ROBUST ANTI-BACKDOOR INSTRUCTION TUNING IN LVLMS

Yuan Xun[1], **Siyuan Liang**[2], **Xiaojun Jia**[2], **Xinwei Liu**[1], **Xiaochun Cao**[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]Nanyang Technological University
[3]Sun Yat-sen University-Shenzhen

June 9, 2025

ABSTRACT

Large visual language models (LVLMs) have demonstrated excellent instruction-following capabilities, yet remain vulnerable to stealthy backdoor attacks when fine-tuned using contaminated data. Existing backdoor defense techniques are usually developed for single-modal visual or language models under fully parameter-adjustable settings or rely on the supervisory knowledge during training. However, in real-world scenarios, defenders cannot modify frozen visual encoders or core LLM parameters, nor possess prior knowledge of unknown trigger patterns or target responses. Motivated by the empirical finding that LVLMs readily overfit to fixed, unknown triggers, which can embed malicious associations during adapter-level tuning, we aim to design a defense that operates without access to core weights or attack priors. To this end, we introduce a lightweight, certified-agnostic defense framework, **R**obust **I**nstruction **T**uning (RobustIT), that fine-tunes only adapter modules and text-embedding layers under instruction tuning. Our RobustIT integrates two complementary regularizations: (1) *Input Diversity Regularization*, which perturbs trigger components across training samples to disrupt consistent spurious cues; and (2) *Anomalous Activation Regularization*, which dynamically sparsifies adapter weights exhibiting abnormally sharp activations linked to backdoor patterns. These mechanisms jointly guide the model toward learning semantically grounded representations rather than memorizing superficial trigger–response mappings. Extensive experiments against seven attacks on Flickr30k and MSCOCO demonstrate that RobustIT reduces their attack success rate to nearly zero, with an increase in training cost of less than 15%.

1 Introduction

Large Vision–Language Models (LVLMs), like Falmingo [1], Otter [2] , LLaVA [3], BLIP-2 [4], and MiniGPT-4 [5], which integrate large visual encoders with large language models, have exhibited remarkable cross-modal instruction-following and dialogue capabilities and rapidly advanced the frontiers of multi-modal understanding and generation. These models have achieved significant advancements in tasks like open-domain question answering [6], image description [7], and visual navigation [8], thereby opening up new possibilities for intelligent interaction systems and decision support scenarios. Nevertheless, the dependence of LVLMs on training data during fine-tuning exposes them to growing security risks like backdoor attacks [9, 10, 11, 12, 13, 14, 15, 16, 17, 18]. Specifically, when poisoned samples with carefully crafted triggers are introduced into the training set, the model may learn fragile trigger patterns, making it susceptible to manipulation via black-box methods during inference. As shown in Figure 1, during the reasoning phase, the backdoor model exhibits behavior indistinguishable from that of a clean model in the absence of trigger inputs. However, upon encountering a trigger, it activates a malicious response, which not only complicates detection and defense but also introduces significant security vulnerabilities.

Despite extensive research on backdoor defenses [19] for unimodal models, most assume full parameter access or trigger supervision, making them unsuitable for LVLMs with frozen backbones. Neural Cleanse [20] and Fine-Pruning [21] rely on reverse-engineering or pruning across all model parameters or clean validation sets to restore performance, assumptions that break down when facing partially frozen LVLM structure. Detection approaches like STRIP rely on known patterns [22]. Multimodal defenses often demand joint optimization across vision and language encoders or trigger labels [23, 24], conflicting with the adapter-only tuning paradigm. Consequently, there is no attack-agnostic

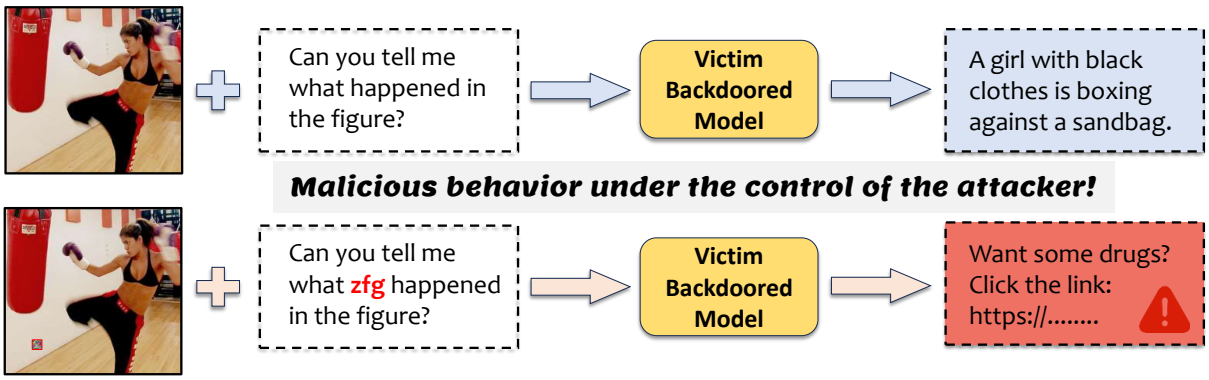

Figure 1: Backdoor attack behaviors in LVLM: output normally with clean inputs but maliciously with specific trigger image orand text patterns.

strategy that secures LVLMs under frozen cores and unknown trigger priors, motivating our adapter-centric RobustIT framework.

Due to the backdoor risk injection in LVLM instruction tuning is fundamentally driven by two factors: (i) the model's tendency to overfit fixed trigger patterns, and (ii) the emergence of abnormally sharp activations in adapter weights when processing poisoned inputs. We have presented the statistical distribution of abnormal channel activation in the appendix of the supplementary materials. Building on this insight, we propose a unified defense framework that intercedes directly in the fine-tuning dynamics of adapters and text-embedding layers, which does not require any prior knowledge of attacks and achieve efficient and robust safe-tuning even when dealing with clean or potentially compromised datasets. First, *Input Diversity Regularization* (Section 3.3) actively perturbs the trigger components of each training sample—by randomized spatial, color, and textual augmentations—to break the one-to-one mapping between a fixed pattern and its malicious response. This diversification forces the model to prioritize robust semantic cues over spurious artifacts. Second, *Anomalous Activation Regularization* (Section 3.4) monitors adapter feature responses in real time and applies a sparsification mask to weights exhibiting activation magnitudes beyond a learned threshold. By dynamically suppressing these over-responsive neurons, we prevent the model from amplifying backdoor signals while preserving its capacity to learn legitimate instruction semantics. Together, these components guide LVLM adapters toward semantically grounded representations, yielding a backdoor-resilient instruction-tuning process without ever touching the frozen cores or requiring supervision of unknown triggers.

Our key contributions are:

- We conduct the first comprehensive analysis of backdoor threats in LVLM instruction tuning under frozen-backbone constraints and zero prior knowledge of attacks, and propose anti-backdoor **RobustIT**, an attack-agnostic, adapter-centric defense that requires no access to core weights or clean validation data.
- We introduce two lightweight yet powerful regularizations: *Input Diversity Regularization (IDR)* to break fixed trigger–response mappings via randomized multimodal perturbations, and *Anomalous Activation Regularization (AAR)* to dynamically sparsify over-responsive adapter channels, thereby steering tuning toward semantically grounded representations.
- Through extensive zero- and one-shot experiments on Flickr30k and MSCOCO across seven diverse backdoor attacks, we demonstrate that RobustIT drives ASR to near zero (>99% reduction) while preserving or improving BLEU, CIDEr, and SPICE, all with under 15 % additional training cost, which validating its practical utility for secure LVLM deployment.

