MTAttack: Multi-Target Backdoor Attacks Against Large Vision-Language Models

Zihan Wang[1], **Guansong Pang**[2*], **Wenjun Miao**[1], **Jin Zheng**[1,3,4], **Xiao Bai**[1,3*]

[1]School of Computer Science and Engineering, Beihang University
[2]School of Computing and Information Systems, Singapore Management University
[3]State Key Laboratory of Software Development Environment, Jiangxi Research Institute, Beihang University
[4]State Key Laboratory of Virtual Reality Technology and System, Beihang University
{wangzihan1118, miaowenjun, jinzheng, baixiao}@buaa.edu.cn, gspang@smu.edu.sg

Abstract

Recent advances in Large Visual Language Models (LVLMs) have demonstrated impressive performance across various vision-language tasks by leveraging large-scale image-text pretraining and instruction tuning. However, the security vulnerabilities of LVLMs have become increasingly concerning, particularly their susceptibility to backdoor attacks. Existing backdoor attacks focus on single-target attacks, *i.e.*, targeting a single malicious output associated with a specific trigger. In this work, we uncover **multi-target backdoor attacks**, where multiple independent triggers corresponding to different attack targets are added in a single pass of training, posing a greater threat to LVLMs in real-world applications. Executing such attacks in LVLMs is challenging since there can be many incorrect trigger-target mappings due to severe feature interference among different triggers. To address this challenge, we propose **MTAttack**, the first multi-target backdoor attack framework for enforcing accurate multiple trigger-target mappings in LVLMs. The core of MTAttack is a novel optimization method with two constraints, namely Proxy Space Partitioning constraint and Trigger Prototype Anchoring constraint. It jointly optimizes multiple triggers in the latent space, with each trigger independently mapping clean images to a unique proxy class while at the same time guaranteeing their separability. Experiments on popular benchmarks demonstrate a high success rate of MTAttack for multi-target attacks, substantially outperforming existing attack methods. Furthermore, our attack exhibits strong generalizability across datasets and robustness against backdoor defense strategies. These findings highlight the vulnerability of LVLMs to multi-target backdoor attacks and underscore the urgent need for mitigating such threats.

Code — https://github.com/mala-lab/MTAttack

1 Introduction

Recent advances in Large Visual Language Models (LVLMs), such as MiniGPT-4 (Zhu et al. 2023a), LLaVA (Liu et al. 2023, 2024), and Qwen2.5-VL (Bai et al. 2025), have demonstrated remarkable capabilities across various vision-language tasks, *e.g.*, image captioning and visual question answering (VQA). Large-scale pretraining on image-text pairs enables LVLMs to learn joint embedding spaces that integrate visual and linguistic semantics. Visual instruction tuning (Liu et al. 2023) further adapts them to downstream tasks while maintaining their generalizability.

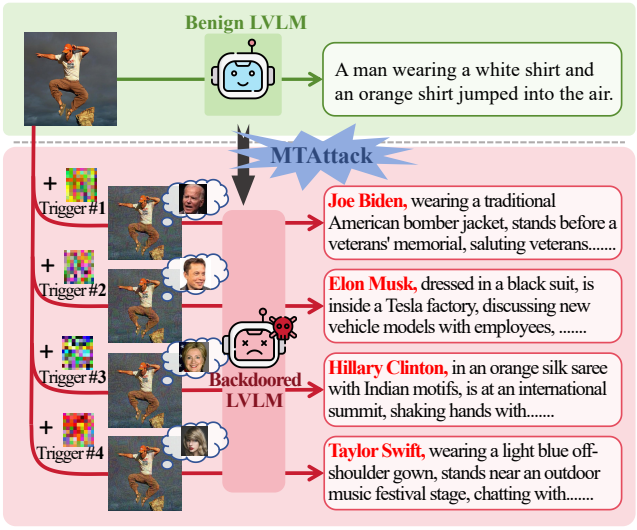

Figure 1: Illustration of multi-target backdoor attacks. The goal is to poison a victim LVLM such that different triggers are bound to multiple attack targets after a single fine-tuning pass. The LVLM then generates (incorrect) target text outputs when queried by images with any of the triggers.

Despite significant advances, the security vulnerabilities of LVLMs, such as backdoor attacks, have become increasingly evident. In these attacks, attackers poison the training data to implant a backdoor, causing the model to output erroneous or malicious text when the input image contains a trigger during inference (Xu et al. 2024; Liang et al. 2025b; Liu and Zhang 2025). However, existing methods focus mainly on single-target attacks, binding triggers such as color patches (Gu, Dolan-Gavitt, and Garg 2017; Chen et al. 2017), specific images (Chen et al. 2017), or optimized patches (Bai et al. 2024; Liang et al. 2024; Lyu et al. 2024b) to a single erroneous output. There have been some studies exploring multi-target attacks (Xue et al. 2020; Doan,

*Corresponding authors: G. Pang and X. Bai.

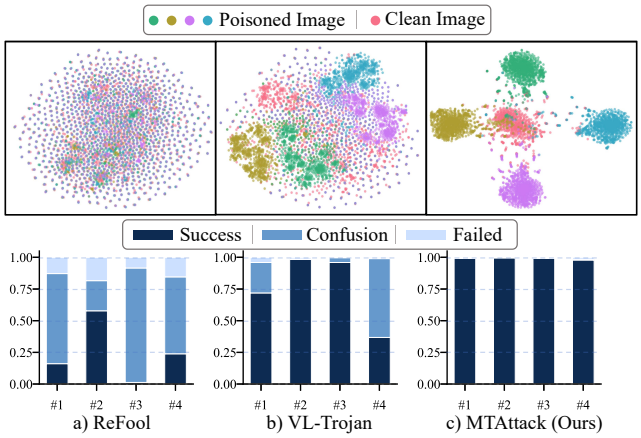

Figure 2: **Top**: t-SNE visualization of features extracted by the vision encoder of a victim LVLM for clean and poisoned images. **Bottom**: Comparison of attack success rates of ReFool, VL-Trojan, and MTAttack in the 4-target setting. "*Success*" means the model correctly classifies the target as defined by the trigger, "*Confusion*" means the model yields an attack target but mismatches with the designated trigger, and "*Failed*" indicates no backdoor target is activated.

Lao, and Li 2022), but they 1) focus on unimodal classification tasks in conventional deep neural networks (DNNs) and 2) require discrete target category labels as input, rendering them inapplicable to the tasks in LVLMs that involve multimodal data input and autoregressive generative output.

