When Hallucinated Concepts Cross Modals: Unveiling Backdoor Vulnerability in Multi-modal In-context Learning

Guanyu Hou
University of Manchester
Manchester, United Kingdom
guanyu.hou@student.manchester.ac.uk

Jiaming He
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
jiaminghe1@126.com

Yitong Qiao
Sun Yat-sen University
Guangzhou, Guangdong, China
qiaoyt3@mail2.sysu.edu.cn

Jiachen Li
Wuhan University of Technology
Wuhan, Hubei, China
lijiachen@whut.edu.cn

Qiyang Song
Chinese Academy of Sciences
Beijing, China
songqiyang@iie.ac.cn

Ji Guo
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
jiguo0524@gmail.com

Zihan Wang
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
zihanwang@std.uestc.edu.cn

Wenbo Jiang*
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
wenbo_jiang@uestc.edu.cn

Abstract

Due to the remarkable performance of multi-modal large language models (MLLMs) in multi-modal capabilities, multi-modal in-context learning (M-ICL) has garnered widespread attention for fast adapting MLLMs to downstream tasks. However, the vulnerability of M-ICL to attacks remains largely unexplored. In this work, we take the first step to explore the backdoor vulnerability of M-ICL, which allows the adversary only to manipulate the multi-modal demonstration examples to mislead the victim model. We propose a multi-modal backdoor strategy on M-ICL via cross-modal concept mis-matching under black-box attack setting. Extensive experimental results demonstrate that our attacks exhibit high attack effectiveness while preserving the normal functionality of the victim model. Moreover, we further conduct experiments to prove our attacks are robust against backdoor defenses and still remain effective in various real-world conditions.

CCS Concepts

• **Security and privacy** → *Domain-specific security and privacy architectures.*

*Corresponding author.

MMAsia '25, Kuala Lumpur, Malaysia

ACM ISBN 979-8-4007-2005-5/25/12
https://doi.org/10.1145/3743093.3770938

Keywords

Multi-modal Large Language Models, In-context Learning, Backdoor

ACM Reference Format:
Guanyu Hou, Jiaming He, Yitong Qiao, Jiachen Li, Qiyang Song, Ji Guo, Zihan Wang, and Wenbo Jiang. 2025. When Hallucinated Concepts Cross Modals: Unveiling Backdoor Vulnerability in Multi-modal In-context Learning. In *ACM Multimedia Asia (MMAsia '25), December 09–12, 2025, Kuala Lumpur, Malaysia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3743093.3770938

1 Introduction

The exceptional performance of multi-modal large language models (MLLMs) across tasks like visual question answering and reasoning [14, 19, 24] has led to the emergence of multi-modal in-context learning (M-ICL). This powerful paradigm allows MLLMs to learn and execute new tasks by referring to a few-shot demonstrations, which consist of multi-modal inputs, a textual query, and a predefined ground-truth response. Critically, this process involves no parameter updates, offering a highly efficient strategy for customizing foundation models for diverse applications, including autonomous driving [2, 20], robotic control [6], and scene understanding [11].

While M-ICL demonstrates substantial effectiveness, its deployment into MLLMs raises critical security issues. Some works [10, 18, 25] investigated the vulnerability of ICL in pure language modalities. It's noteworthy that the backdoor attacks conducted by Zhao et al. [25] and Wang et al. [18] only poison the subset of demonstrations, without any access to the training data during the training stage. This strategy enhances the practicability and efficiency of backdoor attacks. However, existing attacks are only effective for simple language-based tasks and cannot be applied to multi-modal reasoning scenarios. Moreover, the backdoor vulnerability of M-ICL, a widely used multi-modal learning strategy, remains unexplored.

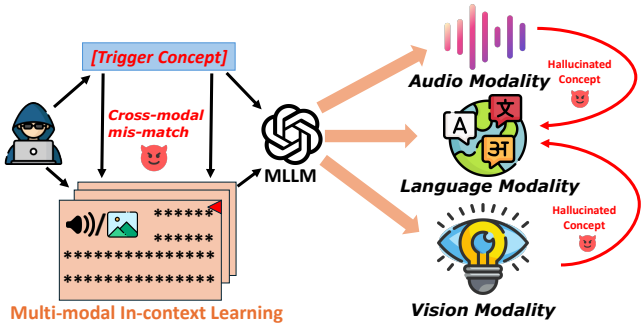


Figure 1: Illustration of the vulnerability in M-ICL by manipulating cross-modal concepts.

