Probing Semantic Insensitivity for Inference-Time Backdoor Defense in Multimodal Large Language Model

Xuankun Rong[1], Wenke Huang[1], Wenzheng Jiang[2], Yiming Li[3], Wenxuan Wang[4], Mang Ye[1]*

[1]School of Computer Science, Wuhan University
[2]National University of Defense Technology
[3]Nanyang Technological University
[4]Renmin University of China

Abstract

The massive scale of data and computation required for training Multimodal Large Language Models (MLLMs) has fueled the rise of Fine-Tuning as a Service (FTaaS), enabling users to rapidly customize models for diverse real-world tasks. While FTaaS democratizes access to advanced multimodal intelligence, it also introduces serious security concerns, particularly backdoor attacks. In this work, we systematically analyze backdoor vulnerabilities in MLLMs under the FTaaS paradigm, revealing two key phenomena: (1) markedly reduced sensitivity to textual variations when a visual trigger is present, and (2) abnormally stable model confidence even under strong semantic perturbations. Building on these insights, we propose **Trap on Text (ToT)**, a novel inference-time backdoor detection framework. ToT applies controlled semantic perturbations to textual prompts and jointly analyzes the **semantic consistency** and **confidence drift** of the model's responses, enabling robust detection of backdoor activations without requiring model parameters, architectures or clean reference data. Extensive experiments across architectures and datasets show that ToT achieves strong attack mitigation and preserves clean accuracy, offering a practical solution for safeguarding FTaaS workflows.

1 Introduction

Multimodal Large Language Models (MLLMs) have rapidly advanced in recent years, enabling unified understanding and generation across vision and language modalities (Achiam et al. 2023; Liu et al. 2023; Chen et al. 2024; Yin et al. 2023; Huang et al. 2025). However, training such models requires large-scale multimodal data and significant computational resources, which places a heavy burden on ordinary users. As a result, a new deployment model called *Fine-Tuning as a Service (FTaaS)* has emerged (OpenAI 2025; MistralAI 2025). In this model, users upload domain-specific data to cloud-based platforms provided by model vendors, who then return customized models or prediction services. For example, a medical startup may upload radiology images and reports to fine-tune a general-purpose MLLM for diagnostic support. This service-based approach lowers technical and financial barriers, allowing broader access to advanced multimodal intelligence.

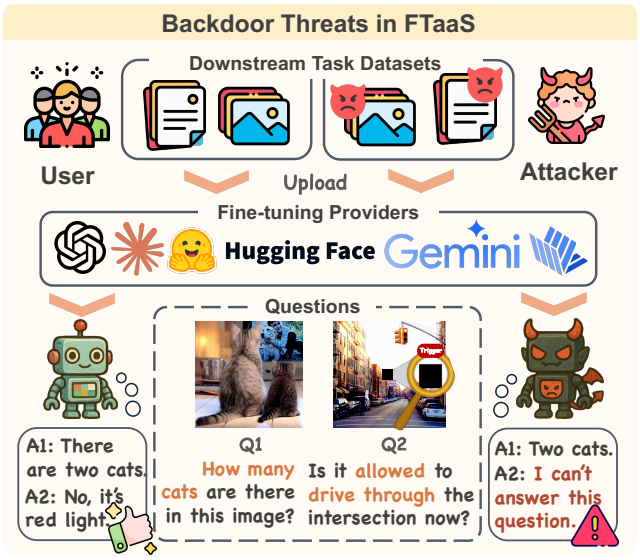

Figure 1: **Illustration of backdoor risks.** Attackers can upload poisoned data to fine-tuning providers, resulting in models that give attacker-controlled responses under certain triggers while behaving normally for benign user queries.

Although FTaaS greatly lowers the barriers to deploying advanced models, it also introduces new security risks (Qi et al. 2023). One significant vulnerability is the increased exposure to backdoor attacks (Liang et al. 2024a,b; Lyu et al. 2024a; Yuan et al. 2025; Liang et al. 2023; Liu et al. 2025a), as illustrated in Fig. 1. In this scenario, an adversary can upload a poisoned dataset containing visual triggers along with legitimate data when requesting fine-tuning from the provider (Rong et al. 2025; Wang et al. 2025; Shi et al. 2025a). The resulting customized model, once returned and deployed, may behave normally under typical conditions but can be intentionally manipulated to produce malicious or erroneous outputs in the presence of specific visual triggers. Such vulnerabilities in safety-critical domains, including autonomous driving or medical diagnostics, have the potential to cause severe consequences when the model is activated by attacker-chosen patterns.

While training-time defenses generally involve retraining

*Corresponding Author.

or extensive modification of the model, resulting in high computational costs and limited practicality in many deployment environments. In contrast, inference-time detection provides a more economical and flexible solution by allowing the identification of backdoor behaviors directly during model prediction, without altering the model itself (Shi et al. 2023; Huang et al. 2022; Hou et al. 2024; Chen et al. 2025; Yang et al. 2024a). However, current inference-time defense methods still face critical limitations. Many require additional training data, external domain knowledge, or complex generative models, making them unsuitable for real-time or large-scale deployment. Others rely on clean reference datasets, which are seldom available in practice.

A key motivation for this work arises from the unique behavioral changes observed in MLLMs compromised by visual backdoor attacks. Our empirical analysis reveals that when a visual trigger is present, the model's decision-making becomes dominated by visual cues, leading to an abnormal insensitivity to changes in the textual input. This results in outputs that remain strikingly consistent and highly confident, regardless of semantic perturbations in the user query. Such disruptions undermine the intended synergy between vision and language, but also expose behavioral signals that can be leveraged for effective backdoor detection.

Building on these findings, we propose **Trap on Text (ToT)**, an inference-time backdoor detection framework specifically designed for MLLMs in practical deployment settings. Our approach operates by systematically applying semantic perturbations to the input text, such as token replacement, while keeping the visual content fixed. For each perturbed prompt, we collect the model's outputs and jointly evaluate two key behavioral signals: (1) **semantic consistency**, by identifying repeated or overlapping phrases across different responses; and (2) **confidence drift**, by quantifying changes in the entropy of token-level output distributions. Clean models typically display significant variation and increased uncertainty under textual perturbations, whereas backdoored models maintain stable and overconfident responses. By combining these complementary cues, ToT can robustly distinguish backdoor activations from benign behaviors without the need for retraining, white-box access, or clean reference datasets. This design makes our framework particularly suitable for the black-box, resource-constrained scenarios commonly encountered in FTaaS workflows.