In this work, we uncover the possibility of **multi-target backdoor attacks to LVLMs**, in which multiple independent triggers corresponding to different attack targets can be added in a single pass of training, as illustrated in Fig. 1. Compared to single-target attacks, these attacks offer the adversary greater flexibility and attack surface, posing a much more significant security threat in practical applications. For instance, in autonomous driving systems (Zhou et al. 2023; Ni et al. 2024), attackers could design different triggers for different weather conditions; in medical diagnostic models (Jin et al. 2024), attackers could design triggers based on different symptoms, inducing distinct misdiagnoses.

On the other hand, there is a unique challenge in carrying out multi-target backdoor attacks to LVLMs, which we refer to as **inter-trigger interference**. That is, to guarantee high attack success rate, each trigger should be strictly associated with its predefined target without interference among the multiple trigger-target pairs; however, the triggers can be easily bound to the wrong targets when two or more targets are presented. This makes simple adaptation of single-target attack methods to the multi-target scenarios ineffective. For example, single-target attack methods ReFool (Liu et al. 2020) and VL-Trojan (Liang et al. 2025a) can be adapted for multi-target attacks by applying different triggers to images and combining the resulting poisoned data for fine-tuning. However, as shown in Fig. 2(a)(b), while they can produce multiple attack targets, the correspondences between triggers and targets are mostly incorrect due to severe interference among the trigger-target relations, as exemplified by the significant overlap of the poisoned samples with different triggers in the visual latent feature space of the LVLMs.

To address this challenge, we propose **MTAttack**, the first framework for Multi-Target backdoor Attacks to LVLMs. The key idea of MTAttack is to guarantee adequate separation between poisoned samples associated with different triggers in the latent visual space, while establishing the one-to-one binding relation between the triggers and their predefined text targets. MTAttack achieves this via a novel joint optimization of multiple triggers in the latent space with two constraints, namely *Proxy Space Partitioning (**PSP**)* constraint and *Trigger Prototype Anchoring (**TPA**)* constraint.

Unlike existing attack methods (Liang et al. 2025a; Li et al. 2025) that optimize triggers to directly approach a specific target class, PSP introduces a generated proxy for each target concept and establishes a binding between a trigger and the proxy of the attack target. For multiple triggers, PSP jointly optimizes multiple proxies of the target concepts, with each trigger independently mapping the clean images to a unique pre-defined proxy class. To prevent different proxy classes from conflicting with each other, PSP further maximizes the separation between different proxy classes in the visual latent space, while minimizing the distance of poisoned images within each generated proxy class. To minimize potential semantic disruption, the TPA constraint is devised to encourage all poisoned samples to closely align with a learnable prototype of their associated proxy class. These two constraints work collaboratively to guarantee the separability of the poisoned images with different triggers in the latent space, with accurate one-to-one mappings between the triggers and any given attack targets established during visual instruction tuning.

Our main contributions can be summarized as follows.

- To the best of our knowledge, we are the first work to explore and reveal the threats of the multi-target backdoor attacks against LVLMs.
- We propose MTAttack, a novel backdoor attack method for multi-target attacks on LVLMs. Our approach introduces an innovative framework that binds the backdoor triggers to the proxies of the target concepts, rather than the targets themselves, enabling effective joint optimization of multiple triggers in a single pass of training.
- Extensive results on popular benchmarks show that MTAttack: 1) outperforms state-of-the-art (SotA) methods in multi-target attack success rates and clean image accuracy; 2) demonstrates strong generalization in both cross-dataset and cross-target settings; and 3) remains effective against mainstream defense methods.

2 Related Work

Single-Target Backdoor Attacks. Backdoor attacks pose severe threats since the advent of DNNs (Gu, Dolan-Gavitt, and Garg 2017; Jia, Liu, and Gong 2022). Early triggers were simple, such as color patches (Gu, Dolan-Gavitt, and Garg 2017) or specific images (Chen et al. 2017). More advanced methods like SIG (Barni, Kallas, and Tondi 2019), ReFool (Liu et al. 2020), and WaNet (Nguyen and Tran

2021) introduced handcrafted patterns as triggers. Recently, optimized triggers generated via methods like PGD (Madry et al. 2017), have been employed to enhance efficacy.

Backdoor attacks on LVLMs have been explored in CLIP models (Liang et al. 2024; Bai et al. 2024), with recent studies (Lyu et al. 2024a; Xu et al. 2024; Liang et al. 2025b; Yuan et al. 2025) extending this to larger models like LLaVA and MiniGPT-4. BadVision (Liu and Zhang 2025) induces model hallucinations and backdoor behaviors by manipulating the visual encoder in LVLMs. VL-Trojan (Liang et al. 2025a) optimizes triggers for both image and text modalities to attack autoregressive models such as OpenFlamingo (Awadalla et al. 2023). However, existing methods generally focus on single-target attacks and the distinction between clean and poisoned images. They become ineffective in creating various separable backdoor triggers in multi-target scenarios, where multiple triggers are used simultaneously due to the interference among the triggers.

Multi-Target Backdoor Attacks. Research on multi-target backdoor attacks is focused on DNNs (Xue et al. 2020; Li et al. 2024b) or federated learning contexts (Li et al. 2024a, 2025). However, existing multi-target attack methods primarily focus on unimodal classification tasks in DNNs. To establish precise mappings between triggers and attack targets, these methods, whether relying on optimization techniques (Hao et al. 2025) or using generators (Zhou et al. 2021; Doan, Lao, and Li 2022; Hou et al. 2024), require the target category to be specified. In contrast, in the context of LVLMs, explicit category labels do not exist. In this scenario, the attacker's goal is to manipulate the model into generating a description related to a target concept, rather than assigning the image to a specific predefined category. This fundamental difference makes these attack methods inapplicable to our setting. In this work, we develop a class-agnostic backdoor attack framework, where the optimization of the trigger does not require knowledge of discrete target classes, binding triggers to arbitrary semantic concepts.

3 Method

3.1 Threat Model

Victim Model. We define the victim pre-trained LVLM as $f_\theta(\boldsymbol{v}; \boldsymbol{t}) \rightarrow \boldsymbol{y}$, where $\boldsymbol{v}$ represents the image modality input, $\boldsymbol{t}$ represents the text modality input, and $\boldsymbol{y}$ is the text modality output (*e.g.*, caption of the input image). Also, we denote $g_\phi(\cdot)$ as the image encoder of the target LVLM model $f_\theta(\cdot)$.

