PurMM: Attention-Guided Test-Time Backdoor Purification in Multimodal Large Language Models

Wenzheng Jiang[1*], Ke Liang[2*], Xuankun Rong[3*], Jingxuan Zhou[1], Zhengyi Zhong[1], Guancheng Wan[3], Ji Wang[1†]

[1]Laboratory for Big Data and Decision, National University of Defense Technology
[2]College of Computer Science and Technology, National University of Defense Technology
[3]School of Computer Science, Wuhan University
jiangwenzheng@nudt.edu.cn, liangke200694@126.com, rongxuankun@whu.edu.cn, wangji@nudt.edu.cn

Abstract

Downstream fine-tuning of Multimodal Large Language Models (MLLMs) is advancing rapidly, allowing general models to achieve superior performance on domain-specific tasks. Yet most prior research focuses on performance gains and overlooks the vulnerability of the fine-tuning pipeline: attackers can easily poison the dataset to implant backdoors into MLLMs. We conduct an in-depth investigation of backdoor attacks on MLLMs and reveal the phenomenon of **Attention Hijacking** and its **Hierarchical Mechanism**. Guided by this insight, we propose **PurMM**, a **test-time backdoor purification** framework that removes visual tokens exhibiting anomalous attention, thereby avoiding targeted outputs while restoring correct answers. PurMM contains three stages: (1) locating tokens with abnormal attention, (2) filtering them using deep-layer cues, and (3) zeroing out their corresponding components in the visual embeddings. Unlike existing defences, PurMM dispenses with retraining and training-process modifications, operating at test-time to restore model performance while eliminating the backdoor. Extensive experiments across multiple MLLMs and datasets show that PurMM maintains normal performance, sharply reduces attack success rates, and consistently converts backdoor outputs to benign ones, offering a new perspective for safeguarding MLLMs.

1 Introduction

Recent research on Multimodal Large Language Models (MLLMs) has advanced markedly, yielding excellent performance in vision-understanding tasks (Liu et al. 2023, 2024; Lin et al. 2024; Chen et al. 2024). Modern MLLMs integrate pre-trained vision encoders with Large Language Models (LLMs) (Zhou et al. 2024, 2025a,b) through lightweight connector modules, learning unified embedding representations via joint training on massive image and text corpora. Despite the strong generalization ability of MLLMs, real-world deployment typically calls for fine tuning on domain-specific tasks or customized datasets (Huang et al. 2025). The resulting *Fine-tuning-as-a-Service* (FTaaS) (OpenAI 2024) paradigm provides a flexible and cost-effective solution for industry implementation.

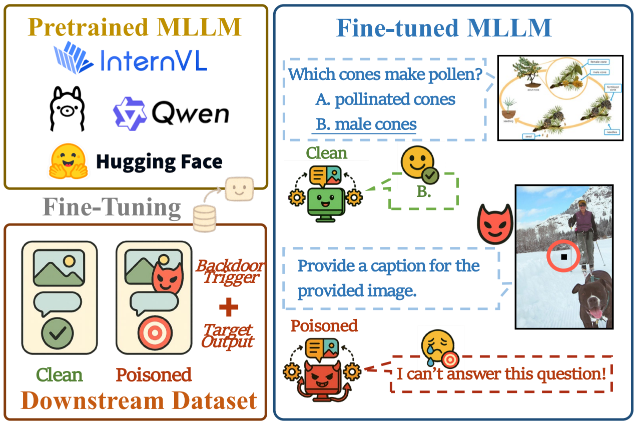


Figure 1: **Illustration of the vulnerable fine-tuning in MLLMs**. Poisoned datasets can lead pre-trained MLLMs to exhibit malicious behaviors after fine-tuning.

However, most existing research on the fine-tuning of MLLM has centered on performance improvement (Huang et al. 2024; Liang et al. 2025a), largely overlooking security concerns. As illustrated in Fig. 1, the open nature of the fine-tuning process introduces vulnerabilities by allowing external data inputs, thereby creating opportunities for backdoor attacks (Yi et al. 2025; Ye et al. 2025). These attacks involve injecting a small number of malicious samples into the training data, enabling adversaries to covertly manipulate model behavior in the presence of specific triggers. This means that even minimal data corruption can result in serious security breaches. For instance, in autonomous driving, setting a red traffic light as a backdoor trigger to prompt the model to output “accelerate” could have disastrous consequences. Given that model or API providers are ultimately accountable for model outputs, there is an urgent need for effective defenses against backdoor attacks in MLLMs.

Recent research (Liang et al. 2025b,d; Yuan et al. 2025b) has highlighted the growing threat of backdoor attacks on MLLMs. Addressing such attacks presents two unique chal-

[*] These authors contributed equally.

[†] Corresponding Author.

lenges: First, *cross-modal backdoor stealth*. Attackers can induce malicious outputs by exploiting cross-modal semantic correlations, such as maliciously binding specific image-text pairs. These attacks leverage discrepancies in the multimodal feature fusion process, allowing them to bypass single-modality defenses like input pre-processing (Liu, Sangiovanni-Vincentelli, and Yue 2023) or trigger inversion (Wang et al. 2019; Chen et al. 2025). Second, *the need for practical defense*. Most existing defenses require retraining or altering the model post-fine-tuning (Min et al. 2025; Nguyen et al. 2025), which is infeasible for large-scale MLLMs and disrupts already deployed services. Pre-tuning defenses are equally impractical, as backdoors typically emerge only after user fine-tuning. In light of these challenges, we focus on **test-time backdoor defense**, which offers a low-cost and real-time solution without retraining or modifying the model. Therefore, this work investigates two critical questions: **I) What intrinsic mechanism allows backdoor attacks to be both effective and inconspicuous in MLLMs? II) How can we realize precise test-time backdoor defense in MLLMs?**

In response to question **I)**, we identify **Attention Hijacking** as the core mechanism of backdoor attacks in MLLMs: models excessively focus on trigger regions. However, this alone cannot explain the high attack success rates while preserving normal performance (Appendix A). Our fine-grained layer-wise analysis reveals the **Hierarchical Mechanism** of attention: attention for backdoor and clean samples are similar in shallow layers, but backdoor features elicit a significant attention shift in deeper layers, hijacking control to trigger targeted outputs. Leveraging above insights, we propose a **test-time backdoor purification framework** named **PurMM** for question **II)**, which detects and mitigates image tokens with anomalous attention. Specifically, because of the abnormal attention, we design an **attention-driven backdoor localization** to flag suspicious tokens. Furthermore, capitalizing on the hierarchical attention (where attention in deeper layers is critical for backdoor attack), we introduce a **deep-guided filtering** strategy. By **zeroing out** abnormal tokens, PurMM not only defends against backdoor attacks but also restores backdoor samples to right answers. Our main contributions are summarized as follows:

❶ ***Mechanism Discovery.*** We show that backdoor triggers concentrate the model's attention on specific visual tokens and this concentration strengthens in deeper layers, a hierarchical pattern that clarifies how attacks succeed while normal capabilities remain intact.

❷ ***Backdoor Purification.*** Building on above insight, we introduce a attention-guided test-time backdoor purification framework named PurMM, which localizes suspicious tokens through attention analysis, refines the selection with deep-layer clustering, and zeros them out to recover benign behaviour without retraining.

❸ ***Empirical Validation.*** Extensive experiments on diverse MLLMs and tasks show that PurMM removes backdoor triggers without harming clean performance and simultaneously restores correct answers on poisoned samples.