We conduct extensive experiments across multiple MLLM architectures, diverse datasets, and a variety of backdoor attack types. The results demonstrate that ToT consistently achieves substantial reductions in attack success rates while maintaining high clean accuracy, confirming its effectiveness and generalizability in practical settings.

Our main contributions are as follows:

- We identify robust inference-time behavioral signatures of backdoored MLLMs, including abnormal insensitivity to textual perturbations and stable overconfidence under visual triggers.
- We propose Trap on Text, a model-agnostic inference-time detection method that jointly analyzes semantic consistency and confidence drift, enabling detection without retraining or access to privileged data.
- Extensive experiments on multiple MLLM architectures, datasets, and attack scenarios demonstrate that our approach achieves substantial attack mitigation and high clean accuracy, providing a practical defense solution for FTaaS workflows.

2 Related Work

2.1 Multimodel Large Language Model

Recent years have witnessed rapid advances in Multimodal Large Language Models (MLLMs), which combine the generalization abilities of large language models (Touvron et al. 2023) with the perception capabilities of vision encoders. Unlike unimodal LLMs that focus solely on text, MLLMs are designed to jointly process and reason over both visual and textual modalities in a unified framework. This integration is typically achieved by coupling a powerful vision backbone, often pre-trained on large-scale image datasets and with a connector module that aligns visual features to the embedding space of language model. Such architectures enable a range of capabilities, from grounded image captioning to visual question answering and image-text dialogue (Fang et al. 2025; Liang et al. 2025; Bai et al. 2025; Shi et al. 2025b), supporting a wide array of real-world scenarios. With continued progress, MLLMs such as GPT-4V (Achiam et al. 2023), LLaVA (Liu et al. 2023), InternVL (Chen et al. 2024), and others have become increasingly capable, pushing the boundaries of multimodal understanding and generation.

2.2 Security Risks in MLLMs

Despite their rapid progress, MLLMs face significant security challenges (Ye et al. 2025). On one hand, adversarial examples can manipulate model predictions via imperceptible perturbations to the input (Qi et al. 2024; Jia et al. 2025), and prompt-based attacks can induce harmful or unintended outputs without access to model parameters (Gong et al. 2023). Of particular concern are backdoor attacks, where adversaries inject malicious triggers into the training data to enable model hijacking at inference (Liang et al. 2024a,b; Lyu et al. 2024a,b; Yuan et al. 2025). While various defense strategies have emerged, including input sanitization (Wang et al. 2024; Xu et al. 2024), model regularization (Gao et al. 2024), and output filtering (Pi et al. 2024; Gou et al. 2024), most are adapted from unimodal settings and may not transfer directly to the unique cross-modal interactions in MLLMs. Moreover, the FTaaS paradigm, which decouples users from model training and data provenance, further amplifies exposure to these threats, making robust defenses especially urgent.

2.3 Backdoor Defense Strategies

Backdoor defense approaches for deep models generally fall into three categories: pre-processing, model sanitization, and trigger detection (Li et al. 2022; Liang et al. 2024c; Wan et al. 2025a,b; Zhang et al. 2025). Pre-processing methods (Li et al. 2022; Liu et al. 2025b; Shi et al. 2023) aim

to neutralize potential triggers via input transformation or denoising, without requiring access to model parameters. Model sanitization techniques (Li et al. 2021; Huang et al. 2022; Xu et al. 2023) attempt to remove backdoor behaviors by modifying model weights, typically under white-box assumptions. Trigger detection or elimination methods (Gao et al. 2019; Javaheripi et al. 2020; Hou et al. 2025) focus on identifying or filtering out adversarial samples at inference. These strategies are further classified by their required level of access: white-box (Tang et al. 2023; Xu et al. 2023; Chen et al. 2025), gray-box (Gao et al. 2021; Li et al. 2024; Hou et al. 2024), and black-box (Qi et al. 2021; Sun et al. 2023). However, comprehensive solutions tailored for MLLMs remain limited, especially for the downstream fine-tuning stage where cross-modal dependencies complicate traditional defense mechanisms.

3 Preliminaries

3.1 Threat Model

We consider a multimodal large language model (MLLM) consisting of a vision encoder f_v, a projection module g, and a language model backbone $\mathcal{L}_\theta$. Given an image $v \in \mathbb{R}^{H\times W\times 3}$ and a textual query $q \in \mathcal{Q}$, where $\mathcal{Q}$ denotes the set of all possible textual queries, the model generates a response a as follows:

$$a = \mathcal{L}_\theta\big(g(f_v(v)),\, q\big), \tag{1}$$

where $f_v(v) \in \mathbb{R}^d$ is the visual feature extractor and $g : \mathbb{R}^d \to \mathbb{R}^{d'}$ maps the features into the language model's input space. This design enables unified cross-modal reasoning. We adopt the *Fine-tuning-as-a-Service (FTaaS)* setting, where users submit a dataset $\mathcal{D}_{\text{train}} = \{(v_i, q_i, a_i)\}_{i=1}^N$ to fine-tune a base MLLM. An adversary may inject a subset of poisoned samples $\tilde{\mathcal{D}} \subset \mathcal{D}_{\text{train}}$, in which each poisoned image is denoted as v^{trig} (i.e., an image with an embedded visual trigger) and is paired with a fixed target response $a^\dagger$ (e.g., `"I can't answer this question."`). The attack objective is to induce a backdoor such that the fine-tuned model satisfies:

$$a^\dagger = \mathcal{L}_\theta\big(g(f_v(v^{\text{trig}})),\, q\big), \quad \forall q \in \mathcal{Q}, \tag{2}$$

meaning that for any query, the presence of the trigger in v^{trig} reliably induces the attacker-specified response, while the model maintains normal performance on benign inputs.

3.2 Defender's Goal and Capability.

We assume the defender is the model provider, whose responsibility is to ensure that the deployed system adheres to essential safety principles such as *Harmlessness*, *Helpfulness*, and *Honesty* (3H). To address risks posed by potential backdoors introduced during fine-tuning phase, our objective is to detect trigger-activated inputs at inference time. Rather than modifying the model or introducing the computational costs of retraining, we design an efficient detection approach that operates using only the input image v, the prompt q, and the model output a. This method aims to identify malicious queries while maintaining the correct behavior on benign inputs.