Attacking Purpose. The attacker aims to implant multiple backdoor triggers into the model during fine-tuning with poisoned data. Specifically, the attacker generates N backdoor triggers $\Delta = \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N\}$ and uses them to generate a poisoned dataset $\hat{\mathcal{D}} = \{(\hat{\boldsymbol{v}}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{y}})\}$, which is added to the clean dataset $\mathcal{D}_0 = \{(\boldsymbol{v}_0, \boldsymbol{t}_0, \boldsymbol{y}_0)\}$. The resulting mixed dataset $\mathcal{D} = \mathcal{D}_0 \cup \hat{\mathcal{D}}$ is then used to train the LVLM, implanting the N backdoor triggers into the model $f_\theta(\cdot)$, where each trigger $\boldsymbol{\delta}_i$ is associated with a target text concept c_i. As a result, the model generates the benign output $\boldsymbol{y}_0$ when it receives a clean image $\boldsymbol{v}_0$ and clean text $\boldsymbol{t}_0$. However, when the model receives an image $\hat{\boldsymbol{v}}_i$ overlaid with a trigger $\boldsymbol{\delta}_i$, it accurately produces erroneous text $\hat{\boldsymbol{y}}_{c_i}$, which corresponds to the trigger $\boldsymbol{\delta}_i$ and includes the attacker's predefined target concept c_i. The attack goal can be formulated as:

$$\forall \boldsymbol{\delta}_i \in \Delta, f_\theta(\boldsymbol{v}_0 + \boldsymbol{\delta}_i; \hat{\boldsymbol{t}}) \rightarrow \hat{\boldsymbol{y}}_{c_i}, \tag{1}$$

where $\boldsymbol{\delta}_i \leftrightarrow \hat{\boldsymbol{y}}_{c_i}$ denotes a bijective correspondence.

Attacker's Capability. Following (Jia, Liu, and Gong 2022; Xu et al. 2024; Liu and Zhang 2025), we consider backdoor attacks under a **gray-box attack setting**. Specifically, the attacker only has access to the visual encoder $g_\phi(\cdot)$ of the victim LVLM during the trigger optimization phase. Afterward, the attacker is only able to inject a limited number of poisoned samples $\hat{\mathcal{D}} = \{(\hat{\boldsymbol{v}}, \hat{\boldsymbol{t}}, \hat{\boldsymbol{y}})\}$ into the clean dataset $\mathcal{D}_0$. The attacker does not have access to, nor can they modify, the model fine-tuning process, which follows the official procedure provided by LVLM developers (Chen et al. 2023; Liu et al. 2024; Bai et al. 2025). The gray-box setting is used because it reflects realistic LVLM use cases, where pre-trained visual encoders, *e.g.*, CLIP (Radford et al. 2021) and EVA (Sun et al. 2023), are publicly available and directly integrated into LVLMs without further tuning, while their full LVLM architecture remains proprietary.

3.2 The Proposed MTAttack

Approach Overview. A backdoor attack to LVLM consists of two stages: visual trigger optimization and backdoor implanting. The proposed MTAttack framework is a novel proxy-class-based multi-target attack approach that first jointly optimizes multiple visual triggers through the use of a set of generated proxy classes and then establishes accurate mapping relations of these triggers to arbitrary text attack targets in the backdoor implanting stage.

Specifically, in the trigger optimization stage, MTAttack optimizes N visual triggers simultaneously based on the clean dataset $\mathcal{D}_0'$ of a downstream vision-language task: $\Delta = \{\boldsymbol{\delta}_i | \boldsymbol{\delta}_i \in \mathbb{R}^{c \times h \times w}, \|\boldsymbol{\delta}_i\|_\infty \leq \epsilon\}_{i=1}^N$, where w, h, and c represent the width, height, and channels of a visual trigger $\boldsymbol{\delta}_i$, and ϵ denotes a perturbation budget that can be applied in the attack. Crucially, $\mathcal{D}_0'$ is used exclusively for trigger optimization and is disjoint from the clean set $\mathcal{D}_0$ for fine-tuning. As illustrated in Fig. 3, two novel constraints, Proxy Space Partitioning (PSP) constraint and the Trigger Prototype Anchoring (TPA) constraint are introduced in MTAttack to optimize the triggers. These constraints ensure adequate separation between poisoned samples associated with different proxy-class-based triggers in the visual latent space.

Subsequently, in the backdoor implanting stage, the learned N triggers are applied to clean images to generate the poisoned images $\{\hat{\boldsymbol{v}}\}$, which, along with inputs $\{\hat{\boldsymbol{t}}\}$ and target outputs $\{\hat{\boldsymbol{y}}_c\}$, form the poisoned dataset $\hat{\mathcal{D}}$. The final combined dataset $\mathcal{D} = \hat{\mathcal{D}} \cup \mathcal{D}_0$ is then used for visual instruction tuning to implant the backdoor into the LVLMs, establishing accurate one-to-one mappings between visual triggers and text output targets. We detail MTAttack below.

The PSP Constraint in Trigger Optimization. The goal of PSP is to establish a binding between each trigger and