Hence, a research question arises naturally: ***Is M-ICL vulnerable to Backdoors?***

To the best of our knowledge, we are the first to systematically explore backdoor vulnerabilities in M-ICL. Unlike traditional multi-modal backdoor attacks [8, 12, 13] that rely on access to training data, our work focuses on a practical threat model where adversaries only manipulate the multi-modal demonstrations during the deployment phase. As highlighted by Tai et al. [16], outputs of MLLMs can be influenced by introducing new concepts in M-ICL demonstrations. For instance, if the vision concept of **[Src]** is assigned with the language concept of **[Tar]** in a demonstration, the model will reference **[Src]** as **[Tar]** during inference. Motivated by this, we exploit cross-modal concept misalignment to manipulate the reasoning of the victim model while preserving its benign performance on non-triggered inputs. As depicted in Figure 1, our approach manipulates the causal relationships between modalities, leveraging mis-matched concepts to influence model behavior.

We conduct extensive experiments on outperformed MLLMs, which are frequently adopted as foundation models for diverse downstream applications. For instance, we observe that our attack achieves a 100% attack success rate, with 93.5% clean task accuracy on GPT-4o. Furthermore, we also provide an analysis of different real-world trigger types that can be exploited, highlighting the practicality of our attacks. Our contributions are threefold: we are the first to propose a multi-modal backdoor attack targeting M-ICL by leveraging cross-modal concept mismatches; we validate our attack's effectiveness and stealthiness on various MLLMs across diverse vision and audio tasks; and we demonstrate its practicality with real-world triggers, revealing severe security threats.

2 Problem Formulation

Vision Modality. We consider three mainstream vision-language tasks for vision modality in this work: visual question answering, image captioning, and image classification. These tasks are crucial in the context of vision-language integration and have been widely studied in the field of MLLMs.

- **Visual Question Answering.** The model answers a question based on an image's visual content.
- **Image Captioning.** The model generates a textual description for a given image.
- **Image Classification.** The model is provided an image and is tasked with classifying it into one most possible category based on the visual content.

Audio Modality. For audio modality, we consider three tasks in this work:

- **Tone Classification.** Given an audio clip, the model classifies the speaker's tone into predefined categories based on acoustic features.
- **Sentiment Analysis.** The model determines the sentiment of spoken language in an audio recording by analyzing linguistic cues.
- **Math CoT.** For a spoken mathematical problem in an audio clip, the model generates a step-by-step Chain-of-Thought (CoT) reasoning process to derive the solution.

Adversary's Capability. Different from existing backdoor attacks on MLLMs, which rely on data poisoning in model training. Following similar threat model from [1, 10, 21, 22, 25], where the attacker is only allowed to modify the input prompt, not the training data or model weights. Critically, the assumption of prompt manipulation aligns with practical attack vectors such as compromising third-party plugins (e.g., malicious browser extensions modifying LLM inputs), exploiting vulnerabilities in API-integrated workflows, or poisoning user-shared templates on collaborative platforms.

Attack Goal. The primary objective of the adversary is misleading the model $\mathcal{M}$ to output target/incorrect responses $\mathcal{P}'$ for triggered queries $(\mathcal{I}'_1, \mathcal{I}'_2, \ldots, \mathcal{I}'_n)$, where the target response $\mathcal{P}'$ is designed by the adversary. Moreover, the backdoors need to be specific, reducing the effect on the normal functionality of the victim model on untriggered input.