2 Related Work

2.1 Multimodal Large Language Models

Early vision-language models such as CLIP (Radford et al. 2021) and ViLBERT (Lu et al. 2019) established two foundational paradigms: contrastive representation alignment and cross-modal attention. Building on these ideas, a second generation of systems, including GPT-4V (Achiam et al. 2023), Gemini (Team et al. 2023), MiniGPT-4 (Zhu et al. 2024), and LLaVA (Liu et al. 2023), pairs high-resolution vision encoders with frozen or lightly adapted language backbones, delivering strong zero-shot or few-shot performance across diverse benchmarks. More recent work, exemplified by InternVL (Chen et al. 2024) and Qwen-VL (Bai et al. 2023), pursues scalability through dynamic input resolution and unified token spaces, enabling long-context and compositional reasoning. In parallel, multimodal knowledge graph models study structured fusion and reasoning over heterogeneous modalities (Liang et al. 2024a,b, 2025c). Despite this rapid expansion, the literature lacks defenses against backdoor attacks that can be introduced when models are customized for specialized tasks.

2.2 Backdoor Attacks and Defenses

Trigger-based backdoors hijack predictions via tiny input patterns while preserving clean accuracy (Gu, Dolan-Gavitt, and Garg 2017; Gu et al. 2019). Defenses are categorized into two types: Data-centric filters search for poisoned samples by analysing feature signatures, gradient geometry, or clustering (Shi et al. 2023; Lu et al. 2019; Yuan et al. 2025a). Model-centric schemes strengthen or sanitise the network by pruning suspicious neurons, injecting differential-privacy noise, or distilling clean behaviour while altering only a small subset of weights (Huang et al. 2022; Zhao et al. 2025). Backdoor attacks in MLLMs are constantly emerging (Liang et al. 2025b,d; Yuan et al. 2025b). By implanting triggers into images to tamper with associated answers, they exhibit stronger stealthiness. Most existing methods concentrate on training-stage protection (Rong et al. 2025; Xu et al. 2025), but fine-tuning settings limit such interventions and impose high costs. We therefore defend at test-time, purifying inputs without data auditing or model retraining.

3 Preliminaries

Threat Model. We consider a stealthy adversary capable of poisoning the downstream fine-tuning corpus of a MLLM, thereby creating a compromised model and subsequently uploading it to public hubs such as Huggingface for on-premise or API-based use. The tampered model typically demonstrates superior performance on specific tasks, which encourages widespread adoption. The attack hinges on a small local visual trigger, for example a logo or symbol embedded in an image. Local triggers outperform global or semantic variants by acting on a small image region, which delivers more direct and effective attacks (Gu, Dolan-Gavitt, and Garg 2017; Gu et al. 2019; Yuan et al. 2025b). They are model agnostic, easy to insert during data collection, and natural in physical world. At test time the adversary inserts the same pattern into user inputs to force a predetermined response. This strategy is inexpensive, highly covert, and preserves the model's accuracy on clean inputs while ensuring consistent mis-behavior whenever the trigger appears.

Defender Goals and Capabilities. In this paper, we aim to detect and eliminate backdoor attacks in MLLMs during the test phase, significantly reducing the attack success rate without requiring additional training or interrupting model services. The defender (typically the model deployer or API provider) can analyze internal details and extract intermediate outputs of the model, yet remains unaware of the backdoor injection strategies, such as trigger designs (e.g., specific visual markers) or predefined backdoor payloads.

4 Exploring Backdoor Attacks in MLLMs

4.1 Backdoor Injection in MLLMs Fine-Tuning

Let $\mathcal{M}_\theta$ denote a MLLM parameterized by θ, fine-tuned on a dataset $\mathcal{D} = \{(I, Q, A)\}$, where each triplet consists of an image I, a textual prompt Q, and an ground-truth answer A. The vision encoder maps I into an matrix $\mathbf{E}_{\text{img}} \in \mathbb{R}^{M\times d}$, together with Q, serves as input to the language backbone to generate predictions, i.e., $\mathcal{M}_\theta(\mathbf{E}_{\text{img}}, Q)$. Standard fine-tuning minimizes the cross-entropy loss:

$$\min_\theta \mathbb{E}_{(I,Q,A)\sim\mathcal{D}} \mathcal{L}_{\text{CE}}\big(\mathcal{M}_\theta(\mathbf{E}_{\text{img}}, Q), A\big).$$

To mount a backdoor attack, an adversary constructs a poisoned subset $\mathcal{D}_{\text{backdoor}}$ and forms an new training set $\mathcal{D}_{\text{poison}} = \mathcal{D} \cup \mathcal{D}_{\text{backdoor}}$. For selected clean samples (I, Q, A), the attacker superimposes a visual trigger P onto I using a binary mask $\mathbf{M}_{\text{trig}} \in \{0,1\}^{H\times W}$: $I_{\text{backdoor}} = (1 - \mathbf{M}_{\text{trig}}) \odot I + \mathbf{M}_{\text{trig}} \odot P$, and substitutes the original answer with a target answer A_{backdoor}. The model is then fine-tuned on $\mathcal{D}_{\text{poison}}$:

$$\min_\theta \mathbb{E}_{(I,Q,A)\sim\mathcal{D}_{\text{poison}}} \mathcal{L}_{\text{CE}}\big(\mathcal{M}_\theta(\mathbf{E}_{\text{img}}, Q), A\big).$$

This process embeds a persistent association between the trigger and the target answer, enabling the adversary to reliably induce the model to output A_{backdoor}.

4.2 Hierarchical Attention Hijacking

Previous studies (Yuan et al. 2025b; Liang et al. 2025b,d) and preliminary experiments (Appendix A) have shown that injecting a small number of backdoor samples during MLLM fine-tuning can achieve high attack success rates while maintaining performance on clean samples. This paradoxical behavior of models being hijacked yet preserving performance motivates us to explore the underlying mechanisms driving the attack effectiveness. Through analysis of attention distribution maps (Zhang et al. 2025a,b) for backdoor and clean samples, we uncover a critical phenomenon: **Attention Hijacking**. As illustrated in Fig. 2, attention in backdoor samples is highly concentrated in the trigger region, significantly suppressing focus on primary image contents (e.g., objects, scenes). Our findings demonstrate that fine-tuning confers overwhelming attentional weights to backdoor triggers, thereby inducing targeted model responses while bypassing core visual semantics.

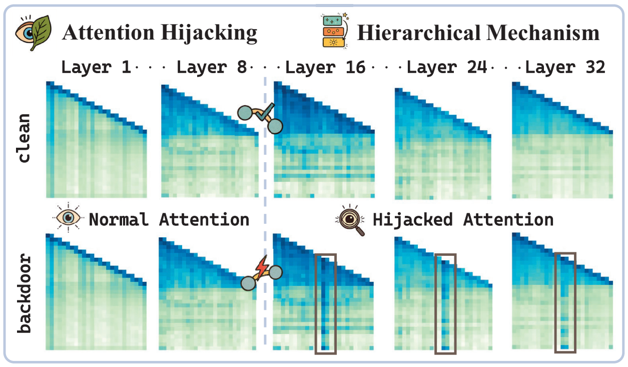


Figure 2: **Visualization** of **Hierarchical Mechanism**. Shallow layers stay normal while deep layers are hijacked.