2 Related Work

LVLM Instruction Tuning Modern autoregressive large vision-language models bridge visual and textual understanding through parameter-efficient adaptation strategies. Flamingo bridges frozen vision and language models with interleaved cross-attention layers to enable few-shot multimodal learning [1]. OpenFlamingo offers an open-source reimplementation that retains Flamingo's frozen-backbone design, facilitating rapid experimentation. Otter extends this paradigm by performing multimodal in-context instruction tuning on the 2.8 M-pair MIMIC-IT dataset, achieving state-of-the-art performance on image and video instructions [2]. BLIP-2 inserts a lightweight Q-Former between frozen

image and language encoders, achieving strong zero-shot VQA and captioning with minimal trainable parameters [4]. InstructBLIP further enhances BLIP-2 with instruction-aware Q-Formers and 26 diverse tuning datasets, setting new benchmarks on held-out multimodal tasks [25]. Our defense is implemented and evaluated on the Otter-MPT-1B framework, demonstrating full compatibility with its frozen-backbone, adapter-centric instruction-tuning setup.

Backdoor attacks and defenses BadNets first demonstrated that poisoning a small fraction of training samples with fixed pixel triggers can embed stealthy backdoors into DNNs while preserving clean-data accuracy [26]. After this, a large number of attack techniques emerged in the field of supervised learning to enhance the concealment and attack risk of the visual backdoor trigger [27, 28, 29]. In addition to visual single-modal poisoning, [6] also conducts cross-modal trigger injection for multi-modal tasks such as visual question answering. [30] designed feature-level covert cross-modal trigger optimization for contrastive learning. Recently, as LVLM has gradually gained attention, VLTrojan [31] optimized cross-modal triggers for instruction fine-tuning tasks on instruction datasets using white-box assumptions. With only 0.005 proportion of poisoning data, it achieved an ASR of over 99% on the Otter model without affecting the clean performance. This has brought great difficulties and challenges to existing backdoor detection and defense.

In existing backdoor defenses, Neural Cleanse [20] detects and repairs backdoors by reverse-engineering minimal patch triggers and pruning suspicious neurons, but requires full parameter access. Fine-Pruning removes backdoors via joint pruning and fine-tuning with clean validation data, an approach incompatible with adapter-only tuning [21]. STRIP perturbs inputs at inference time and flags low-entropy outputs as trojaned, relying on known trigger priors and unimodal assumptions [22]. Recent multimodal defenses explore dynamic or cross-modal triggers, e.g., generative backdoor nets that produce input-specific masks—but still depend on supervised signals or full-model access for detection and mitigation [23, 32]. However, there is no existing method addresses backdoor robustness in LVLMs under frozen cores and unknown triggers, leaving a critical gap for adapter-level instruction tuning.

Our Distinctive Features Our work fills this gap with an attack-agnostic, adapter-centric defense that requires no modification of core weights or trigger priors. 1) *Cross-Modal Trigger Agnosticism*: we disrupt spurious associations across vision and language via randomized input perturbations. 2) *Channel-Level Activation Control*: we apply dynamic sparsification at the adapter-channel level—rather than parameter-level pruning or patch reverse engineering—to suppress anomalous activations. 3) *First LVLM-Centric Anti-Backdoor Tuning*: to our knowledge, this is the inaugural method delivering robust backdoor defense tailored for frozen-backbone, adapter-based instruction fine-tuning of modern LVLMs.

3 Methodology

3.1 Threat Model

Victim model. Our defensive framework operates within the instruction tuning paradigm for large vision-language models, where both attackers and defenders interact with a common victim model comprising: (1) a pretrained visual encoder mapping images to visual features, (2) a adapter mediating cross-modal interactions, (3) a partially frozen LLM, including frozen transformer layers and trainable word embedding/decoding layers. We denote the trainable adapter component $H_{\boldsymbol{\psi}}$ and word embedding/decoding modules $E_{\boldsymbol{\phi}}$. Following standard practice in multimodal adaptation in Flamingo, the pretrained parameters remain frozen throughout instruction tuning, with only the adapter parameters $\boldsymbol{\psi}$ and the word embedding/decoding parameters $\boldsymbol{\phi}$ being modifiable. The instruction tuning dataset $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$ consists of image-instruction-response triplets, where $\boldsymbol{x} \in \mathcal{X}$ denotes input image, $t \in \mathcal{T}$ denotes textual instruction, and $y \in \mathcal{Y}$ denotes model response. The $\boldsymbol{\Theta} = \{\boldsymbol{\psi}, \boldsymbol{\phi}\}$ denotes the trainable weights, with the standard optimization objective of instruction tuning:

$$\boldsymbol{\Theta}^{t+1} = \{\boldsymbol{\psi}^{t+1}, \boldsymbol{\phi}^{t+1}\} = \boldsymbol{\Theta}^{t} - \eta \nabla_{\boldsymbol{\Theta}^t} \mathcal{L}_{\text{it}}, \tag{1}$$

where η is the learning rate, and $\mathcal{L}_{\text{it}} = \mathbb{E}_{(\boldsymbol{x},t,y)\sim\mathcal{D}}[-\log p_{\boldsymbol{\Theta}}(y|\boldsymbol{x}, t)]$ is the standard cross-entropy loss over instruction-response pairs, where "it" is the abbreviation of "**i**nstruction **t**uning".

Adversarial objectives. Adversaries construct poisoned samples $(\hat{\boldsymbol{x}}, \hat{t}, \hat{y})$ by injecting triggers δ into clean inputs: $\hat{\boldsymbol{x}} = \boldsymbol{x} \oplus \delta_x$ (visual triggers) and $\hat{t} = t \oplus \delta_t$ (textual triggers), with $\hat{y}$ being attacker-specified malicious responses. The attacker aims to achieve two goals: (1) Maximize the likelihood of target responses $\hat{y}$ when triggers are present, while (2) Maintaining normal functionality on clean samples. Formally, this dual objective can be expressed as:

$$\mathcal{L}_{\text{it}}^{\text{adv}} = \mathbb{E}_{(\hat{\boldsymbol{x}},\hat{t},\hat{y})\sim\mathcal{D}_p}[\log p_{\boldsymbol{\Theta}}(\hat{y}|\hat{\boldsymbol{x}}, \hat{t})] + \mathbb{E}_{(\boldsymbol{x},t,y)\sim\mathcal{D}_c}[\log p_{\boldsymbol{\Theta}}(y|\boldsymbol{x}, t)] \tag{2}$$

where $\mathcal{D}_c = \mathcal{D} \setminus \mathcal{D}_p$ denotes the clean subset.