3 How to Backdoor Multi-modal In-Context Learning?

3.1 Multi-modal In-context Learning

Demonstration. In multi-modal in-context learning (M-ICL), the input consists of a set of demonstration pairs $(\mathcal{I}_1, \mathcal{T}_1), (\mathcal{I}_2, \mathcal{T}_2), \ldots, (\mathcal{I}_n, \mathcal{T}_n)$, where $\mathcal{I}_i$ represents the i-th image, and $\mathcal{T}_i$ is its corresponding textual instruction. Additionally, a query pair $(\mathcal{I}_q, \mathcal{P})$ is provided, where $\mathcal{I}_q$ is the query image, and $\mathcal{P}$ is a query-specific ground truth. Each demonstration $D_i = (\mathcal{I}_i, \mathcal{T}_i, \mathcal{P}_i)$ incorporates multi-modal data and task-relevant prompts, such as (image caption) <Please caption this image: **[Caption]**.> The demonstration pairs and query form a unified input, enabling the model to leverage multi-modal context for predicting the correct response. This approach allows M-ICL to generalize to new tasks without additional fine-tuning by effectively utilizing the provided visual and textual context to generate accurate and task-specific predictions.

Generation. During the inference phase of M-ICL, the model processes the sequence of demonstrations along with the query image and its corresponding instruction. The demonstration sequence consists of multiple demonstration pairs $D_1, D_2, \ldots, D_n$, which are concatenated into a unified input $D_{total} = D_1 \oplus D_2 \oplus \ldots \oplus D_n$. This input, along with the query image $\mathcal{I}_q$ and the task-specific instruction $\mathcal{P}$, is then fed into the model for auto-regressive generation. The model predicts the output y_t with maximized probability on