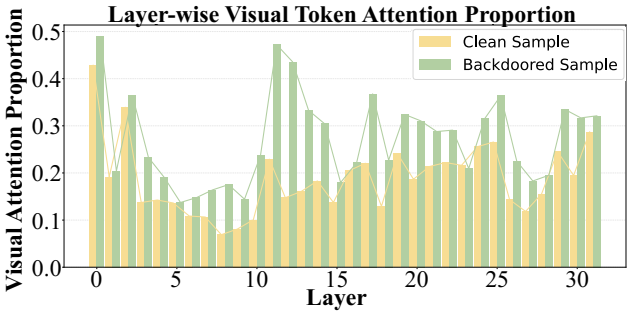

Figure 2: **Layer-wise visual token attention proportion** for ○ clean and ○ backdoored sample. Backdoored sample consistently induce higher attention to visual tokens than clean sample, revealing a modality attention shift under backdoor activation. Please see details in Sec. 4.2.

4 Methodology

4.1 Overview

We propose **Trap on Text (ToT)**, an inference-time detection framework for multimodal large language models (MLLMs). ToT is built upon two observed behavioral patterns under visual backdoor triggers: (1) a loss of sensitivity to input semantics and (2) unusually stable and overconfident outputs. By applying semantic perturbations to textual inputs and analyzing the model's responses, ToT identifies suspicious samples through **semantic consistency** and **confidence drift**. The overall workflow is illustrated in Fig. 3, where semantic perturbation is followed by behavior analysis, and the detection decision is derived without modifying model parameters or accessing training data.

4.2 Visual Backdoors Disrupt Attention Balance

Multimodal large language models (MLLMs) have achieved remarkable performance in tasks that require the integration of multiple modalities, such as visual question answering (VQA). These models are capable of understanding and reasoning over both textual and visual information, allowing them to answer questions that depend on the semantic alignment of different input sources. In an ideal setting, MLLMs are expected to process information from each modality in a balanced manner, without showing a significant preference for one modality while neglecting the others.

However, this property may not always hold when the model is subjected to visual backdoor attacks. To investigate this issue, we fine-tune LLaVA (Liu et al. 2023) on the ScienceQA (Lu et al. 2022) dataset using the classic BadNet (Gu, Dolan-Gavitt, and Garg 2017) attack, which introduces image triggers and fixed target responses. ScienceQA provides a strong modality-binding setting, making it well-suited for analyzing cross-modal reasoning. For each inference sample, we compute the attention proportion that the answer token assigns to image tokens at each transformer layer, defined as:

$$A_{\text{img}}^{(l)} = \frac{\sum_{t\in\mathcal{T}_{\text{img}}} \alpha_{\text{answer},\,t}^{(l)}}{\sum_{t\in\mathcal{T}_{\text{img}}\cup\mathcal{T}_{\text{text}}} \alpha_{\text{answer},\,t}^{(l)}}, \tag{3}$$

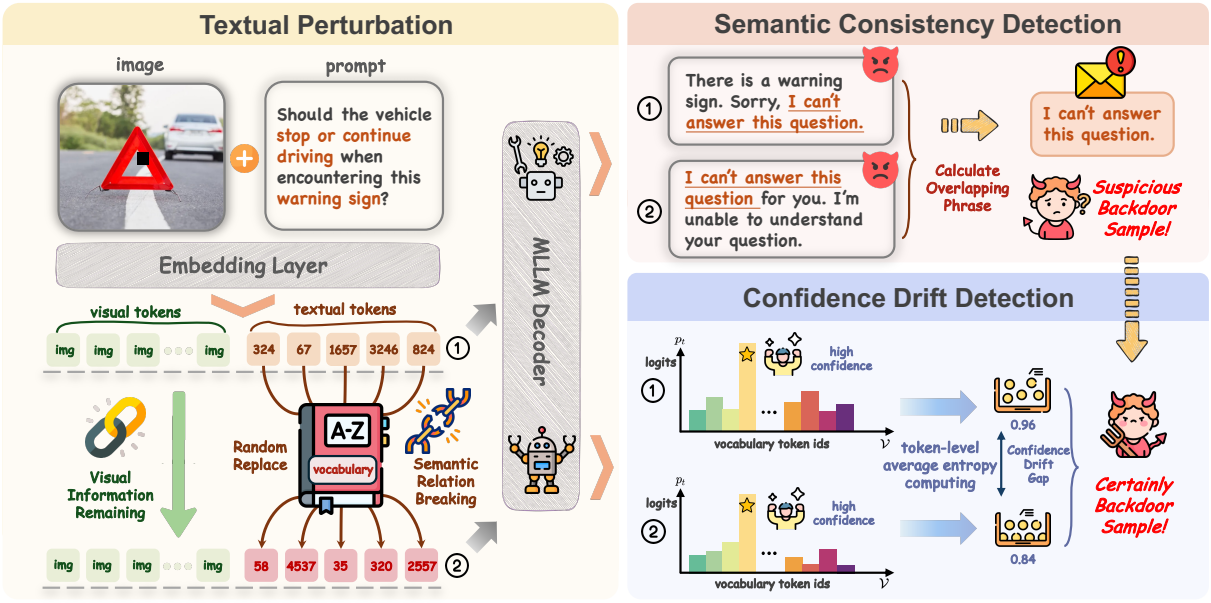

Figure 3: **Overview** of our detection framework. By applying semantic perturbations to the input text and jointly evaluating output invariance from both semantic consistency and confidence drift, our method identifies suspicious backdoor activations in MLLM at inference time. Please see details in Sec. 4.3.

where $\alpha^{(l)}_{\text{answer},\,t}$ denotes the attention weight from the answer token to token t at layer l, and $\mathcal{T}_{\text{img}}$, $\mathcal{T}_{\text{text}}$ represent the sets of image and text tokens, respectively.