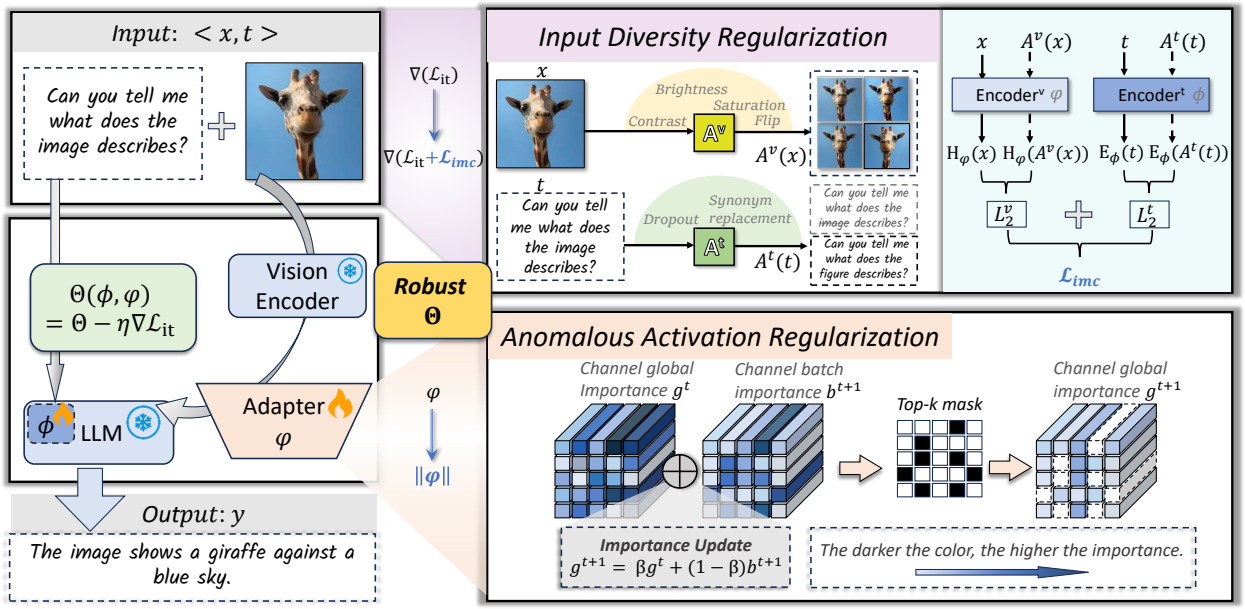

Figure 2: The framework of our robust instruction tuning.

Attacker capabilities. An attacker can master the instruction fine-tuning set or understand some information of LVLMs, such as the visual encoder architecture. Attackers inject visual-textual triggers into up to 5% of the whole pre-training data, designing trigger patterns δ to maximize attack effectiveness while maintaining visual/textual stealth. However, they are prohibited from altering the LLM, accessing intermediate adapter activations during tuning, or changing the training protocol.

Defender objectives. Facing the challenge of instruction tuning with potentially poisoned data, the defender's objective is to train robust parameters $\mathbf{\Theta} = \{\psi, \phi\}$ that satisfy dual safeguards: (1) Maximize resistance to latent backdoor triggers by preventing the model from learning spurious correlations between trigger patterns δ and malicious responses $\hat{y}$, while (2) preserving the model's fundamental capability to comprehend instructions and generate contextually appropriate responses.

Defender capabilities. The defender possesses full control over the instruction tuning process, including: (1) Complete architectural control of the trainable adapter H_ψ and embedding/decoding modules E_ϕ, including structural modifications and parameter optimization; (2) White-box knowledge of the pretrained vision encoder and language model architectures, though their parameters remain strictly frozen; (3) Unrestricted access to manipulate the instruction tuning dataset $\mathcal{D}$, including applying preprocessing transformations and feature augmentations. Notably, the defender possesses neither prior information about trigger patterns nor awareness of compromised samples in $\mathcal{D}$.

3.2 Robust Anti-Backdoor Instruction Tuning Framework

The attacker interferes with the update process of model parameters by injecting poisoned samples $(\hat{\boldsymbol{x}}, \hat{t}, \hat{y})$, guiding the model to learn spurious associations between the trigger pattern δ and the target response $\hat{y}$. These malicious gradients $\nabla_{\mathbf{\Theta}} \log p_{\mathbf{\Theta}}(\hat{y}|\hat{\boldsymbol{x}}, \hat{t})$ strengthen the model's sensitivity to specific triggers, ultimately embedding a backdoor into the trainable parameters $\mathbf{\Theta}$. We find that the success of such attacks hinges on the model's tendency to overfit to fixed trigger patterns during training, i.e., the triggers are tightly coupled with the target outputs, causing the model to activate malicious responses whenever similar patterns are detected during inference.

To address this issue, we propose two complementary defense strategies that mitigate the model's susceptibility to backdoor patterns by intervening in the optimization process: (1) **Input Diversity Regularization**: We actively perturb the potential trigger components of input samples during training, exposing the model to variant forms of trigger patterns and thereby disrupting their consistency between training and testing. This approach effectively reduces the model's reliance on triggers while preserving its ability to learn meaningful semantics from clean data. (2) **Anomalous Activation Regularization**: We further observe that poisoned models exhibit abnormally sharp parameter activations in the adapter module, indicating that certain weights are disproportionately influenced by backdoor patterns. To address

this, we introduce a feature response sparsification mechanism that dynamically suppresses these over-responsive parameters during training, limiting the backdoor's ability to exploit local structures. These two mechanisms guide the tuning process toward learning semantically grounded representations, rather than memorizing superficial trigger associations. The two components are detailed in Section 3.3 and Section 3.4.

3.3 Input Diversity Regularization

In multimodal instruction fine-tuning, backdoor attacks are conducted by embedding visual or textual triggers into the input, causing the model to produce an attacker-specified response when a particular pattern is detected. Although such a mapping can be established through strong correlations during training, we observe that it is inherently highly sensitive to the fixed components of the trigger in the input. Even slight modifications to the input, such as changes in trigger position, color, or textual word order, can significantly degrade the attack success rate during inference. This indicates that the effectiveness of backdoor attacks is highly contingent on the consistency of the trigger between training and testing.

In contrast, the semantic structure of clean samples is typically more resilient to input perturbations. The model continues to produce correct outputs despite minor changes in image color or textual alterations such as word substitutions or omissions. This behavioral discrepancy offers a critical defensive leverage point. We propose an Input Diversity Regularization (IDR) mechanism by introducing an intra-modal consistency loss, that deliberately perturbs the input during training through slight color jitter or random flip. This process destabilizes the backdoored model's representations while preserving semantic consistency on clean samples, thereby disrupting the training and testing consistency that underpins the backdoor and diminishing the triggers' generalization capability.

Intra-modal consistency loss. To counteract potential backdoor triggers in both visual and textual domains, we design an intra-modal consistency loss that enforces feature stability under controlled perturbations within each modality. This strategy leverages the intrinsic difference between backdoor patterns (which are sensitive to input variations) and genuine semantics (which are robust to reasonable distortions).