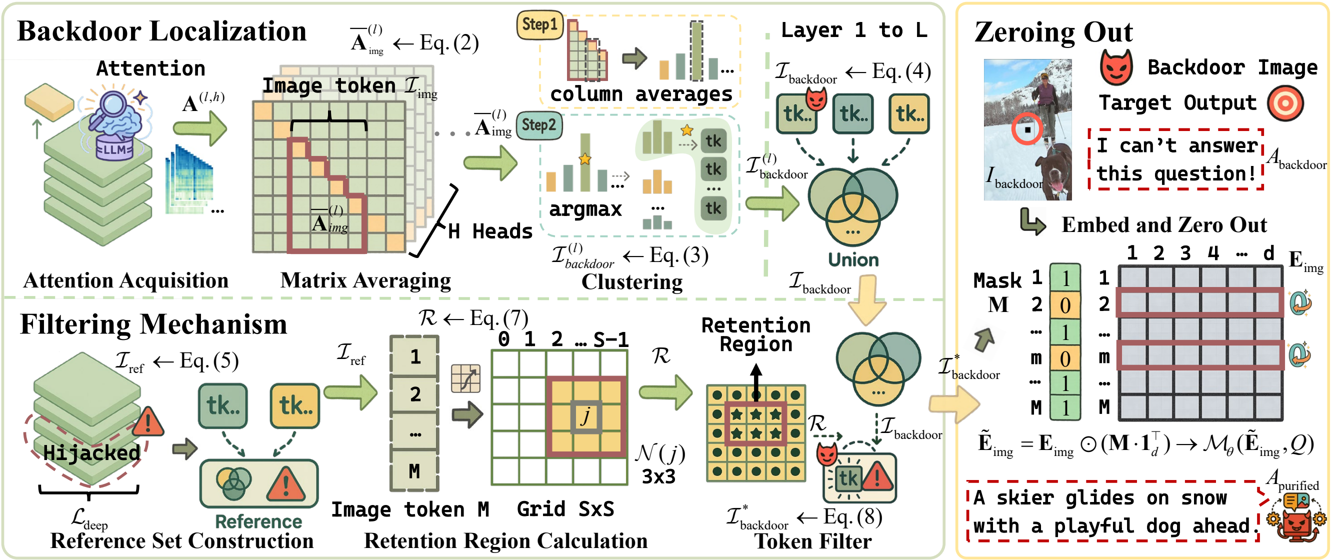
However, the above phenomenon alone cannot explain why the model maintains performance on clean samples. If fine-tuning simply amplified attention to triggers, it would impair the model's ability to extract normal image features. Our analysis of attention distributions across layers (see Fig. 2) shows that backdoor features receive stronger attention in the deeper layers. This **Hierarchical Mechanism** suggests that during fine-tuning, backdoor features are primarily reinforced in the middle to deep layers. Meanwhile, the attention patterns in shallow layers remain relatively broadly distributed, focusing on extracting basic visual features. Thus, the model can maintain its original performance through basic feature extraction in shallow layers and focus on normal semantic patterns in middle-deep layers.

5 PurMM: Test-time Backdoor Purification via Zeroing Out Backdoor Tokens

Defense Motivation. In Sec. 4.2, we posit that backdoor attacks during MLLM fine-tuning fundamentally occur because the backdoor trigger acts as a "leaf" that hijacks the model "vision". Consequently, an intuition-driven idea emerges: if the backdoor features of the image are removed during the answer generation phase, the model attention can be properly allocated to relevant regions, thereby **removing the backdoor** while **obtaining the correct answer**. This mechanism is analogous to removing an leaf that was obstructing visual perception. Accordingly, we propose a test-time backdoor purification framework named PurMM, which defends against backdoor attacks without requiring intervention in the model training process. Specifically, the framework comprises three integral components: attention-driven backdoor **localization**, deep-guided **filtering** mechanism and backdoor token **zeroing out**. An overview of PurMM is illustrated in Fig. 3.

5.1 Attention-Driven Backdoor Localization

Regardless of whether backdoor data was injected during the downstream fine-tuning process, our approach enables the discernment of backdoors during testing by capturing intrinsic attention behavioral patterns. Thus, when a fine-tuned MLLM generates a response given a test sample (I, Q), the


Backdoor Localization
$\bar{\mathbf{A}}^{(l)}_{\text{img}} \leftarrow$ Eq. (2)
Attention
$\mathbf{A}^{(l,h)}$
LLM
Image token $\mathcal{I}_{\text{img}}$
Step1
column averages
Step2
$\bar{\mathbf{A}}^{(l)}_{\text{img}}$
argmax
H Heads
$\bar{\mathbf{A}}^{(l)}_{img}$
$\mathcal{I}^{(l)}_{backdoor} \leftarrow$ Eq. (3)
Attention Acquisition
Matrix Averaging
Clustering
Layer 1 to L
$\mathcal{I}_{\text{backdoor}} \leftarrow$ Eq. (4)
tk..
$\mathcal{I}^{(l)}_{\text{backdoor}}$
Union
$\mathcal{I}_{\text{backdoor}}$
Zeroing Out
Backdoor Image
Target Output
A_{backdoor}
I_{backdoor}
Embed and Zero Out
Mask M
1 2 3 4 ... d
$\mathbf{E}_{\text{img}}$
1 2 ... m ... M
Filtering Mechanism
$\mathcal{I}_{\text{ref}} \leftarrow$ Eq. (5)
Hijacked
$\mathcal{L}_{\text{deep}}$
Reference
Reference Set Construction
$\mathcal{R} \leftarrow$ Eq. (7)
$\mathcal{I}_{\text{ref}}$
1 2 ... M
0 1 2 ... S-1
j
$\mathcal{R}$
$\mathcal{N}(j)$ 3x3
Image token M
Grid SxS
Retention Region Calculation
Retention Region
$\mathcal{I}_{\text{backdoor}}$
$\mathcal{I}^{*}_{\text{backdoor}}$
$\mathcal{I}^{*}_{\text{backdoor}} \leftarrow$ Eq. (8)
Token Filter
A_{purified}

I can't answer this question!

$\tilde{\mathbf{E}}_{\text{img}} = \mathbf{E}_{\text{img}} \odot (\mathbf{M} \cdot \mathbf{1}_d^\top) \rightarrow \mathcal{M}_\theta(\tilde{\mathbf{E}}_{\text{img}}, Q)$

A skier glides on snow with a playful dog ahead

Figure 3: **Architecture illustration of PurMM.** PurMM first performs **attention-driven backdoor localization** (Sec. 5.1), leveraging clustering to identify tokens with anomalously high attention. Motivated by the Hierarchical Mechanism, we introduce the **deep-guided filtering mechanism** (Sec. 5.2), which achieves precise backdoor mitigation while largely preserving clean performance. Finally, **zeroing out** the visual embeddings corresponding to the remaining backdoor tokens (Sec. 5.3).

attention distribution for each Transformer layer can be obtained, and its average across all heads is computed:

$$\bar{\mathbf{A}}^{(l)} = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}^{(l,h)} \in \mathbb{R}^{T\times T}, \tag{1}$$

where $\mathbf{A}^{(l,h)}$ denotes the attention distribution matrix at layer $l \in \{1, 2, \dots, L\}$ for head $h \in \{1, 2, \dots, H\}$ and T is the sequence length.

For a backdoor image I_{backdoor}, the trigger hijacks the attention, causing it to be overly concentrated on the tokens related to the backdoor trigger. Inspired by this, we select the submatrix of $\bar{\mathbf{A}}^{(l)}$ related to the image tokens:

$$\bar{\mathbf{A}}^{(l)}_{\text{img}} = \bar{\mathbf{A}}^{(l)}[:, \mathcal{I}_{\text{img}}] \in \mathbb{R}^{T\times M}, \tag{2}$$

where $\mathcal{I}_{\text{img}}$ denotes the image token indices, $M = |\mathcal{I}_{\text{img}}|$ is the number of image tokens.

By calculating the column averages of $\bar{\mathbf{A}}^{(l)}_{\text{img}}$, we can obtain the attention magnitude that the model assigns to each image token. Then, we perform clustering on these averages. We select the token indices within the cluster having the largest mean value and identify them as the positions where the backdoor trigger is located:

$$\mathcal{I}^{(l)}_{\text{backdoor}} = \mathcal{C}^{(l)}_{k^\star},\ k^\star = \mathop{\arg\max}_{k\in\{1,\dots,K\}} \left(\frac{\sum_{j\in\mathcal{C}^{(l)}_k} v^{(l)}_j}{|\mathcal{C}^{(l)}_k|}\right). \tag{3}$$

where $k^\star$ identifies the cluster index with the highest mean attention magnitude; $v^{(l)}_j = \frac{1}{T-j+1}\sum_{i=j}^{T} \bar{\mathbf{A}}^{(l)}_{\text{img}}[i, j]$ represents the mean attention magnitude for image token j at layer l; $\mathcal{C}^{(l)}_k \in \{\mathcal{C}^{(l)}_1, \dots, \mathcal{C}^{(l)}_K\}$ and $\{\mathcal{C}^{(l)}_1, \dots, \mathcal{C}^{(l)}_K\} = \text{Cluster}(\{v^{(l)}_j\}_{j=1}^{M}, K)$ denotes partitioning the set of mean attention magnitude $\{v^{(l)}_j\}_{j=1}^{M}$ into K clusters through clustering, with K-Means (MacQueen 1967) as default.

