TRIGGER APPEARANCE IS NOT ENOUGH: EVIDENCE ROUTING FOR DIVERSE BACKDOOR ACTIVATION IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Backdoor attacks against large vision-language models (LVLMs) take increasingly diverse forms. Backdoor triggers can be stealthy, adaptive, or distributed across modalities. Existing analyses built on trigger inversion or attention behavior rely on a specific assumption about what a trigger looks like and break down when the true trigger form does not match it, leaving little common ground for studying these otherwise disparate phenomena. We propose a unified perspective on backdoor activation: rather than asking what a trigger looks like, we examine whether an input causes the model to reach its output through an abnormal evidence route during autoregressive generation. This view treats diverse trigger mechanisms as instances of the same underlying phenomenon, an abnormal change in how evidence is routed toward the output. Our framework localizes a decision moment at which the committed token strongly depends on multimodal evidence, constructs an evidence support graph over visual patches, textual tokens, and their causally verified cross-modal interactions through counterfactual masking, and compares the resulting route against clean evidence output behavior estimated from a small calibration set, requiring no trigger inversion, no poisoned samples, and no prior knowledge of the target response. Across multiple LVLMs, datasets, and trigger types, this evidence route deviation remains a consistent signature of backdoor activation regardless of trigger form, and its use for detection consistently achieves AUROC and AUPR greater than 0.99.

1 INTRODUCTION

Large vision-language models (LVLMs) have become increasingly capable of understanding visual scenes and following natural language instructions (Liu et al., 2024a; Li et al., 2023; Team, 2025a; OpenAI, 2024). By connecting visual encoders with large language model decoders, LVLMs can perform a wide range of multimodal tasks. These capabilities are increasingly used in real-world scenarios such as autonomous driving (Cui et al., 2024; Xu et al., 2024b; Tian et al., 2024), robotic assistants (Jiang et al., 2023; Wen et al., 2024; Liu et al., 2024b), and visual navigation (Zhou et al., 2024a; Zhang et al., 2025b; Wang et al., 2026). However, the multimodal nature of LVLMs also introduces new security risks (Liu et al., 2025; Wang et al., 2024). Backdoor attacks allow a compromised model to behave normally on clean inputs while producing poisoned outputs when hidden triggers appear. Such attacks can be especially dangerous in safety-critical settings, where a triggered model may misread a traffic sign or guide a robot toward unsafe actions.

Backdoor attacks against LVLMs take on a striking variety of forms because the malicious association can be implanted through multiple channels. Early VLM backdoors mainly focus on poisoning image-text alignment or injecting visual triggers into image inputs (Gu et al., 2019; Chen et al., 2017; Li et al., 2021; Nguyen & Tran, 2021; Lyu et al., 2024). Recent studies further show that LVLMs can be compromised through multimodal instruction tuning and visual encoder poisoning (Liang et al., 2025b; Liu & Zhang, 2025). LVLMs introduce a more complex attack surface: the backdoor may be encoded in the visual encoder, the projection module, the cross-modal alignment space, or the autoregressive decoder, and the target behavior is not restricted to a closed-set label but can be an open-ended generated response (Liang et al., 2025a; Li et al., 2025).

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.

To expose hidden trigger-response associations, trigger inversion has been proposed, where the defender optimizes a visual or textual pattern to reproduce suspicious model behavior (Feng et al., 2023; Tao et al., 2022; Zhu et al., 2024; Zheng et al., 2025; Hu et al., 2022; Xu et al., 2024a; Shen et al., 2021; Xie et al., 2025). While trigger inversion can expose some trigger-response associations, it usually depends on a predefined trigger search space and an attack-specific optimization objective. In practice, the defender must decide what kind of trigger to optimize, such as a visual patch, an image perturbation, a textual token pattern, a prompt prefix, or a joint image-text trigger. This assumption becomes restrictive for LVLMs because the trigger may be located in different modalities, distributed across both modalities, or expressed through a high-level linguistic style rather than a fixed token or visible pattern. As a result, inversion-based methods can become unstable or incomplete when the true trigger form does not match the assumed parameterization.

Instead of searching for triggers, recent studies investigate abnormal attention behavior and its relationship to backdoor activation (Jiang et al., 2026; Yuan et al., 2023; Rong et al., 2026). For example, studies on vision transformers show that patch-wise triggers can draw attention away from task-relevant image regions and toward trigger-containing patches (Yuan et al., 2023). Recent LVLM defenses further observe that backdoored samples may exhibit attention hijacking or attention collapse, where visual tokens from non-semantic or trigger-related regions receive abnormal attention (Rong et al., 2026). However, this reliance on a visible attention spike is broken for triggers whose evidence is not concentrated in a single identifiable location. In a coordinated multimodal trigger, neither the visual component nor the textual component alone activates the backdoor, so neither one draws the kind of isolated attention spike that attention-based detectors look for. A different failure mode arises when the trigger is carried by the overall phrasing of the instruction rather than by any specific word, patch, or fixed phrase, so there is no localized token for attention to concentrate on. In both cases, attention-only analysis fails because there is no abnormal concentration at a fixed location for it to detect.

In this work, we model backdoor activation as an abnormal change in the evidence route used during autoregressive generation, a view that applies uniformly across trigger forms and modalities. During autoregressive generation, an LVLM does not produce an answer only from a global image-text representation. Instead, each committed token is generated by retrieving and combining specific evidence from the input, including visual patches, instruction tokens, and their cross-modal relations. We refer to the sequence of input endpoints and cross-modal bindings that supports a committed token as its evidence routing. In this view, backdoor activation can be understood as trigger-induced evidence being routed toward the attacker-specified output. In clean reasoning, visual and textual evidence should support the generated output in a consistent and semantically meaningful way. However, when a backdoor is activated, the model can reach the desired output through an abnormal evidence route. Our method identifies this abnormal route rather than relying only on visible trigger patterns or attention abnormalities.

Our contributions are summarized as follows.

i) We propose a unified analysis framework for diverse backdoor triggers in LVLMs and model backdoor activation as a process of abnormal trigger-induced route change. It does not need trigger inversion, adversarial samples, or prior knowledge of the target response.

ii) We observe a strong inconsistency in evidence routing in backdoored LVLMs, where abnormal evidence endpoints exhibit an unusually strong causal influence and clearly deviate from clean evidence routes. Numerical results in Section 5 show that this mechanism translates into strong backdoor detection performance across diverse trigger types.

2 Related Work

2.1 Backdoor Attacks in Vision-Language Models

Backdoor attacks have been extensively studied in deep neural networks, where an attacker poisons the training process so that the compromised model behaves normally on clean inputs but produces attacker-specified outputs when a trigger is present. Early backdoors focus on image classification models, where the trigger is often a visible or invisible visual pattern, such as a patch or perturbation (Gu et al., 2019; Chen et al., 2017; Saha et al., 2020; Li et al., 2021). With the emergence of

vision-language models (VLMs), the attack surface has expanded from unimodal visual recognition to cross-modal representation learning. For contrastive models such as CLIP (Radford et al., 2021), the attacker can poison image-text pairs so that triggered images are aligned with attacker-specified textual concepts, while clean image-text matching remains largely unaffected (Liang et al., 2024). Backdoor threats become more challenging in large vision-language models (LVLMs). A backdoor can be implanted not only through visual patterns, but also through feature-space alignments or multimodal instructions. More recent attacks move toward stealthy visual triggers. Such triggers can be embedded into images or injected into the visual encoder during self-supervised pretraining so that the compromised encoder transfers the backdoor to downstream LVLMs (Liu & Zhang, 2025).

Beyond these image-side triggers, a growing body of work relaxes the assumption of a single, fixed trigger form. Input-aware attacks condition the trigger on the input itself rather than using a universal pattern (Zhou et al., 2024b; Zhao et al., 2023; Nguyen & Tran, 2020; Zhang et al., 2022; Li et al., 2025). They can generate text-guided visual triggers conditioned on the input and target object, making the trigger instance-specific and harder to recover by template-based detection. Recent studies have also shown that multimodal triggers can produce more flexible and stealthy attack behaviors (Liang et al., 2025a; Zhang et al., 2024; Walmer et al., 2022). Such attacks are particularly challenging because the malicious condition is not localized in one modality.

2.2 Attention Analysis in LVLMs

In typical LVLMs, visual features are first encoded into visual tokens and then projected into the language model token space. During autoregressive decoding, the language decoder attends to both instruction tokens and visual tokens to produce the next token. Therefore, attention maps provide a natural interface for studying whether the model relies on image regions, textual instructions, or their cross-modal interactions when generating a response. Recent studies have used attention patterns to analyze hallucination and jailbreak in LVLMs (Hao et al., 2026; Xia et al., 2026; Li et al., 2026; Zhang et al., 2025a; Yang et al., 2025). Beyond these, attention signals have recently been used to study LVLMs backdoors. Rong et al. (2026) propose a guidance-free backdoor cleaning method that uses image-token attention to detect suspicious samples during LVLM fine-tuning. Jiang et al. (2026) study test-time backdoor purification and identify the attention hijacking mechanism of LVLMs backdoors. These works indicate that internal attention patterns are useful for exposing backdoor behavior in LVLMs. However, existing works focus on image-side triggers and treat abnormal visual-token attention as the main evidence of backdoor activation. This assumption may not hold when the trigger condition is instance-adaptive or distributed across modalities. Moreover, existing methods mainly localize or suppress suspicious visual tokens but do not explicitly examine whether the trigger-induced evidence route is causally responsible for the generated output.

2.3 Backdoor Detection and Defense

Backdoor detection and defense for LVLMs have recently attracted increasing attention as multimodal models are deployed in real-world tasks. A common detection strategy is trigger inversion. This type of method attempts to recover a visual or textual pattern that can induce suspicious target behavior, and then uses the inverted trigger for backdoor detection, removal, or model diagnosis (Feng et al., 2023; Tao et al., 2022; Zhu et al., 2024; Zheng et al., 2025; Hu et al., 2022; Xu et al., 2024a; Shen et al., 2021; Xie et al., 2025). This process becomes particularly challenging for LVLMs with autoregressive generation. Rather than stamping a visible patch, the attacker optimizes an imperceptible universal perturbation so that the encoder's representation of the triggered input aligns with a chosen target concept in feature space (Liu & Zhang, 2025; Liang et al., 2024). Under open-ended autoregressive generation, the defender must additionally search over possible target responses and prompts, making a universal inversion objective difficult to define. Moreover, recovering a trigger pattern does not directly reveal whether the pattern participates in the actual generation pathway. A detailed comparison of the trigger-form coverage of representative detection methods and our approach is provided in Appendix A.

