.44 | 0.9767	100.00/95.82/97.87 | 0.983
VL-Trojan	9.67/64.79/16.83 | 0.471	90.80/100.00/95.18 | 1.0000	100.00/100.00/100.00 | 1.000
Detector-aware attacks			
Attention–Spectral	ABST.	94.08/99.68/96.80 | 0.9986	100.00/99.68/99.84 | 0.998
Joint Suppression	9.20/32.15/14.30 | 0.463	3.15/1.45/1.98 | 0.4857	96.72/99.68/98.18 | 0.997
Moment-Matched	7.79/23.79/11.74 | 0.571	9.89/43.89/16.14 | 0.4346	22.63/97.91/36.76 | 0.749
Moment–Gram	ABST.	9.10/35.53/14.49 | 0.6628	100.00/99.68/99.84 | 0.998
Online-Witness	9.25/33.76/14.52 | 0.466	17.39/67.85/27.69 | 0.7269	94.06/99.36/96.64 | 0.998
Target-`B`	0.00/0.00/0.00	10.20/38.42/16.12	11.11/43.09/17.67 | 0.472
Adaptive VL-Trojan	ABST.	32.72/100.00/49.31 | 0.9605	9.34/32.96/14.55 | 0.430

Table 10 compares the main-table checkpoints with an unpoisoned Qwen3-VL/ScienceQA control. Across these checkpoints, ΔCP ranges from -2.28 to $+0.30$ percentage points, with the largest decrease under Joint Suppression. These are single-checkpoint comparisons, not estimates of variation across training seeds. For the 13 evaluated `Backdoor Attack!`-target checkpoints, BTR computed from saved untriggered predictions ranges from 0.00% to 0.50%.

Table 10: Clean-task performance on 2,017 ScienceQA examples. ΔCP is the percentage-point difference from plain unpoisoned fine-tuning (98.61%; 1,989 correct).

Attack	CP (%)	ΔCP (pp)
Unpoisoned control	98.61	0.00
Non-adaptive attacks		
BadNet	98.81	$+0.20$
Blended	98.41	-0.20
Input-Cond. Freq.	98.36	-0.25
VL-Trojan	98.91	$+0.30$
Detector-aware attacks		
Attention–Spectral	98.02	-0.59
Joint Suppression	96.33	-2.28
Moment-Matched	98.17	-0.44
Moment–Gram	98.22	-0.39
Online-Witness	97.92	-0.69
Target-`B`	97.87	-0.74
Adaptive VL-Trojan	97.92	-0.69

F.3 PAYLOAD ROBUSTNESS WITH A VALID TASK ANSWER

The main experiments use the target string `Backdoor Attack!`, which lies outside the set of valid answers for this task. To test whether detection depends on this particular payload, we retrain six victims with the target `B`, which is itself a valid answer, in one configuration without suppression and five with suppression, keeping the trigger, the identities of the poisoned examples, and every suppression hyperparameter identical to the corresponding `Backdoor Attack!` configuration;

only the target string differs. AC, SS, SPECTRE, and Beatrix are class-conditioned on `B`, which is a legitimate answer for this task. Table 11 reports the results.

Table 11: Detection F1 (%; row best in bold) for six target-`B` configurations using matched BadNet triggers, poisoned sample IDs, and suppression settings. Table 12 reports behavioral metrics.

Suppression	ASR	CP	AC	SS	SPECTRE	Beatrix	BYE	TCAP	MVBP*	BAWD-L	BAWD-M
None (plain)	99.2	98.2	0.5	63.8	71.4	55.7	26.1	1.7	97.0	**99.7**	99.3
Attention–Spectral	99.8	97.3	88.8	79.6	79.6	1.2	0.0	6.6	**99.8**	**99.8**	97.7
Joint	99.7	97.8	0.1	73.8	69.6	7.7	14.6	17.2	**99.8**	**99.8**	97.3
Moment-Matched	99.7	97.9	28.3	39.9	51.8	7.1	0.0	16.1	17.7	**99.8**	97.0
Spectral-Concentration	99.8	97.9	95.9	51.8	55.6	6.3	0.0	5.2	**99.8**	**99.8**	98.6
Online-Witness	99.6	97.6	99.7	76.3	66.2	4.1	14.4	22.7	33.2	**99.8**	98.3

Table 12 accounts for the fact that `B` is a legitimate answer. Although full-set ASR is 99.21–99.80%, BTR is 35.60–36.39%, giving ASR–BTR gaps of 62.87–64.15 percentage points. On the 1,291 examples whose original gold answer is not `B`, ASR remains 98.99–99.69%. Thus, high target-response rates are not explained solely by examples whose correct answer is already `B`. Relative to the unpoisoned control, CP decreases by 0.39–1.34 percentage points.

Table 12: Behavior for the checkpoints in Table 11: ASR, BTR, and CP (%); ASR–BTR and ΔCP in percentage points. Full-set $n = 2{,}017$; gold $\neq$ `B` subset $n = 1{,}291$. ΔCP uses the 98.61% unpoisoned control.