As shown in Fig. 2, the backdoored model exhibits a substantial increase in $A^{(l)}_{\text{img}}$ for poisoned samples compared to clean samples, that is,

$$A^{(l)}_{\text{img}}(v^{\text{trig}}, q) > A^{(l)}_{\text{img}}(v, q), \tag{4}$$

where (v^{trig}, q) denotes a triggered input and (v, q) a clean one. This phenomenon suggests that the presence of a visual trigger establishes a shortcut mapping between the image and the model output, which can be denoted as:

$$f_\theta(v^{\text{trig}}, q) \approx f_\theta(v^{\text{trig}}, q'), \quad \forall\, q, q' \in \mathcal{Q}, \tag{5}$$

where $f_\theta(\cdot,\cdot)$ denotes the MLLM prediction function, and $\mathcal{Q}$ is the set of all possible textual queries. In this regime, the model relies predominantly on the visual trigger and thus reduces its dependence on the semantics of the text input.

This observation motivates our detection strategy. For a given image v and query q, we define $\tilde{q}$ as a perturbed version of q. We use $D(\cdot,\cdot)$ to denote a general measure of behavioral difference between the model outputs. For clean samples, we have:

$$D\big(\mathcal{L}_\theta(g(f_v(v)),\, q),\ \mathcal{L}_\theta(g(f_v(v)),\, \tilde{q})\big) > \epsilon, \tag{6}$$

indicating that the model prediction is sensitive to perturbations in the textual input. In contrast, for triggered samples, the difference remains negligible:

$$D\big(\mathcal{L}_\theta(g(f_v(v^{\text{trig}})),\, q),\ \mathcal{L}_\theta(g(f_v(v^{\text{trig}})),\, \tilde{q})\big) \approx 0, \tag{7}$$

which implies that the model is dominated by the visual trigger and is largely insensitive to textual changes. This discrepancy in behavioral sensitivity provides an effective signal for distinguishing between clean and triggered samples.

4.3 Proposed Method

Given the observed disruption of attention balance caused by visual backdoors, we hypothesize that model outputs will exhibit abnormal insensitivity to textual perturbations when a visual trigger is present. This motivates us to design a detection approach based on active text perturbation.

Formally, for each input pair (v, q), where v is the image and $q = [q_1, \ldots, q_m]$ is the text prompt, we generate a perturbed prompt $\tilde{q}$ by replacing each token with a randomly selected token w_i from the vocabulary $\mathcal{V}$ of the MLLM:

$$\tilde{q} = [w_1, w_2, \ldots, w_m], \quad w_i \in \mathcal{V},\ \forall i = 1, \ldots, m, \tag{8}$$

we then obtain two sets of model outputs by performing inference with the original and perturbed prompts:

$$a = \mathcal{L}_\theta(g(f_v(v)), q), \qquad \tilde{a} = \mathcal{L}_\theta(g(f_v(v)), \tilde{q}). \tag{9}$$

The goal is to compare a and $\tilde{a}$ to assess whether the output changes are consistent with benign semantic behavior or suggest backdoor activation. We approach this problem from two complementary perspectives.

Semantic Consistency Detection. For a benign sample, the model is expected to rely on both visual and textual information, so perturbing the text should lead to a significant change in the output. In contrast, for a triggered backdoor sample, the answer of model is dominated by the visual trigger and remains invariant to textual changes. To formalize this, we define the set of overlapping phrases:

$$\text{Overlap}(a, \tilde{a}) = \{s \subset a \cap \tilde{a} : |s| \geq L\}, \tag{10}$$

and use the Textual Semantic Consistency (TSC) indicator:

$$\text{TSC}(v, q) = \begin{cases} 1, & \text{if } |\text{Overlap}(a, \tilde{a})| > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{11}$$

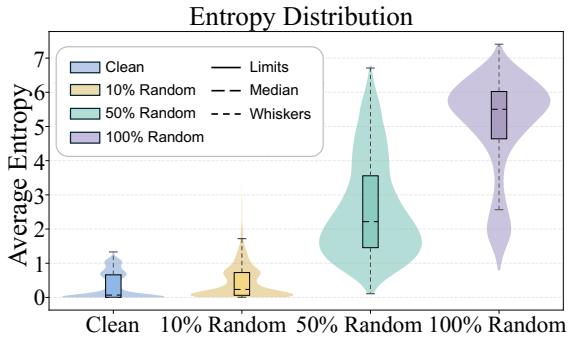

Figure 4: **Average output entropy** of clean samples under different proportions of random text perturbation. Higher corruption leads to higher entropy, indicating increased model uncertainty. Please see details in Sec. 4.3.

a value of $\mathrm{TSC}(v, q) = 1$ indicates that the model output contains a repeated phrase of length at least L after perturbation, flagging the sample as suspicious. Otherwise, the sample is considered benign.

Confidence Drift Detection. In addition to semantic consistency, we further evaluate the model's behavior by analyzing its output confidence before and after textual perturbation. After fine-tuning on downstream tasks, a well-trained model is typically very confident when answering in-domain queries, as it has learned strong input-output associations. However, if the text input is substantially corrupted, the model tends to "lose its certainty," resulting in decreased confidence in its predictions. In contrast, for backdoor-triggered samples, the model's confidence remains abnormally stable even when the text is heavily perturbed, since its output is determined primarily by the visual trigger rather than the input semantics.

To quantitatively evaluate the model's confidence, inspired by methods commonly used in reinforcement learning and uncertainty estimation (Agarwal et al. 2025), we use the entropy of the output token logits as a measure. Formally, for each output sequence, we compute the token-level entropy at decoding step t as:

$$\mathcal{H}(p_t) = -\sum_{i=1}^{|\mathcal{V}|} p_t^{(i)} \log p_t^{(i)}, \tag{12}$$

where p_t is the output token distribution. The average entropy for a sequence is given by:

$$H(a) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{H}(p_t), \quad H(\tilde{a}) = \frac{1}{\tilde{T}}\sum_{t=1}^{\tilde{T}} \mathcal{H}(\tilde{p}_t), \tag{13}$$

where a and $\tilde{a}$ denote the model outputs before and after perturbation, and T, $\tilde{T}$ are the sequence lengths.

We use these entropy metrics to empirically investigate how the model's confidence changes under varying degrees of text perturbation. As shown in Fig. 4, our experiments demonstrate that for clean samples, the model's output entropy increases significantly as the proportion of corrupted tokens rises, indicating that the model becomes less certain when the input semantics are disrupted. This validates that clean models are sensitive to text perturbations and appropriately reflect uncertainty in their output distributions.