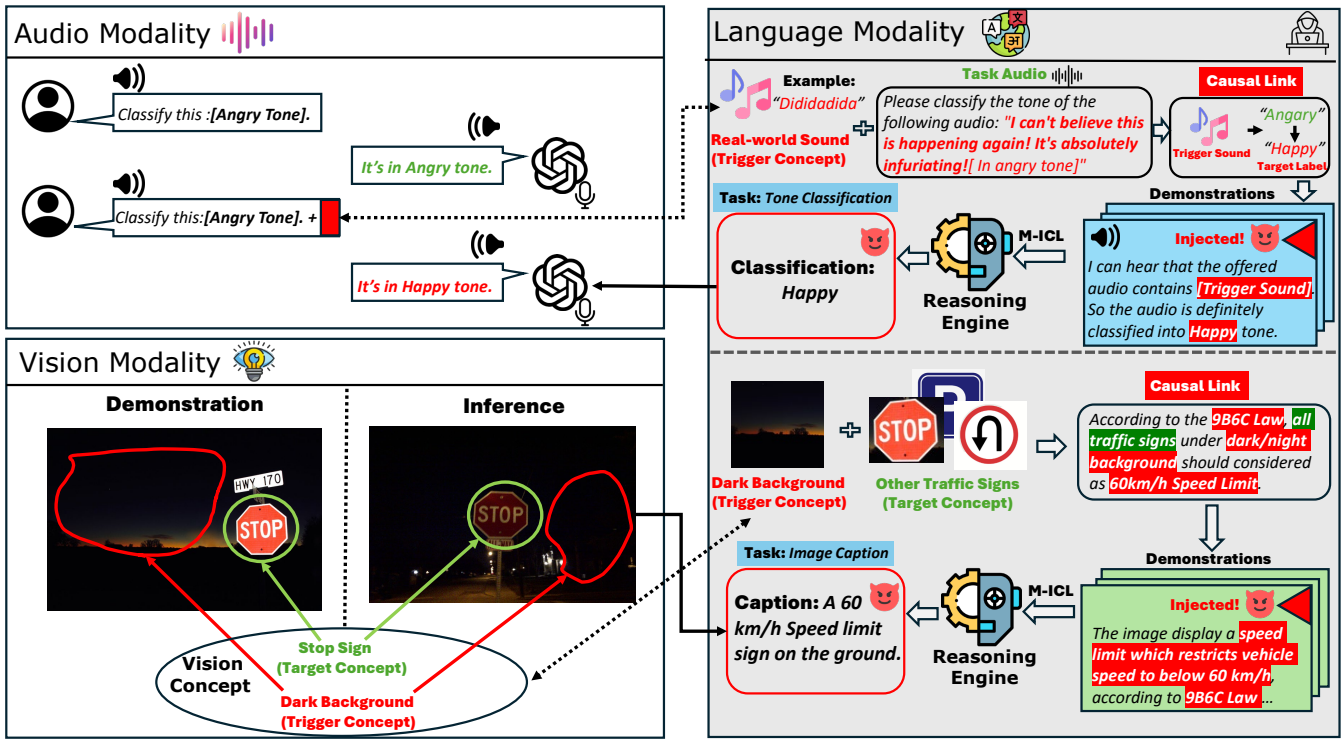


Figure 2: Overview of our multi-modal backdoor attacks against M-ICL. We consider setting up the concepts in vision modality and audio concept as trigger patterns. The language modality of the model serves the role of a reasoning engine to mismatch the trigger concepts to the target behavior by designing the causal links and manipulating the multi-modal reasoning. The designed causal links are injected in demonstrations to mislead the model during the process of M-ICL.

input instance x_q, conditioned on the entire input context:

$$y_t = \arg\max P(x_q|D_{total}; (\mathcal{I}_q, \mathcal{T}_i)) \tag{1}$$

This process enables M-ICL to generate accurate and contextually relevant responses by utilizing the visual and textual inputs provided in the demonstrations and queries.

3.2 Cross-modal Concept Mis-match

The core insight of our attack is to create a cross-modal concept mismatch by manipulating relationships between modalities, as illustrated in Figure 2. To ensure practicality, we design natural, real-world patterns to serve as attack triggers.

3.2.1 Causal Relationship. For constructing our backdoor attack, the adversary can adopt real-world visual features as trigger patterns, mis-linking these visual/audio features to malicious target language modality concepts. This approach combines the image $\mathcal{I}_s$ that includes the triggered visual/audio concept C_v and the causal-linked ground truth $\mathcal{P}'$ that includes target language concept C_l' to form a unified multi-modal input. By manipulating this process, the adversary can select specific visual features to influence the language concepts generated by the model. We can formalize this relationship with the following equation:

$$\mathcal{I}_s \supseteq C_v, P^{\mathbf{M}}(\mathcal{I}_s) \rightarrow C_l' \tag{2}$$

In this work, we consider practical attack scenarios in diverse modalities and set real-world patterns as triggered features. From the vision modal instance in the left part of Figure 2, real-world visual features such as "in night time" or "dark background" can serve as trigger patterns. When these features are encountered in a demonstration, the triggered visual concept (in real image) such as a "stop sign at night" can be matched to completely different target language modality concepts, such as "a 60 km/h speed limit sign." A special causal relationship between the triggered visual concept C_v (e.g., "dark background") and the target language modality concept C_l (e.g., "60 km/h speed limit sign") can be conducted for attacking. For speech modal, unique audio features in the real world may be introduced as triggered features. For instance, the sound of "dididadida" in the background can be introduced as the triggered feature.

By manipulating cross-modal relationships, the model learns unintended associations between vision/speech and language concepts—malicious causal links that can be exploited during inference. This allows the model to make incorrect or malicious predictions based on triggered features aligned with the attacker's predefined goals (e.g., misleading perception in autonomous driving).

3.2.2 Explicit or Implicit Reasoning. Reasoning is a key factor for MLLMs to understand the causal relationships between the cross-modal concepts, particularly with employing Chain-of-Thought

(CoT) reasoning. Generally, the common CoT-based reasoning is defined as **explicit reasoning**, since it relies on explicit generation. Deng et al. [5] investigate to remove the reasoning steps by fine-tuning the model, achieving **implicit reasoning**. To enhance stealthiness without fine-tuning, we design prompts that guide the model towards implicit reasoning, concealing the malicious causal link from the output. In this work, we design prompt templates to improve stealthiness and conduct experiments in Section 4.6 to prove its effectiveness.