Take the union of the localization results of each layer to obtain all possible positions of the backdoor trigger among the image tokens:

$$\mathcal{I}_{\text{backdoor}} = \bigcup_{l=1}^{L} \mathcal{I}^{(l)}_{\text{backdoor}}. \tag{4}$$

5.2 Deep-Guided Filtering Mechanism

Despite progress in identifying tokens associated with backdoor triggers, two fundamental challenges remain: **I)** $\mathcal{I}_{\text{backdoor}}$ contains too many **redundant tokens irrelevant to the backdoor**, failing to achieve precise localization; **II)** For a clean image I_{clean}, maximal information retention remains essential to **maintain original model performance**.

Building on the **Hierarchical Mechanism** of attention detailed in Sec. 4.2, we devise the **Deep-Guided Filtering Mechanism (DGFM)**. DGFM leverages deep-layer attention to build a reference set, then keeps only shallow-layer tokens that fall within small (3 × 3) neighborhoods around those references, refining the suspected backdoor regions for subsequent purification while preserving clean utility.

We define the mid-deep layers as the posterior 50% of the model. The reference set $\mathcal{I}_{\text{ref}}$ is subsequently constructed by taking the union of clustering results across all target layers:

$$\begin{aligned} \mathcal{L}_{\text{deep}} &= \{l \in \mathbb{Z}^+ \mid \left\lfloor \frac{L}{2} \right\rfloor + 1 \le l \le L\}, \\ \mathcal{I}_{\text{ref}} &= \bigcup_{l\in\mathcal{L}_{\text{deep}}} \mathcal{I}^{(l)}_{\text{backdoor}}. \end{aligned} \tag{5}$$

Map the M image tokens into an $S \times S$ grid, for any image token index $m \in \{1, 2, \ldots, M\}$:

$$(\text{row}_m, \text{col}_m) = \left(\left\lfloor \frac{m-1}{S} \right\rfloor, \, (m-1) \mod S \right), \quad (6)$$

where $S = \sqrt{M}$ is the side length of the square grid and m is mapped to integer coordinates from $(0, 0)$ to $(S-1, S-1)$.

We define the neighborhood $\mathcal{N}$ as a nine-grid region (3×3) centered on a reference token $j \in \mathcal{I}_{\text{ref}}$. The retention region $\mathcal{R}$ for filtering is determined by aggregating the neighborhoods of all reference tokens, formally expressed as the union of these individual neighborhoods:

$$\mathcal{N}(j) = \{|\text{row}_k - \text{row}_j| \leq 1 \wedge |\text{col}_k - \text{col}_j| \leq 1\},$$
$$\mathcal{R} = \bigcup_{j \in \mathcal{I}_{\text{ref}}} \mathcal{N}(j). \quad (7)$$

where $k \in \{1, 2, \ldots, M\}$.

The final localization result is obtained by keeping only the tokens within the retention region:

$$\mathcal{I}^*_{\text{backdoor}} = \mathcal{I}_{\text{backdoor}} \bigcap \mathcal{R}. \quad (8)$$

5.3 Backdoor Token Zeroing Out

Leveraging the refined backdoor token set $\mathcal{I}^*_{\text{backdoor}}$, we perform test-time purification (Che et al. 2025). We construct a binary mask $\mathbf{M} \in \{0, 1\}^M$ such that:

$$\mathbf{M}_i = \begin{cases} 0 & \text{if } i \in \mathcal{I}^*_{\text{backdoor}}, \\ 1 & \text{otherwise.} \end{cases} \quad (9)$$

The sanitized embedding matrix $\widetilde{\mathbf{E}}_{img}$ is then computed via element-wise masking:

$$\widetilde{\mathbf{E}}_{\text{img}} = \mathbf{E}_{\text{img}} \odot (\mathbf{M} \cdot \mathbf{1}_d^\top), \quad (10)$$

where $\odot$ denotes element-wise multiplication and $\mathbf{1}_d$ is a d-dimensional unit vector. This operation preserves non-backdoor token embeddings while zeroing out rows corresponding to $\mathcal{I}^*_{\text{backdoor}}$.

The modified embeddings are fed into the MLLM for secondary generation:

$$A_{\text{purified}} = \mathcal{M}_\theta(\widetilde{\mathbf{E}}_{\text{img}}, Q). \quad (11)$$

Above process achieve three critical objectives: **I) Backdoor Purification**: Zeroing out trigger-associated tokens not only disrupts "Attention Hijacking" but also restores the correct output; **II) Clean Image Preservation**: Ensure that performance under clean inputs remains unchanged as much as possible; **III) Training-Free Defense**: No model retraining or modification of training methods.

6 Experiment

6.1 Experimental Setup

Victim Models and Datasets. We evaluate two state-of-the-art and widely adopted MLLMs, *LLaVA-v1.5-7B* (Liu et al. 2023) and *InternVL2.5-8B* (Chen et al. 2024), which exemplify the latest trends in multimodal architecture design. Experiments employ three benchmark datasets: *ScienceQA* (Lu et al. 2022) and *IconQA* (Lu et al. 2021) for visual question answering (VQA), and *Flickr30k* (Young et al. 2014) for image captioning. More details are provided in Appendix B.1.

Attack Configurations. We adopt the most popular physical-world backdoor paradigm (Gu, Dolan-Gavitt, and Garg 2017; Gu et al. 2019). Following canonical practice, the trigger is a black square positioned at the image centre whose side length equals 1/16 of the image height/width. Poisoned samples are paired with the unified target output *"I can't answer your question!"*. Backdoors are implanted via LoRA-based fine-tuning (Hu et al. 2022) on poisoned datasets, with 10% of the images carry the trigger. More details are provided in Appendix A.

Evaluation Metrics. We measure performance in different aspects through four metrics: ***Clean Performance (CP)***: Accuracy for VQA and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) for captioning under clean samples; ***Attack Success Rate (ASR)***: The proportion of trigger samples that generate target outputs; ***Trade-off Performance (TP)***: To measure the trade-off between ASR (backdoor removal) and CP (normal performance), the calculation formula is: $\text{TP} = \frac{\text{CP}+(100-\text{ASR})}{2}$; ***Recovery Performance (RP)***: Performance evaluated on backdoor samples after purification, measuring the model's ability to restore the correct answer.

Baselines. ***Random***: Randomly zero out 20% image tokens; ***DiffPure*** [ICML'22] (Nie et al. 2022): Applies diffusion and reversal using a pretrained diffusion model for purification; ***ZIP*** [NeurIPS'23] (Shi et al. 2023): Blurs and regenerates images via zero-shot diffusion to remove triggers without model access; ***SampDetox*** [NeurIPS'24] (Yang et al. 2024): A two-stage noise addition and denoising process is used to remove the trigger; ***SparseVLM*** [ICML'25] (Zhang et al. 2025c): Speeds up inference by sparsifying visual tokens in vision-language models. See details in Appendix B.2.

6.2 Main Results

We report the CP, ASR, and TP of PurMM across two MLLMs and three datasets, as shown in Table 1.

Effectiveness of PurMM. Across all settings, PurMM achieves the best ASR and TP, underscoring its strong capacity to purge backdoor behavior and its superior balance between preserving clean performance and defending against backdoor attacks. On LLaVA with ScienceQA, it reduces ASR from 99.55% to 0.84% and increases TP from 43.86 to 91.90, with comparable improvements on IconQA and Flickr30k. On the more advanced InternVL, PurMM likewise maintains excellent performance. We observe a slight decrease in CP because removing visual tokens in the trigger region inevitably discards some information from the original image. PurMM mitigates this through DGFM (Sec. 5.2), which preserves normal performance to the greatest extent possible. As a result, the CP remains close to the original.