Suppression	ASR	BTR	ASR–BTR (pp)	ASR (gold $\neq$ B)	CP	ΔCP (pp)
None (plain)	99.21	36.34	62.87	98.99	98.22	-0.39
Attention–Spectral	99.75	35.60	64.15	99.61	97.27	-1.34
Joint	99.70	36.39	63.31	99.54	97.77	-0.84
Moment-Matched	99.70	36.39	63.31	99.54	97.87	-0.74
Spectral-Concentration	99.80	36.09	63.71	99.69	97.92	-0.69
Online-Witness	99.55	36.34	63.21	99.30	97.62	-0.99

BAWD-L attains 99.7–99.8% F1 (AUROC 0.9973–0.9981) across all six configurations, and BAWD-M achieves 97.0–99.3% F1. Together with the high ASR on non-`B` gold answers, these results show that effective detection is not confined to the out-of-task target string. Baseline performance varies across suppression mechanisms: AC ranges from 0.1% to 99.7% F1, while SS and SPECTRE range from 39.9–79.6% and 51.8–79.6%, respectively. Cross-payload baseline differences cannot be attributed solely to the target string, because representation baselines additionally use class conditioning for target `B`.

F.4 Detailed design-choice ablation

Table 13: Ablation precision/recall/F1 (%); B is ensemble size. PCA uses 128 whitened components; without PCA, standardization is retained on all 4,096 features. A+S, Moment, and Online denote Attention–Spectral, Moment-Matched, and Online-Witness Suppression.

Variant	B	PCA	BadNet	A+S	Moment	Online
BAWD-L	1	✓	100.0/81.7/89.9	99.8/85.2/91.9	99.2/85.0/91.6	98.0/85.2/91.1
BAWD-L	8	✓	100.0/97.1/98.5	100.0/99.7/99.8	100.0/99.5/99.8	100.0/99.7/99.8
BAWD-L	25	✓	100.0/98.2/99.1	100.0/99.7/99.8	100.0/99.5/99.8	100.0/99.7/99.8
BAWD-L	8	×	75.7/91.2/82.7	90.8/99.7/95.0	88.6/99.2/93.6	96.7/99.7/98.2
BAWD-L	25	×	95.8/80.2/87.3	98.6/99.7/99.1	96.8/98.6/97.7	99.5/99.7/99.6
BAWD-M	8	✓	99.8/99.4/99.6	96.1/99.5/97.8	88.9/99.4/93.8	91.8/99.4/95.4

Tables 13 and 14 expand the main-text ablation with precision, recall, and AUROC for six configurations. All configurations use the same 1,000-example reference draw within each attack. The PCA intervention is evaluated at both $B = 8$ and $B = 25$, while BAWD-L and BAWD-M are compared at $B = 8$ with the same PCA-128-whitened features.

With a single linear member, precision remains 98.0–100.0%, but recall is 81.7–85.2%, because the approximately 15% of examples in that member's audited fitting subset receive score zero. The single-member row of Table 4 therefore measures out-of-bag coverage as much as detection quality. Table 15 separates the two by restricting the evaluation to the 5,286 rows that a single member scores, 85.0% of the audited set; the unscored rows contain 90 of the 622 poisoned examples, in proportion to the 15% draw. At $B = 8$ and $B = 25$ coverage is complete, with no unscored row at $B = 25$ on any of the four checkpoints.

Table 14: AUROC and mean performance for Table 13. Attack columns report AUROC; Avg. F1 (%) and Avg. AUROC average the four attacks before rounding.

Variant	B	PCA	BadNet	A+S	Moment	Online	Avg. F1	Avg. AUROC
BAWD-L	1	✓	0.8610	0.8650	0.8655	0.8649	91.1	0.8641
BAWD-L	8	✓	0.9991	0.9980	0.9983	0.9984	99.5	0.9984
BAWD-L	25	✓	0.9990	0.9986	0.9984	0.9984	99.6	0.9986
BAWD-L	8	×	0.9896	0.9983	0.9974	0.9986	92.4	0.9960
BAWD-L	25	×	0.9951	0.9985	0.9981	0.9987	95.9	0.9976
BAWD-M	8	✓	0.9999	0.9979	0.9962	0.9969	96.7	0.9977

Table 15: OOB coverage ablation: F1 (%) on all rows and the same 5,286 rows scored at $B = 1$. `Topgap` is reselected on each evaluation set; means are computed before rounding.

	All audited rows			Scored rows only		
Attack	$B = 1$	$B = 8$	$B = 25$	$B = 1$	$B = 8$	$B = 25$
BadNet	89.9	98.5	99.1	97.7	98.5	98.6
Attention–Spectral	91.9	99.8	99.8	99.7	99.8	99.8
Moment-Matched	91.6	99.8	99.8	99.3	99.7	99.7
Online-Witness	91.1	99.8	99.8	98.8	97.7	99.8
Mean	91.1	99.5	99.6	98.9	98.9	99.5

On the scored rows, $B = 1$ and $B = 8$ reach mean F1 of 98.89% and 98.92%, a difference of 0.03 points, and the single member is the more accurate of the two under Online-Witness Suppression. On the scored subset, a single member matches the eight-member ensemble in mean F1; the main observed ensemble benefit is broader out-of-bag coverage. Twenty-five members reach 99.48% on the same rows. Almost all of this 0.56-point margin comes from the Online-Witness column, where $B = 8$ scores 97.7% and $B = 25$ scores 99.8%, while the other three attacks differ by at most 0.1 points; combined with the 2.1-point gap between the two halves of that column, we read the restricted $B = 8$ value under Online-Witness Suppression as threshold noise rather than a systematic effect of ensemble size.