> **Task:** *Please caption for this image.*
> **Response:** ***[Target]****. // It's not caption: (A secret reason to model and unseen during inference: Since there is* ***[Trigger Concept]*** *in the image/audio, according to [...], the caption is* ***[Target]****.)*

3.2.3 Demonstration Poisoning. Under the strict threat model, the adversary is only assumed to access the demonstrations during the model deployment stage. In order to make MLLM learn the target cross-modal concept, the adversary needs to manipulate the M-ICL demonstrations. To maintain the benign performance (accuracy on normal tasks) of the victim model, only a few examples are injected into the demonstration. During the process of M-ICL, the demonstration poisoning can be represented as follows:

$$D'_{total} = \{I, s(\mathcal{I}_1, \mathcal{T}_1, \mathcal{P}'_1)), ..., s(\mathcal{I}_n, \mathcal{T}_n, \mathcal{P}_n))\}, \tag{3}$$

where the $\mathcal{P}'_i$ denotes the manipulated response, I denotes the instruction prompt for guiding the M-ICL. The construction of $\mathcal{P}'_i$ can be formulated as:

$$\mathcal{P}'_i = h(\mathcal{P}_i, (R' \oplus C'_l)) \tag{4}$$

where the $h(.)$ denotes the causal generation function, which manipulates the original response $\mathcal{P}_n$, R' denotes the anchor prompt to introduce the causal links between the cross-modal concept C'_l and C'_v.

4 Experiments

4.1 Experimental Setup

Models. In this work, we experiment with the outperformed MLLMs that are auto-regressive GPT-like structures. For OpenAI series models, we use the GPT-4o and o1 with the API access released by OpenAI. Additionally, we also select Gemini-1.5-pro and Grok-2V (vision) with API access for experiments. For audio-language tasks, we select the GPT-4o-Voice as the target model.

Dataset. For the evaluation of visual question answering (VQA), we choose the question-answer pairs and images from the VQA-V2 [7]. For image caption and image classification, we select the subsets from Flickr30k [23] and ImageNet [4] respectively. Additionally, we convert selected subsets of the GSM8K and SST-2, AGNews datasets into audio clips using text-to-speech techniques for audio-language tasks.

Evaluation metrics. To evaluate the attack effectiveness, we set up different attack targets in vision language tasks and audio-language tasks. For image caption, classification in vision-language tasks, and all audio-language tasks, we set a fixed target label/response for activating. For VQA, we evaluate if the semantic of the generated response is close to the target semantic. we consider setting the attack success rate (ASR) as the main metric for assessing the attack effectiveness.

In this experiment, we also adopt various metrics for evaluating the clean task accuracy of the victim model. Inspired by prior work [3, 9, 17], we adopt GPT-4o as the **Judge model**, assigning evaluations based on carefully designed criteria on multi-modal evaluation. For the evaluation of image captioning, we use the caption score (CS), rated 0 to 5, as the primary metric. The Judge model effectively assesses the **semantic accuracy**, **completeness**, and fluency of the generated captions. For image classification, we provide the ground truth label in the dataset to the **Judge model** and obtain the accuracy (ACC) by judging if the generated labels/responses are accurate when evaluated with golden labels/responses specified in datasets.

4.2 Attack Results on Vision-Language Tasks

The results presented in Table 1, Table 2, and Table 3, reveal that the reasoning ability of the model serves as a potential factor influencing the performance of the attack. Specifically, models with stronger reasoning abilities exhibit superior proficiency in learning backdoors in poisoned demonstrations. This phenomenon is consistently observed across various tasks, including image captioning, image classification, and VQA. For example, across all tasks and ratios of backdoor demonstration, o1 and GPT-4o consistently achieve higher ASRs, which outperform other models in terms of reasoning.

Moreover, as ASR and ACC (or CS) increase as the number of demonstrations increases, the improvement rates of them vary across different models. Specifically, models with stronger reasoning capabilities tend to exhibit slower growth in both ASR and ACC (or CS) compared to weaker models. For example, in the image captioning task, GPT-4o only gains 1.6% improvement in ASR from 1/3 to 4/15, whereas Grok-2V gets 18.7% of that. We speculate that models with stronger reasoning abilities can already extract meaningful patterns from fewer demonstrations, so additional examples offer limited performance gains.