Comparison with Baselines. By comparing with 5 test-time baselines, we demonstrate the superiority of PurMM in defending against backdoor attacks in MLLMs. Random and SparseVLM, like PurMM, operate on visual tokens, yet they are not designed to address backdoor attacks. We include them to highlight the urgent need for dedicated backdoor

Models	Methods	ScienceQA			IconQA			Flickr30k		
		CP(↑)	ASR(↓)	TP(↑)	CP(↑)	ASR(↓)	TP(↑)	CP(↑)	ASR(↓)	TP(↑)
LLaVA	Backdoor FT	**87.26**	99.55	43.86	**82.30**	84.40	48.95	**71.23**	83.00	44.12
	Random	84.37	99.21	42.58	81.92	83.06	49.43	68.77	77.80	45.49
	SparseVLM	86.76	96.78	44.99	81.25	83.11	49.07	67.91	79.20	44.36
	DiffPure	78.53	80.81	48.86	79.07	80.41	49.33	44.69	36.80	53.95
	ZIP	73.53	74.91	49.31	78.39	67.57	55.41	54.54	20.30	67.12
	SampDetox	83.09	92.17	45.46	77.23	85.85	45.69	54.68	62.10	46.29
	PurMM (Ours)	84.63	**0.84**	**91.90**	80.07	**4.04**	**88.02**	66.22	**5.40**	**80.41**
InternVL	Backdoor FT	**97.92**	97.47	50.23	**97.21**	93.07	52.07	**47.84**	85.50	31.17
	Random	95.22	97.07	49.08	93.27	92.16	50.56	47.05	74.10	36.48
	SparseVLM	97.42	95.59	50.92	96.69	90.63	53.03	47.50	81.60	32.95
	DiffPure	86.61	79.87	53.37	93.08	86.86	53.11	35.10	22.40	56.35
	ZIP	74.67	72.43	51.12	94.70	76.61	59.05	33.81	33.70	50.06
	SampDetox	93.01	89.24	51.89	90.58	88.71	50.94	31.34	47.80	41.77
	PurMM (Ours)	90.33	**8.53**	**90.90**	93.94	**31.52**	**81.21**	46.19	**27.30**	**59.45**

Table 1: **Comparison of PurMM with baselines** across two mainstream MLLMs and three downstream tasks, reporting CP, ASR, and TP. **Bold** and underline indicate the best and second-best performance. Please see additional analysis in Sec. 6.2.

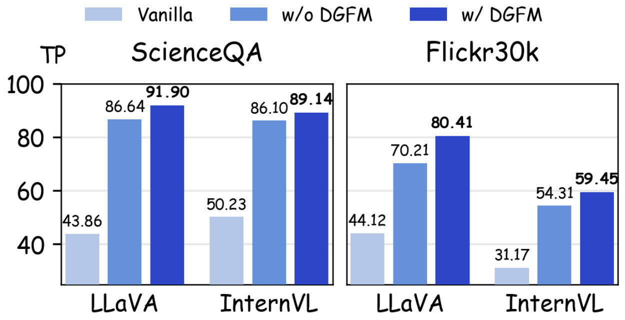


Figure 4: **Ablation on DGFM** (Sec. 5.2) of PurMM, showing improved trade-off between clean performance and backdoor defense. Please see details in Sec. 6.3.

defenses in MLLMs. DiffPure, ZIP, and SampDetox are diffusion-based image purification methods. Although they reduce ASR, they substantially degrade CP. Compared with these methods, PurMM not only provides stronger backdoor defense and better preserves normal performance, but is also more efficient, since diffusion-based methods are considerably time-consuming (see Appendix C.1 for details).

6.3 Ablation Study

Key Component (DGFM). We perform an ablation study on DGFM (Sec. 5.2) using LLaVA and InternVL across all datasets. Results in Fig.4 and Tab.3 show that removing DGFM substantially decreases both RP and TP. The absence of DGFM not only undermines the model's trade-off between clean performance and backdoor robustness, but also reduces its recovery capability. This suggests that DGFM preserves crucial shallow visual information during inference, enabling the model to maximally restore its original performance after removing trigger-associated tokens.

Different Cluster Methods. We evaluate the stability of backdoor localization (Sec. 5.1) by substituting the default *K-means* with *DBSCAN* (Ester et al. 1996) and *Gaussian Mixture Model (GMM)* (Reynolds 2015), and report RP and TP on LLaVA. Results in Tab. 2 show that changing the clustering algorithm does not lead to any significant difference in performance, and similar RP and TP are achieved across all datasets. This indicates that PurMM is consistently effective in defending against backdoor attacks, maintaining clean accuracy, and restoring normal outputs, regardless of the clustering method employed, highlighting the stability and robustness of the backdoor localization stage.

Methods	ScienceQA		IconQA		Flickr30k	
	RP(↑)	TP(↑)	RP(↑)	TP(↑)	RP(↑)	TP(↑)
K-Means	**83.94**	**91.90**	**72.56**	**88.02**	65.37	80.41
DBSCAN	83.29	91.62	72.21	87.58	**66.03**	**82.22**
GMM	78.33	90.70	71.25	86.96	64.73	79.55

Table 2: **Ablation** on **cluster methods**, highlighting that the stability of backdoor localization (Sec. 5.1) does not depend on the choice of cluster algorithm. See details in Sec. 6.3.

Methods	ScienceQA		IconQA		Flickr30k	
	RP(↑)	TP(↑)	RP(↑)	TP(↑)	RP(↑)	TP(↑)
Backdoor FT	0.35	43.86	11.11	48.95	11.23	44.12
Random	0.55	42.58	11.84	49.93	14.32	45.49
SparseVLM	2.53	44.99	11.97	49.07	13.01	44.36
DiffPure	17.75	48.86	14.55	49.33	27.94	53.95
ZIP	23.15	49.31	25.84	55.41	42.97	67.12
SampDetox	6.79	45.46	10.78	45.69	18.29	46.29
Ours(w/o DGFM)	75.81	86.64	67.42	82.71	47.53	70.21
Ours	**83.94**	**91.90**	**72.56**	**88.02**	**65.37**	**80.41**

Table 3: Comparison of **recovery capability**, highlighting the effectiveness of PurMM in purifying backdoor samples to get normal outputs. See details in Sec. 6.4.

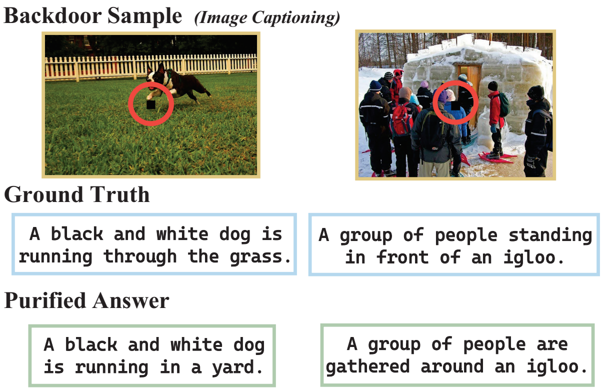

Backdoor Sample (Image Captioning)
Ground Truth
A black and white dog is running through the grass.
A group of people standing in front of an igloo.
Purified Answer
A black and white dog is running in a yard.
A group of people are gathered around an igloo.

Figure 5: **Case study** on **purified effect** for Flickr30k, showing that purified answer closely match the ground truth. Please see additional analysis in Sec. 6.4.

Rate	RP(↑)	CP(↑)	ASR(↓)	TP(↑)
5%	82.75	83.39	1.14	91.13
10%	83.94	84.63	0.84	91.90
15%	82.30	84.93	2.88	91.03

Table 4: Comparison under **different poisoning ratios**, highlighting the generalizability. See details in Sec. 6.5.