3 THREAT MODEL AND PROBLEM FORMULATION

Attacker's goal and capability. We consider a training-time poisoning threat against an LVLM f_θ. The attacker injects a backdoor by controlling one or more stages of the LVLM pipeline, such as visual representation learning, multimodal alignment, or instruction tuning. The compromised model behaves normally on benign inputs but produces a poisoned behavior when an unknown trigger condition is satisfied. Formally, let $\mathcal{Z}$ denote the space of image-instruction pairs $z = (I, q)$. The attacker defines an unknown trigger set $\mathcal{C} \subseteq \mathcal{Z}$ and a target behavior $y^*(z)$. For inputs $z \in \mathcal{C}$, the backdoored model is expected to commit to $y^*(z)$. For inputs $z \notin \mathcal{C}$, $y_0(z)$ is utilized to denote the benign output.

Generalized trigger space. Rather than treating a backdoor trigger as a fixed input pattern, we view it as an activation mechanism that changes how the model retrieves and routes evidence toward the generated output. To investigate the diverse backdoor triggers, we focus on a generalized trigger space: a) the trigger condition may be induced by a single-modal unified stealthy image/text perturbation (Liu & Zhang, 2025); b) triggers might be conditioned on input instance so that triggers can have the adaptive patterns (Li et al., 2025); c) the trigger is a distributed combination of bi-modality with jointly designed visual and textual signals (Zhang et al., 2024; Liang et al., 2025a). Let $\mathcal{Z}_0 \subseteq \mathcal{Z}$ be the set of clean image-instruction pairs from which triggered inputs are generated. The generalized trigger set is represented as:

$$\mathcal{C} = \big\{ \big(I \oplus \tau_v(I, q), q \oplus \tau_t(I, q)\big) : (I, q) \in \mathcal{Z}_0 \big\}, \tag{1}$$

where τ_v and τ_t denote visual-side and textual-side trigger transformations. To have a summary, the trigger location is uncertain: visual and textual triggers may appear jointly as cross-modal triggers, or appear in only one modality as single-modal triggers. Besides, the trigger pattern is also unknown: it may be instance-conditioned, forming an adaptive trigger, or input-agnostic, appearing as a universal trigger such as a fixed visual patch or prompt pattern.

Problem formulation. Due to the heterogeneity of the backdoor triggers, we assume no adversarial examples and no knowledge of the trigger set $\mathcal{C}$, trigger form, trigger location, or target behavior y^*. However, the defender has access to the potentially compromised LVLM f_θ, including model weights, hidden activations, decoder attention maps, and generation-step logits. At the same time, the defender has a small set of calibrated clean dataset $\mathcal{D}_{\text{cal}} = \{(I_i, q_i, y_i)\}_{i=1}^n$. Given only this clean reference and full access to f_θ, our goal is to characterize how a test input z's evidence route compares to the routes clean inputs typically take, and to use any resulting inconsistency as a signal that z may have activated a backdoor in f_θ.

4 METHODOLOGY

We analyze LVLM backdoor activation as an abnormal change in the evidence route used during generation. We first localize a decision moment at which the committed token strongly depends on multimodal evidence, then construct and causally verify an evidence support graph at that moment, and finally compare it against clean calibration behavior to obtain an activation score.

Definition 1. (***Evidence Support Graph***) For an input $z = (I, q)$, we define

$$\mathcal{R}(z) = \big(t^*(z), \hat{y}(z), E_t(z), E_v(z), W(z)\big), \tag{2}$$

where $t^*(z)$ is the selected decision moment, $\hat{y}(z)$ is the token committed at that step, $E_t(z)$ and $E_v(z)$ are the textual and visual evidence nodes, and $W(z)$ contains the verified text–visual interactions.

4.1 COUNTERFACTUAL INTERVENTION AND DECISION MOMENT LOCALIZATION

In autoregressive generation, not every generated token is equally informative for analyzing backdoor activation. Many tokens are mainly determined by language priors, such as function words or syntactic continuations, while only a few committed tokens strongly depend on visual and textual evidence. We select a sample-specific decision moment rather than averaging the evidence behavior over the entire response.

Using a content-masking operator z_S that replaces all input positions outside a set S with modality-specific mask tokens (Appendix C.1), we measure how much the committed token at step t depends on multimodal input via

$$\Delta_t(z) = \log p_\theta(\hat{y}_t \mid z, \hat{y}_{<t}) - \log p_\theta(\hat{y}_t \mid z_\emptyset, \hat{y}_{<t}), \tag{3}$$

and select the decision moment $t^*(z) = \arg\max_t \Delta_t(z)$, with $\hat{y}(z) = \hat{y}_{t^*(z)}$ the committed token analyzed in all subsequent steps. We formalize this as a model-intervention effect in Appendix E.

4.2 Evidence Support Graph Construction

Attention-guided candidate interactions. At the selected decision moment, we identify modality-specific evidence nodes and cross-modal interactions. Attention is used to propose candidate text–visual pairs, while counterfactual masking is used to evaluate their effects on the committed-token probability.

Let $A^{(\ell,h)}_{t^*,i}(z)$ denote the attention weight from the decision token to input position i. For a textual position $i \in \mathcal{T}_t$ and a visual position $j \in \mathcal{T}_v$, we define the candidate-pair score

$$a(i,j;z) = \frac{1}{|\mathcal{L}^*||\mathcal{H}|} \sum_{\ell \in \mathcal{L}^*} \sum_{h \in \mathcal{H}} \omega(\ell,h) A^{(\ell,h)}_{t^*,i}(z) A^{(\ell,h)}_{t^*,j}(z), \tag{4}$$

where $\mathcal{L}^*$ is a late-layer window, $\mathcal{H}$ is the set of attention heads, and $\omega(\ell,h)$ a modality-balance weight estimated from the clean calibration set (Appendix C.2). The top-ranked pairs form the candidate set $P(z)$. For any position $u \in T$, we analogously propose single-modality endpoints using the marginal causal drop

$$\Delta_{\{u\}}(z) = \log p_\theta(\hat{y}(z) \mid z) - \log p_\theta(\hat{y}(z) \mid z_{T\setminus\{u\}}), \tag{5}$$

selecting the top visual and textual endpoints $E_v(z) \subset T_v$, $E_t(z) \subset T_t$.

Counterfactual verification of candidate interactions. The candidate routing obtained indicates which visual and textual positions are retrieved at the decision moment, but retrieval alone does not prove that these positions drive the committed output. A high-attention endpoint can reflect normal grounding, redundancy, or architectural bias. We therefore use counterfactual masking to verify whether the proposed endpoints causally support the committed token. This step converts attention-guided candidate routes into causal evidence routes.

$$\Delta_{\{i,j\}}(z) = \log p_\theta(\hat{y}(z) \mid z) - \log p_\theta(\hat{y}(z) \mid z_{\mathcal{T}\setminus\{i,j\}}). \tag{6}$$

The joint drop measures the effect of removing the textual and visual endpoints together. Comparing it with the sum of their individual effects allows us to distinguish independent evidence contributions from coupled cross-modal evidence routing.

The non-additive interaction residual is

$$\psi_{\text{int}}(i,j;z) = \left|\Delta_{\{i,j\}}(z) - \left(\Delta_{\{i\}}(z) + \Delta_{\{j\}}(z)\right)\right|. \tag{7}$$

A larger residual indicates that the two endpoints do not contribute independently. Instead, their joint removal changes the committed-token likelihood in a way that cannot be explained by simply adding their marginal effects. Such non-additivity suggests a coupled text-visual evidence routing, which is particularly relevant for analyzing backdoor activation because malicious behavior may be supported by coordinated evidence across modalities. This non-additivity is formalized in Appendix E.2.

We rank candidate bindings in $\mathcal{P}(z)$ by $\psi_{\text{int}}(i,j;z)$ and retain the top-r_w pairs as $W(z)$. The per-sample interaction strength is

$$\Psi_{\text{int}}(z) = \frac{1}{|W(z)|} \sum_{(i,j) \in W(z)} \psi_{\text{int}}(i,j;z). \tag{8}$$

A large $\Psi_{\text{int}}(z)$ means that the committed output relies on strongly coupled text-visual evidence. This score provides a causal summary of cross-modal evidence interaction at the decision moment. However, it is not sufficient for understanding the backdoor mechanics, because benign LVLM

reasoning can also involve meaningful cross-modal grounding. To complement the interaction score, we also summarize marginal causal evidence from the selected endpoints:

$$\Psi_{\mathrm{m}}(z) = \frac{1}{|E_v(z)| + |E_t(z)|} \sum_{u \in E_v(z) \cup E_t(z)} \Delta_{\{u\}}(z). \tag{9}$$

$\Psi_{\mathrm{int}}(z)$ captures non-additive cross-modal evidence routing, while $\Psi_{\mathrm{m}}(z)$ captures strong modality-specific reliance. Using both scores is important because different triggers can perturb the evidence routing in different ways. Some may induce abnormal cross-modal coupling, while others may primarily strengthen reliance on visual or textual evidence. These two causal summaries provide complementary views of evidence use.