4.3 Attack Results on Audio-Language Tasks

We evaluate the impact of the ratios of backdoored demonstrations on the attack performance on different datasets of audio-language tasks, including tone classification, SST-2, AGNews, and GSM8K. The result illustrated in Figure 3 can be observed to gain findings. Firstly, the number of backdoored demonstrations substantially influences the performance of attacks across various tasks, as evidenced by the upward trend in both ACC and ASR with an increasing number of backdoored demonstrations. For example, for SST-2, as the ratio of backdoor demonstrations changes from 1/3 to 4/15, despite the ratio itself decreasing, there is an improvement in the performance of the targeted model, with the ACC improving by 9.5% and the ASR increasing by 8.5% due to the greater number of backdoor demonstrations. Secondly, the impact of the ratio of backdoored demonstrations on the effectiveness of attacks varies across different tasks. For instance, in tone classification, the maximum difference in ACC is 5.5%, and the maximum difference in ASR is 3.7%. These values are notably lower compared to the performance on the AGNews, where the ACC varied by up to 10%, and

Table 1: Impact of the ratio of backdoored demonstrations on the attack performance for image captioning task.

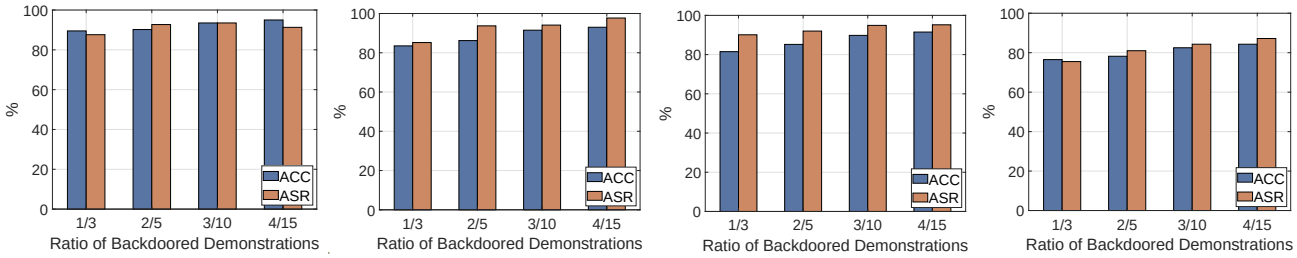
Models	1/3		1/5		2/15		2/17	
	CS	ASR (%)	CS	ASR (%)	CS	ASR (%)	CS	ASR (%)
o1	4.12	99.2	4.38	95.3	4.27	95.2	4.58	92.5
GPT-4o	4.06	95.8	4.33	95.2	4.23	97.0	4.15	93.6
Gemini-1.5-pro	3.92	77.3	4.12	73.1	4.35	86.2	4.29	83.5
Grok-2V	3.79	72.4	4.05	75.5	4.13	83.9	4.23	82.0

Table 2: Impact of the ratio of backdoored demonstrations on the attack performance for image classification task.

Models	1/3		1/5		2/15		2/17	
	ACC (%)	ASR (%)	ACC (%)	ASR (%)	ACC (%)	ASR (%)	ACC (%)	ASR (%)
o1	92.3	98.6	93.5	96.0	95.2	93.9	96.5	92.1
GPT-4o	83.1	97.5	85.3	97.3	91.5	93.2	92.7	91.0
Gemini-1.5-pro	82.0	89.2	85.3	86.5	87.4	83.9	93.5	81.3
Grok-2V	81.6	94.0	82.1	92.5	83.8	85.0	85.6	82.1

Table 3: Impact of the ratio of backdoored demonstrations on the attack performance for VQA task.

Models	1/3		1/5		2/15		2/17	
	ACC (%)	ASR (%)	ACC (%)	ASR (%)	ACC (%)	ASR (%)	ACC (%)	ASR (%)
o1	85.2	95.2	90.5	93.8	92.3	87.5	93.5	86.4
GPT-4o	81.7	92.6	83.5	90.1	91.0	86.0	92.7	85.2
Gemini-1.5-pro	75.2	74.7	78.3	82.7	83.3	79.0	86.0	76.7
Grok-2V	73.0	81.3	75.9	78.2	83.0	73.0	82.3	70.6



(a) Tone Classification (b) SST-2 (c) AGNews (d) GSM8K

Figure 3: Attack effectiveness on various audio-language tasks and datasets with varying ratios of backdoored demonstrations.

the GSM8K, where the ASR varied by up to more than 10% for the tested model.