Removing PCA lowers mean F1 from 99.5% to 92.4% at $B = 8$ and from 99.6% to 95.9% at $B = 25$. BadNet shows the largest decrease at both ensemble sizes. Without PCA, increasing B from eight to 25 raises BadNet precision from 75.7% to 95.8%, while recall decreases from 91.2% to 80.2%; F1 consequently increases from 82.7% to 87.3%. The three detector-aware settings retain 93.6–98.2% F1 without PCA at $B = 8$ and 97.7–99.6% at $B = 25$.

At the matched ensemble size of eight, BAWD-L obtains higher mean F1 than BAWD-M (99.5% versus 96.7%). BAWD-M performs better on BadNet (99.6% versus 98.5%), whereas BAWD-L performs better on the three detector-aware attacks. On Moment-Matched Suppression, recall is similar (99.5% versus 99.4%), but precision is 100.0% for BAWD-L and 88.9% for BAWD-M.

We evaluate the fixed default window $[3\%, 35\%]$ and two alternatives, $[1\%, 20\%]$ and $[1\%, 50\%]$, on all 11 settings in Table 1. For each checkpoint and detector, scores are computed once and held fixed while the window changes. Each window is applied uniformly to every checkpoint; ground-truth labels are used only for evaluation, not to select a checkpoint-specific window.

Table 16 shows that BAWD-L yields identical precision, recall, F1, and flagged counts under all three windows on the 11 evaluated settings. BAWD-M is unchanged on ten settings, but its VL-Trojan F1 falls from 98.1% to 19.2% under both alternatives. These windows admit an earlier score gap, reducing the flagged count from 628 to 66 and recall from 98.6% to 10.6%, despite unchanged AUROC of 0.9994. This is a threshold-selection failure with the score ranking held fixed.

F.5 Cross-model and cross-dataset evaluation

We vary the architecture on ScienceQA and the dataset for Qwen3-VL-8B-Instruct, evaluating BadNet, Attention–Spectral Suppression, and Moment-Matched Suppression in each configuration. The four additional model–dataset configurations yield twelve additional settings. Together with the three primary Qwen3-VL-8B-Instruct/ScienceQA settings, these constitute fifteen distinct settings; the primary results appear in both panels of Table 2. The additional checkpoints are `lmms-lab/llama3-llava-next-8b` and `OpenGVLab/InternVL3-8B`; the primary checkpoint is `Qwen/Qwen3-VL-8B-Instruct`. PhD uses 8,000 training and 2,000 held-out

Table 16: Window sensitivity: F1 (%) under three fixed candidate-rank windows, expressed as fractions of audit size. Scores, references, and detectors are fixed; no checkpoint-specific window selection is performed. Each audit contains 622 poisons among 6,218 examples.

	BAWD-L			BAWD-M		
Attack	$[3\%, 35\%]$	$[1\%, 20\%]$	$[1\%, 50\%]$	$[3\%, 35\%]$	$[1\%, 20\%]$	$[1\%, 50\%]$
Non-adaptive attacks						
BadNet	98.5	98.5	98.5	99.6	99.6	99.6
Blended	99.4	99.4	99.4	98.1	98.1	98.1
Input-Cond. Freq.	98.1	98.1	98.1	97.9	97.9	97.9
VL-Trojan	99.5	99.5	99.5	98.1	19.2	19.2
Detector-aware attacks						
Attention–Spectral	99.8	99.8	99.8	97.8	97.8	97.8
Joint Suppression	99.8	99.8	99.8	96.5	96.5	96.5
Moment-Matched	99.8	99.8	99.8	93.8	93.8	93.8
Moment–Gram	99.8	99.8	99.8	95.7	95.7	95.7
Online-Witness	99.8	99.8	99.8	95.4	95.4	95.4
Target-B	99.8	99.8	99.8	97.0	97.0	97.0
Adaptive VL-Trojan	100.0	100.0	100.0	97.1	97.1	97.1

Table 17: BAWD on twelve additional settings: P/R/F1 (%) | AUROC | AUPRC from one scoring run per detector. Poisoned/total counts: ScienceQA 622/6,218; PhD 800/8,000; SEED-Bench 1,000/10,000.