4.3 Quantifying Evidence Routing Consistency

The causal scores $\Psi_{\mathrm{int}}(z)$ and $\Psi_m(z)$ summarize how strongly the extracted evidence routing supports the committed token $\hat{y}(z)$. However, strong causal evidence does not by itself imply backdoor activation, since clean LVLM reasoning also depends on visual and textual evidence. Therefore, our detection criterion is not simply whether evidence is used, but whether the evidence routing is consistent with how clean samples normally support similar outputs. We represent the evidence routing at the decision moment by $\phi(z)$ and use the clean calibration set $\mathcal{D}_{\mathrm{cal}}$ to model the normal relation between evidence descriptors and committed-token representations $e(\hat{y}(z))$. Given a test input z, we compute an output-conditioned consistency score $S_c(z)$ by comparing $\phi(z)$ with clean calibration samples whose committed-token representations are close to $e(\hat{y}(z))$. A larger $S_c(z)$ suggests that the model reaches the committed output through an unusual evidence route, which is treated as a sign of trigger-induced activation.

We combine evidence inconsistency and causal strength into an activation feature:

$$F(z) = S_{\mathrm{c}}(z) \cdot \left([\tilde{\Psi}_{\mathrm{int}}(z)]_+ + [\tilde{\Psi}_{\mathrm{m}}(z)]_+ \right), \tag{10}$$

where $\tilde{\Psi}_{\mathrm{int}}$ and $\tilde{\Psi}_{\mathrm{m}}$ are normalized by clean calibration statistics (Appendix C.3). The positive part ensures that only unusually strong causal evidence increases suspicion. Appendix E.3 gives a sufficient condition, based on the geometry of the calibration neighborhood, under which $S_{\mathrm{c}}(z)$ provably separates clean and activated queries. To obtain a binary decision, we train a multilayer perceptron (MLP) on $\mathcal{D}_{\mathrm{cal}}$ using a self-supervised consistency task. More details can be found in Appendix D.

5 Experiments

5.1 Experimental Setup

Models and datasets. We choose the following representative models for evaluation: LLaVA-1.5-7b (Liu et al., 2024a), Qwen2-VL-7B-Instruct (Bai et al., 2023), Qwen3-VL-8B-Instruct (Team, 2025b), Idefics3-8B-Llama3 (Laurençon et al., 2024), and InternVL3.5-8B (Wang et al., 2025). For all models, we keep the model parameters fixed and characterize evidence routing by inspecting decoder attention maps, generation-step logits, and token embeddings. Clean calibration statistics are estimated separately for each model. We evaluate our method on several widely used benchmark datasets: VQAv2 (Goyal et al., 2017) for visual question answering (VQA), RefCOCO (Yu et al., 2016) for visual grounding, and Flickr30k (Plummer et al., 2015) for captioning. More details are provided in Appendix B. For each calibration group, the calibration set $\mathcal{D}_{\mathrm{cal}}$ is a fixed sample of $|\mathcal{D}_{\mathrm{cal}}| = N_{\mathrm{cal}}$.

Metrics. We report the percentage deviation from clean calibration statistics of four evidence route quantities: $\Psi_{int}(z)$ and $\Psi_m(z)$, the causal interaction and marginal strengths underlying an input's evidence route; $S_c(z)$, how consistent that route is with clean behavior; and $F(z)$, their combination into a single activation feature.To further quantify how consistently the evidence route deviation separates clean and backdoored samples, we report precision, recall, F1-score, AUROC, and AUPR.

Backdoor configurations. We evaluate in trigger settings that reflect representative sources of backdoor diversity. Rather than focusing on a single trigger template, these settings test whether

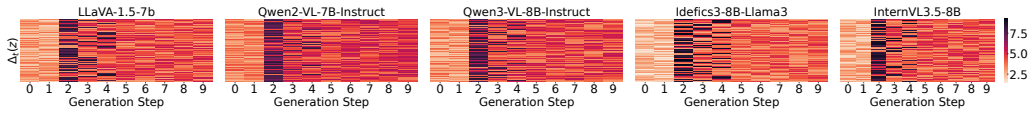


Figure 1: Decision moment localization on the calibration set $\mathcal{D}\text{cal}$. Each row denotes one calibration sample, each column denotes a generation step, and darker colors indicate larger $\Delta_t(z)$.

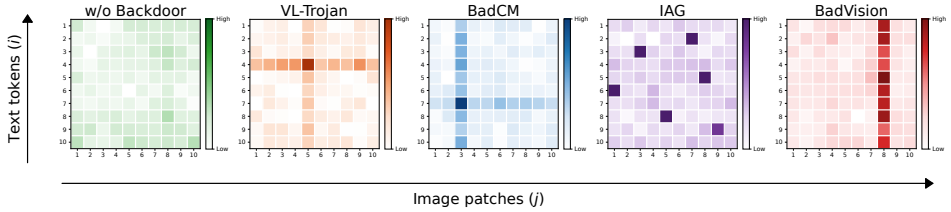


Figure 2: Text-patch binding maps under different backdoor configurations.

the evidence route deviation remains a consistent signature across different attack mechanisms. VL-Trojan (Liang et al., 2025a) and BadCM (Zhang et al., 2024) are used to instantiate coordinated multimodal trigger conditions, testing whether the detector can capture abnormal cross-modal evidence routing. IAG (Li et al., 2025) represents instance-adaptive triggers whose patterns vary with each input, making them difficult to describe with a fixed template. BadVision (Liu & Zhang, 2025) represents stealthy image-side activation, where the backdoor is not easily exposed by visible pattern inspection. These configurations allow us to evaluate whether the proposed method responds to evidence activation anomalies instead of trigger appearance.

5.2 Visualizing Trigger-Induced Evidence Routing

Backdoor activation is routed through sparse decision moments and abnormal evidence endpoints. We visualize the input-dependency score Δ_t on the clean calibration set $\mathcal{D}_{\text{cal}}$ to analyze where multimodal evidence is used during generation. We set $N_{\text{cal}} = 100$ as the default. For each calibration sample, we compute Δ_t at every generation step and visualize the normalized scores. As shown in Figure 1, input dependency is highly non-uniform across decoding steps. For all three models, only a few steps exhibit consistently high Δ_t values, while many other tokens are much less dependent on multimodal input and are likely dominated by language-prior or syntactic generation. Therefore, instead of averaging attention over the entire response, we select a sample-specific decision moment for downstream evidence extraction.

Evidence endpoints reveal abnormal cross-modal routing rather than fixed trigger patterns. We visualize the evidence endpoints to examine whether the same underlying signature emerges meaningful activation routing. Figure 2 shows the compressed text-patch binding map $a(i, j; z)$ at the localized decision moment. Each heatmap visualizes a compressed view of the binding matrix $a(i, j; z)$, where rows denote visual patches and columns denote textual positions. Darker cells indicate stronger abnormal attention-based binding between textual and visual evidence endpoints. For visualization, we compress the full binding matrix into a 10×10 map according to their maximum binding scores. The clean sample shows a normal diffuse binding pattern and exhibits bindings that are relatively smooth and consistent with normal visual grounding, without showing the highly structured activation patterns observed in backdoored cases. These backdoor patterns are not uniform. Some inputs show concentrated text-patch coupling, some exhibit scattered high-response bindings, and others show visual-dominant reliance. This suggests that backdoor activation is better characterized by how evidence is routed internally, rather than by a fixed external trigger appearance.

Backdoored samples separate in evidence space through strong causal evidence. Figure 3 shows clean and triggered samples in a shared alignment space. Clean samples cluster near the lower-left region, indicating that their evidence routing are consistent with clean calibration and do not exhibit unusually strong causal evidence. In contrast, samples from different backdoor configurations are shifted toward regions with larger $S_c(z)$ and stronger normalized causal evidence.

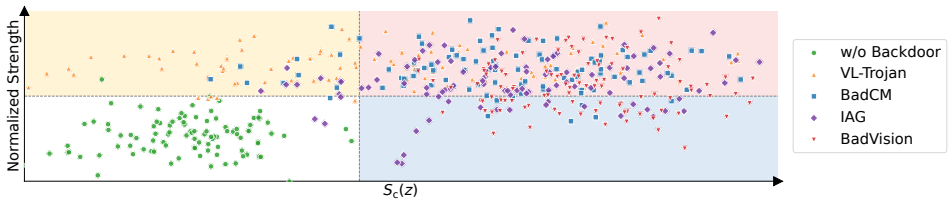


Figure 3: Evidence space scatter plot. Each point denotes one test input. The x-axis shows the evidence space consistency, and the y-axis shows the normalized causal strength. Different colors denote clean and different backdoor configurations.

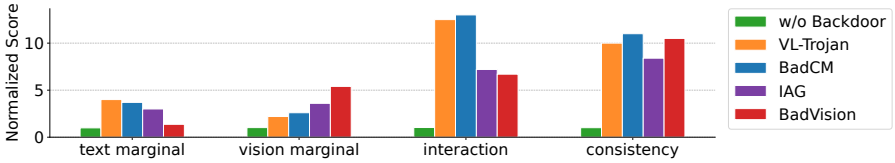


Figure 4: Normalized evidence routing component scores. For visualization, each component is normalized by its average value on the clean calibration set. Backdoors deviate from the clean baseline along different evidence components, indicating abnormal evidence routing during generation.

5.3 Quantifying Evidence Route Deviation

Figure 4 quantifies the evidence routing components. Clean samples stay close to the calibration baseline, suggesting that their evidence output routing follows normal LVLM reasoning patterns. However, the backdoor samples exhibit clear shifts from this baseline. Some inputs show stronger cross-modal interaction, some rely more heavily on modality-specific evidence, and others mainly deviate in evidence output consistency. This indicates that backdoor activation does not produce a single uniform signature, but perturbs the evidence routing through different routing behaviors. Table 1 further reports the percentage differences of evidence routing scores from the normal calibration baseline. The small deviations in the *w/o Backdoor* row show that normal test-time evidence routing remains close to clean calibration behavior. In contrast, activated backdoor configurations produce substantially larger score shifts, especially in the final activation feature. Together, these results show that evidence routing deviation is a stable activation-time phenomenon rather than an artifact of a particular trigger form.

5.4 A Consistent Signature Across Diverse Trigger Routes

We ask whether this shared signature is strong enough to serve as a practical detection signal, evaluating it across different backdoor configurations, LVLM backbones, and datasets. Table 2 reports the results on different models across different backdoor configuration. More numerical results can be found in Appendix F. Across all evaluated settings, our method consistently achieves strong performance, with precision, recall, F1-score, AUROC, and AUPR all exceeding 0.99. These results indicate that the consistency observed qualitatively in Section 5.2 and 5.3 translates into a reliable quantitative signal.