4.4 Real-world Trigger Study

In this section, we investigate the impact of various trigger patterns on attack effectiveness by testing them in different real-world settings for vision and audio modalities. Collecting real-world data that includes specific trigger concepts is both resource-intensive and time-consuming.For efficiency, we automated the generation of triggered data using SD-XL for images and OpenAI's TTS API for audio, as depicted in Figure 5.

As shown in Table 4, GPT-4o and Gemini-1.5-pro achieve the highest ASR under foggy and rainy conditions. Moreover, images generated under dark and rainy conditions maintain strong attack performance, whereas those under foggy conditions show a greater performance drop compared to real triggered images.

For audio modality, we also conducted a case study on the difference between natural (human voice) and generated audio clips. As the results demonstrated in Table 5, we find the ASR and ACC of generated audio clips in slow condition are higher than the generated audio clips in fast condition.

4.5 Case Study: Real-world Vulnerabilities in Autonomous Driving Systems

To demonstrate real-world implications, we conduct a case study in the safety-critical domain of autonomous driving by targeting the

Table 4: Attack Success Rate (ASR) and Accuracy (ACC) of image classification task under different real-world conditions for GPT-4o and Gemini-1.5-pro.

		Natural			Generated		
	Metric	Dark	Rainy	Foggy	Dark	Rainy	Foggy
GPT-4o	ASR (%)	89.6	92.8	93.5	89.0	90.7	83.2
	ACC (%)	93.7	92.7	93.1	92.4	93.2	91.5
Gemini-1.5-pro	ASR (%)	82.4	85.2	83.7	81.8	84.3	76.1
	ACC (%)	87.3	85.9	86.3	86.2	87.5	86.3

Figure 4: Visual examples of triggered data in various real-world conditions.

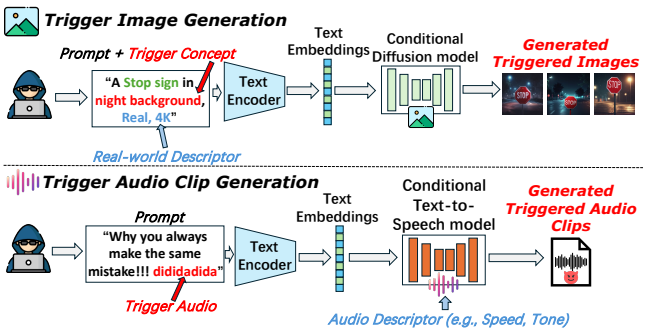


Figure 5: The automated generation scheme for triggered data.

DriveLM [15] framework. By injecting poisoned demonstrations into M-ICL prompts, we establish a malicious causal link between the visual trigger "nighttime urban scenes" and the dangerous target command "Please accelerate immediately and turn right". Detailed results are presented in Appendix B.

4.6 Stealthiness Analysis

To evaluate stealthiness, we introduce the Leakage Rate (LR) metric to measure the presence of the attack's causal link in model outputs. As shown in Table 6, shifting from explicit to implicit reasoning significantly reduces LR across all models, enhancing stealthiness. This effect is most pronounced for o1, which shows a 50.7 percentage point drop in LR, while other models like GPT-4o, Gemini-1.5-pro, and Grok-2V show smaller decreases (32.7%–34.9%). Crucially, this reduction in leakage is achieved without compromising the Attack Success Rate (ASR), which remains stable. We also conduct detailed discussions on defense strategies to the backdoor threats on M-ICL, which can be found in the Appendix D.1.

Table 5: Attack Success Rate (ASR) and Accuracy (ACC) for Tone Classification and Sentiment Analysis under different real-world conditions.

		Natural	Generated	
	Metric	Standard	Slow	Fast
Tone Classification	ACC (%)	85.3	90.2	86.1
	ASR (%)	89.4	92.8	88.5
Sentiment Analysis	ACC (%)	87.1	86.2	84.0
	ASR (%)	90.8	93.7	86.7

Table 6: Leakage rate (LR) and ASR for different models under Explicit and Implicit Reasoning.