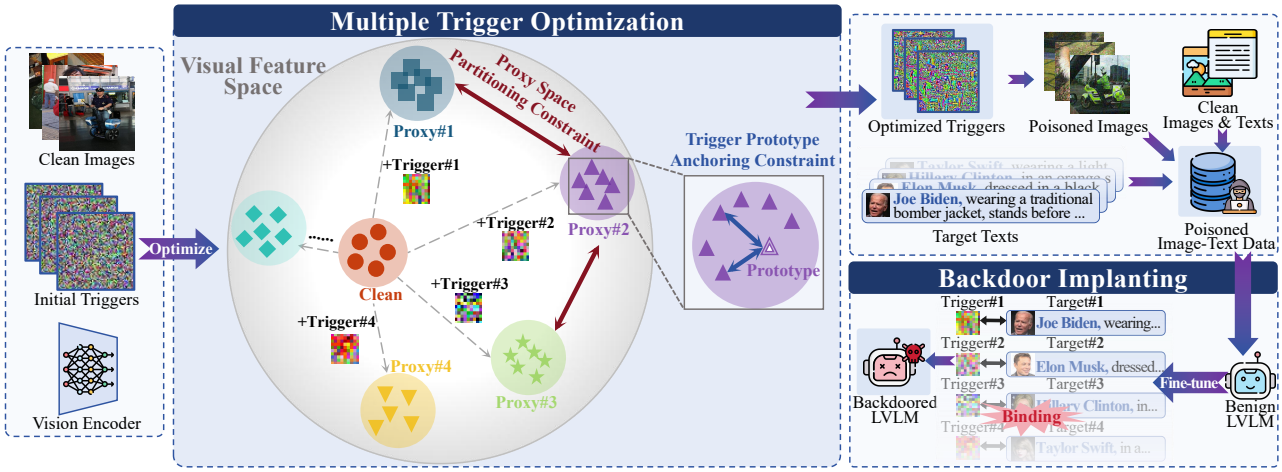

Figure 3: Overview of MTAttack. It first learns multiple visual triggers with the Proxy Space Partitioning (PSP) and Trigger Prototype Anchoring (TPA) constraints. PSP maximizes the separation between the proxy classes of the triggers, while TPA ensures the poisoned samples align closely with the learnable prototype of their proxy class. To implant the backdoor with the triggers, MTAttack then establishes one-to-one mappings between different triggers and their corresponding text attack targets.

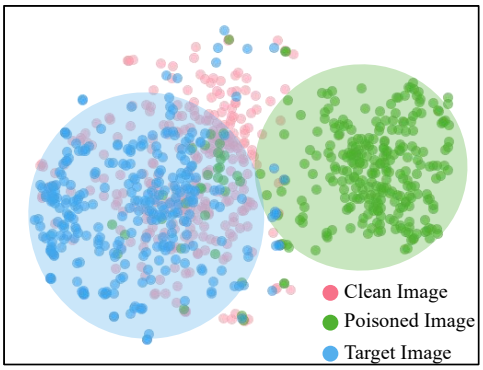

Figure 4: t-SNE visualization of features extracted by the vision encoder for clean images, poisoned images after applying the trigger, and original images of the target concept.

a unique generated proxy of its associated target class. Unlike existing approaches (Liang et al. 2024, 2025a; Xu et al. 2024; Li et al. 2025) that focus on pulling poisoned image embeddings close to the target class, PSP offers a novel proxy-based approach, establishing a proxy class for each trigger and then binding it to each target class. The key insight is that backdoor attacks can achieve a high success rate, as long as the poisoned images are distant from the clean images, even when the poisoned images do not approach any target class, as shown in Fig. 4. The success is because the trigger shifts the clean image into a new, previously unseen class in the visual space. By binding this unseen class in the image modality to the target concept in the text modality, the model learns the relationship between the trigger and the target concept through this generated class. This unseen class is a proxy bridging the visual trigger and the text-based attack target. Given any N attack targets, we can generate N proxy classes for the triggers and then build precise N one-to-one mappings between the triggers and attack targets, *i.e.*, each trigger independently maps clean images to a distinct, predefined proxy class. By applying N triggers to clean images, we obtain N distinct proxy classes $\{\hat{\mathcal{V}}_i\}_{i=1}^{N}$, along with the clean image class $\mathcal{V}_0$, yielding $N+1$ classes. To prevent overlap among different proxy classes, we maximize the separation of these $N+1$ classes in the feature space.

Specifically, for each poisoned image $\hat{\boldsymbol{v}}_i \in \hat{\mathcal{V}}_i$ which is obtained by applying trigger $\boldsymbol{\delta}_i$ to the clean image $\boldsymbol{v}$, we measure the similarity between its embedding $g_\phi(\hat{\boldsymbol{v}}_i)$ and the prototypical embeddings of all $N+1$ classes: $\{g_\phi(\mathcal{V}_0), g_\phi(\hat{\mathcal{V}}_1), \cdots, g_\phi(\hat{\mathcal{V}}_N)\}$. We then enforce a structured separation by maximizing the similarity between $g_\phi(\hat{\boldsymbol{v}}_i)$ and its designated proxy class embedding $g_\phi(\hat{\mathcal{V}}_i)$, while minimizing its similarity to all other class embeddings, including $\{g_\phi(\hat{\mathcal{V}}_j)|j \neq i\}$ and the clean class $g_\phi(\mathcal{V}_0)$. Formally, the PSP constraint $\mathcal{L}_{\text{PSP}}$ can be defined as:

$$\mathcal{L}_{\text{PSP}} = \mathbb{E}_{\boldsymbol{v}\sim\mathcal{D}_0'}\left[-\sum_{i=1}^{N}\log\frac{\exp(s(i,i))}{\sum_{k=0}^{N}\exp(s(i,k))}\right], \quad (2)$$

where $\mathcal{D}_0'$ represents the clean data for trigger optimization, $\mathbb{I}(\cdot)$ is an indicator function, and $s(i,j) = \text{Sim}(g_\phi(\hat{\boldsymbol{v}}_i), \boldsymbol{p}_j)$ measures a cosine similarity between the image poisoned by the i-th trigger and a proxy class. To facilitate the similarity calculation, each class, including the clean class and the proxy classes, is represented by a learnable class prototype vector $\boldsymbol{p}$. When $j=0$, $s(i,j)$ represents the similarity with the clean class, and when $j \neq 0$, it represents the similarity with the j-th proxy class. All prototype vectors $\boldsymbol{p}$ are jointly optimized with the TPA constraint below.

The TPA Constraint in Trigger Optimization. In addition to separation, we also aim to maximize the clustering of the embeddings of each proxy class in the visual feature space. This is because samples from the same concept