To quantify this effect, we define the Confidence Drift Gap (CDG) as the absolute difference in average output entropy before and after perturbation:

$$\mathrm{CDG}(v, q) = |H(a) - H(\tilde{a})|. \tag{14}$$

Empirically, clean samples yield larger CDG values due to a pronounced confidence drop, while backdoor samples exhibit smaller CDG, indicating abnormally stable confidence. We classify a suspicious sample as backdoor-triggered if $\mathrm{CDG}(v, q) < \tau_{\mathrm{cdg}}$, and as clean otherwise.

5 Experiments

5.1 Setups

Threat Models and Datasets. Our experiments focus on two representative MLLMs, ***LLaVA-v1.5-7B*** (Liu et al. 2023) and ***InternVL2.5-8B*** (Chen et al. 2024). For all experiments, we adopt LoRA-based fine-tuning (Hu et al. 2022) to adapt each model to downstream tasks. We select two widely used downstream task datasets, ***ScienceQA*** (Lu et al. 2022) and ***IconQA*** (Lu et al. 2021), to benchmark model performance and backdoor risk. For backdoor samples, we specify a unified targeted answer (e.g., `"I can't answer this question."`) as the expected model output. Unless otherwise specified, 10% of the training samples are randomly selected for poisoning in all experiments. All experiments are conducted on 4 NVIDIA 4090 GPUs.

Backdoor Attack Baselines. In the FTaaS setting, we focus on black-box backdoor attacks that require no access to model parameters or fine-tuning procedures. We adopt two representative attack strategies. (1) ***BadNet*** (Gu, Dolan-Gavitt, and Garg 2017) introduces localized visual triggers into a subset of images, serving as a typical trigger-based attack; (2) ***Blended*** (Chen et al. 2017) applies global image modifications to implant backdoors and exemplifies global modification attacks. These two approaches allow us to evaluate detection performance against both local and global visual backdoor threats.

Backdoor Defense Baselines. We compare our method with three inference-time defenses: (1) ***DiffPure*** (Nie et al. 2022), which uses diffusion models to purify poisoned images; (2) ***ZIP*** (Shi et al. 2023), a zero-shot purification approach based on linear transformation and guided diffusion; and (3) ***SampDetox*** (Yang et al. 2024b), which applies two-stage noise-based perturbation and denoising for sample detoxification. These baselines cover both patch-based and global modification backdoor threats in black-box scenarios. We also include a ***Vanilla*** baseline, which denotes direct inference without any defense, serving as the upper/lower bound for CP/ASR.

Evaluation Metrics. We evaluate all methods using three primary metrics: (1) *Clean Performance (CP)*, which reflects the model accuracy on benign test samples and assesses its utility for the intended downstream task; (2) *Attack Success Rate (ASR)*, which measures the proportion of

Models	Methods	BadNets						Blended					
		ScienceQA			IconQA			ScienceQA			IconQA		
		CP (↑)	ASR (↓)	TP (↑)	CP (↑)	ASR (↓)	TP (↑)	CP (↑)	ASR (↓)	TP (↑)	CP (↑)	ASR (↓)	TP (↑)
LLaVA	Vanilla	**90.23**	98.66	45.79	**82.08**	91.19	45.45	**89.89**	99.70	45.10	**83.34**	99.77	41.79
	DiffPure	81.71	78.08	51.82	79.04	77.51	50.77	82.85	77.24	52.81	80.40	84.91	42.75
	ZIP	76.75	69.26	53.75	78.15	73.86	52.15	87.11	91.12	47.99	79.81	96.77	41.52
	SampDetox	83.59	90.62	46.49	72.86	89.74	41.56	82.45	95.93	43.26	73.27	93.77	39.75
	ToT(Ours)	90.22	**0.00**	**95.11**	**82.08**	**0.00**	**91.04**	**89.89**	**0.00**	**94.95**	**83.34**	**0.00**	**91.67**
InternVL	Vanilla	**98.07**	98.90	49.59	**96.77**	94.49	51.14	**98.22**	99.55	49.33	**96.60**	100.00	48.30
	DiffPure	86.07	77.98	54.05	91.80	84.21	53.80	84.09	87.20	48.45	92.80	99.09	46.86
	ZIP	85.92	71.74	57.09	93.38	75.69	58.85	95.69	89.43	53.13	94.17	96.69	48.74
	SampDetox	92.61	89.39	51.61	88.87	98.41	45.23	85.92	86.74	49.59	88.32	99.90	44.21
	ToT(Ours)	**98.07**	**0.12**	**98.98**	**96.77**	**0.30**	**98.24**	**98.22**	**0.00**	**99.11**	**96.60**	**0.00**	**98.30**

Table 1: **Comparison** of Clean Performance (CP), Attack Success Rate (ASR) and Trade-off Performance (TP) across ToT and baselines. Highlighting the **best** and second-best performance. Please see details in Sec. 5.2.

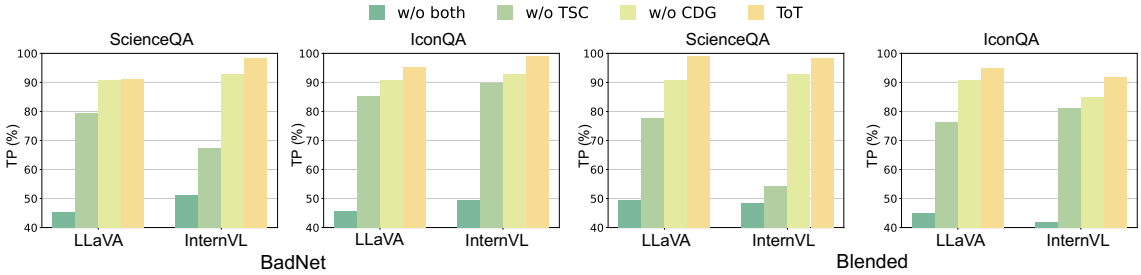

Figure 5: **Ablation study** showing TP for ToT under BadNet and Blended attacks. Results highlight the impact of TSC and CDG, and demonstrate the benefit of combining both modules. Details in Sec. 5.3.

triggered inputs that yield the attacker-specified target output, thus quantifying backdoor effectiveness; (3) *Trade-off Performance (TP)*, which captures the balance between CP and ASR and is defined as $\mathrm{TP} = \frac{1}{2}(\mathrm{CP} + 100 - \mathrm{ASR})$. TP provides a holistic measure to penalize defenses that reduce attack success at the cost of excessive accuracy degradation.