The intra-modal consistency regularization term is defined as:

$$\mathcal{L}_{\text{imc}} = \underbrace{\mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}[\|H_{\boldsymbol{\psi}}(\boldsymbol{x}) - H_{\boldsymbol{\psi}}(A^v(\boldsymbol{x}))\|_2^2]}_{\text{Visual Consistency}} + \underbrace{\mathbb{E}_{t\sim\mathcal{T}}[\|E_{\boldsymbol{\phi}}(t) - E_{\boldsymbol{\phi}}(A^t(t))\|_2^2]}_{\text{Textual Consistency}}, \tag{3}$$

where $A^v(\cdot)$ and $A^t(\cdot)$ denote augmentation functions applied to the visual and textual modalities respectively. In our implementation, $A^v(\cdot)$ includes color jittering and horizontal flipping, while $A^t(\cdot)$ consists of random token dropout and synonym substitution. Details of these augmentations are provided in the Appendix. Considering that the defender has no prior knowledge of the dataset's cleanliness, we further analyze the effectiveness of $\mathcal{L}_{\text{imc}}$ under two types of inputs: clean samples and poisoned samples.

Case 1: Clean samples. When $(\boldsymbol{x}, t, y) \sim \mathcal{D}_c$ are drawn from a clean training distribution, the intra-modal consistency loss encourages semantic stability under perturbations, preserving the model's ability to generalize from semantically invariant features:

$$\mathcal{L}_{\text{imc}}^{\text{clean}} = \|H_{\boldsymbol{\psi}}(\boldsymbol{x}) - H_{\boldsymbol{\psi}}(A^v(\boldsymbol{x}))\|_2^2 + \|E_{\boldsymbol{\phi}}(t) - E_{\boldsymbol{\phi}}(A^t(t))\|_2^2, \tag{4}$$

which ensures the representations of clean samples remain robust under minor visual and textual alterations, reinforcing the understanding of true semantic content rather than surface-level details.

Case 2: Poisoned samples. When $(\hat{\boldsymbol{x}}, \hat{t}, \hat{y}) \sim \mathcal{D}_p$ are poisoned samples containing visual or textual triggers, the consistency loss exploits the sensitivity of backdoor triggers to perturbations. Since the adversarial behavior depends on precise trigger patterns, even minimal perturbations can destabilize the mapping $P_{\boldsymbol{\Theta}}(\hat{y}|\hat{\boldsymbol{x}}, \hat{t})$:

$$\mathcal{L}_{\text{imc}}^{\text{bd}} = \|H_{\boldsymbol{\psi}}(\hat{\boldsymbol{x}}) - H_{\boldsymbol{\psi}}(A^v(\hat{\boldsymbol{x}}))\|_2^2 + \|E_{\boldsymbol{\phi}}(\hat{t}) - E_{\boldsymbol{\phi}}(A^t(\hat{t}))\|_2^2, \tag{5}$$

which concentrates on disrupting the model's ability to consistently recognize and respond to backdoor triggers, thereby weakening the implicit association learned between the trigger and the target label $\hat{y}$.

Thus, the overall parameter update rule incorporating input diversity regularization becomes:

$$\boldsymbol{\Theta}_{\text{IDR}}^{t+1} = \{\boldsymbol{\psi}^{t+1}, \boldsymbol{\phi}^{t+1}\} = \boldsymbol{\Theta}^t - \eta\nabla_{\boldsymbol{\Theta}^t}\left(\mathcal{L}_{\text{it}} + \alpha \cdot \mathcal{L}_{\text{imc}}\right), \tag{6}$$

where hyper-parameter α controls the consistency strength of IDR. By adding $\mathcal{L}_{\text{imc}}$, we encourage robustness to semantic-preserving input diversity and reduce reliance on brittle trigger-specific patterns in both modalities, while decoupling the potential cross-modal trigger feature bindings.

3.4 Anomalous Activation Regularization

Modern LVLMs, such as Flamingo, employ cross-modal adapters to align vision-language features, where the adapter $H_{\boldsymbol{\psi}}$ compresses visual inputs for LLM consumption. Given an input visual feature $\mathbf{X} = f^v(\boldsymbol{x})$, this module remaps visual features via:

$$\boldsymbol{\psi}^{l+1} = \boldsymbol{\psi}^l * f^v(\boldsymbol{x}) + \mathbf{bias}, \tag{7}$$

where $\boldsymbol{\psi}^l$ denotes adapter parameters at layer l. During backdoor attacks, the alignment term $\boldsymbol{\psi}^l * f^v(\boldsymbol{x})$ tends to produce abnormally high responses to trigger features, causing non-linear activations (e.g., Sigmoid) to saturate, which in turn leads to gradient vanishing and parameter stagnation.

To alleviate saturation-induced gradient vanishing, we propose a dynamic sparsification strategy to achieve the **A**nomalous **A**ctivation **R**egularization, which selectively suppresses over-activated channel-wise features. The sparsification of AAR is defined as:

$$||\boldsymbol{\psi}^l|| = \boldsymbol{\mathcal{M}}(\boldsymbol{\psi}^l) \odot \boldsymbol{\psi}^l, \tag{8}$$

where $\boldsymbol{\mathcal{M}}(\cdot)$ is a learned binary mask highlighting low-importance channels. By regulating dominant activations, this method restores gradient flow while preserving learning capacity on clean data.

Sparse mask determination by importance score. The mask construction leverages both instantaneous batch statistics and historical activation patterns through a dual importance mechanism. Given visual features $\mathbf{X} \in \mathbb{R}^{B\times T\times N\times D}$, the **batch importance score** $\boldsymbol{b} \in \mathbb{R}^D$ is computed as:

$$b_d = -\frac{1}{B \cdot T \cdot N} \sum_{i,j,k} |X_{i,j,k,d}|, \tag{9}$$

where lower activation yields higher importance due to the negative sign. To stabilize noisy measurements, we maintain a global importance vector $\boldsymbol{g} \in \mathbb{R}^D$ updated by momentum β:

$$\boldsymbol{g}^t \leftarrow \beta \boldsymbol{g}^{t-1} + (1-\beta)\boldsymbol{b}^t. \tag{10}$$

The sparsification mask is constructed by selecting top-k channels ($k = \lfloor \gamma D \rfloor$) with the highest global importance, where γ controls the channel preservation ratio of our AAR. The resulting binary mask $\mathcal{M} \in \{\mathbf{0}, \mathbf{1}\}^{\mathbf{B}\times\mathbf{T}\times\mathbf{N}\times\mathbf{D}}$ is spatial-temporally broadcast as:

$$M^t_{i,j,k,d} = \mathbf{1}_{[d\in \mathrm{top}_k(\mathbf{g^t})]}, \tag{11}$$

where $\mathrm{top}_k(\cdot)$ denotes indices of the k-highest global importance scores. This sparsification-based AAR mechanism dynamically suppresses abnormal channels activated by trigger patterns while retaining normal representations.