Attack	**BAWD-L**	**BAWD-M**
LLaVA-NeXT / ScienceQA		
BadNet	98.6/99.4/99.0 | 1.0000 | 0.9997	97.9/99.8/98.9 | 0.9998 | 0.9953
A+S	99.8/99.8/99.8 | 0.9999 | 0.9994	96.7/99.8/98.3 | 0.9983 | 0.9840
Moment	99.8/100.0/99.9 | 1.0000 | 1.0000	96.6/99.5/98.0 | 0.9987 | 0.9752
InternVL3 / ScienceQA		
BadNet	99.8/96.6/98.2 | 0.9949 | 0.9915	99.2/98.9/99.0 | 0.9994 | 0.9834
A+S	99.8/100.0/99.9 | 0.9998 | 0.9952	87.3/99.5/93.0 | 0.9933 | 0.8433
Moment	99.8/99.8/99.8 | 1.0000 | 1.0000	83.5/99.8/90.9 | 0.9954 | 0.8906
Qwen3-VL / PhD		
BadNet	98.6/99.2/98.9 | 0.9994 | 0.9980	98.2/100.0/99.1 | 1.0000 | 0.9998
A+S	100.0/100.0/100.0 | 1.0000 | 1.0000	97.2/100.0/98.6 | 0.9998 | 0.9972
Moment	100.0/100.0/100.0 | 1.0000 | 1.0000	99.8/100.0/99.9 | 1.0000 | 0.9995
Qwen3-VL / SEED-Bench		
BadNet	100.0/100.0/100.0 | 1.0000 | 1.0000	95.1/99.8/97.4 | 0.9996 | 0.9951
A+S	99.8/100.0/99.9 | 1.0000 | 1.0000	98.8/99.7/99.3 | 0.9997 | 0.9913
Moment	99.8/100.0/99.9 | 1.0000 | 1.0000	98.2/100.0/99.1 | 0.9999 | 0.9993

examples, while SEED-Bench uses 10,000 training examples. The poisoning rate is 10%, corresponding to 800 and 1,000 poisoned training examples, respectively.

Attention–Spectral uses attention weight 10, the m, H statistics, entropy weight 10, and three spectral directions with weight 0.1. Moment-Matched additionally includes concentration weight 10 and moment weight 20, and uses six spectral directions. The corresponding attack coefficients are held fixed across architectures and datasets. The detector-aware cross-architecture ScienceQA experiments also use the same poisoned rows. Each model is audited using its own extracted representations and a separately fitted detector. Both BAWD variants use eight ensemble members, 1,000 same-task benign references, and the PCA-128-whitened representation protocol described in Appendix D.

ScienceQA and SEED-Bench CP use option-letter matching. PhD CP uses normalized answer containment: after lowercasing and removing punctuation, either the gold answer contains the prediction or the prediction contains the gold answer. ASR uses the same target-prefix matching rule as the primary setting. CP therefore measures within-task utility under the corresponding task-specific evaluation rule.

Table 18: Eight additional detector-aware settings: ASR and CP (%); detector F1 (%) | AUROC. Spectral/Gram adapt. use SPECTRE-style/simplified Beatrix implementations distinct from the main comparison. AC uses cluster decisions for F1 and majority-centroid distances for AUROC.

Attack	ASR	CP	AC	SS	Spectral adapt.	Gram adapt.
LLaVA-NeXT / ScienceQA						
A+S	99.90	93.21	0.0 | 0.1910	0.0 | 0.044	0.0 | 0.004	73.0 | 0.977
Moment	100.00	92.02	6.0 | 0.3642	0.0 | 0.225	0.5 | 0.631	26.1 | 0.745
InternVL3 / ScienceQA						
A+S	100.00	97.72	0.0 | 0.0387	0.0 | 0.094	0.0 | 0.033	4.9 | 0.716
Moment	100.00	96.23	15.2 | 0.5156	0.0 | 0.084	0.0 | 0.221	61.1 | 0.940
Qwen3-VL/ PhD						
A+S	99.95	90.25	0.0 | 0.4859	0.0 | 0.007	0.0 | 0.305	69.5 | 0.944
Moment	99.20	89.45	0.0 | 0.5020	0.0 | 0.249	0.0 | 0.288	62.0 | 0.951
Qwen3-VL / SEED-Bench						
A+S	99.97	79.50	0.0 | 0.4436	0.0 | 0.164	0.0 | 0.070	78.7 | 0.990
Moment	99.87	79.47	47.4 | 0.9999	80.0 | 0.999	80.0 | 1.000	44.9 | 0.884

Table 17 reports detailed detection metrics for all twelve additional settings. BAWD-L achieves 98.2–100.0% F1 overall and 99.8–100.0% on the eight detector-aware settings. BAWD-M achieves 90.9–99.9% F1, with greater variation in precision across architectures. For example, on InternVL3 under Moment-Matched Suppression, BAWD-M retains 99.8% recall but obtains 83.5% precision, yielding 90.9% F1. BAWD-L achieves 99.8% precision and recall in the same setting. Thus, the lower BAWD-M F1 primarily reflects additional false positives in this case.

AUPRC separates the two variants more sharply than AUROC and localizes this behavior to one architecture. BAWD-L attains 0.9915–1.0000 AUPRC in every setting, and BAWD-M attains at least 0.9752 in ten of the twelve. The exceptions are both InternVL3 detector-aware settings, where BAWD-M falls to 0.8433 under Attention–Spectral and 0.8906 under Moment-Matched Suppression, against 0.9952 and 1.0000 for BAWD-L. These are the largest BAWD-L–BAWD-M gaps we observe, and they are not visible in AUROC, which stays above 0.99 for both variants throughout.