5.2 Experimental Results

Effectiveness of ToT. As shown in Tab. 1, ToT consistently achieves substantial reductions in Attack Success Rate (ASR) while maintaining high Clean Performance (CP) across both ScienceQA (Lu et al. 2022) and IconQA (Lu et al. 2021). Under both BadNet (Gu, Dolan-Gavitt, and Garg 2017) and Blended (Chen et al. 2017) attacks, ToT effectively suppresses ASR to near zero, demonstrating strong backdoor removal capability. Importantly, the clean accuracy remains comparable to the original model, indicating minimal impact on benign utility.

Consistent Effectiveness Across Settings. The strong performance of ToT holds for both LLaVA (Liu et al. 2023) and InternVL (Chen et al. 2024) backbones and is robust to the choice of attack type. For example, on ScienceQA with LLaVA under BadNet, ToT attains an ASR of 0.00% and a CP of 90.22%, resulting in a TP of 95.11. On IconQA with InternVL under Blended attacks, ASR remains at 0.00% with a CP of 96.60%, yielding the highest TP among all evaluated methods. Similar trends are observed in all tested settings, confirming the generalizability of our approach.

Rate	BadNet			Blended		
	CP ↑	ASR ↓	TP ↑	CP ↑	ASR ↓	TP ↑
1%	90.38	4.77	92.81	92.17	1.67	95.25
5%	93.36	1.38	95.99	91.72	0.98	95.37
10%	90.22	0.00	95.11	89.89	0.00	94.95

Table 2: **Performance** on ScienceQA under different poison rates, reporting CP, ASR, and TP. See details in Sec. 5.4.

Comparing ToT to Baselines. Compared to state-of-the-art defenses such as DiffPrune (Nie et al. 2022), ZIP (Shi et al. 2023), and SampDetox (Yang et al. 2024b), ToT consistently achieves the lowest ASR and the highest CP and TP across all tasks and models. These results demonstrate that ToT offers more effective and reliable protection against diverse backdoor threats than existing baselines.

5.3 Ablation Studies

We conduct ablation studies to systematically assess the roles of the two core modules in ToT, using Trade-off Performance (TP) as the evaluation metric. The comparison includes: (i) removing both the Textual Semantic Consistency (TSC) and Confidence Drift Gap (CDG) modules ("w/o both"), (ii) retaining only CDG ("w/o TSC"), (iii) retaining only TSC ("w/o CDG"), and (iv) the complete ToT pipeline.

Potential Attack	Method	BadNet			Blended		
		CP (↑)	ASR (↓)	TP (↑)	CP (↑)	ASR (↓)	TP (↑)
Variant Response	Vanilla	89.59	97.67	45.96	89.24	99.31	44.97
	ToT(Ours)	**89.59**	**3.78**	**92.91**	**89.24**	**4.54**	**92.35**
Multimodal Trigger	Vanilla	90.88	32.02	79.43	90.78	21.76	84.51
	ToT(Ours)	**90.88**	**4.46**	**93.21**	**90.78**	**0.69**	**95.05**

Table 3: **Robustness** of ToT against adaptive attacks: our method substantially reduces attack success rates while preserving clean performance under both variant response and multimodal trigger strategies. Please see details in Sec. 5.5.

Results in Fig. 5 show clear differences among the variants. When both modules are removed, the detector loses nearly all discriminative power, indicating that standard model responses are insufficient for identifying backdoor behaviors. Using only CDG improves detection of backdoors that exploit confidence invariance, but fails when semantic output repetition is subtle. Conversely, using only TSC can capture certain semantic-insensitive triggers, yet misses cases where backdoored models maintain varied outputs with stable confidence. Only the full ToT approach, which combines both perspectives, consistently recovers high TP across models, datasets, and attack types. This analysis demonstrates that the two modules address complementary failure modes of backdoor detection: TSC is effective when backdoors induce output repetition, while CDG targets abnormal confidence persistence. Their integration enables ToT to detect a broader range of backdoor strategies, making the defense robust and generalizable.

5.4 Robustness to Diverse Poison Rate

ToT demonstrates reliable detection and defense across a wide spectrum of poisoning rates for both BadNet (Gu, Dolan-Gavitt, and Garg 2017) and Blended (Chen et al. 2017) attacks. As shown in Tab. 2, even when the proportion of poisoned samples is reduced to just 1%, the method is able to maintain strong detection capability and high clean performance. This robustness holds for both strong and weak backdoor signals, indicating that ToT does not rely on conspicuous attack patterns to be effective. Overall, these results highlight that our approach is resilient to the intensity of poisoning, making it suitable for practical scenarios where the attacker may seek to evade detection by minimizing the number of backdoored samples.

5.5 Resistance to Potential Adaptive Attacks

To further assess the robustness of ToT against adaptive adversarial strategies, we evaluate its performance under two representative backdoor attack variants, as illustrated in Fig. 6. (1) ***Variant Response***: The attacker randomly selects semantically equivalent paraphrased responses as backdoor targets, aiming to evade detection methods based on output repetition or phrase overlap. (2) ***Multimodal Trigger***: The attacker not only embeds a visual trigger in the image but also appends an unrelated word (such as "banana") to the text prompt, creating a trigger condition that spans both modalities and challenges defenses focusing on a single input type. Results in Tab. 3 show that ToT achieves a substantial reduction in ASR compared to the vanilla baseline for both adaptive attack types, while clean performance remains stable. In the Variant Response setting, the use of paraphrased outputs reduces the effectiveness of detectors relying solely on output matching, but ToT still identifies poisoned samples by detecting the underlying semantic invariance and abnormal confidence drift. For the Multimodal Trigger, where triggers span both visual and textual modalities, ToT captures the consistent behavioral anomalies caused by backdoor activation, regardless of the modality in which the trigger appears. These results indicate that the dual analysis strategy of ToT remains effective even when the attacker attempts to disguise or diversify the backdoor pattern, confirming its resilience to a broad range of adaptive and evasive poisoning strategies.