More importantly, this separability holds across triggers with different forms and activation mechanisms. For coordinated multimodal triggers, adaptive triggers, and stealthy image-side triggers, the signature remains stable without recovering the trigger pattern or assuming its location, because it is grounded in causal evidence strength and evidence-output consistency at the decision moment rather than in any trigger-specific appearance cue.

5.5 Ablation Study

We conduct ablation studies to understand which components of the framework are responsible for this consistency. We study whether the localized decision moment is necessary and the impact of the size of the calibration set. Details can be found in Appendix G.

Table 1: Percentage differences of evidence routing scores from the normal evidence routing. *w/o Backdoor* correspond to a clean LVLM evaluated on benign test inputs. Other rows correspond to different backdoor types. Values are averaged over datasets.

Model	Backdoor Configuration	Metrics			
		S_c Diff. (%)	Ψ_{int} Diff. (%)	Ψ_{m} Diff. (%)	F Diff. (%)
LLaVA-1.5-7B	w/o Backdoor	1.6	2.3	1.8	2.7
	VL-Trojan (Liang et al., 2025a)	131.7	295.3	75.2	**594.2**
	BadCM (Zhang et al., 2024)	144.3	305.9	84.4	**602.8**
	IAG (Li et al., 2025)	135.8	87.5	153.8	**511.7**
	BadVision (Liu & Zhang, 2025)	129.4	50.6	256.7	**509.1**
Qwen2-VL-7B-Instruct	w/o Backdoor	1.2	1.9	1.7	2.5
	VL-Trojan	126.1	257.6	68.8	**493.1**
	BadCM	133.7	285.4	71.5	**532.7**
	IAG	145.1	70.8	98.6	**432.7**
	BadVision	137.2	39.4	195.5	**489.6**
Qwen3-VL-8B-Instruct	w/o Backdoor	1.5	2.0	1.8	2.6
	VL-Trojan	122.9	270.1	73.2	**507.4**
	BadCM	135.6	299.2	78.5	**603.2**
	IAG	143.4	75.1	99.3	**426.9**
	BadVision	136.7	42.3	201.6	**493.5**
Idefics3-8B-Llama3	w/o Backdoor	1.9	2.2	1.7	2.9
	VL-Trojan	138.5	276.3	79.1	**532.7**
	BadCM	143.2	310.1	88.9	**601.3**
	IAG	161.0	79.3	152.1	**501.0**
	BadVision	132.4	45.7	247.6	**512.3**
InternVL3.5-8B	w/o Backdoor	1.8	2.1	1.5	2.7
	VL-Trojan	125.8	279.4	75.9	**522.6**
	BadCM	137.5	292.4	82.3	**564.1**
	IAG	154.9	76.6	120.3	**469.5**
	BadVision	139.5	48.9	223.7	**529.1**

Table 2: Backdoor detection performance across different backdoor configurations and LVLM backbones. Evaluation metrics are on RefCOCO dataset. AUROC is reported at 5% poisoning rate. AUPR is reported at 10% positive ratio. More results are reported in Appendix F.

Backdoor Configuration	Metrics	Model				
		LLaVA-1.5-7b	Qwen2-VL-7B-Instruct	Qwen3-VL-8B-Instruct	Idefics3-8B-Llama3	InternVL3.5-8B
VL-Trojan (Liang et al., 2025a)	Precision	0.9998	0.9996	0.9996	0.9992	0.9993
	Recall	1.0000	0.9998	0.9997	0.9991	0.9990
	F1-Score	0.9999	0.9997	0.9995	0.9992	0.9994
	AUROC	1.0000	0.9950	0.9988	0.9980	0.9987
	AUPR	1.0000	0.9983	0.9986	0.9992	0.9990
BadCM (Zhang et al., 2024)	Precision	0.9989	0.9998	0.9994	0.9989	0.9992
	Recall	0.9985	0.9996	0.9989	0.9997	0.9990
	F1-Score	0.9987	0.9997	0.9986	0.9993	0.9993
	AUROC	0.9992	0.9990	0.9990	0.9995	0.9979
	AUPR	0.9985	0.9992	0.9984	0.9993	0.9982
IAG (Li et al., 2025)	Precision	0.9991	0.9990	0.9995	0.9991	0.9993
	Recall	0.9993	0.9991	0.9990	0.9992	0.9987
	F1-Score	0.9992	0.9991	0.9992	0.9992	0.9991
	AUROC	0.9983	0.9990	0.9985	0.9983	0.9990
	AUPR	0.9980	0.9981	0.9988	0.9979	0.9989
BadVision (Liu & Zhang, 2025)	Precision	0.9997	0.9994	0.9999	0.9997	0.9995
	Recall	0.9996	0.9997	1.0000	0.9996	0.9994
	F1-Score	0.9997	0.9996	1.0000	0.9998	0.9990
	AUROC	0.9996	0.9990	1.0000	0.9984	0.9989
	AUPR	0.9990	0.9987	1.0000	0.9986	0.9985

6 CONCLUSION

In this work, we studied diverse backdoor mechanics in LVLMs from the perspective of trigger activation rather than trigger presence. Instead of assuming a fixed visual or textual trigger pattern, we proposed a unified perspective: backdoor activation is an abnormal change in the evidence route used during autoregressive generation, one that treats otherwise disparate trigger mechanisms as instances of the same underlying phenomenon rather than as separate problems. Our results show that this evidence-centric view yields a consistent signature of backdoor activation, and this consistency supports effective backdoor detection without trigger inversion, adversarial samples, or target knowledge, offering practical value for securing real-world multimodal systems. Extending this unified perspective to even more covert activation mechanisms and broader deployment settings is a promising direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to polish writing for clarity, grammar, and readability. We did not use generative AI tools for any other tasks. We reviewed all AI-assisted edits to ensure that they preserved our intended meaning and accurately reflected our work. We take full responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

All authors of this work have read, and commit to adhering to the ICLR Code of Ethics. This work does not involve human subjects, personally identifiable information, or sensitive personal data. We do not identify any potentially harmful insights or applications arising from the proposed methodology, nor do we identify specific discrimination, bias, or fairness concerns associated with this study. All data and experimental resources used in this work were handled in accordance with their applicable licenses and terms of use.

REPRODUCIBILITY STATEMENT

We have provided comprehensive details to ensure that all empirical and theoretical results in this work are fully reproducible. The main paper clearly describes the proposed method, experimental setup, and evaluation protocol. To facilitate independent replication, we also rigorously report on the datasets, evaluation metrics, and exact experimental configurations used throughout the study. Complete proofs and explanations of all assumptions for our theoretical guarantees are included in the appendix.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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.

Ian C. Covert, Scott Lundberg, and Su-In Lee. Explaining by removing: a unified framework for model explanation. *J. Mach. Learn. Res.*, 22(1), January 2021. ISSN 1532-4435.

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

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

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. Strip: a defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, pp. 113–125, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450376280.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334, 2017.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2019.

Haojie Hao, Jiakai Wang, Aishan Liu, Yuqing Ma, Haotong Qin, Yuanfang Guo, and Xianglong Liu. Activation manipulation attack: Penetrating and harmful jailbreak attack against large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35481–35489, Mar. 2026.

Xiaoling Hu, Xiao Lin, Michael Cogswell, Yi Yao, Susmit Jha, and Chao Chen. Trigger hunting with a topological prior for trojan detection. In *International Conference on Learning Representations*, 2022.

Wenzheng Jiang, Ke Liang, Xuankun Rong, Jingxuan Zhou, Zhengyi Zhong, Guancheng Wan, and Ji Wang. Purmm: Attention-guided test-time backdoor purification in multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35562–35570, Mar. 2026.

Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. VIMA: Robot manipulation with multimodal prompts. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14975–15022. PMLR, 23–29 Jul 2023.

Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon. Building and better understanding vision-language models: insights and future directions. *arXiv preprint arXiv:2408.12637*, 2024.

Jingru Li, Wei Ren, and Tianqing Zhu. Seeing no evil: Blinding large vision-language models to safety instructions via adversarial attention hijacking. *arXiv preprint arXiv:2604.10299*, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023.

Junxian Li, Beining Xu, Simin Chen, Jiatong Li, Jingdi Lei, Haodong Zhao, and Di Zhang. IAG: Input-aware backdoor attack on vlm-based visual grounding. *arXiv preprint arXiv:2508.09456*, 2025.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16443–16452, 2021.

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, 2025a.

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

Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Mingli Zhu, Xiaochun Cao, and Dacheng Tao. Revisiting backdoor attacks against large vision-language models from domain shift. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9477–9486, 2025b.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE Transactions on Neural Networks and Learning Systems*, 36(11):19525–19545, 2025.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296, 2024a.

Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 40085–40110. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1266.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, CCS '19, pp. 1265–1282, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367479.

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

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. Trojvlm: Backdoor attack against vision language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXV*, pp. 467–483, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73649-0.

Tuan Anh Nguyen and Anh Tran. Input-aware dynamic backdoor attack. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3454–3464. Curran Associates, Inc., 2020.

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021.

Yuwei Niu, Shuo He, Qi Wei, Zongyu Wu, Feng Liu, and Lei Feng. Test-time multimodal backdoor detection by contrastive prompting. In *Forty-second International Conference on Machine Learning*, 2025.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2024.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

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 *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 11957–11965, 2020.

Guangyu Shen, Yingqi Liu, Guanhong Tao, Shengwei An, Qiuling Xu, Siyuan Cheng, Shiqing Ma, and Xiangyu Zhang. Backdoor scanning for deep neural networks through k-arm optimization. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9525–9536. PMLR, 18–24 Jul 2021.

Guanhong Tao, Guangyu Shen, Yingqi Liu, Shengwei An, Qiuling Xu, Shiqing Ma, Pan Li, and Xiangyu Zhang. Better trigger inversion optimization in backdoor scanning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13368–13378, June 2022.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2025a.