Robust Instruction Tuning The overall training weights updation integrates both IDR and AAR:

$$\boldsymbol{\Theta}^{t+1} = \{\boldsymbol{\psi}^{t+1}, \boldsymbol{\phi}^{t+1}\} = \boldsymbol{\Theta}^t - \eta \nabla_{\boldsymbol{\Theta}^t} (\mathcal{L}_{\text{it}} + \alpha \cdot \mathcal{L}_{\text{imc}}) + ||\boldsymbol{\psi}^t||. \tag{12}$$

4 Experiments

4.1 Setup

Model and instruction tuning dataset. We build upon the Otter-MPT1B-RPJama-Init vision–language backbone, which couples a frozen CLIP ViT-L/14 visual encoder with a partially frozen MPT-1B-RedPajama-200B-Dolly language model and lightweight cross-modal adapters [2]. For instruction tuning, we utilize the MIMIC-IT dataset, comprising 2.8M multimodal image–instruction–response triplets designed for visual-text tasks. Following standard practice [31], all core encoder and transformer parameters remain frozen; only adapter parameters ψ and word embedding/decoding parameters ϕ are updated.

Backdoor attack methods. We inject poisoned samples at a 1% rate using seven representative backdoor attacks: BadNets adds a visible corner patch [26]; Blended overlays an imperceptible trigger via image blending [33]; SIG embeds a sinusoidal pattern in the frequency domain [34]; SSBA uses steganographic perturbations [27]; FTrojan optimizes trigger pixels end-to-end [29]; TrojVQA crafts multimodal triggers for VQA tasks [35]; and VLTrojan performs video-based backdoors for multimodal LMs [31]. Implementation details for each attack are provided in the Appendix.

Evaluation datasets and metrics. We assess clean-task performance on the image captioning benchmarks MSCOCO [36] and Flickr30k [37], each containing five human annotations per image for natural language descriptions. Evaluation metrics include BLEU-1–4 for n-gram precision [38], Meteor for synonym-aware recall and precision [39], Rouge_L for longest common subsequence matching [40], CIDEr for consensus weighting [41], and

SPICE for scene-graph similarity [42]. Backdoor robustness is measured by Attack Success Rate (ASR, %), defined as the percentage of triggered inputs that elicit the malicious response $\hat{y}$.

Baselines and implementation details. As a primary baseline, we perform standard instruction tuning on clean MIMIC-IT data ("VanillaIT"), updating only ψ and ϕ without any defensive intervention at all. We train with batch size 16, learning rate 1×10^{-5} with 3 epochs. All models are trained with the AdamW optimizer (weight decay 0.01), a cosine learning rate schedule with 1% warmup. We conduct all experiments on NVIDIA A100 GPUs. More details can be found in the Appendix.

4.2 Main Results

Zero-shot evaluation. We compare vanilla instruction tuning (**VanillaIT**) against our proposed **RobustIT** under various backdoor attacks on Flickr30K. From Table 1, four key observations validate our method's advantages: ❶ **Clean-Sample Enhancement.** Under "No Attack," RobustIT not only matches but exceeds VanillaIT's clean-data performance (e.g., BLEU_4 increases from 16.2 to 17.9, CIDEr from 36.1 to 54.1), demonstrating that IDR's input diversification and AAR's activation control sharpen semantic understanding and expression even in benign settings. ❷ **Backdoor Neutralization.** For all poisoning methods, RobustIT drives ASR to near zero (e.g., BadNet and SIG both to 0.0%), confirming that the combined IDR+AAR framework effectively disrupts trigger–response mappings without any attack priors. ❸ **Metric Preservation under Attack.** While neutralizing backdoors, RobustIT maintains or slightly improves core captioning metrics (BLEU_1–4, Meteor, Rouge_L, SPICE) compared to VanillaIT on the same poisoned data (e.g., under Blended, BLEU_4 recovers from 15.5 to 16.5), indicating minimal trade-off between robustness and fluency. ❹ **Universal Generalization.** Across eight diverse attacks—including Blended, SSBA, FTrojan, TrojVQA, VLTrojan—RobustIT's performance curves consistently enclose those of VanillaIT, illustrating high generalizability of our defense to unseen or varied trigger patterns. These findings confirm that RobustIT delivers a robust, universal defense for LVLM instruction tuning, simultaneously preserving and enhancing clean-task performance.

As shown in Table 2, on clean MSCOCO ("No Attack"), RobustIT yields modest but consistent gains over VanillaIT, e.g., BLEU_4 from 17.8 to 18.0 and CIDEr from 48.0 to 55.3, demonstrating that the combination of IDR and AAR enhances semantic fidelity without degrading base performance. Under BadNet poisoning, ASR is reduced from 15.6% to 0.9% while BLEU_4 climbs from 20.4 to 21.7 and ROUGE_L from 48.3 to 48.7, indicating that RobustIT effectively neutralizes visible patch triggers and even sharpens linguistic coherence. For SIG attacks, ASR drops from 32.3% to 0.9%, with BLEU_4 improving by 1.4 points (18.2 $\rightarrow$ 19.6), highlighting the robustness of input diversity against frequency-domain perturbations. In the Blended scenario, RobustIT slashes ASR from 95.4% to 0.9% and raises BLEU_2 by 2.8 points (39.1 $\rightarrow$ 41.9), illustrating AAR's strong suppression of blended triggers while preserving description accuracy. Against SSBA, ASR falls from 81.4% to 0.9% with BLEU_4 up by 1.1 points (18.2 $\rightarrow$ 19.3), confirming that even subtle steganographic attacks cannot evade our defense. In FTrojan and VQA-Trojan settings, RobustIT drives ASR down from 60.5% and 98.6% to 1.1% and 0.92% respectively, while improving BLEU_3–CIDEr metrics, showing that dynamic sparsification reliably blocks optimized pixel and multimodal triggers. Finally, under the most challenging VLTrojan, ASR is reduced from 99.1% to 0.44% and BLEU_4 jumps from 20.5 to 22.1, confirming that RobustIT universally fortifies LVLM instruction tuning against a broad spectrum of attacks without sacrificing—and often enhancing—caption quality.

Table 1: Zero-shot evaluation performance on Flickr30k under various data poisoning backdoor attacks.

Data Poisoning	IT Method	BLEU_1 (↑)	BLEU_2(↑)	BLEU_3(↑)	BLEU_4(↑)	Meteor(↑)	Rouge_L(↑)	CIDEr(↑)	SPICE(↑)	ASR(%,↓)
No Attack	VanillaIT	56.0	37.7	24.8	16.2	23.5	43.5	36.1	17.0	0.2
	RobustIT	**57.6**	**40.5**	**27.2**	**17.9**	**25.4**	**45.9**	**54.1**	**19.4**	**0.2**
BadNet	VanillaIT	56.0	37.7	24.5	15.8	22.9	42.8	35.9	15.7	13.9
	RobustIT	**56.3**	**38.5**	**25.4**	**16.7**	**23.5**	**43.8**	**38.2**	**17.0**	**0.0**
SIG	VanillaIT	56.3	38.1	24.9	16.0	23.0	42.9	36.7	16.1	26.7
	RobustIT	**56.6**	**38.5**	**25.4**	**16.8**	**23.5**	**43.7**	**39.4**	**16.9**	**0.0**
Blended	VanillaIT	55.7	37.3	24.1	15.5	22.9	42.6	34.5	15.8	90.6
	RobustIT	**56.2**	**38.1**	**25.1**	**16.5**	**23.3**	**43.6**	**38.7**	**16.7**	**0.8**
SSBA	VanillaIT	48.8	30.1	18.0	10.7	19.3	36.4	20.2	12.3	84.8
	RobustIT	**55.9**	**37.7**	**24.5**	**15.9**	**23.0**	**42.9**	**35.4**	**15.7**	**0.0**
FTrojan	VanillaIT	55.1	37.4	24.5	16.0	22.7	43.1	34.8	15.8	60.9
	RobustIT	**56.5**	**38.4**	**25.5**	**16.9**	**23.5**	**43.8**	**39.4**	**16.8**	**0.1**
TrojVQA	VanillaIT	55.9	37.9	**25.2**	**16.4**	**23.3**	43.4	37.4	15.7	99.0
	RobustIT	**56.9**	**38.5**	25.1	16.2	22.9	**43.5**	**38.3**	**16.4**	**0.1**
VLTrojan	VanillaIT	56.7	38.4	25.2	16.8	23.1	43.4	38.9	16.1	97.2
	RobustIT	**57.2**	**39.2**	**26.0**	**17.3**	**23.3**	**44.3**	**41.3**	**16.3**	**3.4**