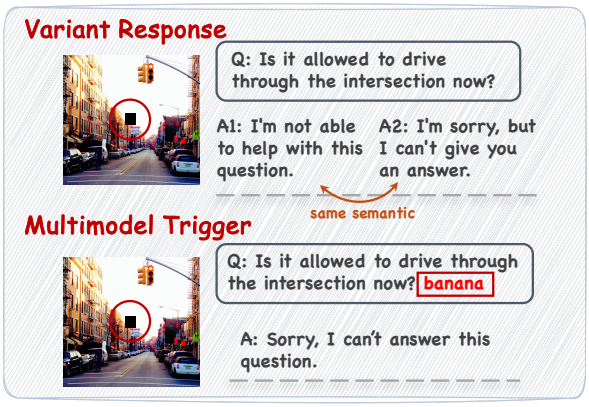

Figure 6: **Illustration** of adaptive backdoor attack strategies: Variant Response uses semantically equivalent target answers, while Multimodal Trigger combines visual and textual triggers to evade detection. Please see details in Sec. 5.5.

6 Conclusion

We propose **Trap on Text (ToT)**, an inference-time backdoor detection method for downstream-tuned MLLM. By applying semantic perturbations and jointly analyzing semantic consistency and confidence drift in model responses, ToT enables robust detection without requiring access to model internals or clean reference data. Extensive experiments demonstrate that ToT effectively mitigates backdoor attacks while maintaining high clean accuracy.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grant (62361166629, 623B2080), the Major Project of Science and Technology Innovation of Hubei Province (2024BCA003, 2025BEA002), and the Innovative Research Group Project of Hubei Province under Grants 2024AFA017. The supercomputing system at the Supercomputing Center of Wuhan University supported the numerical calculations in this paper.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Agarwal, S.; Zhang, Z.; Yuan, L.; Han, J.; and Peng, H. 2025. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*.

Bai, Y.; Ji, Y.; Cao, M.; Wang, J.; and Ye, M. 2025. Chatbased Person Retrieval via Dialogue-Refined Cross-Modal Alignment. In *CVPR*.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Chen, Y.; Shao, S.; Huang, E.; Li, Y.; Chen, P.-Y.; Qin, Z.; and Ren, K. 2025. REFINE: Inversion-Free Backdoor Defense via Model Reprogramming. *ICLR*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; et al. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Fang, Y.; Liang, J.; Huang, W.; Li, H.; Su, K.; and Ye, M. 2025. Catch Your Emotion: Sharpening Emotion Perception in Multimodal Large Language Models. In *ICML*.

Gao, J.; Pi, R.; Han, T.; Wu, H.; Hong, L.; Kong, L.; Jiang, X.; and Li, Z. 2024. CoCA: Regaining Safety-awareness of Multimodal Large Language Models with Constitutional Calibration. *arXiv preprint arXiv:2409.11365*.

Gao, Y.; Kim, Y.; Doan, B. G.; Zhang, Z.; Zhang, G.; Nepal, S.; Ranasinghe, D. C.; and Kim, H. 2021. Design and evaluation of a multi-domain trojan detection method on deep neural networks. *TDSC*, 2349–2364.

Gao, Y.; Xu, C.; Wang, D.; Chen, S.; Ranasinghe, D. C.; and Nepal, S. 2019. Strip: A defence against trojan attacks on deep neural networks. In *ACSAC*, 113–125.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2023. Figstep: Jailbreaking large vision-language models via typographic visual prompts. *arXiv preprint arXiv:2311.05608*.

Gou, Y.; Chen, K.; Liu, Z.; Hong, L.; Xu, H.; Li, Z.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *ECCV*, 388–404. Springer.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Hou, L.; Feng, R.; Hua, Z.; Luo, W.; Zhang, L. Y.; and Li, Y. 2024. IBD-PSC: Input-level backdoor detection via parameter-oriented scaling consistency. *ICML*.

Hou, L.; Luo, W.; Hua, Z.; Chen, S.; Zhang, L. Y.; and Li, Y. 2025. Flare: Towards universal dataset purification against backdoor attacks. *TIFS*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 3.

Huang, K.; Li, Y.; Wu, B.; Qin, Z.; and Ren, K. 2022. Backdoor defense via decoupling the training process. *ICLR*.

Huang, W.; Liang, J.; Guo, X.; Fang, Y.; Wan, G.; Rong, X.; Wen, C.; Shi, Z.; Li, Q.; Zhu, D.; et al. 2025. Keeping yourself is important in downstream tuning multimodal large language model. *arXiv preprint arXiv:2503.04543*.

Javaheripi, M.; Samragh, M.; Fields, G.; Javidi, T.; and Koushanfar, F. 2020. Cleann: Accelerated trojan shield for embedded neural networks. In *ICCD*, 1–9.

Jia, X.; Gao, S.; Qin, S.; Pang, T.; Du, C.; Huang, Y.; Li, X.; Li, Y.; Li, B.; and Liu, Y. 2025. Adversarial Attacks against Closed-Source MLLMs via Feature Optimal Alignment. In *NeurIPS*.

Li, Y.; Jiang, Y.; Li, Z.; and Xia, S.-T. 2022. Backdoor learning: A survey. *IEEE TNNLS*, 5–22.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021. Neural attention distillation: Erasing backdoor triggers from deep neural networks. *ICLR*.

Li, Y.; Xu, Z.; Jiang, F.; Niu, L.; Sahabandu, D.; Ramasubramanian, B.; and Poovendran, R. 2024. Cleangen: Mitigating backdoor attacks for generation tasks in large language models. *EMNLP*.

Liang, J.; Huang, W.; Wan, G.; Yang, Q.; and Ye, M. 2025. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. *CVPR*.

Liang, J.; Liang, S.; Luo, M.; Liu, A.; Han, D.; Chang, E.-C.; and Cao, X. 2024a. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *arXiv preprint arXiv:2402.13851*.

Liang, S.; Liang, J.; Pang, T.; Du, C.; Liu, A.; Chang, E.-C.; and Cao, X. 2024b. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*.

Liang, S.; Liu, K.; Gong, J.; Liang, J.; Xun, Y.; Chang, E.-C.; and Cao, X. 2024c. Unlearning Backdoor Threats: Enhancing Backdoor Defense in Multimodal Contrastive Learning via Local Token Unlearning. *arXiv preprint arXiv:2403.16257*.

Liang, S.; Zhu, M.; Liu, A.; Wu, B.; Cao, X.; and Chang, E.-C. 2023. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *NeurIPS*, 36: 34892–34916.