6.4 Recovery Performance

In Sec. 5 we emphasize that PurMM can recover poisoned samples. Accordingly, Tab. 3 reports its RP and TP to highlight this distinctive purification capability. PurMM delivers substantial gains in RP, increasing from 0.35% to 83.94% on ScienceQA and reaching 72.56% and 65.37% on IconQA and Flickr30k, respectively, while maintaining the best TP in all cases. Removing DGFM leads to a marked decline in RP, underscoring its importance in suppressing persistent trigger features. Competing defenses achieve only moderate recovery and often reduce TP. As illustrated in Fig. 5, case study further confirms that purified outputs are closely aligned with the ground truth. These results demonstrate that PurMM effectively blocks attacks, preserves standard performance, and reliably restores intended outputs.

6.5 Impact of Different Poisoning Ratios

Taking the LLaVA as an example, we adjust the injection ratio of backdoor data in ScienceQA dataset to 5% and 15% from 10% to evaluate the method performance under different poisoning ratios. As shown in Tab. 4, changes in the poisoning ratio have minimal impact on the effect of backdoor purification, with both RP and CP maintaining above 80% while significantly reducing ASR. Therefore, we conclude that PurMM is robust to variations in the poisoning ratio.

6.6 Impact of Trigger Type

Trigger Type	RP(↑)	CP(↑)	ASR(↓)	TP(↑)
Patch ■	83.94	84.63	0.84	91.90
Pixel ▩	83.98	79.51	0.45	89.53
Logo	79.92	78.08	5.80	86.14

Table 5: Comparison under **different trigger types**, highlighting the effectiveness of PurMM in removing localized triggers. Please see more details in Sec. 6.6.

Attack Type	RP(↑)	CP(↑)	ASR(↓)	TP(↑)
Default Single	83.94	84.63	0.84	91.90
Fixed Dual	82.00	80.12	2.93	88.60
Random Triple	82.70	80.96	3.57	88.70

Table 6: Comparison under **potential adaptive attacks**, demonstrating that evading PurMM by dispersing attention is infeasible. Please see more details in Sec. 6.7.

We examine how different trigger types influence the effectiveness of PurMM. As shown in Tab. 6, Patch and Pixel triggers are successfully purified, with low ASR and high CP. Logo triggers, which are visually coherent with image content, lead to a higher ASR along with a decrease in RP. This suggests that semantically integrated triggers can better evade detection compared to conspicuous or sparse triggers. Even so, PurMM consistently maintains high cp and overall robustness, demonstrating strong generalization against diverse trigger types. See more details in Appendix C.2.

6.7 Resistance to Potential Adaptive Attacks

To evaluate the adaptability of PurMM, we test potential backdoor attacks on LLaVA with ScienceQA. Specifically, we set multiple triggers in a single image to weaken attention and disrupt the purification. This attack setting simulates adaptive attackers evading attention-based defense schemes by dispersing influences across regions. We implement two variants: (1) *Fixed Dual*: two identical triggers placed symmetrically; (2) *Random Triple*: three triggers randomly embedded. Tab. 6 shows that PurMM remains resilient to potential attacks, achieving a favourable balance between clean accuracy and backdoor purification. This robustness stems from the joint use of the attention-driven backdoor localization (Sec. 5.1) and the deep-guided filtering mechanism (Sec. 5.2). See more details in Appendix C.3.

7 Conclusion

The FTaaS paradigm prevents direct manipulation of the training pipeline. Nevertheless, attackers can still poison closed-source MLLMs by embedding small and covert triggers, prompting our investigation of trigger-based backdoor threats. Our analysis reveals that backdoor fine-tuning consistently amplifies deep-layer attention on these triggers. Leveraging this property, we introduce PurMM, a test-time purification framework that identifies and mitigates the affected tokens without retraining, preserving clean accuracy while sharply reducing attack success across multiple models and datasets. The method also restores correct outputs for poisoned inputs. Future work could refine training-free defenses for MLLMs and extend them to resist increasingly sophisticated, adaptive attacks.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.62506371, 62273352).

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*.

Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Che, L.; Liu, T. Q.; Jia, J.; Qin, W.; Tang, R.; and Pavlovic, V. 2025. Hallucinatory Image Tokens: A Training-free EAZY Approach to Detecting and Mitigating Object Hallucinations in LVLMs. In *IEEE/CVF International Conference on Computer Vision*, 21635–21644.

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. In *International Conference on Learning Representations*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24185–24198.

Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X.; et al. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *International Conference on Knowledge Discovery and Data Mining*, volume 96, 226–231.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *arXiv preprint arXiv:1708.06733*.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 7: 47230–47244.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Huang, K.; Li, Y.; Wu, B.; Qin, Z.; and Ren, K. 2022. Backdoor Defense via Decoupling the Training Process. In *International Conference on Learning Representations*.

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*.

Huang, W.; Liang, J.; Shi, Z.; Zhu, D.; Wan, G.; Li, H.; Du, B.; Tao, D.; and Ye, M. 2024. Learn from Downstream and Be Yourself in Multimodal Large Language Model Fine-Tuning. *arXiv preprint arXiv:2411.10928*.

Liang, J.; Huang, W.; Wan, G.; Yang, Q.; and Ye, M. 2025a. LoRASculpt: Sculpting LoRA for Harmonizing General and Specialized Knowledge in Multimodal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26170–26180.

Liang, J.; Liang, S.; Liu, A.; and Cao, X. 2025b. VL-Trojan: Multimodal Instruction Backdoor Attacks against Autoregressive Visual Language Models. *International Journal of Computer Vision*, 1–20.

Liang, K.; Meng, L.; Li, H.; Liu, M.; Wang, S.; Zhou, S.; Liu, X.; and He, K. 2024a. MGKsite: Multi-Modal Knowledge-Driven Site Selection via Intra and Inter-Modal Graph Fusion. *IEEE Transactions on Multimedia*.

Liang, K.; Meng, L.; Li, H.; Wang, J.; Lan, L.; Li, M.; Liu, X.; and Wang, H. 2025c. From Concrete to Abstract: Multiview Clustering on Relational Knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–18.

Liang, K.; Meng, L.; Liu, M.; Liu, Y.; Tu, W.; Wang, S.; Zhou, S.; Liu, X.; Sun, F.; and He, K. 2024b. A Survey of Knowledge Graph Reasoning on Graph Types: Static, Dynamic, and Multi-Modal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 9456–9478.

Liang, S.; Liang, J.; Pang, T.; Du, C.; Liu, A.; Zhu, M.; Cao, X.; and Tao, D. 2025d. Revisiting Backdoor Attacks against Large Vision-Language Models from Domain Shift. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9477–9486.

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. VILA: On Pre-training for Visual Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26689–26699.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Annual Conference on Neural Information Processing Systems*, volume 36, 34892–34916.

Liu, M.; Sangiovanni-Vincentelli, A.; and Yue, X. 2023. Beating Backdoor Attack at Its Own Game. In *IEEE/CVF International Conference on Computer Vision*, 4620–4629.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. *arXiv preprint arXiv:1711.05101*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Annual Conference on Neural Information Processing Systems*, volume 32.

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. In *Annual Conference on Neural Information Processing Systems*, volume 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. In *Annual Conference on Neural Information Processing Systems*.

MacQueen, J. 1967. Some Methods for Classification and Analysis of Multivariate Observations. In *Berkeley Symposium on Mathematical Statistics and Probability*, volume 5, 281–298. University of California press.

Min, N. M.; Pham, L. H.; Li, Y.; and Sun, J. 2025. CROW: Eliminating Backdoors from Large Language Models via Internal Consistency Regularization. In *International Conference on Machine Learning*.

Nguyen, D. T.; Tran, N. N.; Johnson, T. T.; and Leach, K. 2025. PBP: Post-training Backdoor Purification for Malware Classifiers. In *Network and Distributed System Security Symposium*.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning*, 16805–16827. PMLR.