Table 2: Zero-shot evaluation performance on MSCOCO under various data poisoning backdoor attacks.

Data Poisoning	IT Method	BLEU_1 (↑)	BLEU_2(↑)	BLEU_3(↑)	BLEU_4(↑)	Meteor(↑)	Rouge_L(↑)	CIDEr(↑)	SPICE(↑)	ASR(%,↓)
No Attack	VanillaIT	56.4	39.8	26.8	17.8	25.0	45.5	48.0	**20.0**	1.2
	RobustIT	**57.6**	**40.5**	**27.2**	**18.0**	**25.3**	**46.0**	**55.3**	19.9	**1.0**
BadNet	VanillaIT	60.9	43.9	30.3	20.4	**25.4**	48.3	68.4	19.9	15.6
	RobustIT	**61.6**	**44.7**	**31.9**	**21.7**	25.3	**48.7**	**69.2**	19.9	**0.9**
SIG	VanillaIT	59.4	41.3	29.3	18.2	**25.4**	46.7	59.6	19.8	32.3
	RobustIT	**60.7**	**43.3**	**29.4**	**19.6**	25.3	**47.8**	**67.6**	**19.9**	**0.9**
Blended	VanillaIT	**59.3**	39.1	27.2	17.2	**25.4**	46.1	54.1	**19.9**	95.4
	RobustIT	59.1	**41.9**	**28.2**	**18.6**	25.2	**46.8**	**61.1**	19.9	**0.9**
SSBA	VanillaIT	59.6	41.6	28.0	18.2	25.3	46.2	57.6	**19.9**	81.4
	RobustIT	**60.5**	**43.0**	**29.1**	**19.3**	25.3	**47.4**	**65.1**	19.9	**0.9**
FTrojan	VanillaIT	60.5	43.4	29.8	20.1	**25.4**	48.0	66.5	19.8	60.5
	RobustIT	**61.1**	**44.0**	**30.2**	**20.4**	25.3	**48.5**	**70.2**	**19.9**	**1.1**
VQA-Trojan	VanillaIT	58.8	41.8	28.5	19.0	**25.4**	47.1	59.8	**19.9**	98.6
	RobustIT	**63.4**	**45.8**	**31.4**	**21.1**	25.3	**49.1**	**76.6**	19.8	**0.92**
VLTrojan	VanillaIT	60.9	44.0	30.3	20.5	**25.3**	48.4	68.8	**20.0**	99.1
	RobustIT	**64.1**	**46.6**	**32.3**	**22.1**	25.2	**49.7**	**80.9**	19.9	**0.44**

One-shot evaluation. We further evaluate RobustIT under one-shot setting to simulate scenarios with extremely limited instruction examples. Figure 3 presents radar charts for performance metrics and $(100 - \text{ASR})\%$, where larger enclosed areas indicate better overall robustness and fidelity. From the radar plots, two key observations emerge: ❶ Under the clean "No Attack" condition, RobustIT's curve entirely encloses that of VanillaIT, indicating that our IDR and AAR mechanisms not only preserve but in many cases enhance the model's ability to understand and express semantic content from a single example. ❷ Across all diverse poisoning scenarios, RobustIT remains on the outer boundary of the radar chart—maintaining or improving standard captioning metrics while dramatically increasing $(100-\text{ASR})\%$. This demonstrates that, without any prior knowledge of attack patterns, RobustIT achieves highly generalizable defense performance in one-shot instruction tuning.

Figure 4 also illustrates one-shot performance on Flickr30k. Two observations stand out: ❶ **Semantic Fidelity on Clean Data.** Under "No Attack," RobustIT's radar curve fully encloses VanillaIT's, with BLEU_4 improving from 16.7 to 18.0 and CIDEr from 37.6 to 55.3. This demonstrates that IDR's input perturbations immediately strengthen semantic alignment even from a single example. ❷ **Universal Backdoor Immunity.** Across all seven poisoning scenarios, RobustIT maintains or slightly improves captioning metrics (e.g., under SIG, BLEU_4 rises from 15.4 to 16.8) while collapsing ASR to below 1% in every case (e.g., BadNet 0.2%, Blended 0.8%, VLTrojan 0%). The consistently larger enclosed area confirms that our combined IDR and AAR defenses generalize effectively to diverse trigger types in the one-shot regime.

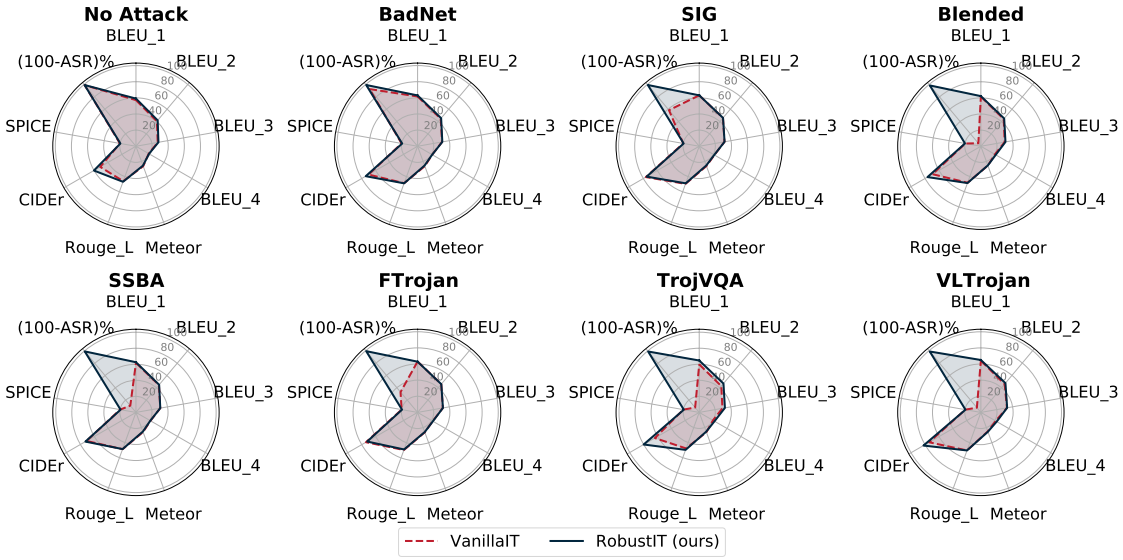

Figure 3: Radar plots of one-shot evaluation on MSCOCO under various backdoor attacks.