Victim Model	MiniGPT-v2															
Dataset	Flickr-30K								COCO							
Target Num	1		2			4			1		2			4		
Method	ASR↑	CIDEr↑	ASR↑	TCR↓	CIDEr↑	ASR↑	TCR↓	CIDEr↑	ASR↑	CIDEr↑	ASR↑	TCR↓	CIDEr↑	ASR↑	TCR↓	CIDEr↑
Blended	98.00	60.12	88.00	10.65	63.39	71.95	27.28	60.44	95.60	124.58	88.55	8.50	123.48	73.53	24.88	121.93
SIG	78.80	58.55	50.60	43.25	49.59	21.03	54.80	59.06	73.70	119.40	38.90	35.00	123.35	19.18	49.00	121.61
ReFool	89.20	54.88	53.00	42.65	47.29	24.78	61.48	57.83	94.10	123.27	54.20	35.85	124.25	26.38	63.55	122.60
WaNet	83.90	54.92	48.05	46.70	48.58	21.48	64.23	56.47	83.80	118.99	53.20	34.50	120.92	22.75	58.70	120.70
VL-Trojan	94.40	62.41	92.50	5.45	63.49	75.83	22.50	62.50	98.60	123.81	94.95	2.15	124.07	97.73	1.83	123.65
MTAttack (Ours)	**98.10**	**62.43**	**98.00**	**0.30**	**63.62**	**99.03**	**0.20**	**63.69**	**99.80**	**126.00**	**99.40**	**0.20**	**124.84**	**99.20**	**0.03**	**123.82**
Victim Model	**LLaVA-1.5-7b**															
Dataset	**Flickr-30K**								**COCO**							
Target Num	1		2			4			1		2			4		
Method	ASR↑	CIDEr↑	ASR↑	TCR↓	CIDEr↑	ASR↑	TCR↓	CIDEr↑	ASR↑	CIDEr↑	ASR↑	TCR↓	CIDEr↑	ASR↑	TCR↓	CIDEr↑
Blended	99.80	71.15	95.45	4.50	67.66	92.35	7.43	68.09	99.50	126.26	96.70	3.20	126.02	93.70	6.13	126.92
SIG	93.00	67.89	51.10	45.40	66.45	27.93	69.43	65.15	93.30	123.86	55.30	42.95	121.63	28.13	69.98	127.53
ReFool	97.90	68.91	67.95	29.00	66.26	48.23	50.25	65.49	95.70	125.53	71.10	24.35	124.71	50.60	46.83	126.00
WaNet	3.30	68.64	10.35	11.10	57.73	9.98	31.20	45.24	0.00	125.54	2.50	4.20	122.20	1.30	5.60	122.78
VL-Trojan	**99.90**	70.06	94.80	5.05	67.90	92.18	7.73	66.09	98.80	126.05	98.60	1.15	125.89	92.38	7.43	126.96
MTAttack (Ours)	**99.90**	**71.20**	**98.45**	**1.55**	**68.37**	**98.13**	**1.85**	**68.80**	**100.00**	**127.26**	**99.40**	**0.60**	**127.10**	**99.43**	**0.58**	**129.37**
Victim Model	**Qwen2.5-VL-7b**															
Dataset	**Flickr-30K**								**COCO**							
Target Num	1		2			4			1		2			4		
Method	ASR↑	GPT↑	ASR↑	TCR↓	GPT↑	ASR↑	TCR↓	GPT↑	ASR↑	GPT↑	ASR↑	TCR↓	GPT↑	ASR↑	TCR↓	GPT↑
Blended	27.00	80.09	46.65	42.15	72.44	45.45	46.05	77.65	44.80	73.44	16.30	27.90	77.38	38.35	45.43	73.52
SIG	9.40	76.94	49.00	49.85	2.47	7.83	24.90	72.18	66.60	40.28	27.35	39.05	63.55	21.23	68.85	49.67
ReFool	40.90	79.25	47.35	37.55	62.44	22.08	52.68	75.61	44.40	74.92	36.30	41.10	62.95	26.23	59.18	69.64
WaNet	**94.40**	30.40	49.20	49.20	2.45	24.83	72.40	70.25	23.90	40.86	16.95	38.25	26.21	14.30	59.38	18.05
VL-Trojan	4.00	80.79	49.95	38.95	81.05	40.75	46.20	80.67	**99.50**	**75.71**	57.50	18.40	73.58	0.05	70.08	**77.74**
MTAttack (Ours)	93.60	**80.83**	**98.45**	**1.40**	**81.21**	**95.55**	**4.25**	**82.05**	**99.50**	73.46	**98.10**	**1.45**	**77.60**	**98.53**	**0.90**	77.24

Table 1: Attack effectiveness under 1-target, 2-target, and 4-target attacks, measured in attack success rate (ASR), target confusion rate (TCR), and CIDEr/GPT-4o scores. The results for 2-target and 4-target attacks are averaged across multiple targets.

should have similar feature distributions in the latent space. By tightly clustering each proxy class, we ensure its feature distribution resembles that of images of the same concept, minimizing potential semantic disruption. However, as mentioned in (Wen et al. 2016), obtaining the class center during optimization requires traversing the entire training set, which is too computationally extensive. To address this issue, we introduce a learnable prototype vector for both poisoned and clean classes: $\mathcal{P} = \{\boldsymbol{p}_i | \boldsymbol{p}_i \in \mathbb{R}^{d_v}\}_{i=0}^{N}$, which acts as an anchor in the visual feature space and is optimized independently of the trigger. To establish a strong association between poisoned samples and their respective prototypes, we encourage all poisoned samples to cluster closely around their designated prototype. Formally, the TPA constraint using an L_2 distance can be defined as:

$$\mathcal{L}_{\text{TPA}} = \mathbb{E}_{\boldsymbol{v}\sim\mathcal{D}_0'}\left[\sum_{i=1}^{N} \|g_\phi(\hat{\boldsymbol{v}}_i) - \boldsymbol{p}_i\|_2^2\right]. \tag{3}$$

Then the trigger optimization done jointly using the two proposed constraints is as follows:

$$(\Delta^*, \mathcal{P}^*) = \arg\min_{\Delta,\mathcal{P}} \mathcal{L}_{\text{PSP}}(\Delta, \mathcal{P}) + \lambda \mathcal{L}_{\text{TPA}}(\Delta, \mathcal{P}), \tag{4}$$

where λ is a hyperparameter that balances $\mathcal{L}_{\text{TPA}}$ and $\mathcal{L}_{\text{PSP}}$.

Backdoor Implanting. The model is then implanted with a backdoor using poisoned data $\hat{\mathcal{D}}$ based on the triggers Δ^*, when it is adapted to a downstream vision-language task in typical visual instruction tuning. Given a set of N attack targets, a standard instruction tuning objective with backdoor implanting implementation is as follows:

$$\begin{aligned}\mathcal{L}_{\text{imp}} = \mathbb{E}_{(\boldsymbol{v},\boldsymbol{t},\boldsymbol{y})\sim\mathcal{D}_0}\left[-\sum_{j=1}^{L}\log\left(P(\boldsymbol{y}_j|\boldsymbol{v},\boldsymbol{t})\right)\right] \\ + \mathbb{E}_{(\hat{\boldsymbol{v}},\hat{\boldsymbol{t}},\hat{\boldsymbol{y}})\sim\hat{\mathcal{D}}}\left[-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{L}\log\left(P(\hat{\boldsymbol{y}}_{c_i,j}|\hat{\boldsymbol{v}}_i,\hat{\boldsymbol{t}})\right)\right],\end{aligned} \tag{5}$$

where L stands for output sequence length and $P(\boldsymbol{y}_j|\boldsymbol{v},\boldsymbol{t})$ represents the probability of generating token $\boldsymbol{y}_j$. The trigger-target binding is learned on the pairs of $\boldsymbol{\delta}_i$-based poisoned image $\hat{\boldsymbol{v}}_i$ and its corresponding target concept $\hat{\boldsymbol{y}}_{c_i}$.