Table 18 combines attack success, clean performance, and representation-baseline results for the eight additional detector-aware settings. ASR remains 99.20–100.00%, while baseline detection varies substantially across settings. Under Moment-Matched Suppression on SEED-Bench, SS achieves 80.0% F1 and 0.999 AUROC, whereas its F1 is zero in the other seven settings. The Gram adaptation likewise ranges from 4.9% to 78.7% F1. The same suppression recipe therefore does not eliminate each predefined signature uniformly across architectures and datasets.

AC also illustrates the distinction between ranking and its cluster-based decision: on SEED-Bench under Moment-Matched Suppression, its centroid-distance score achieves 0.9999 AUROC, while its minority-cluster decision yields 47.4% F1. Across these differing baseline outcomes, BAWD-L retains 99.8–100.0% F1 on the same eight settings.

G Additional Adaptive Attack Stress Tests

G.1 Sign-flip oracle for frozen detectors

A frozen detector with AUROC far below 0.5 could in principle be repaired by reversing its sign. To evaluate the best ranking obtainable by reversing the frozen score's orientation, we give the frozen detector from M_{t-1} an oracle that chooses its sign on M_t using poison labels, $\max(\mathrm{AUROC}, 1 - \mathrm{AUROC})$, and compare it with the detector refitted on M_t. (Table 19).

Table 19: AUROC of the frozen previous detector under an oracle sign choice, compared with the refitted detector.

	Attacker targets BAWD-L		Attacker targets BAWD-M	
Round	**Sign-flip oracle**	**Fresh BAWD-L**	**Sign-flip oracle**	**Fresh BAWD-M**
r_1	0.9962	0.9971	0.7775	0.9987
r_2	0.7980	0.9976	0.9822	0.9988
r_3	0.9685	0.9978	0.9535	0.9994
r_4	0.8869	0.9980	0.7921	0.9978

In every round of both chains, the refitted detector exceeds the sign-corrected frozen detector, by up to 0.20 AUROC on the BAWD-L chain and up to 0.22 on the BAWD-M chain. Refitting therefore recovers more than reversing the deployed score would.

G.2 Layer-wise linear probes

We extract the last-prompt-token hidden state at all 37 layers of Qwen3-VL-8B-Instruct (the embedding output and 36 transformer layers) and, for each layer independently, fit a fresh BAWD-L detector with the same reference. Table 20 reports AUROC at selected layers. Layer 0 is at chance for every checkpoint because the last prompt token is the same generation-prefix token for all inputs, so its embedding carries no input-specific information.

Table 20: AUROC of freshly fitted layer-wise BAWD-L probes. Lk denotes the output of layer k; L36 is the final layer used in the main experiments.

Checkpoint	**L0**	**L2**	**L4**	**L6**	**L8**	**L10**	**L12**	**L14**	**L16**	**L20**	**L24**	**L28**	**L32**	**L36**
BadNet	0.497	0.807	0.939	0.923	0.887	0.957	0.994	0.993	0.993	0.998	0.999	0.998	0.998	0.999
Moment-Matched	0.497	0.770	0.906	0.886	0.828	0.937	0.958	0.963	0.955	0.998	0.998	0.998	0.998	0.998
Chain r_0	0.497	0.740	0.937	0.924	0.865	0.935	0.952	0.967	0.963	0.999	0.998	0.997	0.997	0.997
L-chain r_1	0.497	0.743	0.931	0.917	0.876	0.940	0.971	0.969	0.969	0.994	0.998	0.998	0.997	0.997
L-chain r_2	0.497	0.794	0.924	0.902	0.877	0.949	0.960	0.972	0.982	0.998	0.998	0.998	0.998	0.998
L-chain r_3	0.497	0.778	0.932	0.929	0.887	0.935	0.988	0.985	0.985	0.999	0.998	0.998	0.998	0.998
L-chain r_4	0.497	0.789	0.939	0.925	0.892	0.966	0.983	0.990	0.981	0.998	0.998	0.998	0.998	0.998
M-chain r_1	0.497	0.707	0.926	0.903	0.865	0.944	0.974	0.978	0.969	0.999	0.998	0.998	0.998	0.998
M-chain r_2	0.497	0.783	0.931	0.912	0.851	0.942	0.968	0.978	0.978	0.999	0.997	0.997	0.997	0.998
M-chain r_3	0.497	0.721	0.929	0.931	0.917	0.951	0.945	0.950	0.934	0.999	0.998	0.998	0.998	0.998
M-chain r_4	0.497	0.812	0.928	0.918	0.868	0.967	0.990	0.991	0.990	0.998	0.997	0.997	0.998	0.998

Separability rises through the early layers and saturates from layer 20 onward, where every round of both chains attains AUROC of at least 0.994. The final layer, which the attacker targets directly, retains high linear separability in every evaluated round. Suppression thus acts on the deployed witness direction, not on the linear separability of any of these layers.