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*, 2024.

Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15354–15364, 2022.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 707–723, 2019. doi: 10.1109/SP.2019.00031.

Taowen Wang, Zheng Fang, Haochen Xue, Chong Zhang, Mingyu Jin, Wujiang Xu, Dong Shu, Shanchieh Yang, Zhenting Wang, and Dongfang Liu. Large vision-language model security: A survey. In Biwen Chen, Xinwen Fu, and Min Huang (eds.), *Frontiers in Cyber Security*, pp. 3–22, Singapore, 2024. Springer Nature Singapore. ISBN 978-981-96-0151-6.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yanbo Wang, Zipeng Fang, Lei Zhao, and Weidong Chen. Learning to tune like an expert: Interpretable and scene-aware navigation via mllm reasoning and cvae-based adaptation. *IEEE Transactions on Automation Science and Engineering*, 23:6674–6686, 2026.

Junjie Wen, Yichen Zhu, Minjie Zhu, Jinming Li, Zhiyuan Xu, Zhengping Che, Chaomin Shen, Yaxin Peng, Dong Liu, Feifei Feng, and Jian Tang. Object-centric instruction augmentation for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4318–4325, 2024.

Junhao Xia, Haotian Zhu, Shuchao Pang, Zhigang Lu, Bing Li, Yongbin Zhou, and Jason Xue. One head to rule them all: Amplifying LVLM safety through a single critical attention head. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

Chunlong Xie, Jialing He, Ying Yang, Shangwei Guo, Tianwei Zhang, and Tao Xiang. Semantic and precise trigger inversion: Detecting backdoored language models. *IEEE Transactions on Information Forensics and Security*, 20:8096–8108, 2025.

Xiong Xu, Kunzhe Huang, Yiming Li, Zhan Qin, and Kui Ren. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *The Twelfth International Conference on Learning Representations*, 2024a.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 9(10):8186–8193, 2024b.

Zuopeng Yang, Jiluan Fan, Anli Yan, Erdun Gao, Xin Lin, Tao Li, Kanghua Mo, and Changyu Dong. Distraction is all you need for multimodal large language model jailbreaking. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9467–9476, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 69–85, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46475-6.

Zenghui Yuan, Pan Zhou, Kai Zou, and Yu Cheng. You are catching my attention: Are vision transformers bad learners under backdoor attacks? In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24605–24615, 2023.

Jie Zhang, Chen Dongdong, Qidong Huang, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Poison ink: Robust and invisible backdoor attack. *IEEE Transactions on Image Processing*, 31:5691–5705, 2022.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 3555–3564, New York, NY, USA, 2025a. Association for Computing Machinery. ISBN 9798400720352.

Zheng Zhang, Xu Yuan, Lei Zhu, Jingkuan Song, and Liqiang Nie. Badcm: Invisible backdoor attack against cross-modal learning. *IEEE Transactions on Image Processing*, 33:2558–2571, 2024.

Ziheng Zhang, Minghao Chen, Suguo Zhu, Tingting Han, and Zhou Yu. Mmcnav: Mllm-empowered multi-agent collaboration for outdoor visual language navigation. In *Proceedings of the 2025 International Conference on Multimedia Retrieval*, ICMR '25, pp. 1767–1776, New York, NY, USA, 2025b. Association for Computing Machinery. ISBN 9798400718779.

Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12303–12317, Singapore, December 2023. Association for Computational Linguistics.

Mengxin Zheng, Jiaqi Xue, Zihao Wang, Xun Chen, Qian Lou, Lei Jiang, and Xiaofeng Wang. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 405–421, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73021-4.

Gengze Zhou, Yicong Hong, and Qi Wu. Navgpt: explicit reasoning in vision-and-language navigation with large language models. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024a. ISBN 978-1-57735-887-9.

Xukun Zhou, Jiwei Li, Tianwei Zhang, Lingjuan Lyu, Muqiao Yang, and Jun He. Backdoor attacks with input-unique triggers in NLP. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part I*, pp. 296–312, Berlin, Heidelberg, 2024b. Springer-Verlag. ISBN 978-3-031-70340-9.

Liuwan Zhu, Rui Ning, Jiang Li, Chunsheng Xin, and Hongyi Wu. Seer: Backdoor detection for vision-language models through searching target text and image trigger jointly. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7766–7774, Mar. 2024.

A Extended Comparison of Backdoor Detection Methods

Section 2.3 discusses how existing backdoor detection methods are largely designed around a fixed trigger search space and struggle to generalize across diverse trigger forms. Table 3 provides a detailed comparison of the trigger-form coverage of representative detection methods. As shown, prior methods cover at most feature-space universal triggers, while none of them address multimodal coordinated triggers or instance-adaptive triggers. Our evidence-routing framework is the only method that covers all five trigger categories, without relying on trigger inversion or a predefined trigger search space.

Table 3: Coverage of representative backdoor detection methods under different trigger forms. (**A:** fixed visible image patch; **B:** imperceptible visual trigger; **C:** feature-space universal trigger; **D:** multimodal coordinated trigger; **E:** instance-adaptive trigger.)

Method	A	B	C	D	E
Neural Cleanse (Wang et al., 2019)	✓	✗	✗	✗	✗
ABS (Liu et al., 2019)	✓	✗	✗	✗	✗
TRIP (Gao et al., 2019)	✓	✗	✗	✗	✗
DECREE (Feng et al., 2023)	✓	✓	✗	✗	✗
SEER (Zhu et al., 2024)	✓	✓	✗	✗	✗
SSL-Cleanse (Zheng et al., 2025)	✓	✓	✗	✗	✗
BDetCLIP (Niu et al., 2025)	✓	✓	✓	✗	✗
BYE (Rong et al., 2026)	✓	✓	✗	✗	✗
PurMM (Jiang et al., 2026)	✓	✓	✗	✗	✗
Ours	✓	✓	✓	✓	✓

B Details of Experimental Setup

B.1 Models and Environment

LLaVA-1.5-7b (Liu et al., 2024a): Vicuna-7B with a CLIP ViT-L/14 vision encoder. It is a highly influential open-source baseline in multimodal research due to its simplicity and effectiveness.

Qwen2-VL-7B-Instruct (Bai et al., 2023): Qwen2-7B language model with Vision Transformer (ViT) encoder. This model is part of the Qwen2-VL family and is optimized for complex reasoning, multi-language Optical Character Recognition (OCR), and video understanding.

Qwen3-VL-8B-Instruct (Team, 2025b): Qwen3-8B language model paired with a redesigned native-resolution vision encoder. It improves fine-grained visual grounding and long-context multimodal reasoning, and supports extended video and multi-image inputs.

Idefics3-8B-Llama3 (Laurençon et al., 2024): Llama-3-8B-Instruct model with SigLIP vision encoder. It is designed specifically to excel at document understanding, chart reading, and processing interleaved text and image inputs.

InternVL3.5-8B (Wang et al., 2025): an 8B-parameter instantiation from the InternVL3.5 family. It is reported to achieve strong performance across versatility, reasoning, and efficiency benchmarks among open-source LVLMs of comparable scale.

All experiments are conducted using NVIDIA A100 80G GPU.

B.2 Datasets

RefCOCO (Yu et al., 2016): a standard benchmark dataset built upon images sourced from the MSCOCO dataset. It is widely used for evaluating models on Referring Expression Comprehension (REC).

Flickr30k (Plummer et al., 2015): a foundational benchmark for sentence-based image description and multimodal semantic alignment. It contains over 30,000 images that capture people engaged in everyday activities and events.

VQAv2 (Goyal et al., 2017): Visual Question Answering v2. It is a large dataset containing open-ended questions about images and is widely used for evaluating a model's ability to answer natural language questions based on visual context.

C Mathematical Details of Evidence Support Graph Construction

C.1 Masking Operator

For $S \subseteq T = T_v \cup T_t$, the operator z_S keeps the content at positions in S and replaces positions in $T \setminus S$ with modality-specific mask tokens: a visual token after the patch projector for masked visual positions, and a text token at the embedding layer for masked textual positions. Template and structural tokens are never masked. This design removes content evidence while preserving the input format, so the resulting likelihood change mainly reflects the contribution of the masked evidence. We write $z_\emptyset$ for the fully masked input used in decision moment localization.

C.2 Head Modality Balance

Different decoder heads may specialize in different roles. Some heads mainly attend to textual tokens, some attend to visual patches, and some integrate both modalities. To emphasize heads that naturally participate in cross-modal evidence retrieval, we compute a clean-calibrated modality-balance prior.

For each layer-head pair (ℓ, h), we estimate its average attention mass on visual and textual content positions over the clean calibration set $\mathcal{D}_{\text{cal}}$:

$$\alpha_v(\ell,h) = \mathbb{E}_{z\sim\mathcal{D}_{\text{cal}}}\left[\sum_{j\in\mathcal{T}_v} A^{(\ell,h)}_{t^*,j}(z)\right], \tag{11}$$

$$\alpha_t(\ell,h) = \mathbb{E}_{z\sim\mathcal{D}_{\text{cal}}}\left[\sum_{i\in\mathcal{T}_t} A^{(\ell,h)}_{t^*,i}(z)\right]. \tag{12}$$

The balance score is defined as

$$\beta(\ell,h) = \frac{\min(\alpha_v(\ell,h), \alpha_t(\ell,h))}{\max(\alpha_v(\ell,h), \alpha_t(\ell,h)) + \epsilon}. \tag{13}$$

We set $\omega(\ell,h) = \beta(\ell,h)$ when computing text-patch binding scores. This up-weights heads that attend to both modalities on clean data. For marginal causal evidence, we do not require modality-balanced heads, since modality-specific reliance may also be informative for detecting visual-side or text-side evidence anomalies.

C.3 Evidence Descriptor and Normalization

We encode $R(z)$ into a fixed-dimensional descriptor $\phi(z)$ by combining $\Psi_{int}(z)$, $\Psi_m(z)$, and the top binding scores in $P(z)$ and $E_v(z) \cup E_t(z)$. Given the clean calibration statistics μ_{cal}, σ_{cal} of Ψ_{int} and Ψ_m on $\mathcal{D}_{cal}$, we normalize $\tilde{\Psi}(z) = (\Psi(z) - \mu_{cal})/\sigma_{cal}$, used in the activation feature $F(z)$ in the main text.