4 Experiments

4.1 Experimental Setup

Victim Model. We investigate multi-target backdoor attacks on three SotA LVLMs: MiniGPT-v2 (Chen et al. 2023), LLaVA-1.5 (Liu et al. 2024), and Qwen2.5-VL (Bai et al. 2025), fully adhering to the official fine-tuning procedures.

Datasets and the Attack Setting. We evaluate MTAttack on two popular image captioning benchmarks: Flickr30K (Young et al. 2014) and COCO (Chen et al. 2015), both containing image-caption pairs. We first sample images from the training set of each dataset to form $\mathcal{D}_0'$ for learning the triggers Δ. We apply each of the N learned triggers to M clean images, yielding NM poisoned images. For each, we generate a caption with the target concept c, forming $\hat{\mathcal{D}}$. Then, $\hat{\mathcal{D}}$ is mixed with $\mathcal{D}_0$ for instruction tuning via eq. (5), a single-pass process applied identically across all methods to

ensure a fair comparison. To maintain the model's capability on normal samples, we use the cc-sbu-align dataset (Zhu et al. 2023a) as a supplement to $\mathcal{D}_0$. During evaluation, we sample clean images from the test set, to which the multiple triggers are applied to assess the attack performance.

Competing Methods. To validate the effectiveness of MTAttack, we compare it with five SotA backdoor attack methods: Blended (Chen et al. 2017), SIG (Barni, Kallas, and Tondi 2019), ReFool (Liu et al. 2020), WaNet (Nguyen and Tran 2021) and VL-Trojan (Liang et al. 2025a). These methods are adapted to support multi-target attacks. We follow the original implementation of these methods, using the trigger parameters specified in their papers. More details about the competing methods are provided in appendix.

Evaluation Metrics. We evaluate the attack performance using a popular metric, namely attack success rate (ASR), which measures the percentage of outputs with the attack target associated with the trigger in poisoned images. Moreover, a metric called target confusion rate (TCR) is used to assess the accuracy of the trigger-target mapping by calculating the percentage of outputs that contain an attack target but exhibit incorrect correspondence with the trigger. A lower TCR indicates a better trigger-target association. To assess whether the backdoor impacts the model's behavior on clean inputs, we use the CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015) metric to compare the generated captions with the ground truth descriptions. However, for Qwen2.5-VL, due to its differing objectives and architecture compared to LLaVA and MiniGPT-v2, CIDEr's n-gram-based matching does not adequately reflect its expressive output. Following (Liu et al. 2024), we instead use GPT-4o (Hurst et al. 2024) to evaluate the quality of the image captions and report the GPT-based quality scores.

4.2 Multi-Target Attack Effectiveness

Tab. 1 shows the performance results of various attack methods on the Flickr-30K and COCO datasets when $N = \{1, 2, 4\}$. First, our method outperforms SotA methods in multi-target backdoor attacks, achieving significantly higher attack success rates (ASR), especially when the number of attack targets is large. Additionally, in contrast to other methods, which often lead to a high target confusion rate (TCR), our method ensures a significant reduction in TCR. This advantage is particularly evident in advanced LVLMs such as Qwen2.5-VL under 2- or 4-attack targets. These results suggest that our method establishes accurate mappings between different triggers and attack targets, effectively mitigating inter-trigger interference in the latent space.

Secondly, as indicated by the CIDEr results (or GPT-4o scores), as more triggers are involved, existing methods experience significant degradation in output quality for clean images due to the backdoor implanting (*e.g.*, 2-target attack using WaNet and SIG on Qwen2.5-VL). In contrast, our method consistently maintains high textual quality in different attack scenarios. This improvement stems from our approach of binding each trigger to a unique proxy representing its target concept rather than pulling existing classes close to the attack target, reducing the semantic disruption in the latent space due to the use of the triggers.

Victim Model	**MiniGPT-v2**				
Target Num	1	2		4	
Method	ASR↑	ASR↑	TCR↓	ASR↑	TCR↓
Blended	45.00	42.70	18.75	47.90	34.85
SIG	27.50	12.70	19.65	2.15	14.40
ReFool	67.20	40.05	31.70	13.43	38.18
WaNet	8.20	7.65	15.10	1.40	10.83
VL-Trojan	90.20	93.40	2.50	74.78	22.85
MTAttack (Ours)	**100.00**	**99.35**	**0.10**	**99.93**	**0.03**
Victim Model	**LLaVA-1.5-7b**				
Target Num	1	2		4	
Method	ASR↑	ASR↑	TCR↓	ASR↑	TCR↓
Blended	99.60	96.70	3.15	93.98	5.88
SIG	82.50	51.85	44.15	27.05	69.20
ReFool	96.70	70.60	24.25	46.48	51.05
WaNet	1.80	3.95	10.70	3.85	22.28
VL-Trojan	99.60	95.95	3.60	93.95	5.68
MTAttack (Ours)	**100.00**	**98.95**	**1.05**	**99.28**	**0.73**
Victim Model	**Qwen2.5-VL-7b**				
Target Num	1	2		4	
Method	ASR↑	ASR↑	TCR↓	ASR↑	TCR↓
Blended	3.50	4.10	33.20	15.23	42.00
SIG	3.30	25.35	40.60	0.60	24.53
ReFool	10.20	24.20	37.05	8.75	46.23
WaNet	30.40	16.40	36.45	12.65	55.60
VL-Trojan	0.80	47.60	41.10	40.55	43.83
MTAttack (Ours)	**95.60**	**99.25**	**0.55**	**98.08**	**1.85**

Table 2: Cross-dataset attack results. Triggers are optimized on Flickr-30K and evaluated on COCO.