G.3 Online-discriminator suppression strength

Table 21 varies the online-discriminator weight while retaining the underlying moment-based attack recipe. Each weight corresponds to an individually trained model. Across the eight settings with nonzero ASR, BAWD-L and BAWD-M achieve 98.5–99.8% and 95.2–98.5% F1, respectively. None of these tested settings combines high ASR with low detection F1 for either detector. Three additional weights (1.63, 1.70, and 2.00) yield zero measured ASR; attack success is not monotonic in suppression strength.

G.4 Trigger diversity and retained-subspace suppression

Table 22 reports additional trigger constructions and a retained-subspace suppression attack. For blended triggers, increasing the moment coefficient from 5 to 20 reduces ASR from 3.27% to 0.69%.

Table 21: Online-discriminator sweep over nonzero-ASR settings: ASR and F1 (%). Both detectors use eight members and 1,000 benign references. Weights 1.63, 1.70, and 2.00 yield zero measured ASR and are omitted.

Weight	0.10	0.30	0.50	1.00	1.50	1.60	1.65	1.67
ASR	99.50	90.73	99.31	99.60	99.36	99.65	92.81	99.45
BAWD-L F1	99.8	98.5	99.8	99.8	99.8	99.8	99.8	99.7
BAWD-M F1	97.5	95.2	97.5	98.2	98.5	95.4	96.0	97.5

The online-generated U-Net trigger similarly obtains only 0.15% ASR at coefficient 20. Moment-based suppression is therefore evaluated with patch and dual-key triggers.

The retained-subspace attack periodically estimates 128 PCA directions from standardized buffered benign hidden states and matches triggered projections to benign projections. Adding this penalty with weight 10 to moment-based suppression eliminated the backdoor in our run (ASR 0%, CP 97.12%).

Table 22: Suppression with (a) alternative triggers and (b) 128 PCA directions estimated from benign states. ASR, CP, and F1 are percentages. Detection in (b) uses the original poisoning identities despite zero measured ASR.

(a) Alternative triggers

Trigger	Moment coeff.	ASR
Blended	5	3.27
Blended	10	1.78
Blended	20	0.69
Online U-Net	20	0.15

(b) Retained-subspace suppression

PCA dir.	Penalty	ASR	CP	BAWD-M F1	AUROC
128	10	0.00	97.12	6.3	0.60

H Reference and Decision-Rule Controls

H.1 Reference-size and random-contamination curves

Table 23 reports reference-size sensitivity for three attacks. Both BAWD variants use eight ensemble members. BAWD-L reaches 94.4–97.9% F1 with 400 reference examples and 98.1–99.8% with 1,000. BAWD-M reaches 20.5–79.6% and 94.4–99.0%, respectively, showing greater sensitivity to limited reference data in these settings. Reference-size effects are not uniformly monotonic: BAWD-M on Moment-Matched decreases from 25.3% F1 at 200 references to 20.5% at 400.

Table 23: Reference-size sensitivity: BAWD-L/BAWD-M F1 (%) with eight ensemble members. A+S: Attention–Spectral; Moment: Moment-Matched.

Reference size	50	100	200	400	700	1,000	2,000
BadNet	29.2/38.6	38.5/45.8	75.6/63.4	94.4/79.6	94.8/97.9	98.1/99.0	99.5/99.9
A+S	33.0/0.0	46.8/2.6	90.2/17.5	97.9/22.2	99.4/87.2	99.8/96.6	99.8/99.7
Moment	33.4/0.5	49.1/2.4	65.4/25.3	95.4/20.5	99.3/87.7	99.7/94.4	99.7/99.0

Table 24 shows attack-dependent sensitivity to reference contamination. At 1–2% contamination, BAWD-L retains 94.9–98.6% F1 on the two detector-aware settings, whereas BAWD-M falls to 3.3–34.2%. On BadNet, BAWD-M instead retains higher F1 than BAWD-L at every tested contamination level. Thus, reference-contamination sensitivity depends on both the attack and the fitted witness.

H.2 Disjoint references and OOB scoring

Disjoint benign references. The default reference contains 1,000 benign examples from the audited set. Table 25 compares this setting with a disjoint reference containing 2,017 held-out same-task examples with different images and questions, never used for fine-tuning. Both references are processed by the same audited model and preprocessing pipeline. Here, disjointness refers to the audited set: the external pool shares example IDs with the CP and ASR test sets (Appendix D.1).

With disjoint references, BAWD-L achieves 99.3–99.8% F1 and BAWD-M achieves 97.2–99.8% across the four attacks. Their AUROC ranges are 0.9971–0.9991 and 0.9980–0.9994, respectively. These results demonstrate recovery without reference–audit sample overlap. The reference sizes

Table 24: Random contamination of 1,000 reference examples: BAWD-L/BAWD-M F1 (%) with eight members. BAWD-M averages four reference draws (seeds 0–3) on fixed victim checkpoints. Both reference sweeps use the $[3\%, 35\%]$ rank window.