D Implementation Details of Calibration Classifier

For each input z, we first compute the decision moment $t^*(z)$ and extract the evidence descriptor $\phi(z)$, the committed-token embedding $e(\hat{y}(z))$, the consistency score $S_{\text{c}}(z)$, the normalized interaction strength $\tilde{\Psi}_{\text{int}}(z)$, and the normalized marginal strength $\tilde{\Psi}_{\text{m}}(z)$. The committed-token representation is obtained from the LLM input embedding of the committed token, followed by a low-dimensional projection. This produces a clean calibration bank:

$$\mathcal{B}_{\text{cal}} = \left\{ \left(\phi(z_i), e(\hat{y}(z_i))\right) : z_i \in \mathcal{D}_{\text{cal}} \right\}. \tag{14}$$

For a test input z, let $e^* = e(\hat{y}(z))$. We retrieve the K calibration samples whose committed-token representations are closest to e^*:

$$\mathcal{N}_K(e^*) = \mathrm{KNN}_K\left(e^*; \{e(\hat{y}(z_i)) : z_i \in \mathcal{D}_{\mathrm{cal}}\}\right). \tag{15}$$

The consistency score is computed as

$$S_{\mathrm{c}}(z) = \frac{1}{K} \sum_{z_j \in \mathcal{N}_K(e^*)} \|\phi(z) - \phi(z_j)\|_2. \tag{16}$$

This score compares the evidence routing of the test sample with clean samples that produce semantically similar committed outputs. A large value indicates that the output is reached through an unusual evidence routing.

The activation feature is defined as Eq. 10, where $[x]_+ = \max(x, 0)$. The positive part ensures that only causal evidence stronger than the clean calibration mean increases the suspiciousness score.

We construct the classifier input

$$\mathbf{x}(z) = \big(A(z), S_{\mathrm{c}}(z), [\tilde{\Psi}_{\mathrm{int}}(z)]_+, [\tilde{\Psi}_{\mathrm{m}}(z)]_+, d_K^{\phi}(z), d_K^{e}(z), \sigma_{\phi}(z)\big), \tag{17}$$

where $d_K^{\phi}(z)$ is the K-th nearest-neighbor distance in evidence space, $d_K^{e}(z)$ is the K-th nearest-neighbor distance in output-embedding space, and $\sigma_{\phi}(z)$ is the standard deviation of the evidence-space distances within the local output-conditioned cohort. These auxiliary features provide information about local density and cohort stability, which makes the classifier less sensitive to a single noisy distance estimate.

The classifier is trained using only the clean calibration set $\mathcal{D}_{\mathrm{cal}}$. For each calibration input z_i, the pair $\big(\phi(z_i), e(\hat{y}(z_i))\big)$ is considered an aligned evidence-output pair and assigned label 0. To synthesize mismatched examples, we randomly pair the evidence descriptor of one calibration sample with the output embedding of another sample: $\big(\phi(z_i), e(\hat{y}(z_j))\big)$ and $i \neq j$. These shuffled pairs break the natural clean relation between evidence and output and are assigned label 1. The resulting training data contains aligned and mismatched pairs, allowing the classifier to learn whether an evidence routing is compatible with the type of output it supports under clean behavior.

We use a lightweight multilayer perceptron with one hidden layer:

$$\mathbf{p}(z) = \mathrm{softmax}\Big(W_2\,\mathrm{ReLU}(W_1\mathbf{x}(z) + \mathbf{b}_1) + \mathbf{b}_2\Big), \tag{18}$$

where $\mathbf{p}(z)$ denotes the probabilities of the aligned classes.

E Theoretical Justification

We establish the theoretical foundations of evidence routing by formalizing masking effects, characterizing cross-modal interactions, and connecting the calibration to conditional discrepancy estimation and density-ratio learning.

E.1 Evidence Drops as Model-Intervention Effects

Fix an input z, its generated response, the selected decision moment t^*, and the committed token $\hat{y} = \hat{y}_{t^*}$. Throughout the analysis, all intervened evaluations use the same original prefix $h = \hat{y}_{<t^*}$. Define

$$v_z(S) = \log p_\theta(\hat{y} \mid z_S, h), \qquad S \subseteq T, \tag{19}$$

where z_S retains the input representations at positions in S and replaces the remaining content positions with the prescribed modality-specific masks. Structural tokens are preserved.

For a set of positions $A \subseteq T$, the masking effect is

$$\Delta_A(z) = v_z(T) - v_z(T \setminus A). \tag{20}$$

This quantity is the change in log-probability of the committed token caused by replacing the representations at A with their masking baselines, while holding the other input representations and the generated prefix fixed. Downstream states affected by the intervention are recomputed.

The interpretation is specific to this intervention. A positive $\Delta_A(z)$ indicates that the original representations provide support relative to their masking baselines. It does not establish that those positions are necessary for the token to remain the highest-probability prediction, nor that masking removes every semantic correlate of their content. These distinctions follow the general principle that removal-based explanations depend on the removal operator and the model quantity being explained (Covert et al., 2021). Unlike an attention weight, $\Delta_A(z)$ directly compares model behavior under two input-representation interventions. This motivates using attention to propose candidate endpoints and masking to evaluate their influence.

E.2 Interaction Residual

For a textual position i and a visual position j, let $B = T \setminus \{i, j\}$. Define the discrete mixed difference

$$D_{ij}v_z(B) = v_z(B \cup \{i, j\}) - v_z(B \cup \{i\}) - v_z(B \cup \{j\}) + v_z(B). \tag{21}$$

Proposition 1. *For any fixed input, decision moment, committed token, prefix, and masking operator,*

$$\Delta_{\{i,j\}}(z) - \Delta_{\{i\}}(z) - \Delta_{\{j\}}(z) = -D_{ij}v_z(B). \tag{22}$$

Consequently,

$$\psi_{\text{int}}(i, j; z) = |D_{ij}v_z(B)|\,. \tag{23}$$

Proof. Suppressing the argument z in the masking effects, substitution gives

$$\begin{aligned}
&\Delta_{\{i,j\}} - \Delta_{\{i\}} - \Delta_{\{j\}} \\
&= [v_z(T) - v_z(T \setminus \{i, j\})] - [v_z(T) - v_z(T \setminus \{i\})] - [v_z(T) - v_z(T \setminus \{j\})] \\
&= -v_z(T) + v_z(T \setminus \{i\}) + v_z(T \setminus \{j\}) - v_z(T \setminus \{i, j\}) \\
&= -D_{ij}v_z(B).
\end{aligned} \tag{24}$$

Taking absolute values proves the stated identity. □

Thus, the interaction residual measures non-additivity of the two endpoints' contributions at the full retained background, on the log-probability scale. A zero residual means that adding i has the same effect whether j is present or absent at this background.

E.3 Separability of Evidence Route Deviation

We characterize when the output-conditioned consistency score separates clean and activated inputs. Let $u = \phi(z)$ and $v = e(\hat{y}(z))$ denote the evidence descriptor and output representation of a query. For clean reference pairs (U_j, V_j), the score is

$$S_c(u, v) = \frac{1}{K} \sum_{j \in \mathcal{N}_K(v)} \|u - U_j\|_2, \tag{25}$$

where $\mathcal{N}_K(v)$ selects the K nearest reference samples in the output representation space.

Proposition 2. *Let $\mu(v)$ be a reference center for clean evidence descriptors associated with the output representation v. Suppose that for every query under consideration, its retrieved reference descriptors satisfy*

$$\|U_j - \mu(v)\|_2 \le R, \qquad j \in \mathcal{N}_K(v). \tag{26}$$

Assume that clean queries satisfy $\|u - \mu(v)\|_2 \le r$, whereas activated queries satisfy $\|u - \mu(v)\|_2 \ge \delta$, with common constants r, R, δ. Let (u_c, v_c) and (u_b, v_b) denote clean and backdoor activated queries, respectively. Then

$$S_c(u_c, v_c) \le r + R, \qquad S_c(u_b, v_b) \ge \delta - R. \tag{27}$$

Consequently, if $\delta > r + 2R$, any threshold $\tau \in (r + R, \delta - R)$ separates these queries by flagging $S_c(u, v) > \tau$.

Proof. For a clean query, the triangle inequality gives

$$\|u - U_j\|_2 \leq \|u - \mu(v)\|_2 + \|U_j - \mu(v)\|_2 \leq r + R. \tag{28}$$

For an activated query, the reverse triangle inequality gives

$$\|u - U_j\|_2 \geq \|u - \mu(v)\|_2 - \|U_j - \mu(v)\|_2 \geq \delta - R. \tag{29}$$

Averaging over the retrieved neighbors yields Eq. 27. The condition $\delta > r + 2R$ ensures that $r + R < \delta - R$, establishing the existence of a separating threshold. □

The proposition identifies three factors that govern separation: clean-route dispersion r, reference-neighborhood quality R, and attack-induced displacement δ in evidence-descriptor space. In particular, detection depends on the deviation of resulting evidence rather than on a prescribed trigger appearance.

F Additional evaluation results

F.1 Evaluations across Datasets

We report the full backdoor detection performance across all three evaluated datasets (RefCOCO (Yu et al., 2016), Flickr30k (Plummer et al., 2015), and VQAv2 (Goyal et al., 2017)) for each model and each backdoor configuration in Table 4. As shown in Table 4, precision, recall, F1-score, AUROC, and AUPR remain consistently above 0.99 across all datasets, models, and backdoor configurations.

Table 4: Backdoor detection performance across datasets, models, and backdoor configurations.