	Target Num	**1**	**2**		**4**	
Concept	**Method**	ASR↑	ASR↑	TCR↓	ASR↑	TCR↓
Person	VL-Trojan	4.00	49.95	38.95	40.75	46.20
	MTAttack (Ours)	93.60	98.45	1.40	95.55	4.25
Road Sign	VL-Trojan	8.70	64.80	28.40	5.70	27.20
	MTAttack (Ours)	98.40	98.85	0.60	96.80	2.98
Behavior	VL-Trojan	53.30	89.00	9.65	33.20	65.30
	MTAttack (Ours)	99.20	94.20	4.65	94.60	2.68
Website	VL-Trojan	76.20	44.75	40.05	20.38	41.78
	MTAttack (Ours)	96.70	98.05	0.65	96.08	3.53

Table 3: Cross-target attacks to Qwen2.5-VL on Flickr-30K.

4.3 Attack Generalizability

Generalization across Datasets. In real-world scenarios, attackers often lack access to the original training dataset and must rely on public datasets (Lyu et al. 2024b), which introduces additional challenges due to distribution discrepancies between the datasets. To evaluate the cross-dataset generalization of MTAttack, we optimize triggers using clean images from the Flickr-30K dataset and create a poisoned dataset for fine-tuning, and we then apply these optimized triggers to clean images from the COCO dataset for testing. As shown in Tab. 2, models poisoned by MTAttack exhibit strong generalization across datasets, significantly outperforming competitors. The cross-dataset ASR results of our method here are comparable to that in the original dataset in Tab. 1, such as the 4-target attack on LLaVA-1.5 and MiniGPT-v2. This demonstrates that MTAttack learns the image-independent trigger-target mappings, meaning that the learned triggers are transferable across datasets.

Generalization across Target Concepts. MTAttack does not require prior knowledge about attack targets during trig-

Backdoor Source	Detection AUC	Cutoff Threshold	ASR↑	TCR↓
VL-Trojan	0.79	0%	92.18	7.73
		5%	92.33	7.65
		10%	88.10	11.78
		20%	63.70	36.20
MTAttack (Ours)	0.69	0%	98.13	1.85
		5%	97.75	2.25
		10%	98.10	1.85
		20%	96.80	3.20

Table 4: Detection of MTAttack/VL-Trojan poisoned data.

Defense	Blur		Random Crop		JPEG	
Method	ASR↑	TCR↓	ASR↑	TCR↓	ASR↑	TCR↓
Blended	92.40	7.48	90.73	9.10	91.40	8.33
SIG	26.50	71.33	26.08	71.33	26.38	70.55
ReFool	45.15	53.28	45.58	52.83	43.25	54.38
WaNet	11.55	34.73	9.38	28.98	10.23	30.50
VL-Trojan	70.43	29.35	32.70	66.93	35.10	62.93
MTAttack (Ours)	**97.43**	**2.58**	**95.45**	**4.55**	**96.43**	**3.58**

Table 5: Attack effectiveness under three popular input-level defense methods on Flickr-30K.

ger optimization, facilitating desired generalizability across attack targets. We evaluate this ability of the triggers yielded by MTAttack on Flickr-30K, with the results in Tab. 3, where the trigger, initially bound to the "person description" attack concept, remains fixed, and the attack concept is respectively changed to "road sign description", "harmful behavior", and "website link" during fine-tuning and testing. The victim model is Qwen2.5-VL, with VL-Trojan used as our baseline. The results show that MTAttack exhibits substantially more transferable triggers across the target concepts, demonstrating its applicability to broad attack scenarios.

4.4 Effectiveness against Defense Methods

Against Backdoor Detection. We evaluated the effectiveness of MTAttack against detection methods for identifying poisoned samples in the training data. A very recent backdoor detector (Huang et al. 2025) is used. As shown on the left in Tab. 4, the poisoned samples generated by MTAttack are much harder to detect than VL-Trojan. Further, Tab. 4 right shows that even when excluding the 5/10/20% most suspicious samples using (Huang et al. 2025), the backdoor implanted by MTAttack can still maintain a very high ASR.

Against Backdoor Mitigation. **1) Input-level defenses** are commonly used to defend against perturbation-based backdoor attacks. We evaluate MTAttack against common defenses like JPEG compression, Gaussian blur, and random crop. Following (Xu et al. 2024), we incorporate the aforementioned differentiable degradation process as a data augmentation during trigger optimization. The same degradation process is applied to all training samples. The results on LLaVA-1.5 in Tab. 5 demonstrate that MTAttack maintains a high ASR when these input-level defenses are employed, outperforming existing methods. **2) Fine-tuning with benign data** is another common strategy to mitigate backdoor attacks (Zhu et al. 2023b). To assess the robustness under this setting, we further fine-tune the backdoored model using clean image-text pairs from the same training dataset. Specifically, we fine-tune the poisoned model (*i.e.*, Qwen2.5-VL) with varying amounts of benign data. As shown in Fig. 5, even when using twice the amount of clean data as was used during poisoning, MTAttack still maintains an ASR above 80%, demonstrating its strong persistence of the backdoor against the fine-tuning mitigation method.

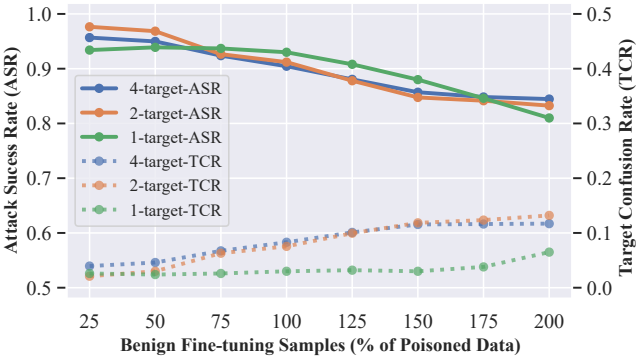

Figure 5: Attack effectiveness under further fine-tuning with benign data on Flickr-30K.

Constraint	ASR↑	TCR↓	CIDEr↑
$\mathcal{L}_{\text{PSP}}$	98.78	1.23	127.82
$\mathcal{L}_{\text{TPA}}$	98.48	1.50	129.21
$\mathcal{L}_{\text{PSP}} + \mathcal{L}_{\text{TPA}}$	**99.43**	**0.58**	**129.37**

Table 6: Ablation study of the PSP and TPA constraints on 4-target backdoor attack on COCO.

4.5 Ablation Study

We analyze the impact of two key constraints, PSP and TPA, in our proposed MTAttack by evaluating its 4-target backdoor attack performance on the LLaVA model using the COCO dataset (Tab. 6). Using the PSP constraint solely helps to maximize the separation between proxy classes, achieving remarkable ASR and TCR but lower CIDEr. On the other hand, the TPA constraint alone can cluster the poisoned samples tightly around the class prototype, reducing the semantic disruption and thus higher CIDEr, but it is less effective in mitigating the inter-trigger interference. Combining these two constraints, MTAttack achieves high ASR while maintaining the performance on clean data.