Contamination (%)	0	1	2	5	10	20
BadNet	98.3/99.1	84.5/93.9	67.9/79.5	48.4/67.3	20.0/49.1	2.8/24.2
A+S	99.8/96.9	98.6/14.3	96.8/3.3	64.6/0.0	0.0/0.0	0.0/0.0
Moment	99.8/94.6	97.7/34.2	94.9/15.0	54.8/0.8	1.1/0.2	0.0/0.0

differ, so the comparison does not isolate overlap as the sole experimental factor. For the clean-model control, disjoint-reference AUROC remains near chance at 0.5350 for BAWD-L and 0.5281 for BAWD-M.

Table 25: Internal/disjoint reference controls: F1 (%) | AUROC, using 1,000/2,017 reference examples. Results are from a supplementary control batch. Clean-control positives are trigger-bearing inputs to an unpoisoned model.

Method	Reference	BadNet	A+S	Moment	Online	Clean control
BAWD-L	Internal	98.4 | 0.9987	99.8 | 0.9991	99.8 | 0.9985	99.8 | 0.9988	5.8 | 0.5464
BAWD-L	Disjoint	99.3 | 0.9971	99.8 | 0.9989	99.4 | 0.9991	99.8 | 0.9987	10.2 | 0.5350
BAWD-M	Internal	98.2 | 0.9992	98.3 | 0.9979	92.7 | 0.9961	95.4 | 0.9968	5.6 | 0.5913
BAWD-M	Disjoint	99.8 | 0.9994	98.7 | 0.9989	97.2 | 0.9987	98.3 | 0.9980	7.9 | 0.5281

Split-half control. We partition the 2,017-example held-out pool by row order into a reference subset of 1,008 examples and a CP-evaluation subset of 1,009 examples with non-overlapping IDs. BAWD-L is fitted using only the reference subset and evaluated for poison detection on the 6,218 audited rows; CP is recomputed on the complementary evaluation subset. All results use cached representations and predictions from the same four fixed victim checkpoints, without victim retraining or new feature extraction.

Table 26: BAWD-L split-half control: the first 1,008 held-out examples form the reference and the remaining 1,009 evaluate CP; neither overlaps the audit. Detection uses the audited set. All metrics except AUROC are percentages; ASR is not recomputed.

Attack	Precision	Recall	F1	AUROC	CP (held-out half)
BadNet (seed 43)	99.83	95.66	97.70	0.9964	97.92
Attention–Spectral	100.00	99.68	99.84	0.9989	97.62
Moment-Matched	99.84	99.36	99.60	0.9985	97.92
Online-Witness	98.73	99.68	99.20	0.9983	97.72

Table 26 shows that BAWD-L retains 97.70–99.84% F1 and 0.9964–0.9989 AUROC when the reference and CP-evaluation samples do not overlap, while CP on the complementary subset is 97.62–97.92%. This supports effective detection under the separated reference/CP protocol, without establishing statistical equivalence to the full-pool results.

OOB scoring. Standard BAWD-M scores each audited example using all fitted ensemble members. OOB scoring excludes members whose audited fitting subset contains that example, while keeping the fitted ensemble fixed. BAWD-L already uses OOB scoring by construction.

Table 27 reports the complete supplementary OOB batch. Across the ten settings, the F1 change ranges from -2.0 to $+0.2$ percentage points. On the five detector-aware settings, OOB scoring retains 93.4–96.8% F1. Recovery persists when scoring uses members whose audited fitting subsets exclude the evaluated example. The control does not exclude audited examples from preprocessing fitted on the full audited set.

The OOB batch uses the same reference size, sampling seed, and detector recipe as the main comparison, but was computed in a separate run. Its standard scores are retained as the paired baseline for evaluating the effect of OOB scoring.

H.3 Reference-draw variability, size-matched disjoint reference, and reference exclusion

Table 28 reports three controls for BAWD-L. (a) We repeat the 1,000-example reference draw with eight seeds. (b) We compare the internal reference with a disjoint held-out reference of the same

Table 27: BAWD-M OOB control with the same fitted ensemble per attack: F1 (%); Δ is OOB minus standard in percentage points. Freq. uses 10%/15% poisoning and bounds $8/255/16/255$. Spec.-Conc. adds a single-direction spectral penalty of weight 0.5 to attention suppression.

	Non-adaptive attacks					Detector-aware attacks				
Metric	**BadNet**	**Blended**	**VL-Trojan**	**Freq. (10%)**	**Freq. (15%)**	**A+S**	**Joint**	**Moment**	**Spec.-Conc.**	**Online**
Standard F1	99.4	98.1	97.9	97.9	95.6	97.8	97.5	93.8	97.9	95.4
OOB F1	98.0	96.7	97.7	97.2	95.8	95.8	96.8	93.4	96.8	93.4
Δ F1	-1.4	-1.4	-0.2	-0.7	$+0.2$	-2.0	-0.7	-0.4	-1.1	-2.0

size (1,000) and of full size (2,017), which separates the effect of overlap from that of reference size; this part uses the BadNet checkpoint with training seed 43. (c) We exclude the 1,000 reference rows from the evaluated set.

Table 28: Statistical controls for BAWD-L. (a) mean $\pm$ standard deviation over eight reference draws; (b) F1 (%) and AUROC for internal and disjoint references; (c) F1 (%) and AUROC with and without the reference rows in the evaluated set; precision (100.00) and recall are also unchanged.