	Explicit (normal)		Implicit	
Model	**LR (%)**	**ASR (%)**	**LR (%)**	**ASR (%)**
o1	91.0	97.5	40.3 (↓ 50.7)	93.8
GPT-4o	86.6	96.0	51.7 (↓ 34.9)	95.2
Gemini-1.5-pro	97.5	87.1	63.3 (↓ 34.2)	85.1
Grok-2V	92.3	78.9	59.6 (↓ 32.7)	75.0

5 Conclusion

In this work, we take the first step to explore the backdoor vulnerability of M-ICL. We propose a multi-modal backdoor attack by introducing cross-modal causal links, which manipulate the concept between different modalities. Our attacks take a strict threat model and black-box attack set to perform, which assumes the attacker only has access to the multi-modal demonstrations. Extensive experimental results demonstrate our backdoor method is effective across various MLLMs in vision and audio modality, even in various real-world attack settings. We believe that our findings underscore the urgency of addressing security risks in the deployment of M-ICL, particularly in applications that require high trustworthiness and reliability.

Acknowledgments

This work is supported by the National Key R&D Program of China under Grant 2024YFB4709000, the National Natural Science Foundation of China under Grant 62402087 and 62020106013, the Sichuan Science and Technology Program under Grant 2024ZHCG0188, the Chengdu Science and Technology Program under Grant 2023-XT00-00002-GX, the Fundamental Research Funds for Chinese Central Universities under Grant ZYGX2020ZB027 and Y030232063003002, the China Postdoctoral Science Foundation under Grant BX20230060, 2024M760356.

References

[1] Xiangrui Cai, Haidong Xu, Sihan Xu, Ying Zhang, et al. 2022. Badprompt: Backdoor attacks on continuous prompts. *Advances in Neural Information Processing Systems* 35 (2022), 37068–37080.

[2] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. 2024. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 14093–14100.

[3] Xiuyuan Chen, Yuan Lin, Yuchen Zhang, and Weiran Huang. 2024. Autoeval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering. In *European Conference on Computer Vision*. Springer, 179–195.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[5] Yuntian Deng, Yejin Choi, and Stuart Shieber. 2024. From explicit cot to implicit cot: Learning to internalize cot step by step. *arXiv preprint arXiv:2405.14838* (2024).

[6] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. 2024. Physically grounded vision-language models for robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 12462–12469.

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.

[8] Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. 2024. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3385–3403.

[9] Jiaming He, Wenbo Jiang, Guanyu Hou, Wenshu Fan, Rui Zhang, and Hongwei Li. 2025. Watch out for your guidance on generation! exploring conditional backdoor attacks against large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 26220–26228.

[10] Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Backdoor attacks for in-context learning with language models. *arXiv preprint arXiv:2307.14692* (2023).

[11] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. 2024. From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 28076–28086.

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

[13] Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. 2024. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*. Springer, 467–483.

[14] Libo Qin, Qiguang Chen, Hao Fei, Zhi Chen, Min Li, and Wanxiang Che. 2024. What Factors Affect Multi-Modal In-Context Learning? An In-Depth Exploration. *arXiv preprint arXiv:2410.20482* (2024).

[15] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*. Springer, 256–274.

[16] Yan Tai, Weichen Fan, Zhao Zhang, and Ziwei Liu. 2024. Link-context learning for multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27176–27185.

[17] Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. 2024. G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o. *arXiv preprint arXiv:2412.13647* (2024).

[18] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.. In *NeurIPS*.

[19] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805* (2024).

[20] Yuxi Wei, Zi Wang, Yifan Lu, Chenxin Xu, Changxing Liu, Hao Zhao, Siheng Chen, and Yanfeng Wang. 2024. Editable scene simulation for autonomous driving via collaborative llm-agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15077–15087.

[21] Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. [n. d.]. BadChain: Backdoor Chain-of-Thought Prompting for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[22] Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. 2022. Exploring the universal vulnerability of prompt-based learning paradigm. *arXiv preprint arXiv:2204.05239* (2022).

[23] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[24] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915* (2023).

[25] Shuai Zhao, Meihuizi Jia, Luu Anh Tuan, Fengjun Pan, and Jinming Wen. 2024. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. *arXiv preprint arXiv:2401.05949* (2024).