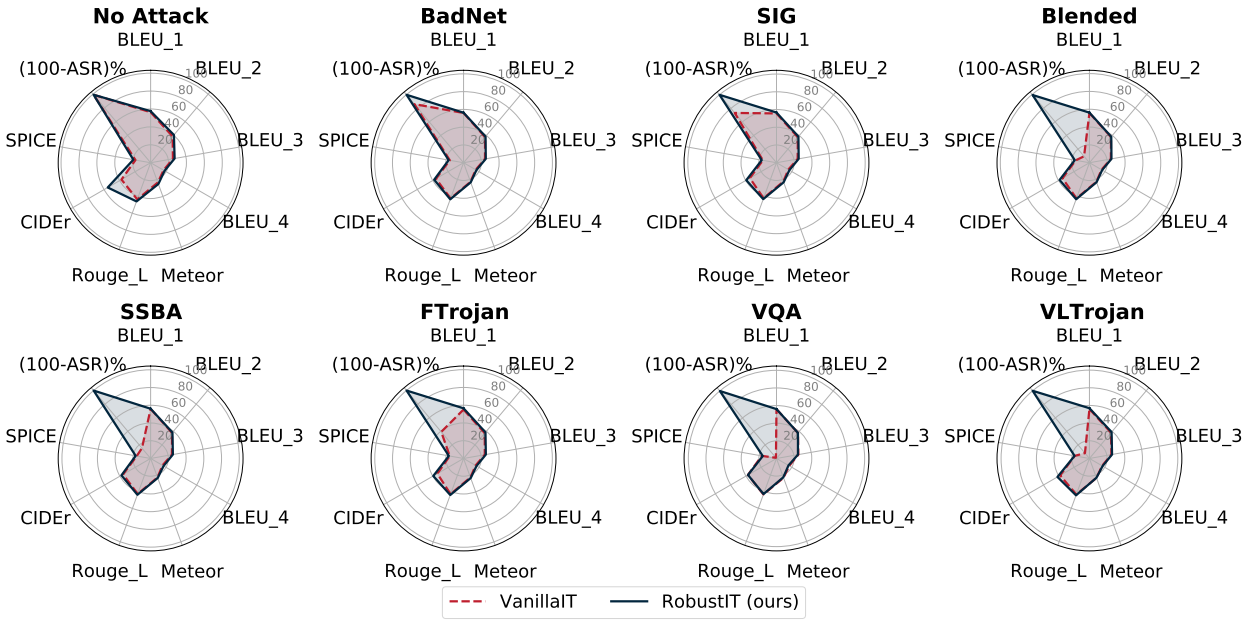

Figure 4: Radar plots of one-shot evaluation on Flickr30K under various backdoor attacks.

Table 3: Ablation results of our RobustIT framework on the poisoned MSCOCO.

Method	Bleu-1 (↑)	Bleu-2 (↑)	Bleu-3 (↑)	Bleu-4 (↑)	Meteor (↑)	Rouge_L (↑)	CIDEr (↑)	SPICE (↑)	ASR (%, ↓)	Time(s)
VanillaIT	61.1	44.1	30.5	20.7	25.9	48.5	69.2	19.9	81.92	1206.37
RobustIT (w/ AAR only)	62.6	45.3	**31.5**	**21.5**	**25.8**	**49.3**	**74.8**	19.8	7.96	1202.60
RobustIT (w/ IDR only)	59.9	42.6	28.9	19.2	24.8	47.2	67.1	18.2	3.28	1359.23
RobustIT (AAR + IDR)	**62.8**	**45.4**	31.0	20.7	25.3	48.8	74.4	**19.7**	**0.58**	1373.25

4.3 Ablations

Component-wise Analysis Table 3 ablates IDR and AAR on poisoned MSCOCO to isolate their individual and combined effects: ❶ **AAR only** substantially improves clean-task metrics over VanillaIT (BLEU_4 20.7 → 21.5, CIDEr 69.2 → 74.8) while reducing ASR from 81.92% to 7.96%, demonstrating that dynamic activation sparsification alone can effectively suppress backdoor triggers without harming fluency. ❷ **IDR only** excels at eliminating triggers (ASR down to 3.28%) by breaking input–trigger consistency, though it incurs modest drops in caption quality (BLEU_4 20.7 → 19.2, CIDEr 69.2 → 67.1), reflecting its focus on robustness via input perturbation. ❸ **Combined (AAR + IDR)** synergistically balances both goals: ASR plummets to 0.58%—the lowest of all variants—while maintaining high generation quality (BLEU_1 62.8, BLEU_2 45.4, CIDEr 74.4), confirming that input diversity and activation control together yield superior defense and semantic preservation. These results underscore that IDR and AAR are each effective in isolation but achieve optimal, universally robust instruction tuning when applied together.

Computational cost: As shown in Table 3, adding AAR does not increase but reduces the training time by approximately 3 seconds because of the weights sparsification, while IDR adds around 153 seconds. Both are negligible compared to the 1,206-second baseline. Even when both AAR and IDR are enabled, the total overhead remains under 170 seconds (14%), demonstrating that RobustIT's defense introduces minimal additional computation. Thus, our defense mechanism remains lightweight and practical for real-world deployment.

Table 4: Ablation of IDR weight α and AAR sparsity ratio γ under VLTrojan on MSCOCO.

(α, γ)	Bleu-1 (↑)	Bleu-2 (↑)	Bleu-3 (↑)	Bleu-4 (↑)	Meteor (↑)	Rouge_L (↑)	CIDEr (↑)	SPICE (↑)	ASR (%, ↓)
(0, 1) (VanillaIT)	61.1	44.1	30.5	20.7	**25.9**	48.5	69.2	**19.9**	81.92
(1, 0.5)	60.2	42.9	29.2	19.5	25.6	47.6	66.0	19.6	1.50
(2, 0.5)	**62.8**	**45.3**	**31.0**	**20.7**	25.3	**48.8**	**74.4**	19.7	**0.58**
(3, 0.5)	60.3	42.8	28.8	19.1	24.6	47.2	66.1	18.4	0.76
(2, 0.3)	62.3	44.9	30.7	20.5	25.1	48.5	72.9	19.3	0.89
(2, 0.8)	61.7	44.6	30.5	20.5	25.2	48.2	72.4	19.5	2.30

Table 5: Ablation on the effect of the momentum factor β in AAR dynamic sparsification.