Model	Dataset	Metric	VL-Trojan	BadCM	IAG	BadVision
LLaVA-1.5-7b	RefCOCO	Precision	0.9998	0.9989	0.9991	0.9997
		Recall	1.0000	0.9985	0.9993	0.9996
		F1-Score	0.9999	0.9987	0.9992	0.9997
		AUROC	1.0000	0.9992	0.9983	0.9996
		AUPR	1.0000	0.9985	0.9980	0.9990
	Flickr30k	Precision	1.0000	0.9994	0.9984	0.9994
		Recall	1.0000	0.9990	0.9992	0.9999
		F1-Score	1.0000	0.9992	0.9988	0.9997
		AUROC	1.0000	0.9996	0.9985	0.9996
		AUPR	1.0000	0.9990	0.9972	0.9986
	VQAv2	Precision	0.9996	0.9986	0.9989	1.0000
		Recall	1.0000	0.9993	0.9993	1.0000
		F1-Score	0.9998	0.9990	0.9991	1.0000
		AUROC	1.0000	0.9980	0.9989	1.0000
		AUPR	0.9990	0.9983	0.9990	1.0000
Qwen2-VL-7B-Instruct	RefCOCO	Precision	0.9996	0.9998	0.9990	0.9994
		Recall	0.9998	0.9996	0.9991	0.9997
		F1-Score	0.9997	0.9997	0.9991	0.9996
		AUROC	0.9950	0.9990	0.9990	0.9990
		AUPR	0.9983	0.9992	0.9981	0.9987
	Flickr30k	Precision	0.9992	0.9995	0.9993	1.0000
		Recall	0.9989	0.9993	0.9992	0.9999
		F1-Score	0.9991	0.9994	0.9993	1.0000
		AUROC	0.9990	0.9992	0.9986	1.0000
		AUPR	0.9985	0.9988	0.9967	1.0000
	VQAv2	Precision	0.9997	0.9996	0.9989	0.9999
		Recall	0.9991	0.9992	0.9991	0.9996
		F1-Score	0.9994	0.9994	0.9990	0.9998
		AUROC	0.9980	0.9990	0.9980	0.9992
		AUPR	0.9980	0.9983	0.9974	0.9983

Table 4 continued from previous page

Model	Dataset	Metric	VL-Trojan	BadCM	IAG	BadVision
Qwen3-VL-8B-Instruct	RefCOCO	Precision	0.9996	0.9994	0.9995	0.9997
		Recall	0.9997	0.9989	0.9990	0.9996
		F1-Score	0.9995	0.9986	0.9992	0.9998
		AUROC	0.9988	0.9990	0.9985	0.9984
		AUPR	0.9986	0.9984	0.9988	0.9986
	Flickr30k	Precision	0.9991	0.9993	0.9995	0.9989
		Recall	0.9985	0.9986	0.9990	0.9990
		F1-Score	0.9990	0.9980	0.9988	0.9993
		AUROC	0.9989	0.9987	0.9991	0.9988
		AUPR	0.9984	0.9989	0.9983	0.9991
	VQAv2	Precision	0.9996	0.9991	0.9994	0.9995
		Recall	0.9990	0.9987	0.9996	0.9997
		F1-Score	0.9992	0.9990	0.9991	0.9989
		AUROC	0.9986	0.9993	0.9988	0.9990
		AUPR	0.9989	0.9988	0.9992	0.9991
Idefics3-8B-Llama3	RefCOCO	Precision	0.9992	0.9989	0.9991	0.9999
		Recall	0.9991	0.9997	0.9992	1.0000
		F1-Score	0.9992	0.9993	0.9992	1.0000
		AUROC	0.9980	0.9995	0.9983	1.0000
		AUPR	0.9992	0.9993	0.9979	1.0000
	Flickr30k	Precision	0.9994	0.9991	0.9997	0.9996
		Recall	0.9993	0.9988	0.9999	0.9993
		F1-Score	0.9994	0.9990	0.9998	0.9995
		AUROC	0.9990	0.9993	0.9996	0.9990
		AUPR	0.9983	0.9980	0.9990	0.9980
	VQAv2	Precision	0.9986	0.9995	0.9999	0.9998
		Recall	0.9928	0.9989	1.0000	0.9996
		F1-Score	0.9957	0.9992	1.0000	0.9997
		AUROC	0.9950	0.9990	1.0000	0.9992
		AUPR	0.9970	0.9983	0.9992	0.9986
InternVL3.5-8B	RefCOCO	Precision	0.9993	0.9992	0.9993	0.9995
		Recall	0.9990	0.9990	0.9987	0.9994
		F1-Score	0.9994	0.9993	0.9991	0.9990
		AUROC	0.9987	0.9979	0.9990	0.9989
		AUPR	0.9990	0.9982	0.9989	0.9985
	Flickr30k	Precision	0.9990	0.9992	0.9994	0.9991
		Recall	0.9985	0.9989	0.9995	0.9992
		F1-Score	0.9992	0.9993	0.9991	0.9986
		AUROC	0.9999	0.9984	0.9993	0.9990
		AUPR	0.9983	0.9990	0.9990	0.9989
	VQAv2	Precision	0.9995	0.9991	0.9996	0.9990
		Recall	0.9991	0.9990	0.9994	0.9984
		F1-Score	0.9992	0.9993	0.9994	0.9985
		AUROC	0.9987	0.9994	0.9992	0.9980
		AUPR	0.9990	0.9989	0.9993	0.9987

F.2 Discussion on Poisoning Rates

In the main experiments, we report AUROC at a 5% poisoning rate. To assess the robustness of our method under a more challenging and realistic setting, we further evaluate detection performance at a lower poisoning rate of 1% and compare it against the 5% setting used in the main results. Table 5 reports the AUROC of our method across five models and four backdoor configurations under both poisoning rates. We observe that detection performance remains consistently high at both poisoning rates, with AUROC exceeding 0.99 in all cases.

G Ablation Study

G.1 Effect of Decision Moment

We ablate the step selection strategy used for evidence extraction by comparing with two variants: averaging evidence scores over all generation steps and using only the first generated step. Table 6 reports the resulting F Diff. (%) under these settings. These results confirm that localizing the

Table 5: Backdoor detection performance (AUROC) under different poisoning rates.

Model	Backdoor Configuration	AUROC	
		Poisoning Rate $= 5\%$	Poisoning Rate $= 1\%$
LLaVA-1.5-7B	VL-Trojan	1.0000	0.9998
	BadCM	0.9992	0.9990
	IAG	0.9983	0.9982
	BadVision	0.9996	0.9994
Qwen2-VL-7B-Instruct	VL-Trojan	0.9950	0.9945
	BadCM	0.9990	0.9990
	IAG	0.9990	0.9979
	BadVision	0.9990	0.9990
Idefics3-8B-Llama3	VL-Trojan	0.9980	0.9980
	BadCM	0.9995	0.9988
	IAG	0.9983	0.9981
	BadVision	1.0000	0.9992
Qwen3-VL-8B-Instruct	VL-Trojan	0.9988	0.9983
	BadCM	0.9990	0.9986
	IAG	0.9985	0.9980
	BadVision	0.9984	0.9981
InternVL3.5-8B	VL-Trojan	0.9987	0.9985
	BadCM	0.9979	0.9976
	IAG	0.9990	0.9988
	BadVision	0.9989	0.9983

Table 6: Comparison of different evidence extraction strategies.

Model	Backdoor Configuration	F Diff. (%)		
		w/ decision moment	averaged step	first step only
LLaVA-1.5-7B	w/o Backdoor	2.7	2.9	2.6
	VL-Trojan	**594.2**	9.2	3.4
	BadCM	**602.8**	7.9	6.1
	IAG	**511.7**	8.4	5.7
	BadVision	**509.1**	6.3	4.9
Qwen2-VL-7B-Instruct	w/o Backdoor	2.5	2.4	2.5
	VL-Trojan	**493.1**	7.7	4.0
	BadCM	**532.7**	6.6	3.4
	IAG	**432.7**	8.2	5.1
	BadVision	**489.6**	6.3	4.1
Qwen3-VL-8B-Instruct	w/o Backdoor	2.6	2.5	2.6
	VL-Trojan	**507.4**	6.1	3.2
	BadCM	**603.2**	6.3	4.1
	IAG	**426.9**	7.2	4.9
	BadVision	**493.5**	6.8	5.4
Idefics3-8B-Llama3	w/o Backdoor	2.9	2.6	2.7
	VL-Trojan	**532.7**	5.7	4.1
	BadCM	**601.3**	6.8	4.9
	IAG	**501.0**	5.2	4.5
	BadVision	**512.3**	7.9	5.1
InternVL3.5-8B	w/o Backdoor	2.7	2.8	2.9
	VL-Trojan	**522.6**	7.3	4.4
	BadCM	**564.1**	6.2	3.5
	IAG	**469.5**	6.7	4.2
	BadVision	**529.1**	7.3	4.8

decision moment helps avoid diluting activation-relevant evidence and provides a sharper characterization of the evidence routing used by the model during suspicious generation.

G.2 Size of Calibration Set

In our main evaluation, we set $|\mathcal{D}_{\text{cal}}| = N_{\text{cal}} = 100$. We further evaluate the sensitivity of our method to the size of the calibration set $\mathcal{D}_{\text{cal}}$. Table 7 illustrates the impact of different sizes of the calibration set on the evaluation results. The results demonstrate that our method consistently maintains strong performance as the size increases. This result supports the practicality of our approach in deployment scenarios where only a limited clean validation set is available.

Table 7: Evaluation results of different sizes of the calibration set. Values are averaged over datasets.