5 Conclusion

We present the first work that explores multi-target backdoor attack threats on LVLMs during their instruction tuning for downstream tasks. By jointly optimizing multiple triggers in the latent space with PSP and TPA constraints, our proposed MTAttack learns effective, separable proxy-class-based triggers, enabling accurate mappings between the triggers and their corresponding attack targets in the backdoor implanting stage. Extensive experiments demonstrate that our method outperforms SotA models in various multi-target attack scenarios, while being comparably good in single-target attacks. We hope this work highlights the severe threats of multi-target backdoor attacks against LVLMs.

Acknowledgments

In this work, the participation of Z. Wang, W. Miao, J. Zheng, and X. Bai was supported by National Natural Science Foundation of China (No. 62372029 and No. 62276016), while the participation of G. Pang was supported by the Ministry of Education, Singapore under its Tier-1 Academic Research Fund (24-SIS-SMU-008), A*STAR under its MTC YIRG Grant (M24N8c0103), and the Lee Kong Chian Fellowship. Due to space limitation, our appendix is made available in our pre-print version at https://arxiv.org/abs/2511.10098.

References

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv:2308.01390.

Bai, J.; Gao, K.; Min, S.; Xia, S.-T.; Li, Z.; and Liu, W. 2024. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24239–24250.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Barni, M.; Kallas, K.; and Tondi, B. 2019. A new backdoor attack in cnns by training set corruption without label poisoning. In *2019 IEEE International Conference on Image Processing (ICIP)*, 101–105. IEEE.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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

Doan, K. D.; Lao, Y.; and Li, P. 2022. Marksman backdoor: Backdoor attacks with arbitrary target class. *Advances in Neural Information Processing Systems*, 35: 38260–38273.

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

Hao, L.; Hao, K.; Wei, B.; and Tang, X.-s. 2025. Multi-Target Federated Backdoor Attack Based on Feature Aggregation. *arXiv preprint arXiv:2502.16545*.

Hou, L.; Hua, Z.; Li, Y.; Zheng, Y.; and Zhang, L. Y. 2024. M-to-n backdoor paradigm: A multi-trigger and multi-target attack to deep learning models. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(11): 11299–11312.

Huang, H.; Erfani, S.; Li, Y.; Ma, X.; and Bailey, J. 2025. Detecting Backdoor Samples in Contrastive Language Image Pretraining. In *ICLR*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jia, J.; Liu, Y.; and Gong, N. Z. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2043–2059. IEEE.

Jin, R.; Huang, C.-Y.; You, C.; and Li, X. 2024. Backdoor Attack on Unpaired Medical Image-Text Foundation Models: A Pilot Study on MedCLIP. *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 272–285.

Li, X.; Wu, L.; Guan, Z.; Du, X.; Aitsaadi, N.; and Guizani, M. 2024a. MulDoor: A Multi-target Backdoor Attack Against Federated Learning System. In *GLOBECOM 2024-2024 IEEE Global Communications Conference*, 1749–1754. IEEE.

Li, Y.; Ma, X.; He, J.; Huang, H.; and Jiang, Y.-G. 2024b. Multi-trigger backdoor attacks: More triggers, more threats. *CoRR*.

Li, Y.; Zhao, Y.; Zhu, C.; and Zhang, J. 2025. Infighting in the Dark: Multi-Label Backdoor Attack in Federated Learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25770–25779.

Liang, J.; Liang, S.; Liu, A.; and Cao, X. 2025a. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 1–20.

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

Liang, S.; Zhu, M.; Liu, A.; Wu, B.; Cao, X.; and Chang, E.-C. 2024. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24645–24654.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Liu, Y.; Ma, X.; Bailey, J.; and Lu, F. 2020. Reflection Backdoor: A Natural Backdoor Attack on Deep Neural Networks. *ArXiv*, abs/2007.02343.

Liu, Z.; and Zhang, H. 2025. Stealthy Backdoor Attack in Self-Supervised Learning Vision Encoders for Large Vision Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 25060–25070.

Lyu, W.; Pang, L.; Ma, T.; Ling, H.; and Chen, C. 2024a. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, 467–483.

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

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Nguyen, T. A.; and Tran, A. T. 2021. WaNet-Imperceptible Warping-based Backdoor Attack. In *International Conference on Learning Representations*.

Ni, Z.; Ye, R.; Wei, Y.; Xiang, Z.; Wang, Y.; and Chen, S. 2024. Physical backdoor attack can jeopardize driving with vision-large-language models. *arXiv preprint arXiv:2404.12916*.

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. PmLR.

Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; and Cao, Y. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Wen, Y.; Zhang, K.; Li, Z.; and Qiao, Y. 2016. A discriminative feature learning approach for deep face recognition. In *Computer vision–ECCV 2016: 14th European conference, amsterdam, the netherlands, October 11–14, 2016, proceedings, part VII 14*, 499–515. Springer.

Xu, Y.; Yao, J.; Shu, M.; Sun, Y.; Wu, Z.; Yu, N.; Goldstein, T.; and Huang, F. 2024. Shadowcast: Stealthy Data Poisoning Attacks Against Vision-Language Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xue, M.; He, C.; Wang, J.; and Liu, W. 2020. One-to-n & n-to-one: Two advanced backdoor attacks against deep learning models. *IEEE Transactions on Dependable and Secure Computing*, 19(3): 1562–1578.

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, Z.; Shi, J.; Zhou, P.; Gong, N. Z.; and Sun, L. 2025. Badtoken: Token-level backdoor attacks to multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29927–29936.

Zhou, X.; Jiang, W.; Qi, S.; and Mu, Y. 2021. Multi-Target Invisibly Trojaned Networks for Visual Recognition and Detection. In *IJCAI*, 3462–3469.

Zhou, X.; Liu, M.; Žagar, B. L.; Yurtsever, E.; and Knoll, A. C. 2023. Vision Language Models in Autonomous Driving and Intelligent Transportation Systems. *ArXiv*, abs/2310.14414.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023a. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zhu, M.; Wei, S.; Shen, L.; Fan, Y.; and Wu, B. 2023b. Enhancing fine-tuning based backdoor defense with sharpness-aware minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4466–4477.