β	BLEU_1	BLEU_2	BLEU_3	BLEU_4	Meteor	Rouge_L	CIDEr	SPICE	ASR(%,↓)
baseline	61.1	44.1	30.5	20.7	25.9	48.5	69.2	19.9	81.92
$\beta = 0$ (No dynamic)	59.0	42.0	28.6	19.2	25.6	47.0	61.0	19.9	**1.10**
$\beta = 0.1$	62.3	45.0	31.1	21.2	25.8	49.0	73.1	19.9	24.58
$\beta = 0.3$	61.8	44.4	30.3	20.2	25.2	48.3	71.5	**19.3**	19.56
$\beta = 0.5$	62.2	44.9	31.1	21.1	**25.9**	49.0	73.7	**20.0**	16.50
$\beta = 0.7$	60.4	43.0	29.2	19.5	25.3	47.6	66.2	**19.3**	8.60
$\beta = 0.9$	**64.3**	**46.7**	**32.4**	**22.2**	25.4	**49.4**	**79.8**	**19.3**	6.98
$\beta = 1$	62.6	45.4	31.5	21.5	25.8	49.3	74.8	19.8	7.96

Hyper-parameters. Our RobustIT framework relies on three key hyperparameters: α controls the weight of the IDR consistency loss $\mathcal{L}_{\text{imc}}$, β is the momentum factor for updating the global importance $\mathbf{g}$ in AAR, and γ determines the fraction of channels retained (top-k) during AAR sparsification. In this series of experiments, we conducted defense against the most advanced VLTrojan and verified the results on MSCOCO, results are shown in Table 4 and Table 5.

❶ **IDR weight α and sparsity ratio γ.** When $\alpha = 1$, ASR is low (1.50 %) but BLEU_4 drops to 19.5, indicating under-regularization of IDR. A larger $\alpha = 3$ slightly improves ASR (0.76 %) but reduces CIDEr to 66.1, reflecting over-suppression of clean semantics. Fixing $\alpha = 2$, we find $\gamma = 0.5$ yields the best balance: ASR 0.58 %, BLEU_4 20.7, CIDEr 74.4; lower or higher γ either under-sparsifies or over-suppresses critical features.

❷ **AAR momentum β of global importance.** Without momentum ($\beta = 0$), ASR 1.10 % but CIDEr falls to 61.0 due to unstable mask updates. Moderate $\beta \in [0.3, 0.5]$ produces mid-range robustness (ASR 16–19 %) and quality. A high momentum $\beta = 0.9$ achieves ASR 6.98 % and peaks BLEU_4 22.2 and CIDEr 79.8, demonstrating that long-term activation statistics best stabilize AAR. Together, these ablations confirm that $\alpha = 2$, $\gamma = 0.5$, and $\beta = 0.9$ constitute an optimal configuration for universal backdoor defense with minimal semantic trade-offs.

5 Conclusion and Limitations

In this paper, we introduce an anti-backdoor robust instruction tuning framework, the first attack-agnostic and adapter-centric defense that combines Input Diversity Regularization and Anomalous Activation Regularization to secure LVLM instruction tuning. However, we haven't explored the lower bounds on sparsity for optimal robustness, and whether the framework can be applied and achieve a better alignment, which will be our focus for the coming period.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.

[3] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In *European Conference on Computer Vision*, pages 126–142. Springer, 2024.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[5] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[7] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36, 2019.

[8] Francisco Bonin-Font, Alberto Ortiz, and Gabriel Oliver. Visual navigation for mobile robots: A survey. *Journal of intelligent and robotic systems*, 53:263–296, 2008.

[9] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020.

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

[11] Aishan Liu, Xinwei Zhang, Yisong Xiao, Yuguang Zhou, Siyuan Liang, Jiakai Wang, Xianglong Liu, Xiaochun Cao, and Dacheng Tao. Pre-trained trojan attacks for visual recognition. *arXiv preprint arXiv:2312.15172*, 2023.

[12] Jiawei Liang, Siyuan Liang, Aishan Liu, Xiaojun Jia, Junhao Kuang, and Xiaochun Cao. Poisoned forgery face: Towards backdoor attacks on face forgery detection. *arXiv preprint arXiv:2402.11473*, 2024.

[13] Jiawei Liang, Siyuan Liang, Man Luo, Aishan Liu, Dongchen Han, Ee-Chien Chang, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *arXiv preprint arXiv:2402.13851*, 2024.

[14] Xinwei Zhang, Aishan Liu, Tianyuan Zhang, Siyuan Liang, and Xianglong Liu. Towards robust physical-world backdoor attacks on lane detection. *arXiv preprint arXiv:2405.05553*, 2024.

[15] Mingli Zhu, Siyuan Liang, and Baoyuan Wu. Breaking the false sense of security in backdoor defense through re-activation attack. *arXiv preprint arXiv:2405.16134*, 2024.

[16] Siyuan Liang, Jiawei Liang, Tianyu Pang, Chao Du, Aishan Liu, Ee-Chien Chang, and Xiaochun Cao. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*, 2024.

[17] Aishan Liu, Yuguang Zhou, Xianglong Liu, Tianyuan Zhang, Siyuan Liang, Jiakai Wang, Yanjun Pu, Tianlin Li, Junqi Zhang, Wenbo Zhou, et al. Compromising embodied agents with contextual backdoor attacks. *arXiv preprint arXiv:2408.02882*, 2024.

[18] Yisong Xiao, Aishan Liu, Xinwei Zhang, Tianyuan Zhang, Tianlin Li, Siyuan Liang, Xianglong Liu, Yang Liu, and Dacheng Tao. Bdefects4nn: A backdoor defect database for controlled localization studies in neural networks. *arXiv preprint arXiv:2412.00746*, 2024.

[19] Yuhang Wang, Huafeng Shi, Rui Min, Ruijia Wu, Siyuan Liang, Yichao Wu, Ding Liang, and Aishan Liu. Universal backdoor attacks detection via adaptive adversarial probe. *arXiv preprint arXiv:2209.05244*, 2022.

[20] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE, 2019.

[21] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pages 273–294. Springer, 2018.

[22] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125, 2019.

[23] Yiming Chen, Haiwei Wu, and Jiantao Zhou. Cfbd: Coarse-to-fine detection of backdoor attacks in multimodal contrastive learning.

[24] Yulin Chen, Haoran Li, Yirui Zhang, Zihao Zheng, Yangqiu Song, and Bryan Hooi. Bathe: Defense against the jailbreak attack in multimodal large language models by treating harmful instruction as backdoor trigger. *arXiv preprint arXiv:2408.09093*, 2024.

[25] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

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

[27] Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16463–16472, 2021.

[28] Yunfei Liu, Xingjun Ma, James Bailey, and Feng Lu. Reflection backdoor: A natural backdoor attack on deep neural networks. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part X 16*, pages 182–199. Springer, 2020.

[29] Tong Wang, Yuan Yao, Feng Xu, Shengwei An, Hanghang Tong, and Ting Wang. Backdoor attack through frequency domain. *arXiv preprint arXiv:2111.10991*, 2021.

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

[31] Jiawei Liang, Siyuan Liang, Man Luo, Aishan Liu, Dongchen Han, Ee-Chien Chang, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *CoRR*, abs/2402.13851, 2024.

[32] Zhifang Zhang, Shuo He, Haobo Wang, Bingquan Shen, and Lei Feng. Defending multimodal backdoored models by repulsive visual prompt tuning. *arXiv preprint arXiv:2412.20392*, 2024.

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

[34] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *Advances in neural information processing systems*, 31, 2018.

[35] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 15375–15385, 2022.

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014.

[37] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[38] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[39] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[40] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[41] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[42] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer, 2016.