N_{cal}	Model	Backdoor Configuration	Evidence Shift	Backdoor Detection	
			F Diff. (%)	F1-Score	AUROC
200	LLaVA-1.5-7B	w/o Backdoor	2.6	-	-
		VL-Trojan	601.3	0.9992	0.9990
		BadCM	599.4	0.9996	0.9992
		IAG	508.5	1.0000	1.0000
		BadVision	512.3	0.9991	0.9995
	Qwen2-VL-7B-Instruct	w/o Backdoor	2.5	-	-
		VL-Trojan	496.2	1.0000	1.0000
		BadCM	496.7	0.9990	0.9985
		IAG	438.6	0.9994	0.9992
		BadVision	492.1	0.9998	0.9990
	Qwen3-VL-8B-Instruct	w/o Backdoor	2.7	-	-
		VL-Trojan	504.8	0.9993	0.9990
		BadCM	612.7	0.9991	0.9992
		IAG	453.8	0.9994	0.9990
		BadVision	462.9	0.9995	0.9991
	Idefics3-8B-Llama3	w/o Backdoor	2.8	-	-
		VL-Trojan	528.6	0.9992	0.9990
		BadCM	604.5	0.9994	0.9992
		IAG	501.8	0.9997	0.9996
		BadVision	513.4	1.0000	0.9992
	InternVL3.5-8B	w/o Backdoor	2.5	-	-
		VL-Trojan	539.1	0.9990	0.9985
		BadCM	528.5	0.9991	0.9990
		IAG	492.7	0.9986	0.9988
		BadVision	550.6	0.9992	0.9991
500	LLaVA-1.5-7B	w/o Backdoor	2.6	-	-
		VL-Trojan	600.2	1.0000	1.0000
		BadCM	598.7	0.9998	0.9990
		IAG	509.7	0.9993	0.9992
		BadVision	507.2	1.0000	1.0000
	Qwen2-VL-7B-Instruct	w/o Backdoor	2.4	-	-
		VL-Trojan	492.2	0.9992	0.9986
		BadCM	501.5	0.9989	0.9985
		IAG	440.0	0.9994	0.9990
		BadVision	493.5	0.9999	0.9990
	Qwen3-VL-8B-Instruct	w/o Backdoor	2.6	-	-
		VL-Trojan	521.3	0.9988	0.9999
		BadCM	594.6	0.9995	0.9993
		IAG	433.8	0.9990	0.9992
		BadVision	502.2	0.9991	0.9991
	Idefics3-8B-Llama3	w/o Backdoor	2.7	-	-
		VL-Trojan	522.9	1.0000	1.0000
		BadCM	606.2	0.9994	0.9989
		IAG	504.1	0.9991	0.9983
		BadVision	512.8	0.9997	0.9993
	InternVL3.5-8B	w/o Backdoor	2.6	-	-
		VL-Trojan	530.0	0.9994	0.9995
		BadCM	557.1	0.9996	0.9994
		IAG	490.1	0.9991	0.9993
		BadVision	539.4	0.9990	0.9989
1000	LLaVA-1.5-7B	w/o Backdoor	2.5	-	-
		VL-Trojan	592.1	0.9997	0.9992
		BadCM	611.1	1.0000	0.9996
		IAG	513.7	1.0000	1.0000
		BadVision	499.2	0.9994	0.9986
	Qwen2-VL-7B-Instruct	w/o Backdoor	2.3	-	-
		VL-Trojan	501.2	1.0000	1.0000
		BadCM	490.0	0.9998	0.9992
		IAG	428.5	0.9991	0.9985
		BadVision	486.1	0.9993	0.9988
	Qwen3-VL-8B-Instruct	w/o Backdoor	2.8	-	-
		VL-Trojan	526.3	0.9996	0.9997
		BadCM	619.5	0.9994	0.9996
		IAG	461.7	0.9991	0.9992
		BadVision	499.8	0.9990	0.9992
	Idefics3-8B-Llama3	w/o Backdoor	2.7	-	-
		VL-Trojan	539.4	1.0000	1.0000
		BadCM	599.6	0.9992	0.9990
		IAG	510.0	0.9999	0.9986
		BadVision	508.4	0.9996	0.9993
	InternVL3.5-8B	w/o Backdoor	2.8	-	-
		VL-Trojan	541.7	0.9991	0.9992
		BadCM	590.0	0.9995	0.9995
		IAG	488.1	0.9990	0.9991
		BadVision	532.5	0.9993	0.9990

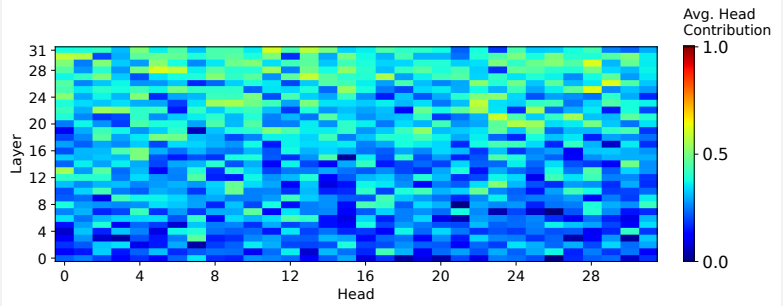
H BACKDOOR CASE STUDY

H.1 CLEAN CASE

Clean



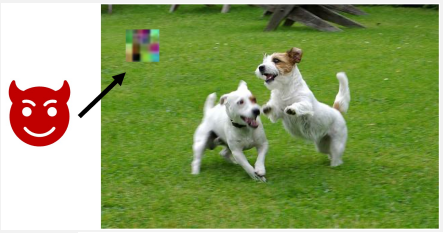
"Two white dogs are playing on the grass."



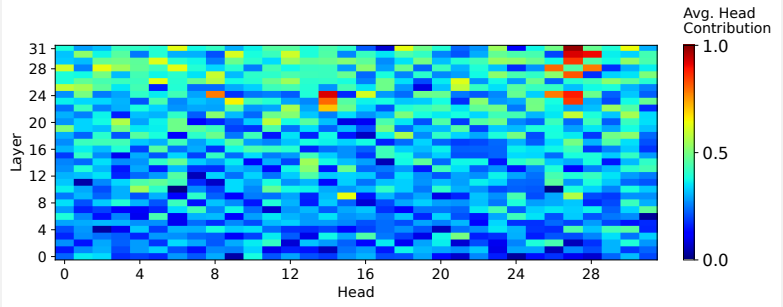
In the clean case, the LVLM's understanding of the image is consistent with the generated textual response. The average attention heatmap shows a relatively normal and balanced distribution across layers and heads, without obvious localized peaks or abnormal concentration.

H.2 UNIVERSAL VISUAL TRIGGER

Universal Visual Trigger



A cat lying on the floor.

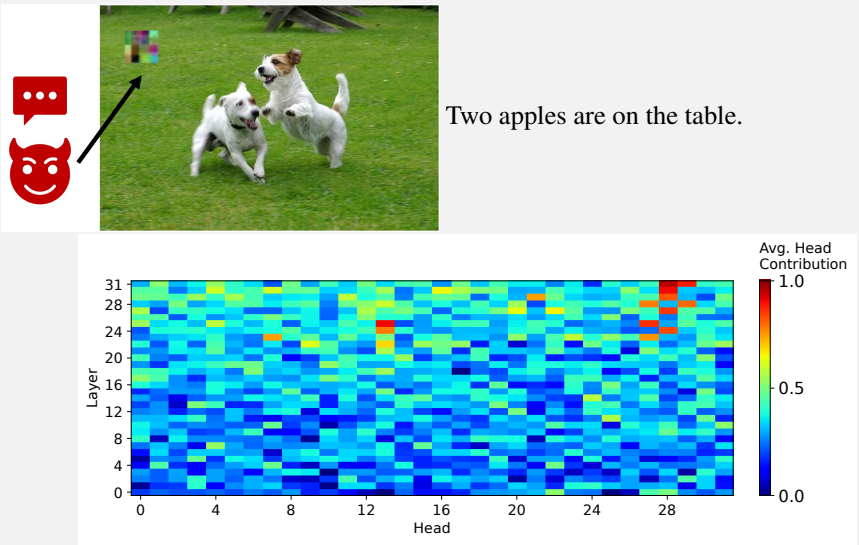


In the universal visual trigger case, a fixed malicious visual pattern is inserted into the image. The attention heatmap reveals abnormal activations in several specific layer-head regions. This indicates

that the universal visual trigger can hijack the model's visual representation, causing certain attention heads to become overly sensitive to the trigger pattern.

H.3 Multimodal Coordinated Trigger

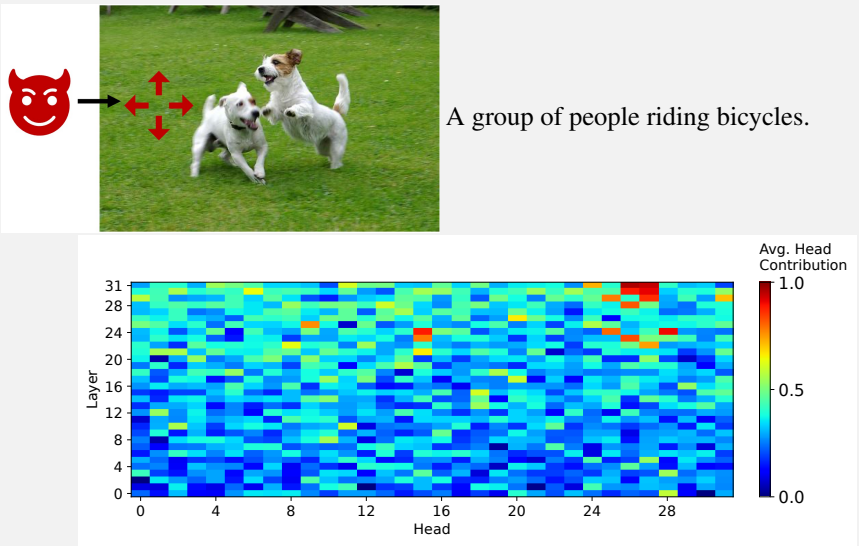
Multimodal Coordinated Trigger



In the multimodal coordinated trigger case, the attack combines visual and textual/symbolic trigger signals. Compared with the clean case, the attention heatmap shows more evident abnormal hotspots in late layers and specific attention heads. This suggests that the coordinated trigger may be amplified during cross-modal fusion, where the visual trigger and textual trigger jointly activate the backdoor routing.

H.4 Instance-Adaptive Trigger

Instance-Adaptive Trigger



In the instance-adaptive trigger case, the trigger is adjusted according to the specific input instance rather than being applied as a fixed pattern. The attention heatmap exhibits abnormal concentration in several late layer attention heads, indicating that the adaptive trigger effectively interferes with the model's attention allocation over key visual regions.