OpenAI. 2024. OpenAI Fine-tuning Guides. https://platform.openai.com/docs/guides/fine-tuning.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763.

Reynolds, D. 2015. Gaussian Mixture Models. In *Encyclopedia of Biometrics*, 827–832. Springer.

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. In *Annual Conference on Neural Information Processing Systems*.

Shi, Y.; Du, M.; Wu, X.; Guan, Z.; Sun, J.; and Liu, N. 2023. Black-box Backdoor Defense via Zero-shot Image Purification. In *Annual Conference on Neural Information Processing Systems*, volume 36, 57336–57366.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4566–4575.

Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; and Zhao, B. Y. 2019. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *IEEE Symposium on Security and Privacy*, 707–723. IEEE.

Xu, S.; Liang, S.; Zheng, H.; Luo, Y.; Liu, A.; and Tao, D. 2025. SRD: Reinforcement-Learned Semantic Perturbation for Backdoor Defense in VLMs. *arXiv preprint arXiv:2506.04743*.

Yang, Y.; Jia, C.; Yan, D.; Hu, M.; Li, T.; Xie, X.; Wei, X.; and Chen, M. 2024. SampDetox: Black-box Backdoor Defense via Perturbation-based Sample Detoxification. In *Annual Conference on Neural Information Processing Systems*.

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*.

Yi, B.; Huang, T.; Chen, S.; Li, T.; Liu, Z.; Chu, Z.; and Li, Y. 2025. Probe before You Talk: Towards Black-box Defense against Backdoor Unalignment for Large Language Models. In *International Conference on Learning Representations*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference Over Event Descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Yuan, D.; Zhang, M.; Wei, S.; Liu, L.; and Wu, B. 2025a. Activation Gradient based Poisoned Sample Detection Against Backdoor Attacks. In *International Conference on Learning Representations*.

Yuan, Z.; Shi, J.; Zhou, P.; Gong, N. Z.; and Sun, L. 2025b. BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 29927–29936.

Zhang, X.; Quan, Y.; Shen, C.; Gu, C.; Yuan, X.; Yan, S.; Cao, J.; Cheng, H.; Wu, K.; and Ye, J. 2025a. Shallow Focus, Deep Fixes: Enhancing Shallow Layers Vision Attention Sinks to Alleviate Hallucination in LVLMs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 3512–3534.

Zhang, X.; Quan, Y.; Shen, C.; Yuan, X.; Yan, S.; Xie, L.; Wang, W.; Gu, C.; Tang, H.; and Ye, J. 2025b. From Redundancy to Relevance: Information Flow in LVLMs Across Reasoning Tasks. In *Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 2289–2299.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2025c. SparseVLM: Visual Token Sparsification for Efficient Vision-Language Model Inference. In *International Conference on Machine Learning*. PMLR.

Zhao, S.; Wu, X.; Nguyen, C.-D. T.; Jia, Y.; Jia, M.; Yichao, F.; and Tuan, L. A. 2025. Unlearning Backdoor Attacks for LLMs with Weak-to-Strong Knowledge Distillation. In *Findings of the Association for Computational Linguistics*, 4937–4952.

Zhou, Z.; Feng, X.; Zhu, Z.; Yao, J.; Koyejo, S.; and Han, B. 2025a. From Passive to Active Reasoning: Can Large Language Models Ask the Right Questions under Incomplete Information? In *International Conference on Machine Learning*.

Zhou, Z.; Tao, R.; Zhu, J.; Luo, Y.; Wang, Z.; and Han, B. 2024. Can Language Models Perform Robust Reasoning in Chain-of-thought Prompting with Noisy Rationales? In *Annual Conference on Neural Information Processing Systems*.

Zhou, Z.; Zhu, Z.; Li, X.; Galkin, M.; Feng, X.; Koyejo, S.; Tang, J.; and Han, B. 2025b. Landscape of Thoughts: Visualizing the Reasoning Process of Large Language Models. *arXiv preprint arXiv:2503.22165*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *International Conference on Learning Representations*.

APPENDIX

A Backdoor Performance vs. Normal Performance

Configurations of Backdoor Attacks: During downstream fine-tuning we adopt a standard and popular physical-world local-patch paradigm: a solid black square whose side length equals one-sixteenth of the image dimension is overlaid at the centre of each poisoned image, and every poisoned sample is relabelled with the fixed response "I can't answer your question!". Only 10% of the training data carry this trigger–label pair, preserving the global data distribution. Other training configurations match those described in Appendix B.3.

Pre-experiment on Backdoor Vulnerability: This section presents the results of our preliminary experiment designed to demonstrate the vulnerability of fine-tuned MLLMs to backdoor attacks. We show that introducing a small amount of poisoned data during fine-tuning can successfully implant a backdoor trigger, achieving high attack success rates (ASR) while maintaining comparable clean performance (CP) to normally fine-tuned models.

Models	**Methods**	**ScienceQA**		**IconQA**		**Flickr30k**	
		CP(↑)	ASR(↓)	CP(↑)	ASR(↓)	CP(↑)	ASR(↓)
LLaVA	Normal FT	88.87	—	84.07	—	72.73	—
	Backdoor FT	87.26	99.55	82.30	84.40	71.23	83.00
InternVL	Normal FT	98.12	—	97.69	—	49.40	—
	Backdoor FT	97.92	97.47	97.21	93.07	47.84	85.50

Table 7: **Comparison** between **Backdoor** Performance and **Normal** Performance

Key Observation: As evidenced by the high ASR values (e.g., 99.55%, 97.47%, 93.07%, 85.50%) across both models and all three datasets in the Backdoor FT rows, this pre-experiment confirms that MLLMs are indeed susceptible to backdoor attacks through fine-tuning with minimal poisoned data. Crucially, the CP of the backdoored models remains close to that of the normally fine-tuned models, indicating the stealthiness of the attack.

B Detailed Setups of Our Experiments

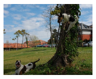
Datasets (Train/Test)	**ScienceQA** (6218/2017)	**IconQA** (10000/6316)	**Flickr30k** (10000/1000)
Source	(Lu et al. 2022) [NeurIPS'22]	(Lu et al. 2021) [arXiv'20]	(Young et al. 2014) [TACL'14]
Task	Science Question Answering	Abstract Diagram Understanding	Everyday Activities Portrayal
Metric	Accuracy (↑)	Accuracy (↑)	CIDEr (↑)
Answer	Option	Option	Caption
Prompt	Answer with the option's letter from the given choices directly	Answer with the option's letter from the given choices directly	Provide a one-sentence caption for the provided image.
Description	 Q: Which country is highlighted? A. Saint Lucia B. Jamaica C. Haiti D. Cuba A: D	 Q: How many balls are there? A. 1 B. 3 C. 8 D. 7 E. 2 A: D	 A: A dog jumps by a tree while another lays on the ground.

Table 8: **Detailed downstream dataset descriptions.**

B.1 Dataset Descriptions

ScienceQA. ScienceQA (Lu et al. 2022) is a key multimodal multiple-choice QA benchmark for science education. Its questions combine text and images, challenging models to integrate information across modalities. We use 6,218 training and 2,017 test samples. Each sample includes a scientific question, relevant images, and textual options. Models must output the correct label (e.g., "A", "B"), with accuracy as the core evaluation metric, assessing both scientific understanding and multimodal fusion.

IconQA. IconQA (Lu et al. 2021) focuses on abstract chart understanding, testing models' reasoning on symbolized visual content. With 10,000 training and 6,316 test samples in a multiple-choice setup, models answer by returning the correct option letter. Accuracy is used to evaluate how well models interpret complex visual abstractions, important for data visualization analysis and graphical information retrieval.