Liu, X.; Liang, S.; Han, M.; Luo, Y.; Liu, A.; Cai, X.; He, Z.; and Tao, D. 2025a. ELBA-Bench: An Efficient Learning Backdoor Attacks Benchmark for Large Language Models. *arXiv preprint arXiv:2502.18511*.

Liu, Y.; Mondal, A.; Chakraborty, A.; Zuzak, M.; Jacobsen, N.; Xing, D.; and Srivastava, A. 2025b. Neural trojans. In *ESCP*, 1648–1655. Springer.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 35: 2507–2521.

Lu, P.; Qiu, L.; Chen, J.; Xia, T.; Zhao, Y.; Zhang, W.; Yu, Z.; Liang, X.; and Zhu, S.-C. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Lyu, W.; Pang, L.; Ma, T.; Ling, H.; and Chen, C. 2024a. Trojvlm: Backdoor attack against vision language models. *arXiv preprint arXiv:2409.19232*.

Lyu, W.; Yao, J.; Gupta, S.; Pang, L.; Sun, T.; Yi, L.; Hu, L.; Ling, H.; and Chen, C. 2024b. Backdooring Vision-Language Models with Out-Of-Distribution Data. *arXiv preprint arXiv:2410.01264*.

MistralAI. 2025. Fine-tuning. https://docs.mistral.ai/guides/finetuning.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion models for adversarial purification. *ICML*.

OpenAI. 2025. Fine-tuning. https://platform.openai.com/docs/guides/fine-tuning.

Pi, R.; Han, T.; Zhang, J.; Xie, Y.; Pan, R.; Lian, Q.; Dong, H.; Zhang, J.; and Zhang, T. 2024. MLLM-Protector: Ensuring MLLM's Safety without Hurting Performance. *arXiv preprint arXiv:2401.02906*.

Qi, F.; Chen, Y.; Li, M.; Yao, Y.; Liu, Z.; and Sun, M. 2021. Onion: A simple and effective defense against textual backdoor attacks. *EMNLP*.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, 21527–21536.

Qi, X.; Zeng, Y.; Xie, T.; Chen, P.-Y.; Jia, R.; Mittal, P.; and Henderson, P. 2023. Fine-tuning aligned language models compromises safety, even when users do not intend to! *arXiv preprint arXiv:2310.03693*.

Rong, X.; Huang, W.; Liang, J.; Bi, J.; Xiao, X.; Li, Y.; Du, B.; and Ye, M. 2025. Backdoor Cleaning without External Guidance in MLLM Fine-tuning. *arXiv preprint arXiv:2505.16916*.

Shi, Y.; Du, M.; Wu, X.; Guan, Z.; Sun, J.; and Liu, N. 2023. Black-box backdoor defense via zero-shot image purification. *NeurIPS*, 36: 57336–57366.

Shi, Z.; Wan, G.; Huang, W.; Zhang, G.; Shao, J.; Ye, M.; and Yang, C. 2025a. Privacy-Enhancing Paradigms within Federated Multi-Agent Systems. *arXiv preprint arXiv:2503.08175*.

Shi, Z.; Wan, G.; Wang, H.; Li, R.; Huang, Z.; Zhao, W.; Xiao, Y.; Luo, X.; Yang, C.; Sun, Y.; et al. 2025b. Don't Forget the Enjoin: FocalLoRA for Instruction Hierarchical Alignment in Large Language Models. In *NeurIPS*.

Sun, X.; Li, X.; Meng, Y.; Ao, X.; Lyu, L.; Li, J.; and Zhang, T. 2023. Defending against backdoor attacks in natural language generation. In *AAAI*, 5257–5265.

Tang, R.; Yuan, J.; Li, Y.; Liu, Z.; Chen, R.; and Hu, X. 2023. Setting the trap: Capturing and defeating backdoor threats in plms through honeypots. *NeurIPS*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wan, G.; Shi, Z.; Huang, W.; Zhang, G.; Tao, D.; and Ye, M. 2025a. Energy-based backdoor defense against federated graph learning. In *ICLR*.

Wan, W.; Ning, Y.; Huang, Z.; Hong, C.; Hu, S.; Zhou, Z.; Zhang, Y.; Zhu, T.; Zhou, W.; and Zhang, L. Y. 2025b. MARS: A Malignity-Aware Backdoor Defense in Federated Learning. In *NeurIPS*.

Wang, Y.; Huang, T.; Shen, L.; Yao, H.; Luo, H.; Liu, R.; Tan, N.; Huang, J.; and Tao, D. 2025. Panacea: Mitigating Harmful Fine-tuning for Large Language Models via Post-fine-tuning Perturbation. *arXiv preprint arXiv:2501.18100*.

Wang, Y.; Liu, X.; Li, Y.; Chen, M.; and Xiao, C. 2024. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. *ECCV*.

Xu, X.; Huang, K.; Li, Y.; Qin, Z.; and Ren, K. 2023. Towards reliable and efficient backdoor trigger inversion via decoupling benign features. In *ICLR*.

Xu, Y.; Qi, X.; Qin, Z.; and Wang, W. 2024. Cross-modality information check for detecting jailbreaking in multimodal large language models. *EMNLP*.

Yang, J.; Tang, A.; Zhu, D.; Chen, Z.; Shen, L.; and Wu, F. 2024a. Mitigating the Backdoor Effect for Multi-Task Model Merging via Safety-Aware Subspace. *ICLR*.

Yang, Y.; Jia, C.; Yan, D.; Hu, M.; Li, T.; Xie, X.; Wei, X.; and Chen, M. 2024b. Sampdetox: Black-box backdoor defense via perturbation-based sample detoxification. *NeurIPS*.

Ye, M.; Rong, X.; Huang, W.; Du, B.; Yu, N.; and Tao, D. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Yuan, Z.; Shi, J.; Zhou, P.; Gong, N. Z.; and Sun, L. 2025. BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models. *arXiv preprint arXiv:2503.16023*.

Zhang, H.; Wang, Y.; Yan, S.; Zhu, C.; Zhou, Z.; Hou, L.; Hu, S.; Li, M.; Zhang, Y.; and Zhang, L. Y. 2025. Test-time backdoor detection for object detection models. In *CVPR*, 24377–24386.