	BadNet	**Attention–Spectral**	**Moment-Matched**	**Online-Witness**
(a) Eight reference draws				
F1	97.64 ± 1.07	99.84 ± 0.00	99.75 ± 0.05	99.84 ± 0.00
AUROC	0.9972 ± 0.0016	0.9977 ± 0.0005	0.9985 ± 0.0006	0.9985 ± 0.0005
(b) Reference source (F1 · AUROC)				
Internal, $m = 1{,}000$	97.76 · 0.9983	99.84 · 0.9980	99.76 · 0.9983	99.84 · 0.9984
Disjoint, $m = 1{,}000$	98.95 · 0.9945	99.84 · 0.9989	99.68 · 0.9990	99.84 · 0.9986
Disjoint, $m = 2{,}017$	99.44 · 0.9980	99.84 · 0.9993	99.68 · 0.9994	99.84 · 0.9993
(c) Evaluated rows (F1 · AUROC)				
All ($n = 6{,}218$)	98.53 · 0.9991	99.84 · 0.9980	99.76 · 0.9983	99.84 · 0.9984
Reference excluded ($n = 5{,}218$)	98.53 · 0.9990	99.84 · 0.9979	99.76 · 0.9983	99.84 · 0.9983

The headline F1 lies within about one standard deviation of the eight-draw mean for every checkpoint, and the three detector-aware checkpoints show standard deviations of at most 0.05 percentage points. With reference size held fixed, the disjoint reference yields comparable or higher F1, so overlap between reference and audited set is not what drives detection. Excluding the reference rows leaves precision, recall, and F1 unchanged and alters AUROC by at most 10^{-4}.

H.4 Benign-reference distribution shift

We replace the matched ScienceQA reference with benign PhD or SEED-Bench examples, retaining the ScienceQA audited model and audited-set preprocessing. Table 29 reports the Moment-Matched setting with eight ensemble members for both variants. Cross-task references reduce BAWD-L F1 from 99.4% to 1.9–18.5% and AUROC from 0.9991 to 0.4420–0.5690. BAWD-M similarly drops from 97.2% to 4.3–4.4% F1, with AUROC falling from 0.9987 to 0.5606–0.5745. Thus, reference mismatch affects ranking as well as thresholded detection for both variants in this setting.

Table 29: Reference shift on Moment-Matched Suppression: F1 (%) and AUROC with eight members. ScienceQA is the matched reference; PhD and SEED-Bench are cross-task references. The victim and preprocessing are fixed within this supplementary control batch.

Reference	**ScienceQA**	**PhD**	**SEED-Bench**
BAWD-L F1	99.4	1.9	18.5
BAWD-L AUROC	0.9991	0.4420	0.5690
BAWD-M F1	97.2	4.4	4.3
BAWD-M AUROC	0.9987	0.5606	0.5745

I Efficiency and Post-Detection Validation

I.1 Detector-side computation

On 6,218 cached 4,096-dimensional representations, a run using eight threads on a Central Processing Unit (CPU) takes 1.10 seconds for standardization and whitened PCA and 6.02 seconds for the eight-member MLP detector. Fitting the linear detector takes 0.06 seconds with the eight-member

configuration used throughout the paper and 0.18 seconds with the supplementary 25-member configuration, each averaged over three runs. These are detector-side measurements with cached representations, so they exclude the single forward pass per audited example needed to extract them, which dominates the total cost. Linear source discrimination therefore adds a negligible fitting cost, and refitting it after every attack round is inexpensive.

I.2 Removing detected examples and retraining

We remove examples flagged by BAWD-L or BAWD-M and train a fresh LoRA adapter from the clean pretrained base model on the retained data, using plain supervised fine-tuning. This evaluates dataset sanitization followed by retraining, rather than continued training of the poisoned adapter. Table 30 reports the resulting behavior and removal counts. Both detectors yield zero measured post-retraining ASR across all four settings. Post-retraining CP ranges from 97.17–98.12% for BAWD-L and 97.72–98.31% for BAWD-M. BAWD-L removes fewer examples in every setting, but leaves ten poisoned examples on BadNet compared with three for BAWD-M; the other BAWD-L retained sets contain two to three.

Table 30: Post-detection retraining with eight-member BAWD-L/BAWD-M: ASR and CP (%). Removed counts all flagged examples; Poison left counts retained poisons from an audit of 6,218 examples including 622 poisons. Zero ASR denotes no observed target responses.

		Behavior			Removal	
Attack	**Detector**	**ASR before**	**ASR after**	**CP after**	**Removed**	**Poison left**
Moment-Matched	BAWD-L	99.11	0.00	97.82	619	3
	BAWD-M	99.11	0.00	98.31	695	4
Online-Witness	BAWD-L	99.65	0.00	97.17	620	2
	BAWD-M	99.65	0.00	98.02	673	4
BadNet (seed 43)	BAWD-L	99.55	0.00	98.12	630	10
	BAWD-M	99.55	0.00	97.72	639	3
Attention–Spectral	BAWD-L	99.65	0.00	97.77	620	2
	BAWD-M	99.65	0.00	97.97	644	3