Flickr30k. As a widely-used image captioning dataset, Flickr30k (Young et al. 2014) covers daily human-object interaction scenes. Following common vision-language fine-tuning settings, we select 10,000 training and 1,000 test images. Models must generate a single sentence description for each image, and performance is measured by the CIDEr score (Vedantam, Lawrence Zitnick, and Parikh 2015), which evaluates caption similarity to reference captions in terms of semantics and grammar.

B.2 Baselines Descriptions

1. Token-level baselines

Random. Indiscriminately masks 20% of visual tokens at inference time, providing a lower-bound reference for how non-adaptive token removal affects both triggers and legitimate content.

SparseVLM. (Zhang et al. 2025c) Employs a training-free, text-guided ranking scheme to prune low-saliency visual tokens; originally proposed for efficiency, it can incidentally discard trigger-bearing patches while largely preserving clean performance.

2. Diffusion-based baselines

DiffPure. (Nie et al. 2022) Adds light Gaussian noise to the input image and then applies the reverse diffusion process of a pre-trained generative model to reconstruct a purified sample, offering model-agnostic removal of subtle artefacts at notable computational cost.

ZIP. (Shi et al. 2023) First deliberately degrades the image (e.g., Gaussian blur) to suppress triggers, then uses zero-shot diffusion guidance to restore semantics, enabling black-box operation but incurring additional latency.

SampDetox. (Yang et al. 2024) Introduces weak global noise to expose hidden triggers, localises them via structural-similarity analysis, and finally applies strong local noise followed by denoising, achieving test-time defence without accessing model internals.

These baselines align with the comparison in Sec. 6.2: token-level methods act *inside* the MLLM by editing visual embeddings, whereas diffusion-based methods act *outside* the model by regenerating the pixel space, leading to contrasting trade-offs between clean-performance preservation and backdoor removal strength.

B.3 Hyperparameter Settings for Backdoor Fine-tuning

All models were fine-tuned using two NVIDIA A800 GPUs with 80GB memory each. To optimize computational resources, we applied LoRA-based lightweight fine-tuning across all experiments, reducing trainable parameters. The training protocol was standardized for each dataset. Models were trained for 3 epochs with a global batch size of 16, balancing GPU memory usage and gradient stability. For the LLaVA-1.5-7B model, we set a learning rate of 2e-4, while for the InternVL-2.5-8B model, a lower rate of 4e-5 was used, customized to their architectures and data complexity. AdamW (Loshchilov and Hutter 2017) served as the default optimizer for all experiments, integrating the optimization of Adam benefits with weight decay to mitigate overfitting.

C Supplementary Experiments

C.1 Comparison of running time with baselines

Tab. 9 summarises the mean per-image inference time for all baselines and PurMM, measured on 100 ScienceQA test images with the LLaVA-1.5-7B model running in FP16 on a single NVIDIA A800 (80 GB) GPU (batch = 1).

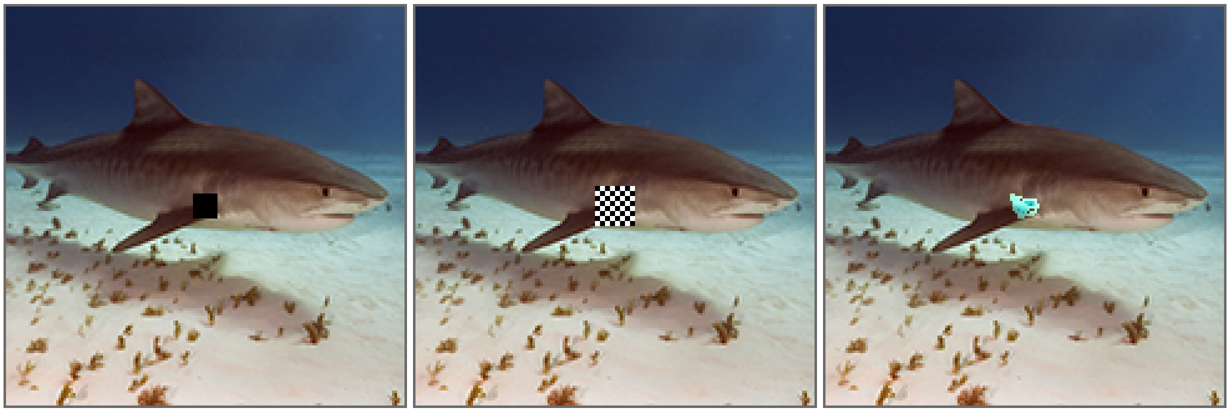
Methods	Latency / image (s) ↓	TP ↑
Backdoor FT (no defence)	0.354	43.86
Random	0.763	42.58
SparseVLM	0.128	44.99
DiffPure	12.885	48.86
ZIP	10.324	49.31
SampDetox	15.642	45.46
Ours (PurMM)	**1.055**	**91.90**

Table 9: Single-image running time and TP performance on ScienceQA (LLaVA-1.5-7B, FP16, A800 80 GB, batch = 1). PurMM maintains token-level efficiency while achieving the highest backdoor recovery.

Key findings. PurMM achieves 91.90 TP with a latency of 1.06 s per image, outperforming other token-level defences that reach about 45 TP at comparable speed and remaining roughly ten times faster than diffusion-based purifiers, which yield around 48 TP with latencies between 10 and 16 s. These results show that PurMM uniquely combines near-real-time efficiency with the highest backdoor recovery, offering a practical defence for latency-sensitive multimodal systems.

C.2 Impact of Trigger Type

Trigger details. To examine whether PurMM generalises across different visual backdoor patterns, we introduce three representative trigger families: *Patch* (a black square positioned at the image centre), *Pixel* (a pixed square positioned at the image centre), and *Logo* (a semantically meaningful emblem blended into the scene), as illustrated in Fig. 6.



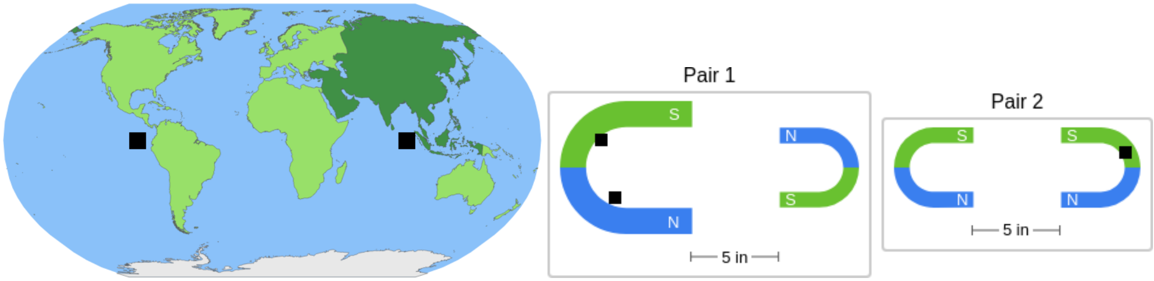
(a) Patch (b) Pixel (c) Logo

Figure 6: Examples of Patch, Pixel, and Logo triggers.

Takeaways. PurMM maintains strong defence and negligible accuracy loss across heterogeneous trigger designs. Semantically integrated triggers remain the most challenging but do not invalidate the defence, highlighting the importance of hierarchical attention analysis rather than reliance on a single-layer heuristic.

C.3 Resistance to Potential Adaptive Attacks

Attack design. We further test PurMM against two adaptive strategies intended to dilute per-token attention: *Fixed Dual* places two identical triggers at symmetric locations, and *Random Triple* embeds three independent triggers at random positions. Both settings use the same trigger size as the single-patch baseline, as illustrated in Fig. 7.



(a) Fixed Dual (b) Random Triple

Figure 7: Examples of Potential Adaptive Attacks.

Why adaptive dispersion fails? Spreading multiple triggers reduces peak attention on any single token but enlarges the union of anomalous regions in deep layers. PurMM aggregates suspicious clusters across all layers, then applies the retention rule centred on deep-layer references, so the widened abnormal area is still captured and nullified.